

UNIVERSITÀ
DI PAVIA

Dipartimento di Biologia e Biotecnologie

"Lazzaro Spallanzani"

Laurea Magistralis in Molecular Biology and Genetics

Insights into Peruvian genetic diversity: a preliminary study of an Andean population from Chachapoyas

Nuove prospettive sulla diversità genetica in Perù: uno studio sulla popolazione Andina di Chachapoyas

Supervisor:
Prof. Alessandro Achilli

Co-supervisor:
Dr Anna Tommasi

Experimental thesis by:
Aqueel Kapadia

Academic Year 2025/2026

TABLE OF CONTENTS

1 – ABSTRACT 5

2 – INTRODUCTION 7

2.1 – Population Genetics 7

2.1.1 – Mitochondrial DNA 9

2.1.2 – Mitochondrial Phylogeny 11

2.2 – From Biological Sample to DNA 13

2.2.1 – The Challenges with DNA Extraction from Hair 13

2.3 – Genetic Variation Analysis 14

2.3.1 – Whole Genome Sequencing (WGS) 14

2.3.2 – Genotyping 14

2.4 – Peopling of the Americas 15

2.4.1 – The South American Context 16

2.4.2 – The Ashaninka 18

2.4.3 – The Chachapoyas Province 19

3 – AIM OF THE WORK 20

4 – MATERIALS AND METHODS 21

4.1 – Peruvian Sample Collection 21

4.1.1 – Chachapoyas 21

4.1.2 – Comparative Datasets 22

4.1.2.1 – Ashaninka 22

4.1.2.2 – South Americans 22

4.2 – DNA Extraction 22

4.2.1 – Maxwell® FSC DNA IQ™ 23

4.2.1.1 – Pre-processing 23

4.2.1.2 – Extraction Phase 24

4.2.2 – QIAamp® DNA Investigator 25

4.2.2.1 – Extraction phase 25

4.2.2.2 – Purification phase 26

4.2.3 – Ancient DNA extraction (aDNA) protocol 27

4.2.3.1 – Extraction phase 28

4.2.3.2 – Purification phase 28

4.3 – Library preparation for Illumina® Sequencing 29

4.3.1 – Blunt-end Fill-in 29

4.3.2 – First Clean-up 29

4.3.3 – Adapter Ligation 30

4.3.4 – Second Clean-up and Adapter Fill-in 30

4.3.5 – PCR Amplification 30

4.4 – Library Purification 33

4.5 – Preparation for Illumina Sequencing 34

4.5.1 – Library Normalization and Pooling 34

4.5.2 – Illumina Whole Genome Sequencing 35

4.6 – Sequence Analysis 37

4.7 – Principal Component Analysis 38

4.8 – mtDNA Analysis 39

4.8.1 – mtDNA Haplogroup Classification 39

4.8.2 – Nucleotide and Haplotype Diversity 40

4.8.3 – Chi-squared Analysis 41

5 – RESULTS AND DISCUSSION 42

5.1 – Comparison of DNA Extraction Protocols 42

5.1.1 – Selection of the Optimal Extraction Method 42

5.1.2 – DNA Yield and Quality Assessment 44

5.2 – Complete Mitogenomes 50

5.2.1 – mtDNA Haplogroups of the Chachapoyas province 50

5.2.2 – Comparative analysis of high- and low-altitude haplogroups 54

5.3 – Complete Genomes 57

5.3.1 – Principal Component Analysis 57

6 – CONCLUSIONS AND FUTURE PERSPECTIVES 60

7 – REFERENCES 63

8 – WEBSITES 70

1 – ABSTRACT

Understanding human genetic diversity sheds light on population history, structure, and adaptations. Despite its complex demographic and genetic history, South America has been underrepresented in population genetic studies for a long time. The Peruvian landscape is marked by significant environmental diversity. This study investigates the genetic diversity of Peruvian populations, with a focus on Andean individuals from the Chachapoyas province and the potential influence of altitude on genetic variation. We evaluated three different extraction protocols and an ancient DNA library-building protocol for implementing a robust workflow to extract and sequence high-quality genomic and mitochondrial DNA from hair samples of individuals from Chachapoyas. Further bioinformatic analyses were conducted to compare the genetic diversity of the Chachapoyas region with the Indigenous Ashaninka population of the Amazon. Whole mitogenome haplogroup classification revealed exclusively Native American maternal lineages of the Chachapoyas individuals, which were determined to be statistically different from those assigned to the Ashaninka population. This suggests a varied population structure inside Peru, further supported by the principal component analysis, which revealed a clear genetic separation between the two populations within a broader South American context. This study provides an initial characterization of the genetic diversity of the Chachapoyas and its relationship to other South American populations, while establishing the potential of DNA extracted from hair for population genetic research. The findings contribute to the growing characterization of genetic diversity in Peru and provide a foundation for future deep-sequencing studies of the extensive Peruvian hair sample cohort.

1 – RIASSUNTO

La caratterizzazione della diversità genetica umana è uno strumento fondamentale per ricostruire la storia demografica, la struttura genetica e i processi di adattamento delle popolazioni. Nonostante la sua complessa storia demografica e genetica, il Sud America rimane ancora relativamente sottorappresentato negli studi di genetica di popolazione. In questo contesto, il Perù riveste particolare interesse per l'elevata diversità ambientale e culturale che lo caratterizza. Il presente studio preliminare si propone di caratterizzare la diversità genetica della popolazione peruviana, con particolare attenzione agli individui di origine andina della provincia di Chachapoyas, e di valutare la possibile influenza della variabilità altitudinale sulla struttura genetica.

A tale scopo, sono stati confrontati tre diversi protocolli di estrazione e utilizzato un protocollo specifico per DNA antico per costruire librerie genomiche al fine di sviluppare un flusso di lavoro efficace per ottenere DNA genomico e mitocondriale di elevata qualità a partire da campioni di capelli raccolti nella popolazione di Chachapoyas. I dati ottenuti sono stati successivamente sottoposti ad analisi statistiche volte a caratterizzare la variabilità genetica della popolazione di Chachapoyas e a confrontarla con quella della popolazione indigena Ashaninka dell'Amazzonia.

L'analisi dell'intero mitogenoma ha evidenziato, negli individui di Chachapoyas, esclusivamente aplogruppi di origine nativa americana. La distribuzione degli aplogruppi è risultata significativamente differente rispetto a quella osservata nella popolazione Ashaninka. Tale evidenza è stata ulteriormente confermata dall'analisi delle componenti principali (PCA), che ha mostrato una netta separazione genetica tra le due popolazioni nel più ampio contesto della diversità genetica sudamericana.

Nel complesso, questo studio fornisce una prima caratterizzazione della diversità genetica della popolazione di Chachapoyas e della sua relazione con altre popolazioni sudamericane, evidenziando al contempo il potenziale dei campioni di capelli come fonte di DNA per studi di genetica di popolazione. I risultati ottenuti costituiscono una base preliminare per futuri studi sull'ampio campione di capelli peruviani a disposizione.

2 – INTRODUCTION

2.1 – Population Genetics

Understanding human origins is a fundamental goal as we constantly ask ourselves the questions, *"who are we?"* and *"where do we come from?"*. Population genetics is one field of science leading the path to answer these questions by investigating the processes that shape genetic diversity in populations. A population is defined as a group of individuals of the same species, occupying the same geographical location and sharing a common gene pool. Over generations, population structures change as they are subjected to the effects of evolutionary processes. Evolution provides the framework for understanding the diversity of life on Earth, from the earliest single-celled organism to the millions of species discovered until today. The concept of evolution revolves around the main molecular driving factor: deoxyribonucleic acid (DNA). Its sequence, composed of four nucleotides (A, T, C, and G), stores the genetic information that determines biological traits. Genetic variation arises through random mutations in the nucleotide sequence of the germline and, together with meiotic recombination, contributes to the formation of new alleles. This is subsequently shaped by evolutionary forces such as natural selection, sexual selection, genetic drift and gene flow. The study of this genetic variation in and between populations is known as population genetics.

As a branch of evolutionary biology, population genetics uses the variation in genomes to elucidate the history of populations, inferring connections between lineages and demographic and geographical differences that account for the wide variations. This is of exceptional importance in the context of human populations as the knowledge of our species' past has been an ongoing venture for millennia.

The study on genetic variation has grown exponentially over the past centuries. Initially, phenotypic traits were the optimal sources to study variation, as observed with the foundational inheritance studies done by Gregor Mendel in the 19th century. The discovery of the DNA structure (Watson & Crick, 1953) and the central dogma of Biology (Crick, 1970) enabled an era of biotechnological advancements. New molecular techniques such as DNA cloning (Cohen et al., 1973) and the Polymerase Chain Reaction (PCR) (Mullis & Faloona, 1987) enabled researchers to manipulate DNA. The development of the first DNA sequencing

technology by F. Sanger further transformed the field by enabling direct determination of nucleotide sequences and replacing analysis of single gene variation with DNA-based markers. The development of molecular markers such as Restriction Fragment Length Polymorphisms (RFLPs), Copy Number Variants (CNVs), and Single Nucleotide Polymorphisms (SNPs) allowed researchers to detect variation throughout the genome and provided the first tools for systematic genetic mapping (Botstein et al., 1980). SNPs in particular have become widely used in genetic studies due to their abundance in the genome, the ease of identification and analysis (Wang et al., 1998), and compatibility with high-throughput genotyping technologies. From 1990, a consortium of worldwide labs was established as the Human Genome Project with the aim of mapping the entire human genome through Sanger sequencing. This global project culminated in the first draft of the human genome published in 2001 (International Human Genome Sequencing Consortium, 2001). Although incomplete, the first assembly provided a reference which enabled comparative analysis of genetic variants among populations, including Genome-Wide Association Studies (GWAS). The emergence of Next Generation sequencing (NGS) technologies has enabled the fast and cost-effective sequencing of entire genomes, becoming the 'gold standard' for population-scale genetic variants studies and improving the characterization of the diversity in the human genome. Highly accurate versions of the human genome are now available with the aid of third generation, long-read sequencing technologies. In particular, the Human Pangenome assembly best captures the genetic diversity amongst populations by replacing the linear reference genome with a diverse collection of genomes, improving the detection of structural variants and reducing biases caused by underrepresentation of certain populations in genomic studies.

Before the introduction of genome-wide data, human population studies relied on the information provided by uniparental markers like the mitochondrial DNA (mtDNA) and Y-chromosome inherited from the maternal and paternal lines, respectively. While autosomal DNA, 22 chromosome pairs excluding the sex chromosomes, are inherited equally from both parents, the uniparental systems are transmitted from only one parent without recombination. The uniparental systems have been utilised to investigate ancestry and population demographic history.

2.1.1 – Mitochondrial DNA

Nearly every eukaryotic organism contains mitochondria, rod-shaped, double-membrane organelles crucial in the process of energy production through aerobic respiration and oxidative phosphorylation. An intermembrane space (IMS) separates the outer membrane (OMM) and inner membrane (IMM), the latter of which folds extensively to form projections referred to as cristae (Figure 2.1).

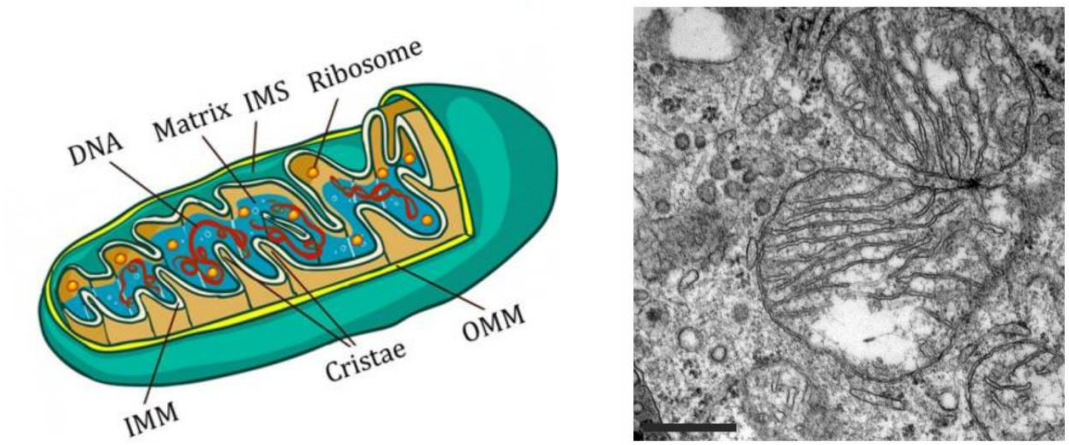


Figure 2.1 *– Left: A schematic representation of a mitochondrion with inner components labelled (Javadov et al., 2020). Right: Transmission Electron Microscope image of two mitochondria with the double membranes and cristae structure clearly visible (Böhler et al., 2018).*

Each mitochondrion contains multiple copies of double-stranded, circular mitochondrial DNA (mtDNA), organized in protein-DNA nucleoids in the matrix. The mtDNA serves as the hereditary unit of the mitochondria. The human mtDNA is a highly conserved double-stranded molecule 16,569 base pairs long, comprising a coding region of 37 genes encoding 13 proteins involved in oxidative phosphorylation, 22 genes coding for transfer RNA (tRNA) molecules, and 2 ribosomal RNA genes. The two strands are distinguished into Heavy (H) and Light (L); the H strand contains a higher proportion of G and A bases and is the primary template for most genes (Figure 2.2). The two strands have different origins of replication. The Heavy strand replicates at a point in the non-coding region called the Displacement loop (D-loop) as there is a triple-stranded conformation due to the invasion of the newly synthesizing strand. The D-loop serves as the control region and regulates the processes of transcription and replication in the molecule. Within the D-loop are sub regions which are known as the hyper-variable regions (HVRs).

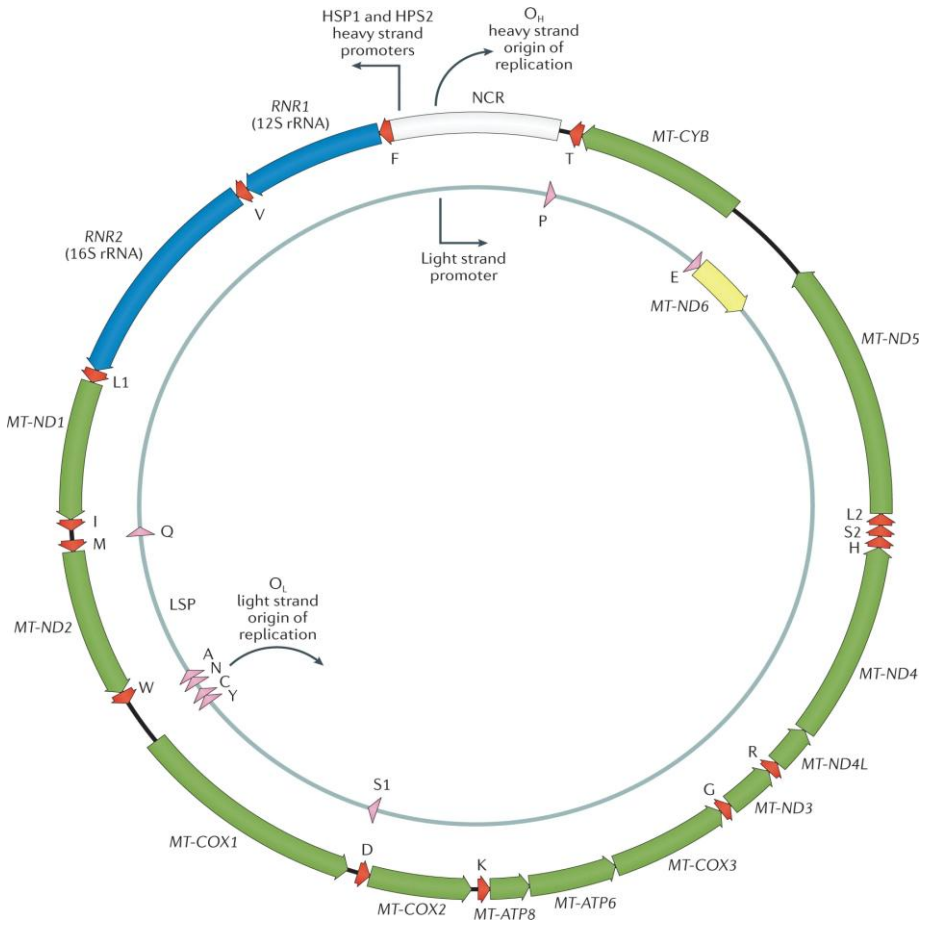


Figure 2.2 *– A schematic representation of the human mitochondrial DNA. The genes of the coding region are represented as green boxes, rRNA genes are marked in blue, and origins of replication are marked on each strand.*

The hypervariable regions (HVRI, HVRII, and HVRIII) mutate at a faster rate than nuclear DNA. Since mtDNA is exclusively passed down through the maternal line with absence of recombination and a limited DNA repair system, accumulations of mutations are the only source of variants. The complete set of variants in the mtDNA is called the haplotype, and characterising mitochondrial haplotypes is a reliable and easy way to reconstruct the maternal lineage. With the advancements in sequencing technologies, the characterisation of the whole mitogenome became more frequent as it increases the resolution and accuracy over HVR characterisation.

