universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Improving HLA Genotyping Accuracy in Ancient DNA Analysis Using the Human Pangenome: A Case Study

verfasst von | submitted by

Nina Stanišić BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 875
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Bioinformatik
Betreut von | Supervisor:	Ass.-Prof. Pere Gelabert Xirinachs PhD

Acknowledgements

I would like to sincerely thank my supervisor, Pere Gelabert, PhD, for his outstanding guidance, seamless collaboration, and empathetic communication throughout the development of this thesis. His expertise, thoughtful feedback, and continuous support were instrumental in shaping the direction and quality of this work.

I am also very grateful to PhD candidate Michelle Hämmerle, MSc , whose project laid the foundation for this thesis. Her generosity in sharing data, providing technical support, and contributing to the conceptual direction of this work was invaluable.

I would like to extend my appreciation to my former company Platomics GmbH for supporting me in pursuing this research, and especially to Frank Ruge and Lukas Geyrhofer for their mentorship.

Lastly, I want to express my heartfelt thanks to everyone who supported this thesis indirectly, especially my family, partner, and friends, for their unwavering support, patience, and understanding throughout this journey.

Without all of you, this thesis would not have been possible!

Abstract

Ancient DNA (aDNA) poses distinct analytical challenges due to fragmentation, chemical damage, and age-related divergence from modern linear reference genomes. These factors contribute to reference bias, where alleles matching the reference are favored and divergent ones overlooked. This can lead to distortion of evolutionary insights and limited variant detection. Graph-based pangenome references offer a promising solution by representing a broader range of genetic diversity through multiple haplotypes in a unified structure. This reduces bias and improves alignment accuracy, particularly in variable regions, making pangenomes a promising tool for aDNA analysis. In this thesis, I compared linear and pangenome-based alignment strategies using high-altitude Inca mummy genomes from Llullaillaco. Leveraging the newly released draft human pangenome, I demonstrated key advantages of pangenome alignment for aDNA analysis. These include broader variant detection, consistent genome-wide regions with improved coverage, and enhanced representation of medically relevant, technically challenging genes. Most notably, I observed significant improvement in the accuracy of HLA class I gene genotyping using a pangenome alignment. This result has strong practical implications that may inform future methodological advances in aDNA research. Overall, these findings underscore the promise of pangenome-based approaches in enabling deeper insights into population structure, disease evolution, and human adaptation through time.

Keywords: ancient DNA (aDNA), reference genome, pangenome, HLA class I genotyping, human genetic diversity

Zusammenfassung

Altertümlich DNA (aDNA) stellt aufgrund von Fragmentierung, chemischen Schäden und altersbedingter Divergenz zu modernen linearen Referenzgenomen besondere analytische Herausforderungen dar. Diese Faktoren führen zu einem Referenzbias, bei dem Allele, die mit der Referenz übereinstimmen, bevorzugt werden, während abweichende Allele übersehen werden. Dies verzerrt evolutionäre Erkenntnisse und schränkt die Entdeckung von Varianten ein. Graphbasierte Pangenom-Referenzen bieten eine vielversprechende Lösung, indem sie durch die Einbeziehung mehrerer Haplotypen in einer einheitlichen Struktur eine breitere genetische Vielfalt abbilden. Dies reduziert den Bias und verbessert die Präzision bei der Zuordnung von Sequenzen, insbesondere in variablen Regionen, was Pangenome zu einem vielversprechenden Werkzeug für die aDNA-Analyse macht. In dieser Arbeit vergleiche ich lineare und pangenombasierte Ausrichtungsstrategien anhand von Inka-Mumiengenomen aus Llullaillaco. Mithilfe des neu veröffentlichten Entwurfs des menschlichen Pangenoms zeige ich wesentliche Vorteile der Pangenom-Alignierung für die aDNA-Analyse auf. Dazu gehören eine erweiterte Variantenerkennung, konsistente genomweite Regionen mit verbesserter Abdeckung sowie eine bessere Darstellung medizinisch relevanter, technisch anspruchsvoller Gene. Besonders hervorzuheben ist die signifikante Verbesserung der Genauigkeit bei der Genotypisierung der HLA-Klasse-I-Gene durch die Pangenom-Alignierung. Dieses Ergebnis hat starke praktische Implikationen und könnte zukünftige methodische Fortschritte in der aDNA-Forschung vorantreiben. Insgesamt unterstreichen diese Erkenntnisse das Potenzial pangenombasierter Ansätze, Referenzbias zu überwinden und tiefere Einblicke in Populationsstrukturen, Krankheitsentwicklung und menschliche Anpassung im Laufe der Zeit zu ermöglichen.