2.1.2 – Mitochondrial Phylogeny

Mitochondrial haplotypes can be further classified under haplogroups, a monophyletic unit containing haplotypes sharing the same combination of mutations. Haplogroups coalesce back to one ancestral mtDNA molecule, known as the most recent common ancestor (MRCA). The definition of mitochondrial haplogroups has led to the construction of a human mitochondrial phylogeny tree (Van Oven and Kayser, 2009) (Figure 2.3). The haplogroup of the matrilineal MRCA, dubbed Mitochondrial Eve, has been denoted as L, and all variants are derivations from this haplogroup. Haplogroups in Africa often fall under the L banner, with L0-L7 representing the lineages centred around the continent (Chen et al., 1995). The L3 haplogroup represents an important event in human history as the first 'out-of-Africa' migration that gave rise to all other haplogroups in the other continents. The N, M and R haplogroups are predicted to have formed with the first out-of-Africa migration 50,000 years ago (50 kya), in the South-west of Asia along the Arabian Peninsula (Kivisild, 2015). Some modern Africans belong to the M1 and U6 haplogroup, suggesting recent back-migrations from the Eurasian populations during the Early Upper Paleolithic era (Pennarun et al., 2012). The N and M haplogroups are highly represented in South Asia supporting the theory that modern humans migrated out of Africa via a south Asian route and came to settle in East Asia between 30-40 kya (Macaulay et al., 2005). These two lineages derived different Asian haplogroups such as the A, B, C, D, F, and Y lines. The phylogenetic relationship of the indigenous American haplogroups with those observed in Asia strongly implies that the peopling of the Americas occurred through migration from northeast Asia and was characterized by a strong founder effect. The prevalent haplogroups among indigenous Americans are indeed based on the four major founding macro-haplogroups A, B, C, and D (Torroni et al., 1993), while European-linked haplogroup X is a minor founding haplogroup exclusively restricted to North America. The majority of Pan-American haplogroups are classified under the A2, B2, C1, and D1 haplogroups, while rarer haplogroups, such as D4h3 and C4c have also been identified (Achilli et al., 2008).

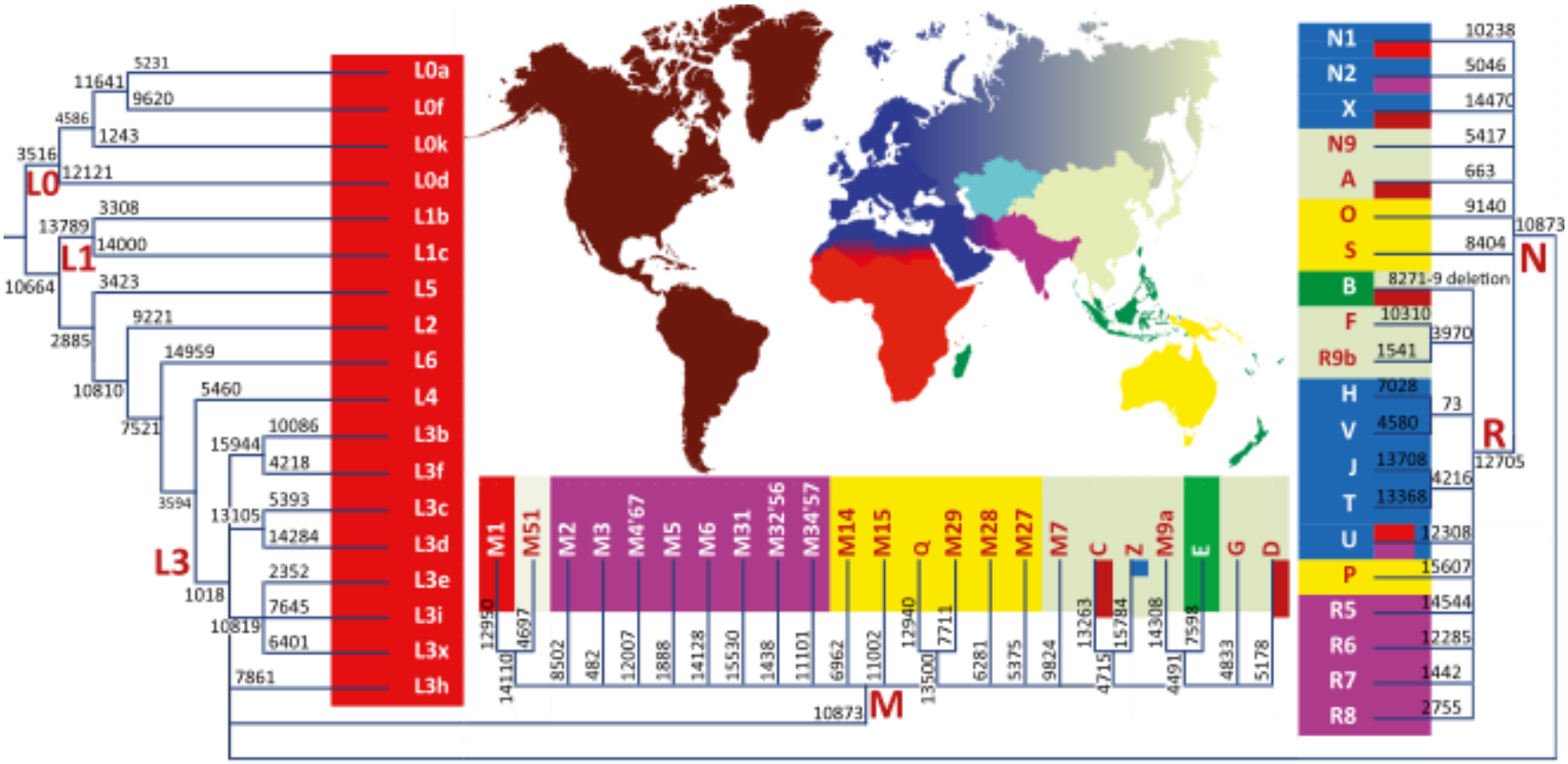


Figure 2.3 *– Worldwide mitochondrial DNA tree showing haplogroup distributions based on https://www.phylotree.org/ (Kivisild, 2015). Major haplogroups found in the Americas as marked in dark red; these include A, B, C, D, and X.*

2.2 – From Biological Sample to DNA

Population genetics studies begin with the extraction of DNA. Blood and saliva are the two most widely used biological sources for human DNA extraction in population genetics studies. Both provide high-quality genomic DNA suitable for a broad range of downstream analyses. The high-quality of extracted DNA is attributed to the presence of a large number of healthy nucleated leukocytes. Countless protocols have been developed to maximise DNA yield and purity (Goode et al., 2014; Gong and Li, 2014; Guha et al., 2017). Blood samples are commonly collected through general phlebotomy practices such as venepuncture, or through the less invasive Dried Blood Spot (DBS) technique, while saliva samples are generally collected through a simple buccal swab or the passive drool method. Alternatively, other biological material, such as hair, semen, and nails, can be sampled for specialised studies.

2.2.1 – The Challenges with DNA Extraction from Hair

Hair is a specialized tissue emerging from the epidermal layer of the skin. Each strand of hair is composed of two parts: the follicle containing the bulb that lies underneath the dermis, and the shaft, a tightly packed structure of keratinized dead cells. Shed hair is one of the most common pieces of biological evidence found at crime scenes and is a vital source of genetic information in forensic investigations. However, extraction of viable DNA is difficult due to the structural nature of the hair tissue (Takayanagi et al., 2003). The first challenge is associated with keratin, a fibrous protein prevalent in hair and nails as a structural molecule. The keratin outer coating on hair tissue poses as a physical barrier to the cells in the internal layer. To degrade keratin, an efficient protease is used along with a strong reducing agent to break the disulfide bridges populating throughout the keratin molecules. The second challenge is the low presence of high-quality DNA. The dead cells of the hair shaft contain little to no genomic DNA, however mitochondrial DNA can still be recovered due to the its persistence through the keratinisation of the shaft. MtDNA recovered from shed hair shafts is often analysed for forensic investigations. The bulb, in contrast, still contains clumps of nucleated cells and providing a source for genomic DNA. However, naturally shed hair exist past the telogen phase where the shaft detaches from the follicle and does not contain the cells from the bulb, making it difficult to recover genomic DNA. Few kits and protocols are available for DNA extraction from hair shafts for forensic purposes (Almeida et al., 2011) and their efficiency to extract

high-quality genomic DNA is not ensured. Hair can be a promising source of genomic DNA due to its ease and of access in humans, provided that a robust extraction protocol which accounts for the difficulties of genomic extraction from hair is implemented.

2.3 – Genetic Variation Analysis

The advancements in high-throughput sequencing technologies, together with the availability of high-quality reference genomes, have allowed researchers to characterize genetic variation at the population scale. These approaches facilitated the transition from studying individual genetic markers to analysing millions of variants distributed across the genome, providing new insights into population history and structure.

2.3.1 – Whole Genome Sequencing (WGS)

The practicality and ease of sequencing using NGS technologies makes whole genome sequencing an appealing option for genetic studies. NGS technologies, such as Illumina®, enable high-coverage, massively parallel sequencing of DNA libraries, generating millions of reliable, low-error sequence reads across the genome. The resulting high coverage provides extensive representation of genomic regions, allowing for genome-wide variant discovery, including SNPs, indels (insertions or deletions), and larger structural variants. The output coverage can be customized to cater to the type of study. For this preliminary or 'screening' study, a low-pass sequencing of an average coverage of 0.5x was preferred. At this coverage level, a lack of variant data can be statistically imputed after variant calling from the reference dataset, making it well-suited for an initial screening of the sample.

2.3.2 – Genotyping

The characterization of an individual's genetic variants at specific genomic loci is referred to as genotyping. Genotyping occurs via microarrays, commonly called 'SNP chips', which contain millions of DNA probes that capture predefined SNPs in the genome, revealing which allelic variants are present at specific loci. The variety of diverse SNP chips facilitates the tailoring of genotype data to preferred downstream analyses (Balagué-Dobón et al., 2022).

However, genotyping encounters several limitations: genotyping cannot result in the identification of novel variants nor can it detect larger, structural variants such as indels (insertions or deletions) or CNVs. Ascertainment bias also often presents an issue, as microarrays can over- or underrepresent variants from certain populations. To counter this bias, a frequently used SNP chip for population genetics studies is the Affymetrix® Human Origins Array which covers global variants heterozygous to 11 modern populations and ancient Neanderthal and Denisovan genomes, providing robust data for genome-wide studies (Fournier et al., 2025). While genotyping remains a popular tool for genomic variants analysis, WGS has become a more powerful tool by allowing the sequencing of whole genomes and capturing both known and novel variants present.

2.4 – Peopling of the Americas

The peopling of the Americas represents the most recent major migration of human populations. Populations from East Asia had ventured as far as Beringia, the land mass connecting Asia to North America (Potter et al., 2017). The Last Glacial Maximum (LGM; ~26-18 kya) was a glaciation period in which the Bering Land Bridge was covered with large ice sheets acting as barriers, preventing Beringian populations from journeying into North America for thousands of years (Tamm et al., 2007). Towards the end of the LGM, however, the Cordilleran and Laurentide ice sheets retracted and allowed passage into the Americas. The exact route into North America is still debated, with the prevailing theories suggesting either a Pacific coastal route or an inland, ice-free corridor (Mandryk et al., 2001), or a combination of both. The expansion of the founding populations is suspected to occur swiftly, with evidence of human civilization in Monte Verde (Chile) dating back to 15 kya (Dillehay et al., 2015). Figure 2.4 shows proposed chronological migrations along Beringia and through the Pacific route to populate the Americas.

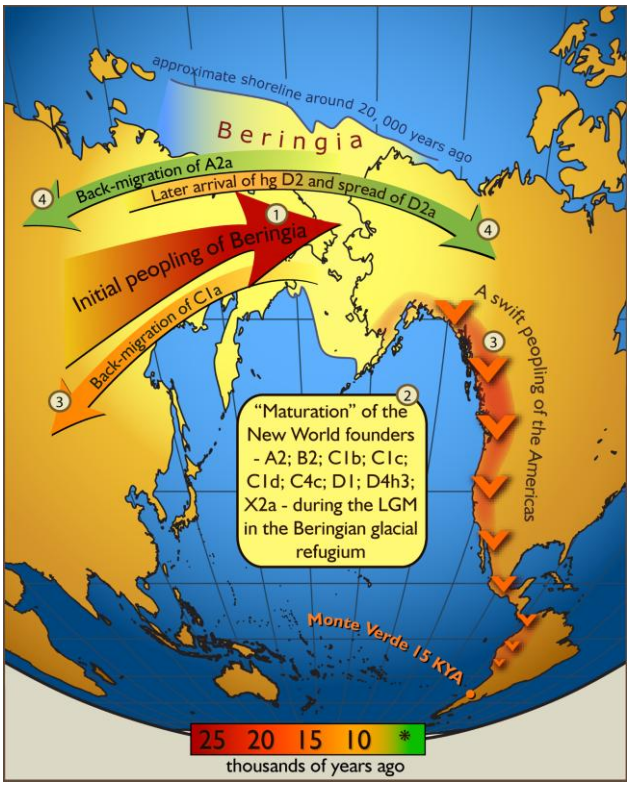


Figure 2.4 *– Schematic illustration of maternal geneflow in and out of Beringia. Colours of the arrows correspond to approximate timing of the events and are decoded in the coloured time-bar (Tamm et al., 2007).*

2.4.1 – The South American Context

As the last continent to be settled, South America experienced several waves of recent migratory events during its initial peopling (Gusareva et al., 2025). Two waves of expansion have been widely suggested, both following the Pacific coastal route through North and Central America between 18-9 kya (Brandini et al., 2018). A recent admixture model suggests a third distinct migration also contributed to the ancestry of present-day Indigenous populations of South America (Castro e Silva et al., 2026).

Following the initial peopling, the next millennia in the continent was characterized by the expansion and diversification of tribal communities. Archaeological discoveries have shed light on the complex civilisations emerging in this period, many preferring the fertile valleys of the Andes and the coast bordering the mountains. The Inca empire, in particular, established itself as the largest civilisation of Southern America, centred around Cusco, Peru. By the mid-

15th century, the empire encompassed modern-day Colombia, Ecuador, Peru, Bolivia, and parts of Argentina and Chile. While the Inca empire facilitated interactions between genetically diverse populations and led to a period of vast cultural and genetic exchange, the expansion of the empire ultimately reduced diversity through conquest and resettlement of the opposing civilisations (Barbieri et al., 2017).

The arrival of Europeans in the late 15th and 16th centuries marked a major shift in the cultural and genetic landscape of the continent (Deneven, 1992). The Spanish colonization introduced long-term genetic changes, with exposure to foreign diseases and violence resulting in severe population bottlenecks for the native populations. Meanwhile, admixture between European colonizers and the Indigenous and the African populations brought through the slave trade introduced diverse, non-native ancestry which characterize many modern South American populations (Guio et al., 2025).

Despite the rich demographic history, South American populations have long been underrepresented in population genetics research. Much of the genetic diversity is uncharacterised and only accounts for a small fraction of the global genome reference bank (Corpas et al., 2024). Recent initiatives such as the Peruvian Genome Project have sought to address this gap by increasing the representation of Indigenous and admixed South American populations in genomic datasets (Guio et al., 2025). The country of Peru captures both the geographic and genetic diversity found across the continent. Figure 2.5 shows the varied terrain of Peru. The country features diverse terrain that ranges between coastal and low-altitude rainforest regions, and the high-altitude Andean valleys. Historically, movement between the regions was geographically restricted, resulting in the prolonged isolation of communities which contributed to the processes of selection and genetic drift to generate unique patterns of genetic variation. The country boasts a large number of Indigenous cultures, with estimated 25-45% of individuals self-identifying as Indigenous (National Population Census Peru, 2017). Currently, there are 55 native communities, with the Quechua being the most represented, while individuals with shared native and Spanish ancestry, referred to as the mestizos, form the largest demographic group of the country. Adaptations to high-altitude environments have become a key area of research to provide valuable information on the genetic adaptation of the populations in the Andes (Julian and Moore, 2019).

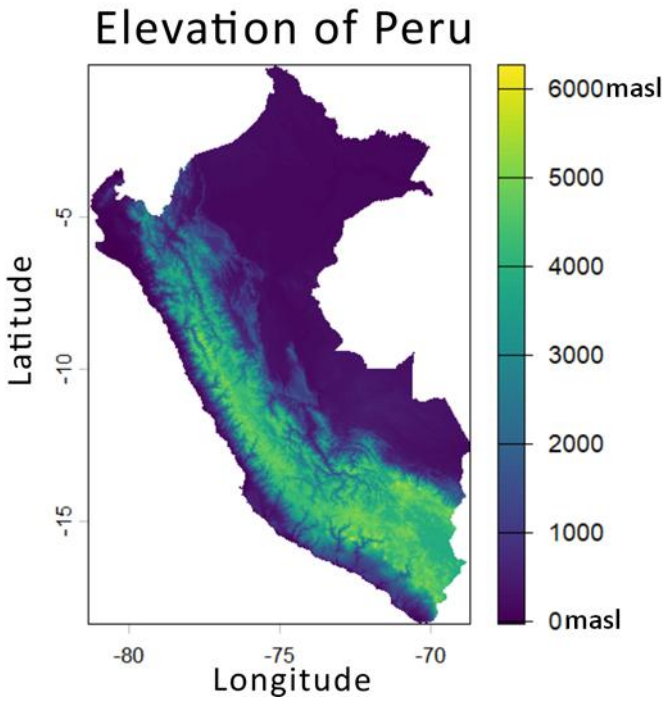


Figure 2.5 – *Map of Peru highlighting the varying altitude.*

2.4.2 – The Ashaninka

The Ashaninka is the largest indigenous group situated in the Peruvian Amazon, at an average elevation between 200 – 800 metres above sea level (m.a.s.l). Population estimates from indigenous rights groups range from 50-80,000 individuals while a national census in 2017 identified 520 Ashaninka communities throughout the Eastern departments, including Pasco, Junin, Ucayali, and Huanuco (Ortega et al., 2020). Historically, there has been evidence of trade between the Ashaninka and Andean populations along the Eastern slopes. A considerable population decline occurred in the 17th century after first contact with the Spanish colonizers due to the exposure to the imported diseases. The Ashaninka have been the subject of several studies on the genetic history of the Americas due to their large population size and the link to pre-colonial communities. Uniparental haplotyping studies have identified maternal haplogroups A2, B2, C1, and D1, which are commonly observed among Indigenous South American populations (Simão et al., 2019), as well as the paternal haplogroup Q (Tineo et al., 2015).

2.4.3 – The Chachapoyas Province

The department of Amazonas is one of the 24 administrative regions of Peru. Despite its name, the department is characterized by mainly Andean topography with elevations starting from 1,000 m.a.s.l. In the south of the department lies the province of Chachapoyas, which has an average elevation of 2,300 m.a.s.l, and shares its name with the ancient civilisation of the Chachapoyas that resided there. The civilisation was active for approximately 1,000 years before being captured and assimilated by the Inca shortly before the Spanish conquest (Barbieri et al., 2017). The area possesses a linguistic mix of Quechua and Spanish. Chachapoyas sits at a divide between the Andes and Amazon environments and previous studies have hinted at rich genetic diversity in the populations of this region. Guevara et al. conducted a haplotyping study on 277 individuals in 2016, revealing approximately 11% of individuals with non-native mtDNA haplogroups, and approximately 40% individuals with non-native Y-chromosome haplogroups. The wider variance in Y haplotypes is a strong indicator of sex-biased admixture, a common occurrence in areas with colonial history (Ongaro et al., 2021). However, the mitochondrial and genome-wide genetic variation of the Chachapoyas populations remains comparatively understudied.

3 – AIM OF THE WORK

South America was the last continent to be settled by modern humans, through a complex peopling process involving multiple migration events. More recently, European colonization and the transatlantic slave trade profoundly reshaped the genetic landscape of the continent. Peru reflects both its remarkable geographic and genetic diversity, making it an ideal setting for investigating population history and adaptation. In particular, the variation in altitude across the country provides a valuable opportunity to study genetic adaptations to different environmental conditions, including those affecting physiological and cardiopulmonary function.

Previous genetic studies on the high-altitude Chachapoyas province in the Andes have begun to characterize the uniparental ancestry of the region and observed high genomic diversity. However, the genetic structure of populations from this region remains relatively unstudied, especially compared to other Peruvian populations. This preliminary project aims to establish a workflow for recovering high quality DNA from hair samples and generating low-coverage, whole-genome sequencing data. The sequences will be used to characterize mitochondrial haplogroups and compare them with the low-land Indigenous Ashaninka population to gain insight into mitochondrial genetic variability across Peru, while the principal component analysis will highlight the genomic position of the Chachapoyas individuals in a South American context.

This project may provide the foundation for future projects utilising the vast hair sample collection across Peru. Deep sequencing studies could provide more comprehensive insights into the genetic diversity, population structure, and evolutionary history of the Peruvians.

4 – MATERIALS AND METHODS

4.1 – Peruvian Sample Collection

4.1.1 – Chachapoyas

The dataset used in this study was a subset of a large-scale sampling carried out in Peru. Hair samples were collected from Peruvian individuals across the country to create a dataset of 10,000 individuals. Upon collection, demographic and genealogical data were recorded, including the collection region, primary language, and birthplaces (BP) of the individual and their parents. For our study, a subset of 94 individuals from the region of Amazonas was initially selected, all from towns/villages situated 1,500 m.a.s.l and above in the province of Chachapoyas (Figure 4.1). Altitudinal data was derived from parental BP locations, using publicly available elevation data as reference.

Hair samples from 47 individuals were selected for DNA extraction after implementation of an exclusion criteria. Samples were excluded when intact hair bulbs were unavailable or when individuals were identified as first-degree relatives (parents, siblings, or offspring) of another participant.

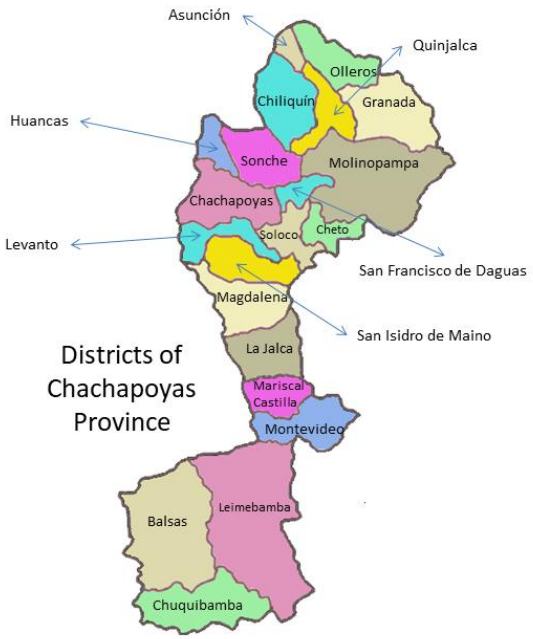


***Figure 4.1** – Map of the districts comprising the Chachapoyas province.*

4.1.2 – Comparative Datasets

4.1.2.1 – Ashaninka

A dataset of 43 Ashaninka individuals was acquired from previously published genotyped data (Capodiferro et al., 2023) generated using the Affymetrix Human Origin 600K array chip. The individuals were healthy, self-identifying Ashaninka adults from the department of Pasco in Peru. The complete mtDNA sequences were not accessible, therefore, the published haplogroup assignments based on mitochondrial SNPs captured by the Human Origin array were used for haplogroup comparison with the Chachapoyas samples.

4.1.2.2 – South Americans

The sequences from the aforementioned datasets were merged with available genotyped data from modern South American populations to create a large comparative dataset for the downstream genomic analyses. These include Indigenous and admixed population groups from Peru, Chile, Colombia, Ecuador, Brazil, Bolivia, and Paraguay.

4.2 – DNA Extraction

Several trials were conducted to find the best extraction method in terms of DNA quality, concentration and integrity due to the difficulty of extracting DNA from hair. Three protocols were selected for this purpose: the Maxwell® FSC DNA IQ™ kit, the QIAGEN® QIAamp DNA Investigator kit, and an extraction method based on a protocol published by Damgaard et al. in 2015 for extracting DNA from ancient specimens. In order to ensure the validity of the results, an equal number of hair strands with attached bulbs were used for each extraction method. The quantity and quality of the extracted DNA was compared to figure out a conclusive method for further extractions. The Qubit 4 fluorometer (Invitrogen, Thermo Fisher Scientific, 2017) was used to measure extracted DNA concentration using a fluorescent DNA-binding dye that emits light upon excitation proportional to the DNA quantity and outputs the concentration in ng/µL. The 4150 TapeStation (Agilent, 2015) instrument was used for capillary electrophoresis

analysis of the quality of the DNA, including average size of the DNA fragments in the extractions. The output of the TapeStation is in the form of an electropherogram (Figure 4.2)

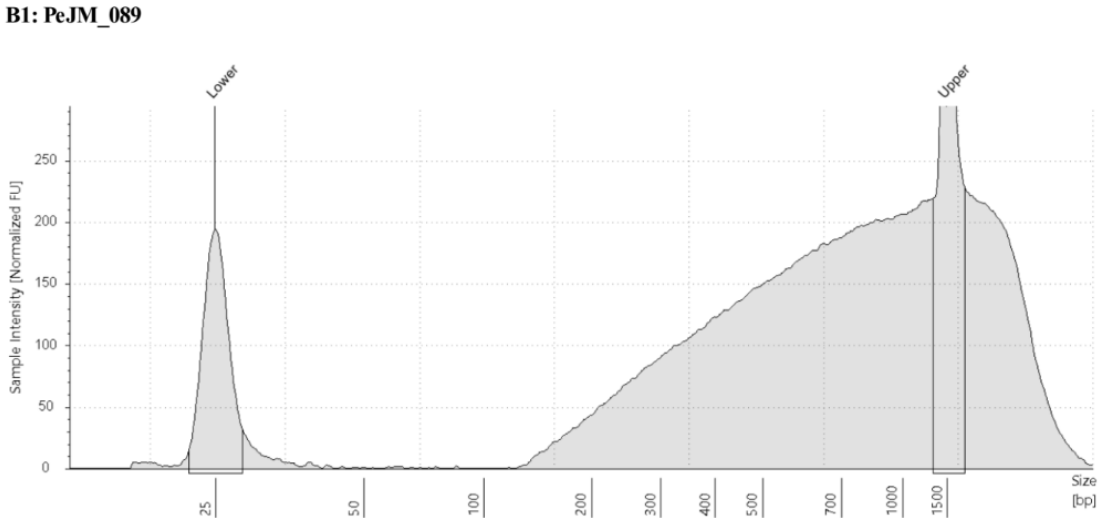


***Figure 4.2** – Electropherogram profile of the extracted DNA sample PeJM_089. The x-axis features fragment size in base pairs, while the y-axis shows the sample intensity in normalized Fluorescent Units (FU). Peaks at 25 and 1500 bp represent the internal library of the TapeStation assays. The D1000 Tape and reagents were used for this quality check.*

4.2.1 – Maxwell® FSC DNA IQ™

The Maxwell® FSC DNA IQ™ kit is a proprietary kit from Promega that utilizes the automated extraction and purification capabilities of the Maxwell® RSC instrument for forensic casework samples. This method can be divided into the pre-processing and extraction phases.

4.2.1.1 – Pre-processing

Hair strands were cut in 1-2 cm pieces from the base using a scalpel, taking care to always include the bulbs; 4-10 hair pieces were cut depending on the overall amount of hair for each sample. Hair samples were incubated in the pre-processing mix provided in the Casework Extraction kit to break down cell membranes and synergistically degrade proteins by destabilizing hydrogen bonding and other non-covalent bonding. This included the proprietary extraction buffer, 1-thioglycerol, and proteinase K. 1-thioglycerol has a thiol functional group, similar to 2-mercaptoethanol or dithiothreitol, which acts as a reducing agent and breaks the disulfide bonds in proteins and can lead to the efficient breakdown of keratin in the hair

(Navone and Speight, 2018). Additionally, Proteinase K enzymatically degrades a broad range of proteins, targeting the peptide bond between hydrophobic residues and is shown to be especially effective for degrading keratin (Ebeling et al., 1974). After adding the pre-processing mix, the samples were incubated at 56 °C for 30 minutes.

4.2.1.2 – Extraction Phase

The samples were partially dissolved after incubation. To separate lysate from remaining hair particles, they were transferred to the DNA IQ™ Spin Baskets (Figure 4.3) and centrifuged for 2 minutes at 13,400 rpm which allowed the lysate to pass through the Baskets. The subsequent addition of a lysis buffer exposed the lysate to chaotropic salts which create a high-ionic environment required for the automatic DNA extraction using the Maxwell® instrument.



***Figure 4.3** – DNA IQ™ Spin Basket used during the extraction phase. When inserted into a tube, the pores on the bottom of the basket allow solution to filter through while retaining undigested material.*

Next, the samples were loaded onto the FSC Cartridges (Figure 4.4). The provided cartridges contained all the reagents required for DNA purification including chaotropic binding buffers, wash buffers, and the magnetic resin. Under high-salt conditions, DNA binds to the magnetic resin. During the automated purification process, the Maxwell® RSC instrument transfers the resin through the different wells of the cartridge using a magnetic plunger, allowing contaminants to be removed during the washing steps. Finally, the DNA is released from the resin under low-salt conditions of the elution buffer and collected in the elution tubes.

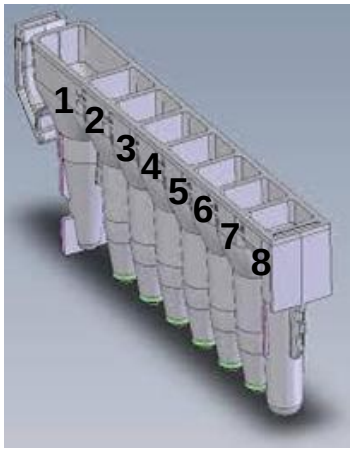


***Figure 4.4** – A Maxwell® FSC cartridge for the DNA IQ protocol. The pre-processed samples are loaded in well #1.*

4.2.2 – QIAamp® DNA Investigator

The QIAamp® DNA Investigator kit is provided by QIAGEN and is specialized in the extraction of DNA from casework samples, including blood, hair, nail clippings, and semen.

4.2.2.1 – Extraction phase

To extract DNA from the selected hair samples, the same number of hair segments with bulbs used in the Maxwell extraction trial was collected. The DNA Investigator kit included multiple buffers containing detergents (ATL buffer) and chaotropic salts (AL buffer) for the cell lysis, these were added to the mixture along with proteinase K and dithiothreitol (DTT) which serve as a disulfide bonds reducing agent instead of 1-thioglycerol (Figure 4.5).

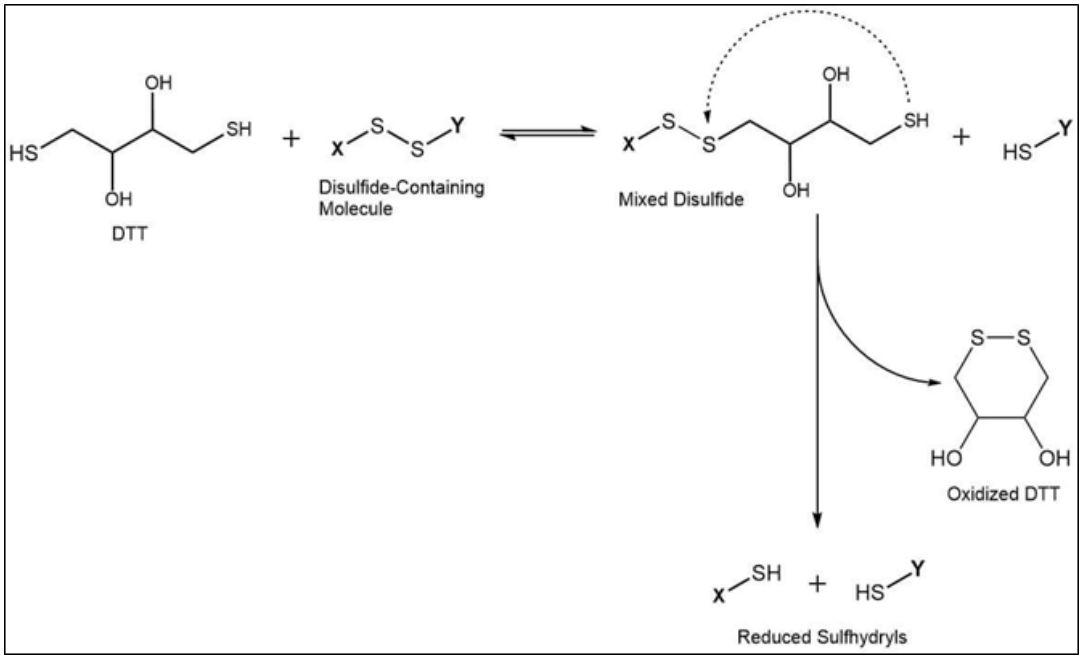


Figure 4.5 – *Mechanism of action of Dithiothreitol (DTT) as a thiol-containing reducing agent, showing the cleavage of a disulfide bond within a disulfide-containing molecule.*

4.2.2.2 – Purification phase

After overnight incubation at 56 °C with agitation at 900 rpm, the hair fragments were partially digested, indicating some level of cell lysis and protein degradation. Cold 100% ethanol was added and the lysate was transferred to MinElute columns for the purification steps. The MinElute column contains a silica gel membrane that selectively binds double stranded DNA under high salt conditions while allowing contaminants such as proteins, salts, and other impurities to pass through (Figure 4.6). The AL buffer's chaotropic salts and ethanol created high-salt conditions that promoted efficient DNA adsorption to the silica membrane. Following centrifugation for 2 minutes, the lysate passed through the silica membrane into the collection tube, while the DNA remained adsorbed to the membrane. The flow-through was then discarded. The next steps involved washing with the AW1 and AW2 buffers containing different concentrations of salt to progressively remove contaminants while maintaining DNA bound to the silica membrane. Each time the flow-through from centrifugation was discarded. A final wash was carried out with 100% ethanol to eliminate any residual salts and contaminants remaining. The column was centrifuged to evaporate the ethanol completely as ethanol can interfere with downstream enzymatic reactions. Finally, the ATE buffer was added

to the column and centrifuged. The low-salt, high pH Tris/EDTA buffer disrupted the interaction between the DNA and the silica membrane, allowing the DNA to elute into the collection tube in a concentrated volume.

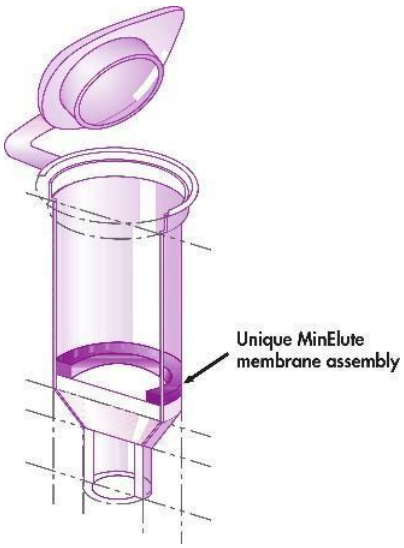


Figure 4.6 – *Cross-section of a MinElute column showing the silica membrane attached to the base. DNA bound by cations in a high-salt solution adsorbs onto the membrane. Source:https://www.qiagen.com/us/products/human-id-and-forensics/investigator-solutions/qiaamp-dna-investigator-kit.*

4.2.3 – Ancient DNA extraction (aDNA) protocol

The ancient DNA extraction protocol used in the archaeogenomic lab at University of Pavia is a protocol adapted from the method described by Damgaard et al. and is routinely employed for the extraction of DNA from powdered bone and teeth samples recovered from archaeological sites. Previous extraction trials showed that the DNA recovered from the hair samples had an average fragment length of approximately 500 bp. Although these fragments are longer than typical ancient DNA, which generally ranges from 150 to 300 bp (Dabney et al., 2013), they are sufficiently fragmented to justify the usage of the aDNA extraction protocol. Therefore, a modified protocol was tested using the same hair samples employed for the other extraction methods. As with the previous protocols, this method comprises both extraction and purification phases but requires more stringent contamination control to ensure successful DNA recovery.