Schlüsselwörter: altertümlich DNA (aDNA), Referenzgenom, Pangenom, HLA-Klasse-I-Genotypisierung, genetische Diversität des Menschen

Table of Content

1 Introduction

The ability to extract DNA from the remains of ancient organisms, including animals, plants, pathogens, and especially humans, referred to as ancient DNA (aDNA) has greatly enhanced our understanding of genetic evolution, history of life as well as human history (Sawchuk, E. A., et al., 2021). This breakthrough led to the emergence of a new scientific discipline known as paleogenomics. Paleogenomics began in the 1980s with a landmark study where DNA from the extinct quagga, a zebra relative, was sequenced (Higuchi, R., et al., 1984). These pioneering efforts demonstrated that genetic material could survive in ancient remains over extended time periods. Crucially, it also demonstrated that such material can be successfully recovered and analyzed , thereby establishing groundwork for the field. However, the field truly entered its transformative phase in the past decade, driven by rapid technological advancements in high-throughput sequencing and bioinformatics (Lan, T., & Lindqvist, C., 2019; Xiang, Q. Y., K. L. Zuo, and H. Liu., 2024). These breakthroughs have enabled discoveries that were once considered science fiction, such as the detailed reconstruction of a 430,000-year-old ancient human genome from the Homo genus (Meyer, M., et al., 2016) or identification of a previously unknown hominin group, the Denisovans (Reich, D., et al., 2010). Thanks to these developments, paleogenomics now allows researchers to trace the origins of human populations, reconstruct ancient migrations and population relationships, and explore the evolutionary history of extinct plants, animals, and hominins (Kerner, G., Choin, J., & Quintana-Murci, L., 2023). The profound impact of this field was recognized in 2022, when the Nobel Prize in Physiology or Medicine was awarded to Swedish geneticist Svante Pääbo for his outstanding contribution in the extinct human genome and illuminating the evolutionary pathways of modern humans (Xiang, Q. Y., K. L. Zuo, and H. Liu., 2024). Despite these remarkable advances and discoveries, significant challenges remain in the field of paleogenomics.

Ancient DNA (aDNA) is typically highly degraded due to prolonged exposure to environmental conditions. It is characterized by fragmentation, chemical modifications, and post-mortem damages (Ottoni, C., Bekaert, B., & Decorte, R. ,2017). These features present significant challenges for computational analyses, particularly during the mapping step, where short and damaged DNA fragments can lead to misalignments or reduced mapping accuracy. One of the main challenges caused by this is a phenomenon known as reference bias, in which sequencing reads that carry the reference allele are more likely to align to the reference genome, while reads containing non-reference alleles are either misaligned or fail to map altogether (Dolenz, S., et al. , 2024; Günther, T., & Nettelblad, C., 2019).

Another problem, associated with the previous one, is that human reference genomes are predominantly derived from individuals of European descent. As a result, sequencing reads from individuals or populations with different genetic variants ancestries, especially those underrepresented in reference panels, are less likely to map correctly (Peterson, R. E., et al., 2019). This leads to systematic underrepresentation of

genetic variation present in non-European populations, which is particularly problematic in paleogenomics where the goal is often to study diverse ancient populations that may differ substantially from modern Europeans. Consequently, the overreliance on a European-centric reference genome exacerbates both the technical challenges in mapping degraded aDNA and the biological bias in interpreting ancient genetic diversity.
Together, these biases limit the global representativeness of genomic studies and obscure a full understanding of human evolutionary history. Utilizing graph-based, pangenome reference genome can be a promising way of overcoming discussed biases. Addressing these challenges will lead to more comprehensive insights into human population genetics and evolution.

1.1 Molecular Properties and Preservation of Ancient DNA

Ancient DNA (aDNA) is characterized by a series of molecular features that distinguish it from modern DNA and present substantial challenges for genomic analysis. These characteristics arise primarily as a consequence of the death of the organism. In living cells, enzymatic repair and maintenance systems, such as base excision repair and mismatch repair, actively preserve the integrity of the genome throughout the cell cycle. Once the organism dies, these systems cease to function, leaving the DNA vulnerable to a range of postmortem processes, including hydrolysis, oxidation, and enzymatic cleavage, that lead to its progressive degradation, chemical damage, and fragmentation (Pääbo, S., et al., 2004).