4.2.3.1 – Extraction phase

A lysing mix of Ethylenediaminetetraacetic acid (EDTA), N-lauryl sarcosine, proteinase K, and 1-thioglycerol was aliquoted. The purpose of EDTA in the original aDNA protocol is two-fold, it is used to decalcify the archaeological fragments and to stabilize extracted DNA by sequestering enzymatic cofactors (Mg^{2+}); it was retained in the modified protocol for its stabilising properties. N-lauryl sarcosine, in contrast, denatures proteins and disrupts the plasma and nuclear membranes (Damgaard et al., 2015); the purpose of proteinase K and 1-thioglycerol have been previously described in section 4.2.1.1. Hair bulbs were incubated in the lysing mix overnight at 50 °C with agitation at 1000 rpm. The resulting solution was clear with no visible hair particles.

4.2.3.2 – Purification phase

The protocol utilized QIAGEN products including the PB, PE, EB buffers, and MinElute columns. While the molecular process of this protocol was similar to the purification phase of the DNA Investigator protocol, a stricter working environment was employed to keep contamination to a minimum. The buffers, columns and tubes were sterilised under UV for 15 minutes before usage under a sanitised biosafety cabinet. The lysate was thoroughly mixed with an excess of the PB buffer which contains a high concentration of chaotropic salts and ethanol to ensure adequate DNA binding, before being loaded into a reservoir connected to the MinElute column and centrifuged to allow the lysate to pass through the silica membrane while the DNA remained bound. After discarding the flow-through, two washes were carried out to thoroughly purify the bound DNA. The PE buffer contains ethanol and a lower salt concentration, which kept the DNA attached to the membrane while the contaminants dissolve and pass through. Following the washes, the columns were centrifuged to evaporate left-over ethanol and the DNA was eluted twice with warmed EBT (Tris-CL and 0.05% Tween 20). In contrast to the high-salt conditions of the DNA binding, the EBT has relatively low ionic strength and weakens the interaction between the DNA and the membrane, allowing elution into the solution. Warmed EBT increased the elution efficiency while the multiple rounds of elution ensured maximum yield quality. Tween 20, a non-ionic surfactant, was included to reduce the adsorption of the recovered DNA to the elution tubes.

4.3 – Library preparation for Illumina® Sequencing

Due to the natural fragmentation of the extracted DNA samples, they were good candidates for Illumina® sequencing. Libraries for the extracts were prepared for Illumina® sequencing following an ancient DNA protocol modified from Meyer and Kircher in 2010.

4.3.1 – Blunt-end Fill-in

The multi-step process began with a blunt-end fill-in of the DNA fragments. Naturally fragmented DNA can have uneven ends, such as single-stranded overhangs, where the two DNA strands do not terminate at the same position. This damaged DNA must be repaired by filling-in of the termini prior to sequencing. The NEBNext® Ultra II™ End Repair reagents contain a combination of enzymes to generate blunt-ended fragments. The fill-in uses a T4 DNA polymerase with proofreading activity to add dNTPs to the 5' end of the fragments and remove protruding 3' ends to create even-ended fragments. A separate polymerase inserts a dA to the 3' end for adapter binding, while the T4 Polynucleotide Kinase (PNK) phosphorylates the 5' hydroxyl ends of the fragments. The resulting blunt-ended fragments possessed phosphorylated 5' ends and a 3' Adenine base (Figure 4.7 A).

4.3.2 – First Clean-up

Following end-repair, a clean-up was performed to remove the enzymes, unincorporated dNTPs, and other reaction reagents. The Qiagen PB, PE, and EBT buffers were used along with QiaQuick columns. An excess of the chaotropic, high-salt PB buffer was added to bind the DNA to the silica membrane, while the contaminants were washed away using the low-salt, high ethanol. PE was added and centrifuged to wash away the enzymes, salts and other impurities, after which the DNA fragments were eluted using EBT. Figure 4.7 B provides a brief schematic overview of the clean-up steps.

4.3.3 – Adapter Ligation

The purified DNA fragments were subsequently ligated to sequencing adapters. Adapters are short DNA oligonucleotides that join to the ends of the DNA fragments and provide the sequences to enable PCR amplification and the anchoring of the fragments to the flow cell during the sequencing run. Adapter ligation was performed using the NEBNext Ultra II ligation reagents (Figure 4.7 C), containing the DNA ligase enzyme and ligation enhancing components. A mix containing double-stranded adapters was aliquoted together with the T4 DNA ligase and the samples were incubated for 30 minutes at room temperature to allow the formation of a phosphodiester bond between the repaired ends of the DNA fragments and the adapter.

4.3.4 – Second Clean-up and Adapter Fill-in

A second purification was carried out following adapter ligation to remove the enzymes, unbound adapters, and remaining reaction components (Figure 4.7 D). The QiaQuick column membranes selectively bind DNA above 100 bp in length, therefore unbound adapters can flow through without binding to the column membrane.

The adapters ligated to the fragments have truncated 3' ends and must be filled-in through a polymerase activity (Figure 4.7 E) to produce fully double-stranded adapter-DNA constructs. The 3' hangs of the adapters were filled in using the Bst polymerase, a truncated polymerase lacking the 3' to 5' proofreading capability. The enzyme also operates in isothermal reactions and is naturally deactivated at high temperatures, therefore avoiding another clean-up step. The incubation consisted of two sequential steps: 20 minutes at 37 °C to allow the elongation of the 3' adapter end, followed by 80 °C for 30 minutes to irreversibly deactivate the enzyme.

4.3.5 – PCR Amplification

Following the fill-in, the DNA fragments were amplified through PCR to increase the copy number of the fragments and incorporate the i7 and i5 Illumina barcoding indexes. Indexes are short sequences, approximately 6-10 bp long, that identify each DNA library, allowing multiple samples to be pooled together and sequenced simultaneously. The incorporation of a unique

combination of the i7 and i5 indexes in both library adapters (double indexing) enabled error-free sample identification after computational demultiplexing of sequences. The indexes are incorporated between the DNA fragment and the adapter.

The Invitrogen Platinum™ SuperFi™ II PCR Master Mix provided the polymerase, buffer, and dNTPs, and bovine serum albumin (BSA) was used at a concentration of 10 mg/mL. The PCR used indexing primers that annealed to the adapter sequences with a 5' tail including the index sequence (Figure 4.7 F). Index addition was carried out through a short 12 cycle PCR enrichment step using the following temperature profile:

Initial denaturation	98 °C	30 sec
Denaturation/cycle	98 °C	10 sec
Annealing/cycle	60 °C	10 sec
Elongation/cycle	72 °C	05 sec
Final extension	72 °C	01 min

Following enrichment, DNA concentration and quality were recorded. A significant level of short fragments, between 25 to 100 bp long, was detected in the electrophoretic profiles, indicating a presence of residual adapters and indexes. Further purification was required before preparing the libraries for sequencing.

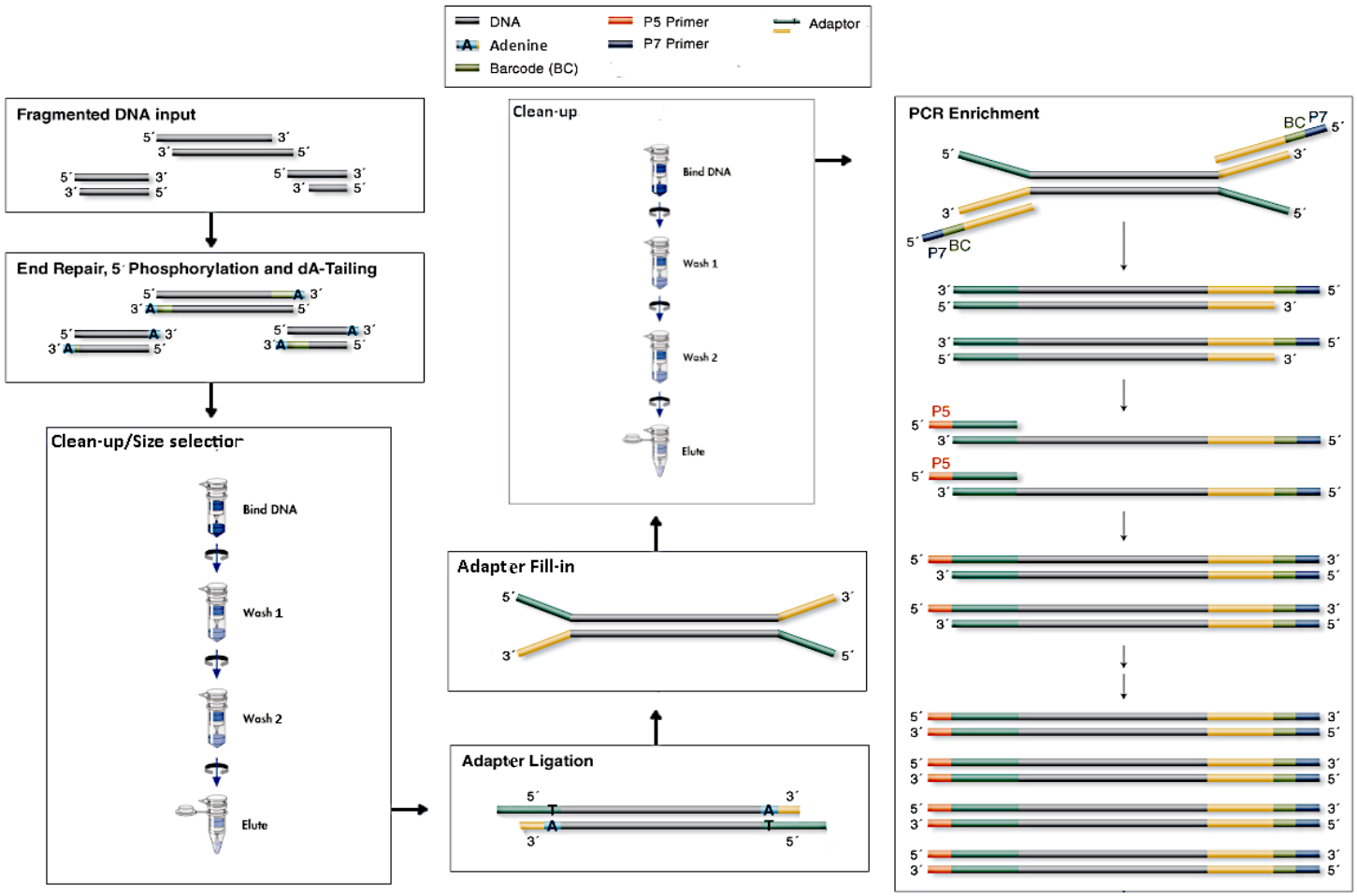


Figure 4.7 *– Workflow of the DNA library prep for Illumina sequencing. A) Blunt-end repair using NebNext Ultra II reagents. B) Clean-up using Qiagen reagents and Qiaquick columns. C-D) Adapter ligation using NebNext Ultra II reagents, and fill-in of 3' ends. E) Second clean-up step with Qiaquick columns. F) PCR enrichment step to incorporate Illumina multiplexing indexes. Adapted from NebNext and Qiagen.*

4.4 – Library Purification

The library purification occurred utilising a process called left-sided size selection. This technique enriches the DNA by selectively removing fragments of a certain length range from the solution. Magnetic beads coated with a polar carboxyl molecule can bind to the negatively charged DNA backbone in a high salt solution. In the presence of a magnet, the beads-DNA complex sticks to the sides of the tube and allows removal of the solution. The ratio of the magnetic bead solution to DNA is crucial to determine the range of the DNA fragments retained during the purification. Figure 4.8 shows the expected electropherogram profiles and corresponding fragment size ranges. The MagMax Pure Bind Beads (Thermo Fisher Scientific) were used for this process. Amplified DNA was brought up to a volume of 80 ul with Nuclease Free Water, after which the magnetic beads solution was added at 1.2 times the DNA volume. The solution was resuspended and incubated at room temperature for 2 minutes to allow binding of the longer fragments, before tubes were placed on a magnetic stand to isolate the beads from the solution. The solution containing the oligonucleotides was discarded, and the beads washed twice with 80% ethanol for further purification before drying. To elute the purified DNA, the beads were resuspended with the Resuspension Buffer (RSB). The change in pH due to low salt-concentration of RSB detached the DNA from the beads. After leaving the tubes on the magnetic stand until the solution turned clear, the solution containing the purified DNA was carefully aliquoted without disturbing the beads.

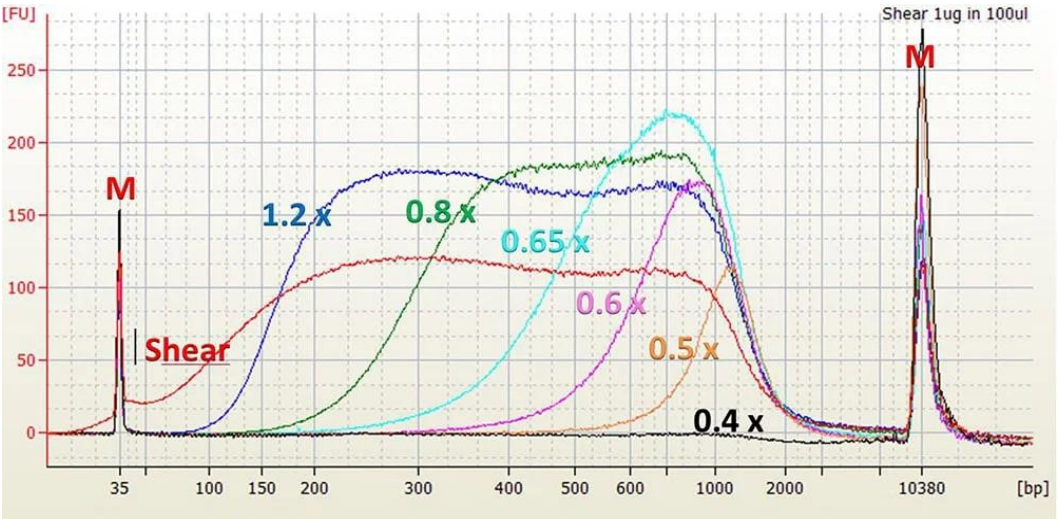


***Figure 4.8** – Electropherogram of left-size selection cases with different sample-bead ratios. For the study purification, a ratio of 1.2 times bead solution to DNA was utilised, allowing fragments in the range of approximately 150-400 base pairs to be recovered. Source:https://www.beckman.it/support/faq/product/size-selection-with-ampure-xp-beads*

Approximately 30 ul of the purified DNA was obtained, and the concentration and quality assessed using the Qubit 4 fluorometer and 4150 TapeStation instruments, respectively, as seen in section 4.2.

4.5 – Preparation for Illumina Sequencing

4.5.1 – Library Normalization and Pooling

The purified libraries were pooled together for the sequencing run. Prior to pooling, a normalization step was performed. The normalization involved equalizing the final concentration of all the libraries to ensure uniform coverage across all samples. Libraries were normalized by diluting with RSB to a final concentration of 2 nM based on the concentration in ng/μL measured via the Qubit 4 Fluorometer and the average fragment size measured via the 4150 TapeStation. The molarity of each library was calculated using the following formula:

$$Concentration\ in\ nM = \frac{Concentration\ in\ ng/\mu L}{660\ g/mol\ \times\ Average\ library\ size\ in\ 150-450\ bp\ range} \times 10^6$$

where 660 g/mol represents the average molar mass of a single base pair (Duewer et al., 2018).

The volume of normalised libraries to be added to the pool was calculated using the following formula:

$$V_{normalised\ library} = Pooling\ factor \times \frac{Final\ pool\ concentration\ (nM)}{Normalised\ library\ concentration\ (nM)}$$

where pooling factor is based on the relative frequency of reads expected for each sample.

Once the libraries were pooled, a final dilution step (8.4 μL pool + 15.6 μL RSB) was performed to prepare the pool to the 700 picomolar concentration required for the Illumina sequencing.

4.5.2 – Illumina Whole Genome Sequencing

Whole Genome Sequencing of the 20 Chachapoyas samples was performed through the Illumina NextSeq 2000 Sequencing System at the Molecular Medicine department, University of Pavia. The NextSeq™ 2000 P3 XLEAP-SBS™ reagent kit was used; 20 µL of the final pool was aliquoted into the cartridge of the kit, which was in turn placed into the NextSeq 2000 instrument.

A sequencing run of 200 cycles for paired end reads was initiated. Upon the start of the run, double-stranded DNA (dsDNA) fragments are denatured into single strands using an alkaline solution within the sequencing cartridge. These single-stranded fragments are then anchored by both ends onto a flow cell, a flat glass surface coated with oligonucleotide primers complementary to the P5 and P7 adapter sequences introduced during library preparation. The fragment forms a 'bridge' and allows bridge amplification to begin as each immobilized fragment is extended by DNA polymerase, after which the newly synthesized strand is denatured, leaving a single strand attached to the surface. The free end subsequently hybridizes to a nearby complementary primer, forming another bridge. Repeated cycles of extension, denaturation, and bridge formation generate thousands of identical copies of each original DNA fragment, producing a localized clonal cluster. Millions of these clusters are formed across the flow cell.

Sequencing is then performed using sequencing by synthesis (SBS). During each cycle, fluorescently labelled reversible-terminator dNTPs are incorporated one at a time into the growing DNA strand. Each nucleotide emits a distinct fluorescent signal, allowing the incorporated base to be identified by the imaging apparatus. The large number of identical molecules within each cluster amplifies the fluorescent signal, improving both the sensitivity and accuracy of base calling. Figure 4.9 gives an overview of the sequencing process.

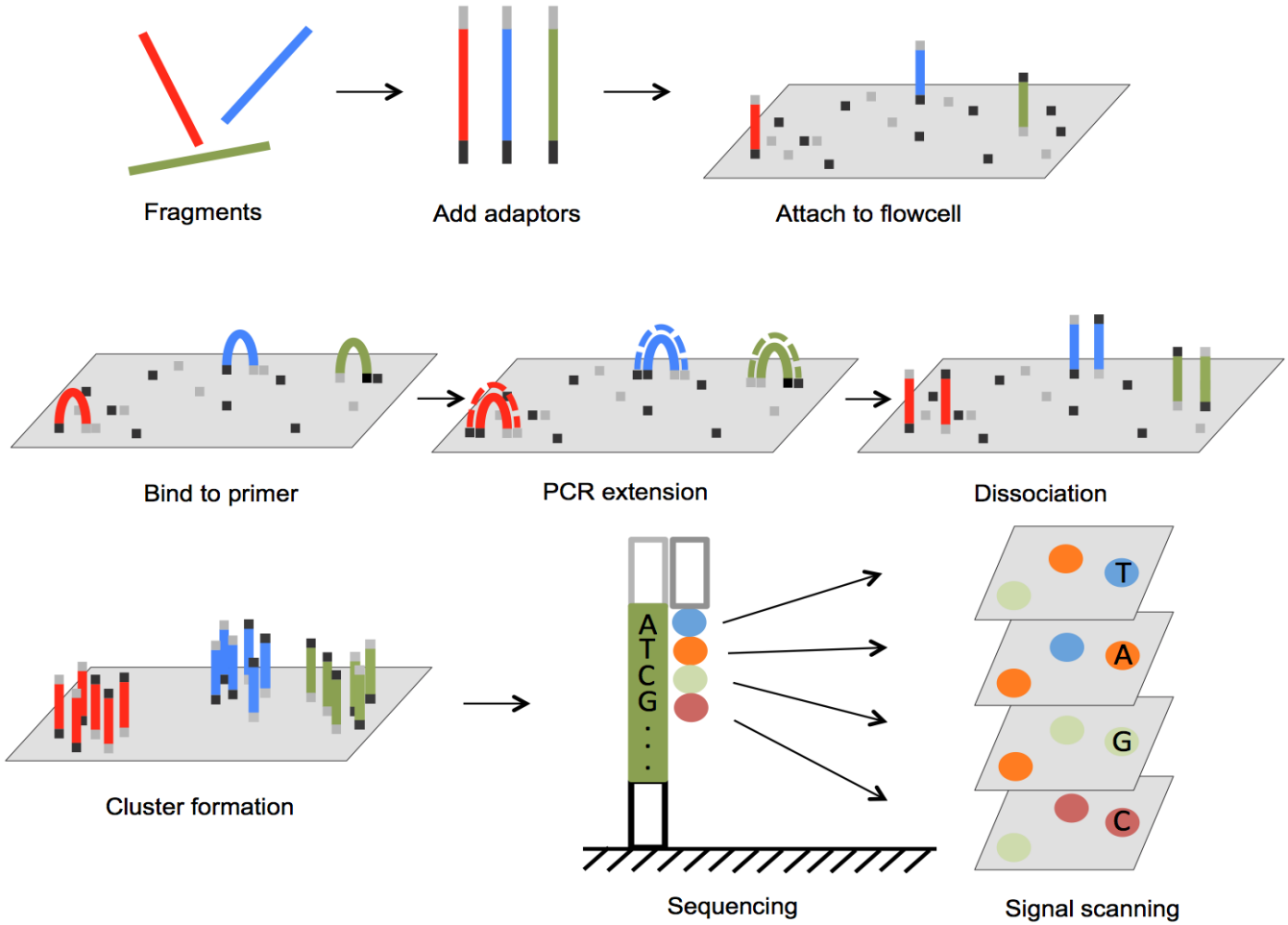


Figure 4.9 *– Overview of the Illumina Sequencing by Synthesis process (Lu et al., 2016).*

4.6 – Sequence Analysis

Paired-end reads were generated in the Binary Base Call (BCL) file format. The outputs were converted to FASTQ files using the Illumina bcl-convert tool (v4.2.4). This software package performs a scanning to identify and separate the reads based on the index combination ligated to the ends of the fragment during library preparation (section 4.3.5). The demultiplexed reads were merged into a single FASTQ file per sample. The FASTQ file is a text file containing sequencing data from filtered clusters and consists of the nucleotide sequence and the quality score of each base (Phred score).

The resulting FASTQ files were analysed by running a dedicated bioinformatics pipeline developed by the lab in charge. This pipeline processes the raw sequence data and produces aligned sequences outputted as Binary Alignment Map (*BAM*) files.

The pipeline assessed raw FASTQ reads for sequencing quality and subsequently trimmed to remove adapter sequences and low-quality bases, with overlapping paired-end reads merged where applicable. The processed reads were aligned separately to the human reference genome (hs37d5) and the revised Cambridge Reference Sequence (rCRS) mitochondrial genome to generate aligned sequences for each type of analysis. The resulting alignments were converted to BAM format and filtered to retain high-quality mapped reads, followed by sorting, duplicate removal, and indexing using SAMtools (v1.23.1; Li et al., 2009). A final quality assessment was performed to verify the integrity of the processed sequencing data prior to downstream genomic and mitochondrial analyses.

After filtering and quality check, the BAM files were converted into genotype data using the plink (v1.9.0-b.8; Purcell et al., 2007) toolset. The software generated the merged comparative dataset in the .bed, .bim, and .fam Plink files.

4.7 – Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate statistical method (Menozzi et al., 1978; Cavalli-Sforza & Feldman, 2003). It uses biallelic polymorphic sites in a quantitative manner by reducing the dimensionality of the data into fewer variables called principal components (PCs). Each PC is a combination of allelic values across the SNPs which decreases the total distance between the data, and its projection on the PC itself. The first PC captures the greatest amount of variance in the dataset, and each subsequent PC is independent of the previous PC. Plotting the PCs against each other can help visualise the variance between individuals.

PCA was performed using the smartpca (version 18140) tool from the Eigensoft (v8.0.0; Price et al., 2006) package. The Plink files generated in section 4.6 required conversion to the Eigenstrat format as a requirement to execute the smartpca command. The conversion was done via another Eigensoft tool, convertf by simply providing a parameter file with the names of the Plink files and the output files. The outputs were in the .geno .snp .ind formats where,

- genotype (.geno) file: contains genotype data for each individual at each SNP;
- snp (.snp) file: contains information about each SNP, including chromosome number, and position in Morgans;
- indiv (.ind) file: contains information about each individual, including sample ID, population ID, and gender.

The smartpca command requires a parameter file providing the inputs, outputs, and a list of the populations to be included in the calculations for the principal components. This is important as populations not included in the population list would be ignored from the computational process, but would be able to be projected onto the PCA itself. The Chachapoyas were included in the PCA in this manner due to their reduced SNP count.

PCA plotting was carried out in R (v4.5.1; R Core Team, 2025) using the tidyverse (v2.0.0; Wickham et al., 2019) and ggplot2 (v4.0.0; Wickham, 2016) packages.

4.8 – mtDNA Analysis

4.8.1 – mtDNA Haplogroup Classification

To classify the haplogroups, the rCRS-aligned BAM files were uploaded for variant calling on the mtDNA-Server2. mtDNA-Server 2 is a pipeline in the mitoverse package (https://mitoverse.i-med.ac.at) dedicated for the analysis of human mtDNA data (Weissensteiner et al., 2024). Several steps occur, including an *Input Validation* step at the start, followed by a quality check through FastQC (Andrews, 2010) and summarised by MultiQC (Ewals et al., 2016). Variant calling is then performed, followed by a contamination report by the mitoverse tool Haplocheck (Weissensteiner et al., 2021). Lastly, the variants are passed through Haplogrep3 (Schönherr et al., 2023) to obtain the haplogroup classification. A fusion of the variant callers Mutect2 and mutserve were used to detect indels and SNPs, while variant detection limit was set to 0.1 or 10%. Thresholds were set for the minimal base quality (>20), mapping quality (>30), and mean coverage (>10), while a subsampling of 2000 reads per sample was selected to standardize the sequencing depth.

A report of the samples was generated, listing the contamination level, mean coverage, average depth, and the haplogroups assigned. A separate file of all variants was also generated, including the detection levels for each mutation. A frequency level ≥85% was considered as a variant, whereas mutations with frequency levels between 15% and 85% were considered as heteroplasmies. A heteroplasmy is defined as the presence of two or more mtDNA variants within a single individual. During sequencing, heteroplasmy can be detected as a mixture of nucleotide bases at the same genomic position. The heteroplasmies detected were compared to the polymorphisms listed in the 'Not Found' column of the HaploGrep output file. These polymorphisms correspond to variants expected for the assigned haplogroup but not detected during the initial classification. If a heteroplasmic variant matched one of these expected polymorphisms, they were included into the final haplotype. Moreover, heteroplasmies not indicated in the 'Not Found' column but within the 15% to 85% range were also incorporated into the final haplotype. For the identification of these heteroplasmic variants, specific ambiguity codes were assigned based on the International Union of Pure and Applied Chemistry (IUPAC) nucleotide code list. These codes are shown in Table 4.1. For example, a nucleotide position containing both Adenine (A) and Guanine (G) would be represented by 'R' to retain the heteroplasmy information.

Table 4.1 – *IUPAC nucleotide codes used for heteroplasmies.*

IUPAC nucleotide code	Base
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C

The final haplotypes were reuploaded to Haplogrep3 to obtain the updated haplogroups.

An exclusion criterion was applied to remove low-quality mitochondrial genomes from downstream analyses. Samples were excluded if they met any of the following criteria:

- Contamination detected by Haplocheck;
- Average depth below 10x;
- Coverage percentage below 90%;
- More than three heteroplasmies detected after final quality filtering.

Based on these criteria, none of the 20 samples were excluded from downstream analyses.

4.8.2 – Nucleotide and Haplotype Diversity

Nucleotide diversity (π) and haplotype diversity (Hd) were calculated to characterize the mitochondrial genetic diversity within the Chachapoyas samples. Nucleotide diversity is defined as the average number of nucleotide differences per site between all possible pairs of sequences in the sample, reflecting the average sequence variation between the mitochondrial genome pairs. Conversely, haplotype diversity is calculated as a measure of the probability that two randomly chosen haplotypes from the population are different, based on the observed haplotype frequencies. Higher Hd values indicate greater mitochondrial haplotype variability in a population (Tommasi et al., 2025). Together, these complimentary measures estimate the

level of mitochondrial diversity by evaluating the number of distinct haplotypes and the amount of sequence variation between them.

The calculation was performed on mitogenome sequences aligned using the MAFFT (v7.526; Katoh & Standley, 2013) software with the *-auto* option which considers the pairwise differences across all sites. The ape (v5.8-1; Paradis & Schliep, 2019) package was used to compute the diversity calculation in R.

4.8.3 – Chi-squared Analysis

A Pearson Chi-squared (χ^2) test of independence was performed to assess whether the distribution of mitochondrial DNA haplogroups differed significantly between the Chachapoyas and Ashaninka populations. A contingency table was constructed using the observed number of haplogroup frequencies in the Chachapoyas samples and the corresponding frequencies reported for the published Ashaninka population. This analysis was used to determine whether the differences in haplogroup frequencies between the two populations were greater than expected. This would provide insight on the haplogroup distributions across the different geographical regions of Peru. The χ^2 test was performed in R using the built-in *chisq.test* command and the corrplot (v0.95; Wei & Simko, 2021) package was used for visualisation of the Pearson residuals to identify the strongest observed differences.

5 – RESULTS AND DISCUSSION

5.1 – Comparison of DNA Extraction Protocols

5.1.1 – Selection of the Optimal Extraction Method

Three trials were performed to extract genomic and mitochondrial DNA using two forensic-level kits (Maxwell FSC DNA IQ and QIAamp DNA Investigator) and a protocol utilised for the extraction of ancient DNA.

The samples PeJM_058, PeJM-062, and PeJM_087 were selected as they possessed the greatest number of hair bulbs. The amount of hair bulbs used for extraction was always kept constant for a fair comparison. Notably only the aDNA protocol resulted in full digestion of the hair material. A comparison of the concentrations of the samples can be seen in Figure 5.1. Each method was successful in extracting DNA from all three samples; the DNA extracted via the Maxwell FSC DNA IQ kit had the highest average concentration, followed by QIAamp and aDNA.

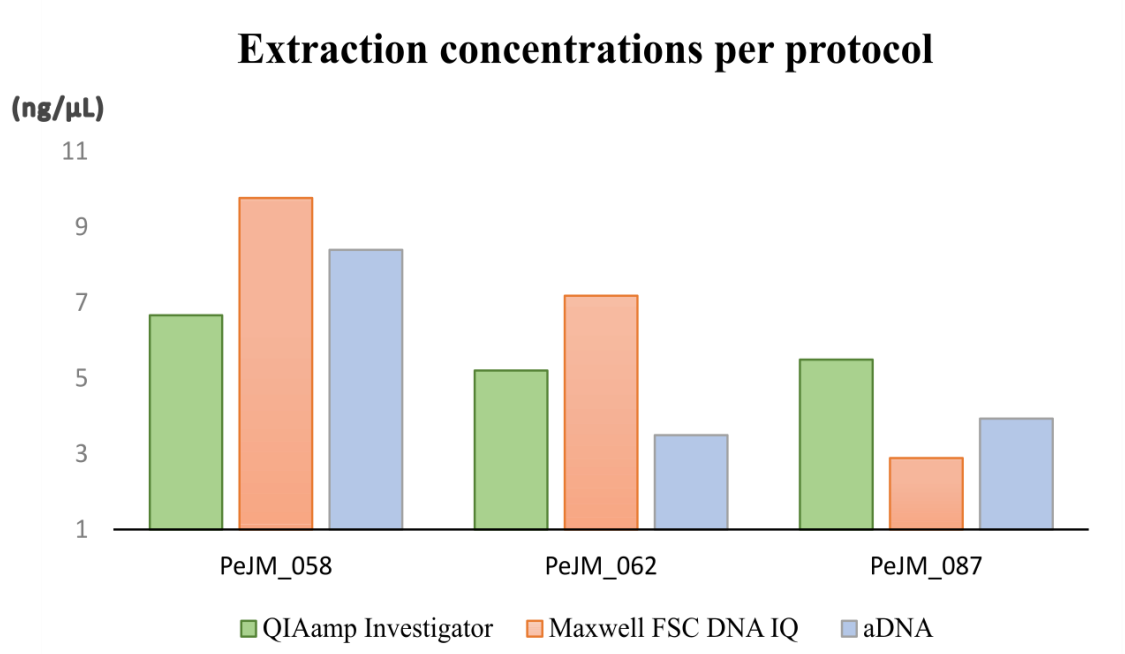


Figure 5.1 *– Extraction concentrations for the samples PeJM_058, PeJM-062, and PeJM_087. Each coloured bar represents the concentration measured for the given protocol.*

A sample extracted via each protocol was selected for Illumina whole-genome sequencing. The sample IDs were updated to include initials denoting the extraction method used; Q – QIAamp DNA Investigator, M – Maxwell FSC DNA IQ, a – aDNA protocol. Libraries were prepared and purified as per sections 4.3 and 4.4, the electropherograms of which are shown in Figure 5.2. The electropherogram for the post-PCR libraries shows two regions of interest, the sample DNA fragments lie between 200 – 700 bp, and the remaining adapters and primer dimer impurities between 25 – 100 bp. After purification, the electropherogram is devoid of the peaks between 25 – 100 bp, indicating successful purification.

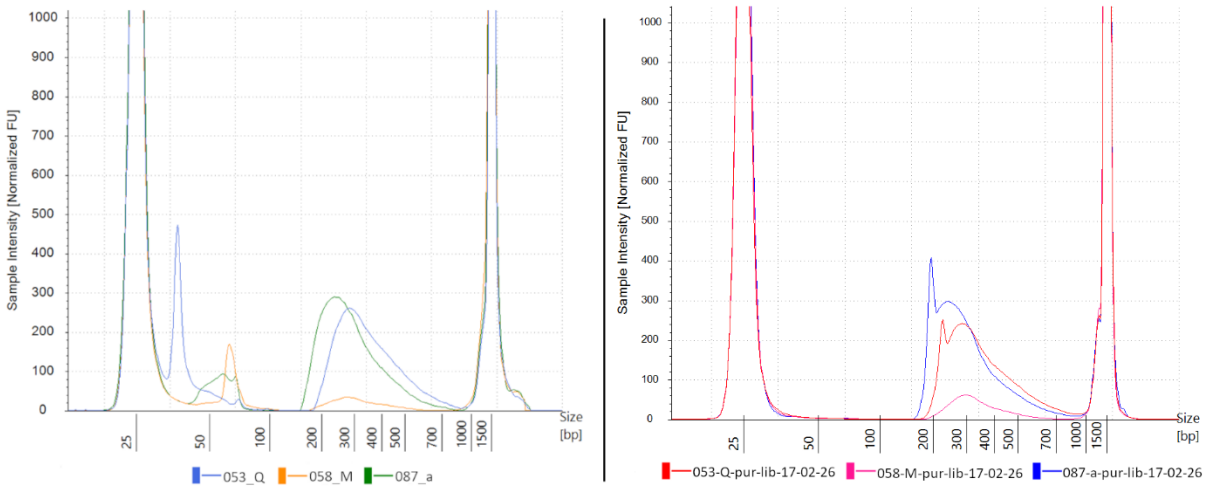


Figure 5.2 *– Right: Electropherogram of post-PCR trial libraries performed via Agilient D1000 ScreenTape. Left: Electropherogram of purified trial libraries performed via the D1000 ScreenTape.*

The sequencing results in Table 5.1 show important parameters to assess the genomic yield of the sequencing, including the mapped reads, endogenous DNA, and the average depth of coverage (DOC). Mapped reads represent the proportion of sequencing reads that align to the reference genome, while the proportion of endogenous DNA is the percentage of reads derived from the sample that successfully mapped to the human reference genome, excluding reads attributable to contamination. The initial endogenous DNA percentage was consistently high across the 3 samples, averaging at 91.5% i.e. in the expected range for typical modern samples (Zhou et al., 2025). However, after filtering out the duplicates, the percentage dropped to an average of 65.12%, indicating the presence of a high number of duplicates introduced during the PCR amplification of the libraries. Moreover, the average DOC represents the sequence

coverage. Trial samples PeJM_053-Q and PeJM_087-a reached the goal of 0.5x coverage or higher, while PeJM_058-M reached 0.477 DOC.