1.1.1 DNA fragmantation

DNA fragmentation is a highly prevalent pattern observed in nearly all ancient biological remains. The mechanistic basis of this degradation follows a well-established two-step process involving hydrolytic depurination followed by β-elimination (Dabney, J., Meyer, M., & Pääbo, S. ,2013; Orlando, L., et al., 2021). In the first step, hydrolytic depurination occurs: the N-glycosidic bond linking a purine base (either adenine or guanine) to the deoxyribose sugar is cleaved through the action of water. This reaction results in the loss of the purine base and generates an abasic site (also called an apurinic site), where the sugar-phosphate backbone remains but the base is missing. In the second step, this unstable abasic site undergoes β-elimination. This chemical reaction involves the removal of a proton (H^+) from the 2' carbon of the deoxyribose sugar, triggering cleavage of the DNA strand at that location (Lindahl, T., 1993). The combined effect of depurination and β-elimination leads to single-stranded breaks. As a result of this process, ancient DNA is typically found in highly fragmented states, with fragment lengths commonly ranging from 40 to 500 base pairs (Dabney, J., Meyer, M., & Pääbo, S. ,2013).

1.1.2 Misincorporation Patterns in Ancient DNA

Another commonly observed pattern of chemical damage in ancient DNA is caused by hydrolytic deamination. This chemical reaction involves the removal of an amino group ($-NH_2$) from a nucleotide base and its replacement with a hydroxyl group (-OH) through the addition of a water molecule. The most prominent example of this is the deamination of cytosine, which converts it into uracil. During sequencing, uracil behaves like thymine and pairs with adenine, resulting in an observed C-to-T substitution in the sequenced data. This misincorporation is most frequently found at the 5' ends of ancient DNA molecules, making it a characteristic damage signature that is widely used as an authentication marker for ancient DNA. In addition, a G-to-A misincorporation pattern is often observed at the 3' ends of sequences. This is the complementary signal to C-to-T damage on the opposite strand and it occurs less frequently than cytosine deamination. Nevertheless, both substitutions serve as critical molecular hallmarks of degraded ancient DNA (Briggs, A. W., et al., 2007; Orlando, L., et al., 2021).

1.1.3 Influence of Internal and External Factors on Preservation

The degradation of ancient DNA is ultimately driven by a complex interplay of external (environmental) and internal (intrinsic) factors that influence the preservation of biological remains over time (Raffone, Caterina, et al., 2021; Eriksen, A. M. H., et al, 2025). External factors include variables such as temperature, humidity, pH, microbial activity, presence of groundwater, and organic matter in the burial environment. High temperatures accelerate chemical reactions like hydrolysis and oxidation, while water promotes enzymatic and microbial attack. Similarly, low pH and rich microbial communities can contribute to faster DNA decay. Optimal preservation occurs in environments that are cool, dry, dark, anaerobic, and slightly alkaline, as these conditions slow down degradation and microbial activity (Bollongino, R., Tresset, A., & Vigne, J. D., 2008). Intrinsic factors refer to the biological and structural properties of the sample itself. These include bone density, collagen content, mineral composition, and the skeletal element. Tissues such as bones and teeth are the most common sources of ancient DNA because of their physical durability and protective internal microenvironments, which reduce microbial infiltration and buffer chemical fluctuations more effectively than soft tissues (Adler, Christina J., et al, 2011). Finally, contamination is a major concern in ancient DNA studies. aDNA extracts often contain less than 1% endogenous DNA, with the majority comprising environmental DNA from microbes or modern human contaminants (Pinhasi, R., et al., 2015). This extremely low proportion of authentic DNA, combined with its degraded state, poses significant challenges for accurate sequencing, mapping, and downstream analyses as well as dramatically increases the cost and efforts of DNA sequencing

1.2 Advancements in Sequencing Techniques for Ancient DNA

Advancements in molecular techniques have significantly improved the recovery and analysis of ancient DNA. The first and crucial step in any aDNA study is the extraction process, which serves as the gateway to maximizing DNA yield from often limited and degraded material. Generally, aDNA extraction involves several key stages: sample homogenization or powdering, lysis/digestion, purification, and elution. Among these, purification is particularly critical, as it largely determines both the efficiency of DNA recovery and the removal of contaminants (Lan, T., & Lindqvist, C., 2019). Following extraction, DNA is prepared for sequencing. While early ancient DNA studies relied on PCR amplification followed by Sanger sequencing, modern research now predominantly employs next-generation sequencing (NGS) technologies, which allow for higher throughput and improved resolution of degraded DNA fragments (Hofreiter, M., et al., 2015).