***Table 5.1** – Sequencing results for the trial extraction samples. Reported values include mapped reads, percentages of endogenous DNA before and after filtering, and average DOC.*

Sample ID	Mapped Reads	Endogenous DNA (%)	Filtered Endogenous DNA (%)	Average Depth of Coverage (DOC)
PeJM_053-Q	17912999	90.96	64.80	0.705
PeJM_058-M	12583685	89.87	65.36	0.477
PeJM_087-a	20428363	93.68	65.20	0.657

Each extraction protocol led to the successful sequencing of genomic and mitochondrial DNA. While the samples fluctuated in performance across the metrics, PeJM_087-a possessed the highest number of mapped reads and percentage of endogenous DNA, aligning with the aim of Damgaard et al. (2015) to increase endogenous DNA content. For this reason and considering the higher quality of the extraction process, the aDNA protocol was chosen for the sample extraction.

5.1.2 – DNA Yield and Quality Assessment

All samples (excluding the trial samples already sequenced) were extracted in batches utilising the aDNA protocol. A total of 43 samples were extracted, out of which 22 proceeded to the library stage based on the concentration and electropherogram results. Two libraries, namely PeJM_330 and PeJM_342 were unable to be amplified, thus they were excluded from the study. The final number of samples included in the study for the Chachapoyas and Ashaninka are indicated in Figure 5.3 A, while Figure 5.3 B highlights the genealogical locations of the Chachapoyas samples. Moreover, Table 5.2 displays the samples that were sequenced, along with the demographic information and the extraction concentration. Genealogical origin was assigned based on the reported birthplaces of the individuals' parents.

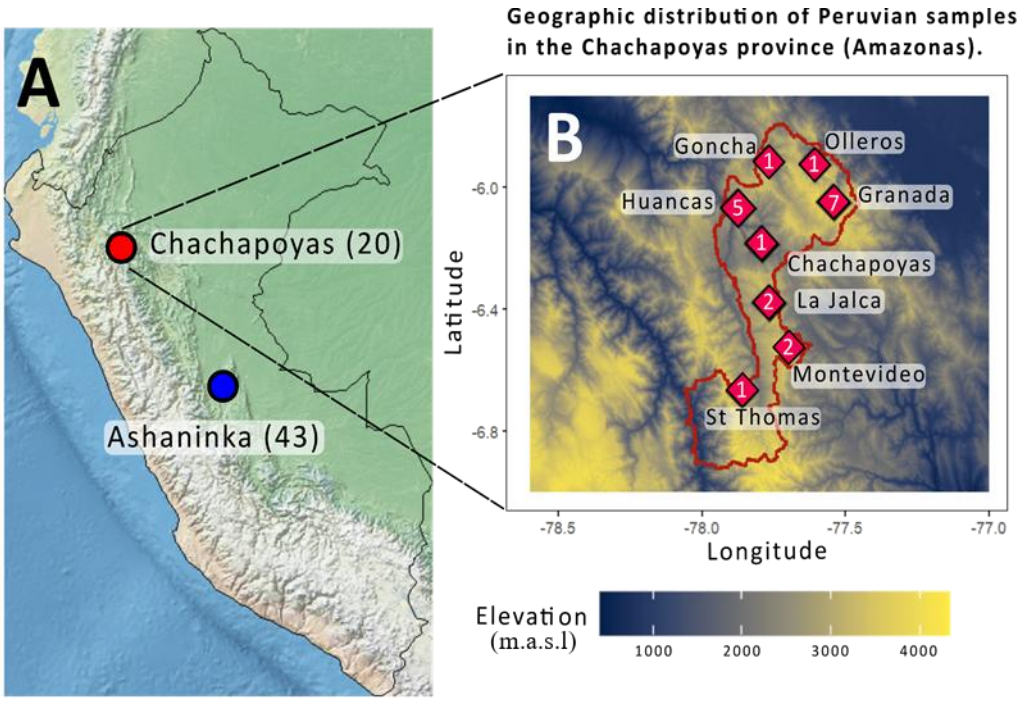


Figure 5.3 *– A) Map of Peru showcasing the diverse topography, accompanied by the locations of the Chachapoyas province in red and the Ashaninka in blue. B) Zoomed raster projection of the Chachapoyas province. The numbers inside the squares indicate the final number of individuals included from each location.*

Table 5.2 – *List of Chachapoyas samples.*

Genealogical Location	Altitude (m.a.s.l)	Sample ID	Extraction concentration (ng/µL)
Goncha	2,820	PeJM_053	3.97
Granada	3,013	PeJM_054	21.7
		PeJM_055	0.874
		PeJM_058	8.40
		PeJM_059	4.07
		PeJM_061	2.87
		PeJM_063	7.58
		PeJM_071	3.49
Olleros	3,050	PeJM_068	3.17
Huancas	2,558	PeJM_080	2.77
		PeJM_081	0.141
		PeJM_083	5.42
		PeJM_087	3.94
		PeJM_088	0.764
La Jalca	2,800	PeJM_098	3.14
		PeJM_099	3.27
Montevideo	2,450	PeJM_323	0.452
		PeJM_324	0.842
St Thomas	2,525	PeJM_333	0.772
Chachapoyas	2,483	PeJM_336	1.97

The concentrations of the post PCR and purified DNA libraries are represented in Figure 5.4. The lowered concentrations of the purified libraries confirm the elimination of adapter/primer in excess, while the electropherogram in Figure 5.5 also indicates that the libraries were successfully purified with fragment lengths between 200 - 500 bp for all samples.

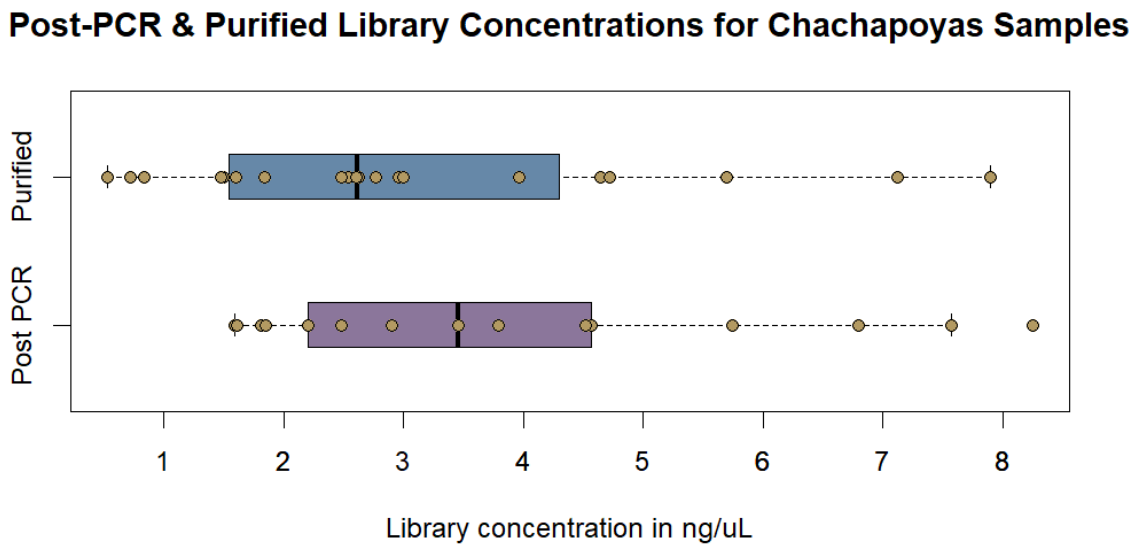


Figure 5.4 *– Boxplot distributions of the concentrations of post PCR and purified libraries.* ***Purified:*** *mean: 3.08, median: 2.61, lower quartile: 1.57, upper quartile: 4.13.* ***Post PCR:*** *mean: 3.95, median: 3.46, lower quartile: 2.20, upper quartile: 4.57.*

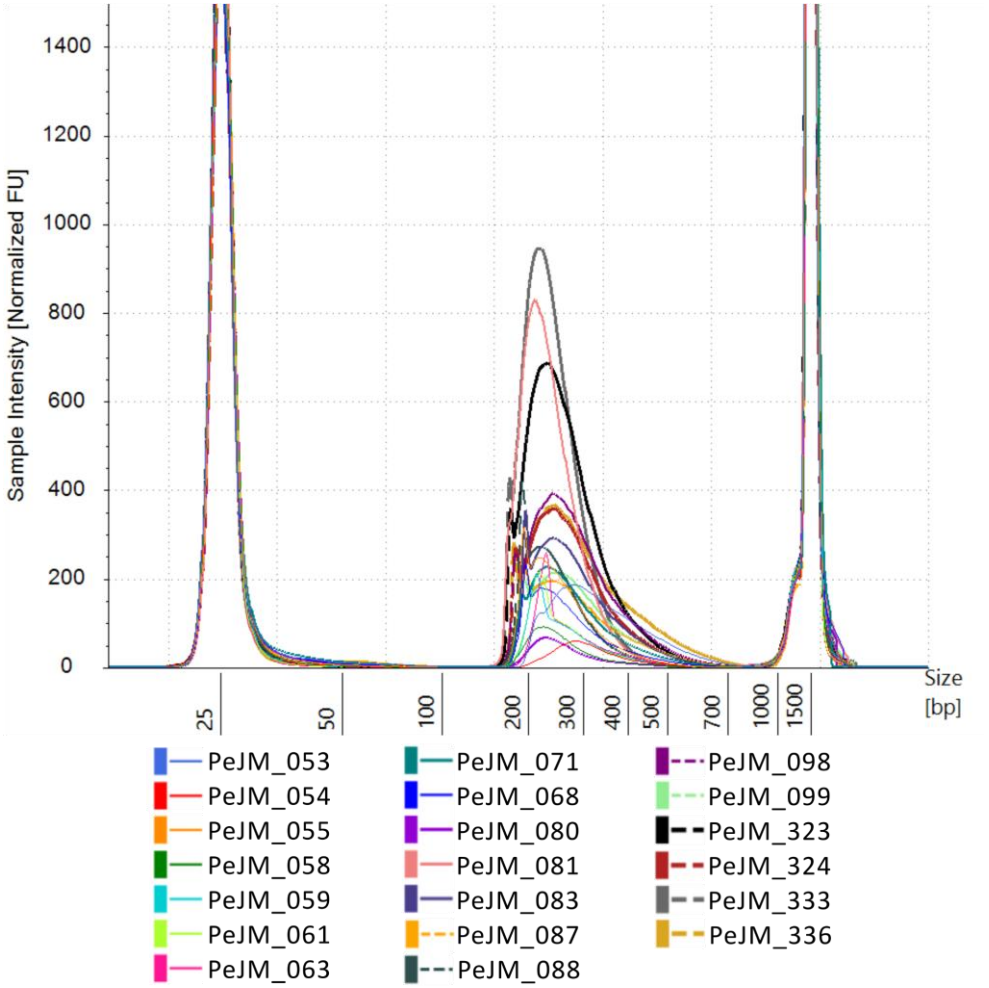


Figure 5.5 *– Electropherogram of purified libraries using the D1000 ScreenTape.*

The twenty samples were eligible for whole genome sequencing based on the quality of the purified libraries. The samples were sequenced using the Illumina NextSeq 2000 platform. Sequencing metrics are listed in Table 5.3, while averages of the metrics are as follows: endogenous DNA content = 94.13%, endogenous DNA content after filtering = 65.25%, and average DOC = 0.80. Overall, these results demonstrate that the selected libraries generated sufficient endogenous DNA for low-coverage whole-genome sequencing and the subsequent downstream genomic analyses.

Table 5.3 – *Sequencing results for the Chachapoyas samples, including mapped reads, percentage of endogenous DNA, and average DOC per sample.*

Sample ID	Mapped Reads	Endogenous DNA (%)	Filtered Endogenous DNA (%)	Average Depth of Coverage (DOC)
PeJM_053	17,912,999	90.96	64.7888	0.7054
PeJM_054	33,935,752	96.57	65.8621	1.0039
PeJM_055	18,742,481	95.35	67.4093	0.5513
PeJM_058	12,583,685	89.87	65.3602	0.4772
PeJM_059	23,293,555	93.73	62.6104	0.7178
PeJM_061	26,461,857	95.40	65.9576	0.8156
PeJM_063	24,266,464	96.0	68.7135	0.7795
PeJM_071	26,270,914	95.90	69.9809	0.8025
PeJM_068	30,154,768	95.68	68.1114	0.8901
PeJM_080	27,108,005	95.81	52.6262	0.837
PeJM_081	33,925,393	95.0	69.655	0.9012
PeJM_083	29,574,750	95.42	68.6928	0.9706
PeJM_087	20,428,363	93.68	65.203	0.6572
PeJM_088	31,027,391	95.21	68.1907	0.9055
PeJM_098	25,327,699	94.22	67.0165	0.8342
PeJM_099	23,108,663	89.70	58.1583	0.7633
PeJM_323	28,989,020	95.07	67.1255	0.9196
PeJM_324	26,584,540	92.11	62.6429	0.8537
PeJM_333	27,229,247	96.19	68.6673	0.7763
PeJM_336	24,591,365	90.78	58.2047	0.8103

5.2 – Complete Mitogenomes

5.2.1 – mtDNA Haplogroups of the Chachapoyas province

The mitogenomes were aligned to the rCRS reference sequence to identify variants and generate haplotype data. The average number of differences per nucleotide position of the Chachapoyas mitogenomes, also known as nucleotide diversity (π), was equal to 0.002. Through haplotype data generated by the mtDNA-Server2, haplotype diversity was calculated to be 0.989. This pattern of diversity suggests the presence of many different haplotypes which are distinguished only by a few mutations.

All sequenced mitogenomes were successfully classified by Haplogrep3. The haplogroups, along with the matching quality and average depth of coverage (DOC), are reported in Table 5.1. Matching quality averaged at 0.92 across all samples, while the average depth was 212 times the genome size.

The Chachapoyas samples were classified into six sub-haplogroups all belonging to Native American haplogroups, except individual PeJM_088 which was classified under the major haplogroup C, likely C1c. The presence of predominantly Indigenous haplogroups in our Chachapoyas samples suggests a strong conservation of Native maternal ancestry in the Andean population. These findings are consistent with haplogroup classification of individuals from the Chachapoyas in previous studies, including those of Guevara et al. (2016 and 2020) and Barbieri et al. (2017), where there was a predominance of the Native American haplogroups A2, B2, D1, and D4 observed in our study. Combining the mtDNA findings with analysis of Y-chromosome ancestral data would provide a more comprehensive assessment of whether the strong conservation of Native American ancestry is reflected in the paternal ancestry.

Table 5.4 – *Haplogroup classification of the Chachapoyas samples via Haplogrep3.*

SampleID	Haplogroup	Quality	Average Depth of Coverage
PeJM_053	B2b	0.952	153.3
PeJM_054	D1	0.866	239.6
PeJM_055	D1	0.935	143.3
PeJM_058	D1	0.891	101.2
PeJM_059	D1	0.935	215.7
PeJM_061	D1	0.926	159.6
PeJM_063	D4h3a9	0.976	215.5
PeJM_068	D1	0.938	163.8
PeJM_071	D1	0.858	206.5
PeJM_080	B2	0.895	149.6
PeJM_081	A2q	0.89	200
PeJM_083	B2	0.907	208.8
PeJM_087	B2	0.934	197.4
PeJM_088	C (likely C1c)	0.907	201
PeJM_098	B2b	0.914	247.4
PeJM_099	D1	0.929	400.5
PeJM_323	D1	0.951	206.1
PeJM_324	B2b	0.914	308.7
PeJM_333	D1	0.912	137.8
PeJM_336	B2b	0.952	394.5

.

The haplogroup distribution is shown in Figure 5.6. Haplogroup D1 was the most represented with 50% of the samples assigned to the haplogroup, followed by B2b (20%) and B2 (15%). Haplogroups A2q, C (likely C1c), and D4h3a9 were reported once for PeJM_081, 088, and 063, respectively. When the haplogroups were distributed across the genealogical locations on the map of the Chachapoyas province, as shown in Figure 5.7, no substructure was noticeable across the genealogical locations. However, the limited sample size makes it difficult to find a proper correlation between location and haplogroup distribution.

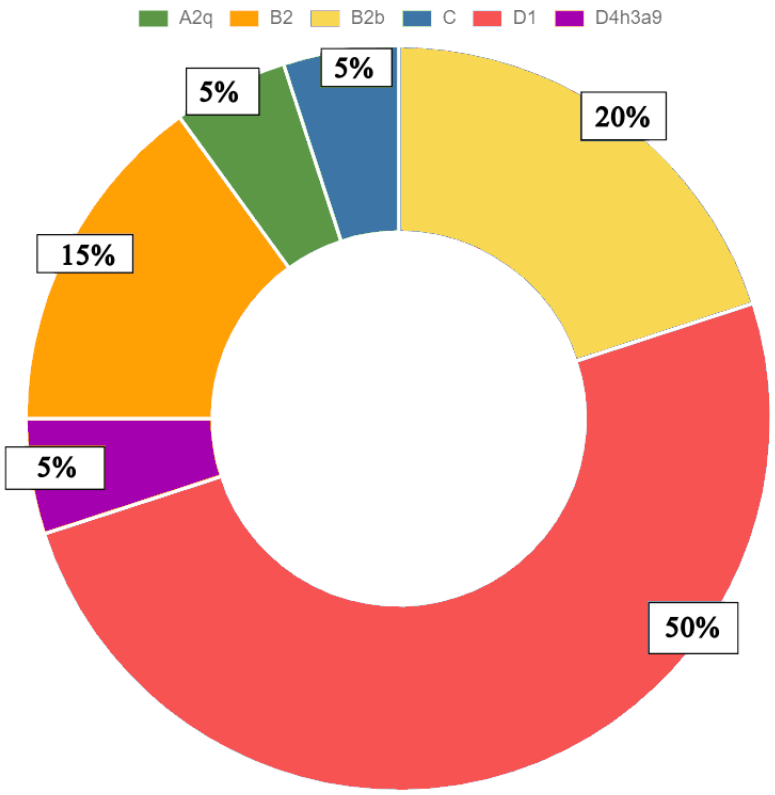


***Figure 5.6** – Haplogroup distribution amongst the Chachapoyas samples.*

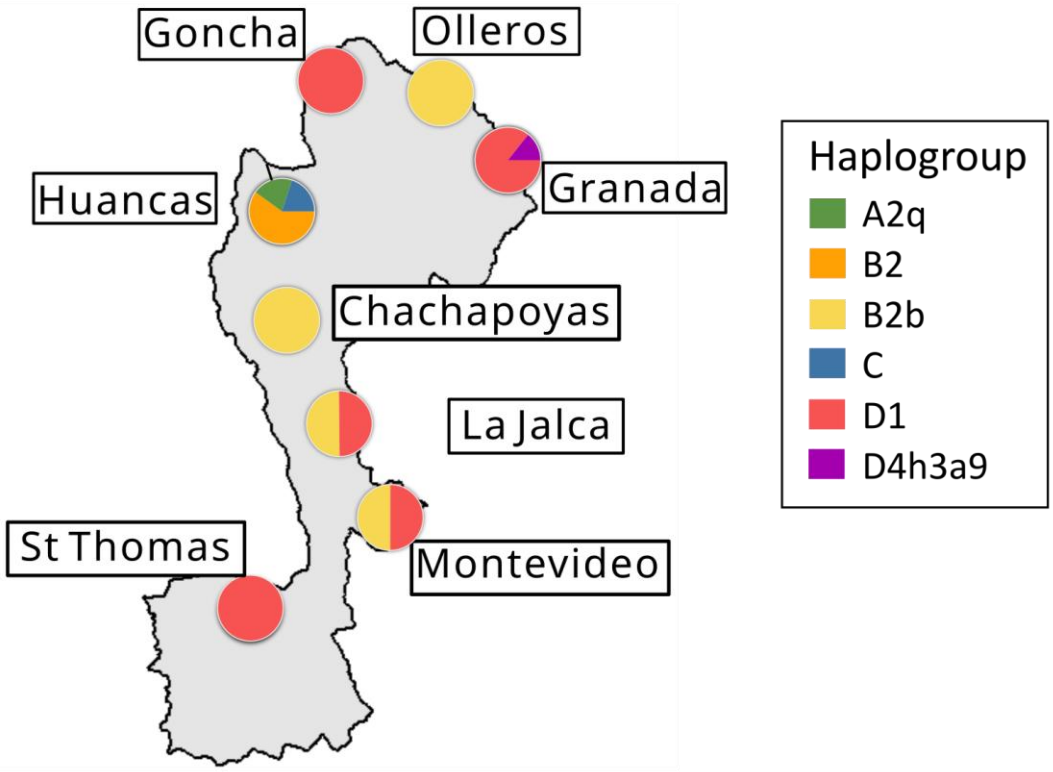


Figure 5.7 *– Haplogroup distribution among genealogical locations. Number of individuals per location; Granada: 7, Olleros: 1, Goncha: 1, Huancas: 5, Chachapoyas: 1, La Jalca: 2, Montevideo: 2, St Thomas: 1.*

5.2.2 – Comparative analysis of high- and low-altitude haplogroups

The Ashaninka population haplogroups were acquired from data published in Capodiferro et al. (2023). Four haplogroups were assigned in the study (Figure 5.8). Of the 43 samples, 15 were assigned to the haplogroup A2af2 (34.9%), 8 individuals (18.6%) were characterized by haplogroup B2, and another 8 (18.6%) by haplogroup C. The remaining 12 individuals (27.9%) were assigned to the haplogroup D4d.

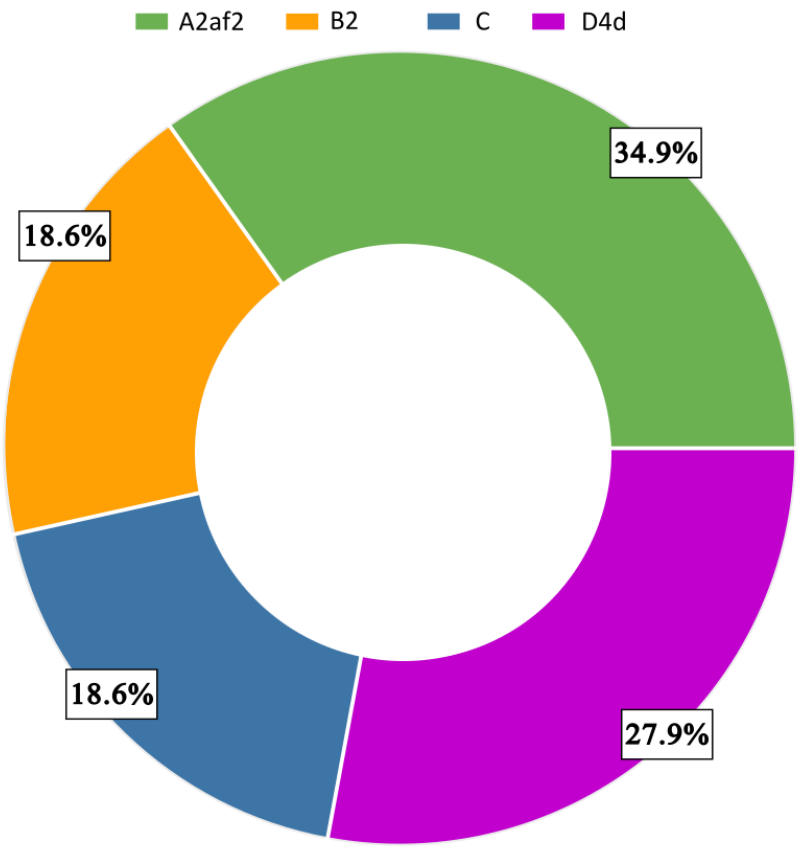


***Figure 5.8** – Haplogroup distribution amongst Ashaninka population.*

To compare the haplogroup distribution between the Chachapoyas sample and published Ashaninka population, a chi-squared test was performed using broader haplogroup categories than initially classified. This categorisation was done due to the limited size of the sample populations (Chachapoyas = 20, Ashaninka = 43) resulting in very low expected frequencies of several of the specific sub-haplogroups of the Chachapoyas sample, such as A2q and D4h3a9. This can reduce the reliability of the chi-squared statistic and make the comparison less sensitive to actual differences. Therefore, related sub-haplogroups were categorised into broader haplogroups to increase the group counts in the contingency table generated. For example, sub-haplogroup A2q observed for the Chachapoyas individual and sub-haplogroup A2af2 obtained from the Ashaninka data were grouped together in the A2 haplogroup category and directly compared with each other. The final haplogroup categories were A2, B2, C, D1, and D4.

The χ^2 test was initially performed on the entire sample populations and resulted in a significant p-value (p-value = 0.0004998). This result suggested the presence of a differential haplogroup distribution between the Chachapoyas and the Ashaninka. To account for the differences in sample sizes, we randomly sampled 20 Ashaninka individuals, replicating the sampling and chi-squared test 10,000 times to increase the reliability of the result. This test also revealed a statistically significant difference between the Chachapoyas and Ashaninka populations (average p-value = 0.004975).

Furthermore, we examined which haplogroups contribute heavily to the difference, observing the mean Pearson residuals across all replicates. The residuals represent the differences in observed versus the expected haplogroup frequencies. Positive residuals indicate a frequency greater than the expected, while lower frequencies are denoted by negatives. For two populations, the residuals for each haplogroup are equal in magnitude but in opposite directions, as shown in Figure 5.9. The A2 and D4 haplogroups showed the strongest positive residuals amongst the Ashaninka, and corresponding negative residuals in the Chachapoyas as these haplogroups were not classified in the sample group. Alternatively, the strongest residual among the Chachapoyas was shown by haplogroup D1, and a corresponding negative residual suggesting that D1 is not prevalent in the Ashaninka population. Other positive residuals among Ashaninka were shown by C, suggesting a higher proportion of these haplogroups among the Ashaninka than Chachapoyas, while B2 was more positive among the Chachapoyas.

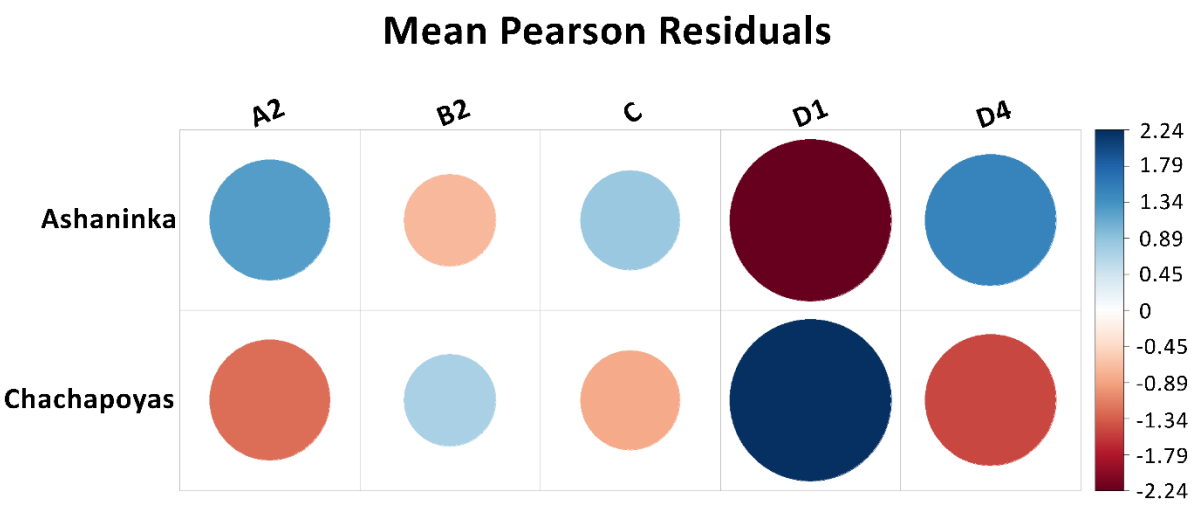


Figure 5.9 *– A representation of the Pearson residuals generated via the chi-squared test. Blue circles represent positive residuals, while red circles indicate negative residuals. The size and intensity of the colour determines the magnitude of the residuals.*

5.3 – Complete Genomes

5.3.1 – Principal Component Analysis

The genomic profiles of the Chachapoyas and Ashaninka were visualised through the principal component analysis. The Chachapoyas samples were simply projected onto the PCA instead of being used in the computation of the principal components; this is a technique utilised for low coverage samples to avoid the skewing of data due to any missed variant calls. Figure 5.10 shows the plotting of the first (PC1) and second (PC2) principal components, explaining 1.16% and 0.73% of the genetic variation, respectively. While these components are relatively small proportions of the total genomic variation, they reveal significant population structure among the South American context. An apparent feature is the isolation of the Ashaninka cluster. The Ashaninka plot at an edge of both PC1 and PC2 axes and form 3 distinct groups. Ashaninka1 is the largest and most isolated group and forms a cline along PC2. Ashaninka2 is located in between Ashaninka1 and the wider South American region, while the two individuals of Ashaninka3 spread along PC1 and cluster with other Peruvian populations. These distinct Ashaninka groups were also reported in Capodiferro et al. 2023 and suggest genetic substructure amongst the Ashaninka. The separation of the Ashaninka from the rest of the indigenous Amazonian populations suggests a relatively conserved genomic profile of the population. The closest cluster to the Ashaninka are the indigenous Amazonian populations of Loreto and Wakyu. The Cocama and Yagua, Amazonian groups that live between Peru, Brazil and Colombia, predictably lie within the large cluster towards the top left of the PCA which includes Brazilian and Colombian Amazonian populations. Interestingly, the Paraguayan populations, although not Amazonian, also lie within the cluster, suggesting the cluster consists mainly of low-land populations. It is important to note that many indigenous populations of the Amazon are geographically spread out between country borders; for the purpose of this PCA, these populations were assigned to a country based on their predominant geographic location.

As PC1 proceeds towards the right, a broad cline forms of the remaining Peruvian populations mixed with the Chilean, Bolivian, and Ecuadorian groups. A cluster of Peruvian populations including the North-western Peruvian groups of Tallan, Sechura, and Eten occupies space between the low-land cluster and the cline.

The Chachapoyas do not cluster into one discrete group, instead following the cline along PC1, overlapping with the Chilean, Bolivian, and other Peruvian populations. These populations, including individuals from populations geographically close to the Chachapoyas province,

predominantly represent the higher altitude regions of the Andes. This creates a strong basis to explain the variation along the first PC as an altitudinal distinction, with the Amazonian/low-land populations on one end, and higher altitude populations on the other.

Another possible explanation of the variation in PC1 would be through ancestry; the Chachapoyas are known to be admixed, therefore the cline could possibly explain the amount of admixture among the South American population. However, PCA alone cannot give evidence of ancestry and must be paired with other analyses, such as ADMIXTURE and F-statistics, to confirm this observation. Interestingly, the individual, PeJM_088, occupies an extreme outlier position from the rest of the Chachapoyas along PC1: this may reflect distinct ancestry patterns, also supported by the mtDNA data (only one within macro-haplogroup C), however the low coverage of the samples leading to missingness of variant calls cannot be ruled out.

Overall, PC1 suggests a separation based on the relative geographical/altitudinal variation, while PC2 separates the Ashaninka from the rest of the Amazonian indigenous populations. There is a case to be made for increased admixture components, however additional analysis would be required to provide better resolution to this point. At the very least, the Ashaninka and Chachapoyas are observed to be widely separated from each other on the basis of both PCs.

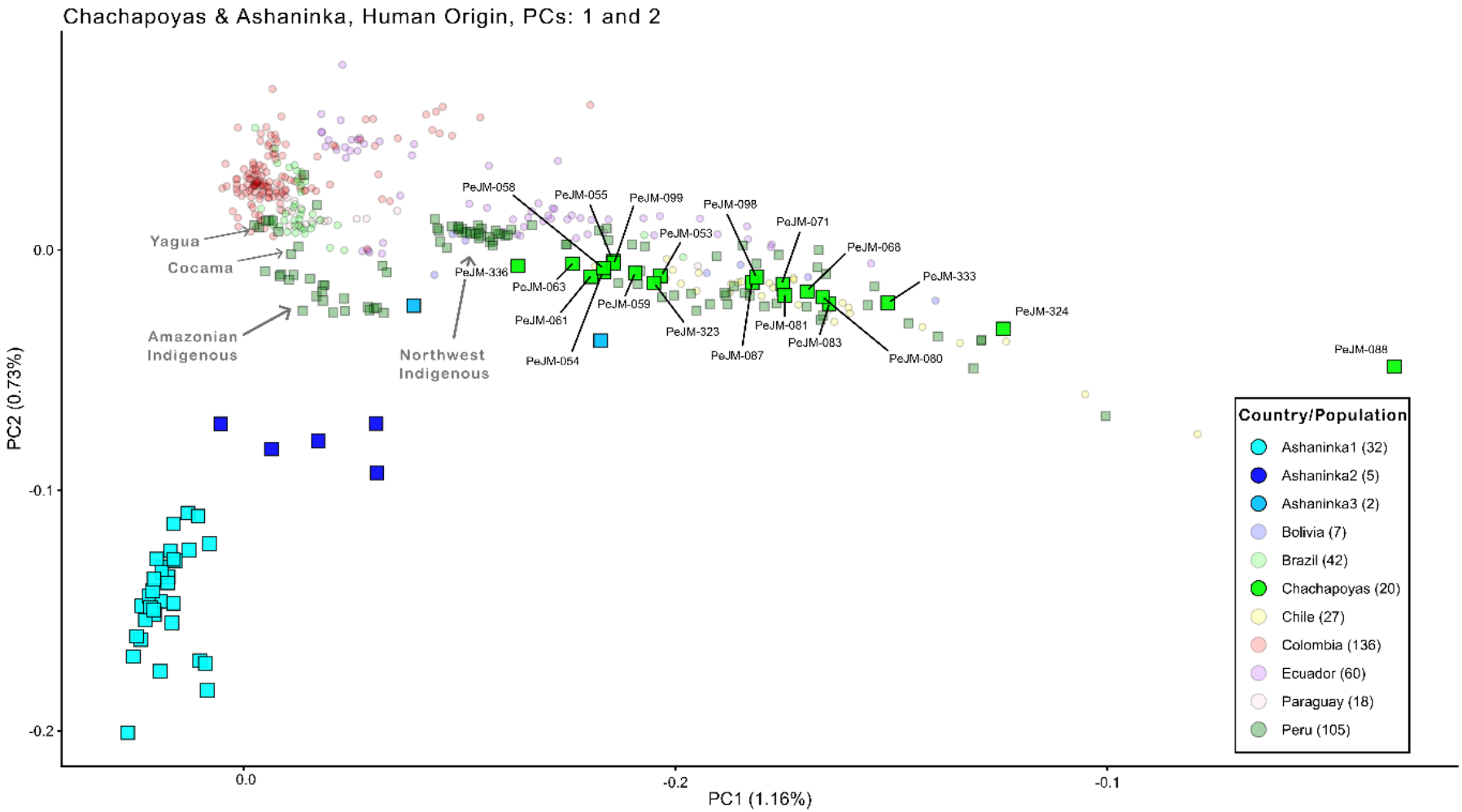


Figure 5.10: *Principal Component Analysis (PCA) of South American populations with projected Ashaninka and Chachapoyas samples. Individuals from the South American reference dataset are represented as coloured circles (except for Peruvian populations, shown as dark green squares). Ashaninka and Chachapoyas individuals are highlighted.*

6 – CONCLUSIONS AND FUTURE PERSPECTIVES

In this study, we assessed the viability of DNA extraction from hair samples collected from across Peru. We selected samples from the Chachapoyas province, a region in northern Andean Peru, and experimented with three different protocols to find the most suitable one. All three protocols generated fragmented DNA, and the aDNA protocol was ultimately considered the optimal method after whole genome sequencing via the Illumina NextSeq 2000 platform. Twenty total Chachapoyas samples were sequenced for the preliminary genetic analyses, with genomic coverage between 0.5 - 1.0x, and over 60% endogenous DNA on average. A high number of PCR duplicates contributed to the lowering of endogenous content.