In general, all currently used next-generation sequencing (NGS) platforms are based on sequencing-by-synthesis technology and require the preparation of sequencing libraries. Library preparation involves the conversion of extracted DNA into a format that adheres to the sequencing flow cell. In the context of ancient DNA, this step poses unique challenges due to three major limiting factors: the extremely low copy number of endogenous DNA, the frequent presence of PCR inhibitors, and the abundance of damaged bases (Lan, T., & Lindqvist, C., 2019). These issues can significantly reduce the efficiency of library construction and downstream sequencing success. To address these limitations, specialized protocols have been developed, such as the single-stranded DNA (ssDNA) library preparation method. This technique is specifically optimized for highly fragmented and chemically damaged DNA and has proven to significantly increase the yield and authenticity of retrievable ancient DNA sequences. One key reason for its efficiency lies in the initial biotinylation of DNA molecules, which are then bound to streptavidin-coated magnetic beads. This setup ensures that the DNA remains immobilized throughout all reaction steps, minimizing loss during purification. Another advantage is the retention of fragmented DNA. During ssDNA preparation, molecules with nicks or breaks on both strands are heat-denatured into single-stranded fragments. Each of these fragments can then be individually recovered and incorporated into the library. Moreover, because ssDNA protocols process each strand independently, damage on one strand does not prevent the complementary strand from being utilized (Gansauge, M. T., & Meyer, M., 2013).

On the other hand, double-stranded (dsDNA) libraries are also commonly used for ancient DNA (aDNA) studies, with one of the most widely adopted protocols being that of Meyer, M., & Kircher, M., 2010. It begins with blunt-end repair, where in enzymatic reactions overhanging ends are converted into blunt ones. This is followed by adapter ligation, and a nicks fill-in step to create continuous DNA strands, which is crucial for stabilizing fragmented and damaged aDNA molecules. Lastly, indexes (barcodes) are

added to one of the adapters, unlike other methods where they are attached directly to template ends. This setup allows greater flexibility, as libraries are first built with universal adapters, and indexes can be added later during PCR with tailed primers. This method has been extensively adapted in aDNA research and demonstrates good efficiency for library construction (Rasmussen, Morten, et al., 2010; Orlando, L., et al, 2013; Raghavan, M., et al., 2014; Mascher, M., et al., 2016).

Lastly, to address the challenge of low endogenous DNA content in ancient samples, a crucial strategy has been developed; targeted enrichment. These methods enable the selective capture of specific genomic regions of interest, such as mitochondrial DNA, exomes, or individual chromosomes (Lan, T., & Lindqvist, C., 2019). Targeted enrichment offers high sensitivity and adaptability, and has demonstrated considerable promise for enhancing the recovery and resolution of ancient DNA sequences, as well as making the in-depth study of ancient genomes affordable (Enk, J. M., et al., 2014).

1.3 Bioinformatic Challenges in Ancient DNA Analysis

1.3.1 Authentication of aDNA

The analysis of aDNA sequencing data presents a distinct set of challenges. First step in a bioinformatic workflows is typically authentication. As discussed earlier, aDNA is often highly fragmented and present in low quantities, and it is frequently contaminated with modern human and microbial DNA. Therefore, distinguishing endogenous aDNA from exogenous sources is essential (Lan, T., & Lindqvist, C., 2019). This process, known as data authentication, relies on characteristic damage patterns, particularly cytosine deamination, which serve as molecular signatures of ancient origin (for more details refer to *1.1 Molecular Properties and Preservation of Ancient DNA*). These damage patterns are modeled and quantified using dedicated software tools and statistical frameworks, such as *mapDamage*, which assess the frequency and distribution of nucleotide misincorporations across sequencing reads (Jónsson, H., et al., 2013). Additionally, mitochondrial genome (mitogenome) analysis is widely employed in aDNA research due to the high copy number of mitochondrial DNA. Tools such as *contamMix* and *Schmutzi* utilize mitochondrial data to estimate levels of contamination, providing further confidence in the authenticity of ancient sequences (Fu, Q., et al., 2013; Renaud, G., et al., 2015).