The second part of the study was spent investigating the genetic characteristics of the Chachapoyas individuals, compared with the Indigenous Amazonian Ashaninka population. By combining the analysis of mitochondrial and genome-wide genetic variation, this work provides insight into the complex genetic history of a population situated at the geographical transition between the Andean highlands and the Amazonian lowlands.

The mitochondrial analysis demonstrated the presence of Native American maternal lineages among the Chachapoyas samples, with a high haplotype diversity indicating substantial variation amongst the population. Through a statistical comparison with the Ashaninka population, we observed a significant difference in haplogroup distribution. The chi-squared test of independence identified enrichment of different haplogroups in each population, such as A2 and D4d in the Ashaninka and D1 in the Chachapoyas. Pairing these results with y-chromosomal data in the future will provide a more comprehensive understanding of ancestral differences between the populations.

The genome-wide principal component analysis demonstrated that the Chachapoyas do not form a single, discrete cluster. Instead, the samples occupy an overlapping region with several Peruvian and other Andean populations. In contrast, the Ashaninka form a particularly distinct and compact cluster, with the three Ashaninka groups occupying different positions within the broader Amazonian genetic structure. The separation of the Ashaninka from the other

Indigenous populations highlights the substantial possibility of substructure amongst the Amazonian populations, rather than supporting a single homogeneous Indigenous genetic profile. A broad geographic pattern is also evident in the PCA. Amazonian and other low-altitude Indigenous populations tended to occupy one side of the first component, whereas populations associated with the Andes extended towards the opposite side in a broad gradient. The Chachapoyas samples lie on this gradient in a pattern consistent with the geographical position in the Andes.

Overall, the mitochondrial and genomic analyses support the view that the Chachapoyas and Ashaninka are genetically distinct. This hints at a larger diversity across the varied regions of Peru, reflecting a complex history of population structure and gene flow.

Several limitations should be considered for this study. First, the number of Chachapoyas individuals sequenced for analysis was relatively limited. While 20 samples are appropriate for a screening study such as this, it still restricts the extent to which the observed patterns can be generalised to the Chachapoyas population. For example, during the haplogroup assignment, haplogroups C and A2q were each observed once in the sample population, whether this is a true representation of the haplogroups or outliers would be resolved with a larger sample size. In the future, larger sample sizes would be better to determine whether the patterns seen represent meaningful population structure or individual outliers. Secondly, the Chachapoyas samples had a relatively low genomic sequencing coverage. While this is another common feature of a screening study, low coverage can reduce the characterisation of the individuals' genetic profiles. Higher coverage genome wide sequencing would improve the genotype resolution during the variant calling step, leading to the better explanation of differences observed in the PCA analysis. Lastly, PCA does not directly explain admixture or resolve ancestry data, nor can it provide evidence of genetic isolation or gene flow. Therefore, while there may be significant admixture in the Chachapoyas population, it is not identifiable with PCA alone. PCA works in tandem with other analytical techniques, such as the ADMIXTURE software, to attribute such patterns to ancestry components.

As mentioned in section 2.5, the genetic adaptations of populations in high- and low-altitude environments is of particular interest. Since populations in Peru live across a dramatic altitudinal range, the adaptation to oxygen availability at high altitudes can be researched at great depth. The biobank of 10,000 hair samples offers an opportunity to investigate the genetics underlying the adaptation to high-altitudes. The Chachapoyas samples used in this

study represent only a fraction of the full cohort. With the implementation of the robust aDNA extraction protocol, this cohort can be utilised in the future for high-throughput, deep-sequencing studies to identify potential genetic variants associated with altitude-related phenotypic adaptations. This can lead to the generation of an important dataset linked to altitude variation worldwide and will not only improve the knowledge gap of South American variation, but also contribute to understanding of cardiopulmonary adaptations.

7 – REFERENCES

Achilli, A. et al. (2008). The phylogeny of the four Pan-American MtDNA haplogroups: Implications for evolutionary and disease studies. *PLoS ONE*, *3*(3), e1764. https://doi.org/10.1371/journal.pone.0001764

Almeida, M. et al. (2011). Efficient DNA extraction from hair shafts. *Forensic Science International: Genetics Supplement Series*, *3*(1), e319–e320. https://doi.org/10.1016/j.fsigss.2011.09.022

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Balagué-Dobón, L. et al. (2022). Fully exploiting SNP arrays: A systematic review on the tools to extract underlying genomic structure. *Briefings in Bioinformatics*, *23*(2), bbac043. https://doi.org/10.1093/bib/bbac043

Barbieri, C et al. (2017). Enclaves of genetic diversity resisted Inca impacts on population history. *Scientific Reports*, *7*(1), 17411. https://doi.org/10.1038/s41598-017-17728-w

Böhler, P et al. (2018). The mycotoxin phomoxanthone A disturbs the form and function of the inner mitochondrial membrane. *Cell Death & Disease*, *9*(3), 286. https://doi.org/10.1038/s41419-018-0312-8

Botstein, D. et al. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, *32*(3), 314–331.

Brandini, S. et al. (2018). The Paleo-Indian entry into South America according to mitogenomes. *Molecular Biology and Evolution*, *35*(2), 299–311. https://doi.org/10.1093/molbev/msx267

Capodiferro, M. R. et al. (2023). The multifaceted genomic history of Ashaninka from Amazonian Peru. *Current Biology*, *33*(8), 1573-1581.e5. https://doi.org/10.1016/j.cub.2023.02.046

Castro E Silva, M. A. et al. (2026). The evolutionary history and unique genetic diversity of Indigenous Americans. *Nature*, *653*(8113), 134–145. https://doi.org/10.1038/s41586-026-10406-w

Cavalli-Sforza, L. L., & Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nature Genetics*, *33*(S3), 266–275. https://doi.org/10.1038/ng1113

Chen, Y. S. et al. (1995). Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *American Journal of Human Genetics*, *57*(1), 133–149.

Corpas, M. et al. (2025). Bridging genomics' greatest challenge: The diversity gap. *Cell Genomics*, *5*(1), 100724. https://doi.org/10.1016/j.xgen.2024.100724

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563. https://doi.org/10.1038/227561a0

Dabney, J. et al. (2013). Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, *5*(7), a012567. https://doi.org/10.1101/cshperspect.a012567

Damgaard, P. B. et al. (2015). Improving access to endogenous DNA in ancient bones and teeth. *Scientific Reports*, *5*(1), 11184. https://doi.org/10.1038/srep11184

Denevan, W. M. (1992). The pristine myth: The landscape of the Americas in 1492. *Annals of the Association of American Geographers*, *82*(3), 369–385. https://doi.org/10.1111/j.1467-8306.1992.tb01965.x

Dillehay, T. D. et al. (2015). New archaeological evidence for an early human presence at Monte Verde, Chile. *PloS One*, *10*(11), e0141923. https://doi.org/10.1371/journal.pone.0141923

Duewer, D. L. et al. (2018). Evaluating droplet digital PCR for the quantification of human genomic DNA: Converting copies per nanoliter to nanograms nuclear DNA per microliter. *Analytical and Bioanalytical Chemistry*, *410*(12), 2879–2887. https://doi.org/10.1007/s00216-018-0982-1

Ebeling, W. et al. (1974). Proteinase K from *Tritirachium album* Limber. *European Journal of Biochemistry*, *47*(1), 91–97. https://doi.org/10.1111/j.1432-1033.1974.tb03671.x

Ewels, P. et al. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fournier, R. et al. (2025). *A SNP panel for co-analysis of capture and shotgun ancient DNA data*. Bioinformatics. https://doi.org/10.1101/2025.07.30.667733

Gong, R., & Li, S. (2014). Extraction of human genomic DNA from whole blood using a magnetic microsphere method. *International Journal of Nanomedicine*, 3781. https://doi.org/10.2147/IJN.S59545

Goode, M. R. et al. (2014). Collection and extraction of saliva DNA for next generation sequencing. *Journal of Visualized Experiments*, (90), 51697. https://doi.org/10.3791/51697

Guevara, E. K. et al. (2016). MtDNA and Y-chromosomal diversity in the Chachapoya, a population from the northeast Peruvian Andes-Amazon divide. *American Journal of Human Biology*, *28*(6), 857–867. https://doi.org/10.1002/ajhb.22878

Guevara, E. K. et al. (2020). Genetic assessment reveals no population substructure and divergent regional and sex-specific histories in the Chachapoyas from northeast Peru. *PloS One*, *15*(12), e0244497. https://doi.org/10.1371/journal.pone.0244497

Guha, P. et al. (2018). A rapid and efficient DNA extraction protocol from fresh and frozen human blood samples. *Journal of Clinical Laboratory Analysis*, *32*(1), e22181. https://doi.org/10.1002/jcla.22181

Guio, H. et al. (2025). The Peruvian genome project: Expanding the global pool of genome diversity from South America. *Frontiers in Genetics*, *16*, 1614021. https://doi.org/10.3389/fgene.2025.1614021

Gusareva, E. S. et al. (2025). From North Asia to South America: Tracing the longest human migration through genomic sequencing. *Science*, *388*(6748), eadk5081. https://doi.org/10.1126/science.adk5081

Hernandez-Rodriguez, J. et al. (2018). The impact of endogenous content, replicates and pooling on genome capture from faecal samples. *Molecular Ecology Resources*, *18*(2), 319–333. https://doi.org/10.1111/1755-0998.12728

International Human Genome Sequencing Consortium et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Javadov, S. et al. (2020). Mitochondria in health and diseases. *Cells*, *9*(5), 1177. https://doi.org/10.3390/cells9051177

Julian, C. G., & Moore, L. G. (2019). Human genetic adaptation to high altitude: Evidence from the Andes. *Genes*, *10*(2), 150. https://doi.org/10.3390/genes10020150

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kivisild, T. (2015). Maternal ancestry and population history from whole mitochondrial genomes. *Investigative Genetics*, *6*(1), 3. https://doi.org/10.1186/s13323-015-0022-2

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H. et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lu, Y. et al. (2016). Next generation sequencing in aquatic models. In J. K. Kulski (Ed.), *Next Generation Sequencing—Advances, Applications and Challenges*. InTech. https://doi.org/10.5772/61657

Macaulay, V. et al. (2005). Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science*, *308*(5724), 1034–1036. https://doi.org/10.1126/science.1109792

Mandryk, C. A. S. et al. (2001). Late Quaternary paleoenvironments of Northwestern North America: Implications for inland versus coastal migration routes. *Quaternary Science Reviews*, *20*(1–3), 301–314. https://doi.org/10.1016/S0277-3791(00)00115-3

Menozzi, P. et al. (1978). Synthetic maps of human gene frequencies in Europeans: These maps indicate that early farmers of the Near East spread to all of Europe in the Neolithic. *Science*, *201*(4358), 786–792. https://doi.org/10.1126/science.356262

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Mullis, K. B., & Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. In *Methods in Enzymology* (Vol. 155, pp. 335–350). Elsevier. https://doi.org/10.1016/0076-6879(87)55023-6

Navone, L., & Speight, R. (2018). Understanding the dynamics of keratin weakening and hydrolysis by proteases. *PLOS ONE*, *13*(8), e0202608. https://doi.org/10.1371/journal.pone.0202608

Ongaro, L. et al. (2021). Evaluating the impact of sex-biased genetic admixture in the Americas through the analysis of haplotype data. *Genes*, *12*(10), 1580. https://doi.org/10.3390/genes12101580

Ortega, J. E. et al. (2020). Overcoming resistance: The normalization of an Amazonian tribal language. *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, 1–13. https://doi.org/10.18653/v1/2020.loresmt-1.1

Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pennarun, E. et al. (2012). Divorcing the late Upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in Africa. *BMC Evolutionary Biology*, *12*(1), 234. https://doi.org/10.1186/1471-2148-12-234

Potter, B. A. et al. (2017). Early colonization of Beringia and Northern North America: Chronology, routes, and adaptive strategies. *Quaternary International*, *444*, 36–55. https://doi.org/10.1016/j.quaint.2017.02.034

Price, A. L. et al. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Purcell, S. et al. (2007). PLINK: A tool set for Whole-Genome Association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Salas, A. et al. (2008). Gender bias in the multiethnic genetic composition of central Argentina. *Journal of Human Genetics*, *53*(7), 662–674. https://doi.org/10.1007/s10038-008-0297-8

Sanger, F. et al. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Schönherr, S. et al. (2023). Haplogrep 3—An interactive haplogroup classification and analysis platform. *Nucleic Acids Research*, *51*(W1), W263–W268. https://doi.org/10.1093/nar/gkad284

Simão, F. et al. (2019). The maternal inheritance of the Ashaninka native group from Peru. *Forensic Science International: Genetics Supplement Series*, *7*(1), 135–137. https://doi.org/10.1016/j.fsigss.2019.09.052

Takayanagi, K. et al. (2003). Investigation of DNA extraction from hair shafts. *International Congress Series*, *1239*, 759–764. https://doi.org/10.1016/S0531-5131(02)00582-4

Tamm, E. et al. (2007). Beringian standstill and spread of Native American founders. *PLoS ONE*, *2*(9), e829. https://doi.org/10.1371/journal.pone.0000829

Tineo, D. H. et al. (2015). Genetic characterization of 27 Y-STR loci in the native population of Ashaninka from Peru. *Forensic Science International: Genetics Supplement Series*, *5*, e220–e222. https://doi.org/10.1016/j.fsigss.2015.09.088

Tommasi, A. et al. (2025). Fifteen millennia of human mitogenome evolution in Sicily. *Science Advances*, *11*(46), eady1674. https://doi.org/10.1126/sciadv.ady1674

Torroni, A. et al. (1993). Asian affinities and continental radiation of the four founding Native American mtDNAs. *American Journal of Human Genetics*, *53*(3), 563–590.

Van Oven, M., & Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, *30*(2), E386–E394. https://doi.org/10.1002/humu.20921

Villani, G. et al. (2026). *From prehistory to present-day: How isolation shaped the distinct genomic makeup of Italian Alpine valleys [Preprint]*. bioRxiv. https://doi.org/10.64898/2026.07.16.738935

Wang, D. G. et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, *280*(5366), 1077–1082. https://doi.org/10.1126/science.280.5366.1077

Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for Deoxyribose Nucleic Acid. *Nature*, *171*(4356), 737–738. https://doi.org/10.1038/171737a0

Weissensteiner, H. et al. (2021). Contamination detection in sequencing studies using the mitochondrial phylogeny. *Genome Research*, *31*(2), 309–316. https://doi.org/10.1101/gr.256545.119

Wei, T. & Simko, V. (2021) Corrplot: Visualization of a correlation matrix. R Package Version 0.92. https://CRAN.R-project.org/package=corrplot

Weissensteiner, H. et al. (2024). mtDNA-Server 2: Advancing mitochondrial DNA analysis through highly parallelized data processing and interactive analytics. *Nucleic Acids Research*, *52*(W1), W102–W107. https://doi.org/10.1093/nar/gkae296

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer International Publishing.

Wickham, H. et al. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Zhou, Y. et al. (2025). Unlocking the potential of animal hair shafts for genomic studies: A comprehensive evaluation of DNA quality. *Biology*, *14*(4), 353. https://doi.org/10.3390/biology14040353

8 – WEBSITES

https://www.agilent.com/en/product/automated-electrophoresis/tapestation-systems/tapestation-instruments/4150-tapestation-system-297322

https://www.beckman.it/support/faq/product/size-selection-with-ampure-xp-beads

https://github.com/taiyun/corrplot

https://www.gob.pe/en/535-check-results-of-the-2017-national-census

https://mitoverse.i-med.ac.at

https://www.neb.com/en/products/e7645-nebnext-ultra-ii-dna-library-prep-kit-for-illumina?srsltid=AfmBOoqHbfVuGOpFNdtCOw70FHj7w155l1SjMrzQSWDifUhpANfnccXL

https://www.phylotree.org/

https://promega.com/en/products/nucleic-acid-extraction/genomic-dna/maxwell-fsc-dna-iq-casework-kit/?tabset0=0

https://www.qiagen.com/us/products/human-id-and-forensics/investigator-solutions/qiaamp-dna-investigator-kit

https://www.thermofisher.com/it/en/home/industrial/spectroscopy-elemental-isotope-analysis/molecular-spectroscopy/fluorometers/qubit/models/qubit-4.html

https://www.thermofisher.com/order/catalog/product/A58521