1.3.2 Reconstructing Genomic Sequences from Ancient DNA

Once endogenous ancient DNA has been authenticated, the reconstruction of its original sequence is required for all subsequent genomic analyses. In theory for this step, two approaches can be employed: de novo genome assembly, which involves assembling overlapping reads into longer contiguous sequences, or reference-based mapping, where sequencing reads are aligned to an existing reference genome. Due to the highly fragmented nature and generally low quality of ancient DNA, de novo assembly is rarely

feasible. Consequently, reference-based mapping has become the standard approach in ancient DNA studies (Hofreiter, M., et al., 2015). Reference-based mapping typically involves three major steps: pre-alignment data processing, mapping/alignment to a reference genome, and post-alignment processing. These steps are specifically adapted for ancient DNA (aDNA) to maximize the recovery of authentic sequences.

In the pre-alignment phase, sequencing adapters are trimmed and overlapping paired-end reads are merged. This step is important in aDNA analyses, due to its highly fragmented nature. Merging overlapping reads helps to reconstruct longer and more accurate sequences, which is especially beneficial in low-coverage regions (Orlando, L., et al., 2021).

For the alignment of ancient DNA (aDNA) reads, two of the most commonly used aligners are BWA and Bowtie 2 (Poullet, M., & Orlando, L. , 2020). When using BWA for ancient DNA, it is best practice to adjust specific parameters, seed regions and gap openings, to optimize mapping specificity and sensitivity (Schubert, M., et al., 2012). By default, BWA uses the first 32 bases of each read (the "seed") to identify candidate mapping locations. Within this region up to two mismatches are allowed, so the overall running time can be reduced. However, due to post-mortem damage, particularly cytosine deamination at the 5′ ends, ancient DNA often contains more than two mismatches per seed. This can cause potential true alignments to be discarded during seeding. Disabling the seed ensures that the entire read is considered during alignment, thereby preventing the potential exclusion of true alignments that might otherwise be missed during the seeding process (Schubert, M., et al., 2012). Another key modification involves permitting more than one gap opening per alignment. By default, BWA applies a high penalty for gap openings (score of 11) and allows only a single gap opening event per alignment. A gap opening in an alignment corresponds to the introduction of an insertion or deletion (indel) to account for differences between the read and the reference sequence. This configuration is optimized for modern DNA sequenced on platforms like Illumina, which typically exhibit low indel rates. However, such settings are suboptimal for ancient DNA, where sequencing errors and post-mortem damage can lead to higher frequencies of indels. Adjusting BWA to tolerate up to two indel events has been shown to substantially improve mapping efficiency (Schubert, M., et al., 2012). Although adjusting both parameters enhances damage-aware and optimized mapping of aDNA, it comes at the cost of increased computational time. However, given the substantial gains in mapping accuracy, this trade-off is widely accepted and has become a best-practice standard in the field of ancient DNA research.

Post-alignment processing steps are often tailored to the specific aims of a study. However, two commonly applied procedures include the removal of duplicate reads and quality-based read filtering. During library preparation, PCR amplification can introduce duplicate reads, artificial copies of the same original molecule, which do not reflect true biological signals. If left unaddressed, these duplicates can lead to overrepresentation of certain sequences, potentially biasing downstream analyses such as variant calling or

gene expression profiling (Oliva, A., et al., 2021). To mitigate this, duplicate reads are typically identified and removed. Additionally, to ensure that only high-confidence data are retained for further analysis, it is standard practice to filter for reads above a defined mapping quality threshold (eg. phred score of 30)(Poullet, M., & Orlando, L. , 2020).

1.3.3 Analytical Frameworks for Ancient DNA Interpretation

Following the reconstruction of ancient genomes, downstream analyses are crucial for interpreting the evolutionary and biological significance of the data. One of the primary steps is the identification of genetic variants, such as single nucleotide polymorphisms (SNPs). Variant calling allows researchers to investigate how ancient genetic variation has contributed to modern human traits and diseases (Lan, T., & Lindqvist, C., 2019). For example, variants associated with immune responses, metabolic disorders, or susceptibility to pathogens can be traced through time to understand their evolutionary trajectories (Kerner, G., Choin, J., & Quintana-Murci, L., 2023). Tools commonly used for variant calling in aDNA datasets include GATK (Genome Analysis Toolkit) and SAMtools (McKenna, A., et al., 2010; Li, H. & Durbin, R., 2009).

Beyond individual-level variant identification, population genetic analyses aim to reconstruct demographic history and population structure (Childebayeva, A., & Zavala, E. I., 2023). These methods help infer past events such as population bottlenecks, expansions, migrations, and admixture. Statistical frameworks, such as those based on Bayesian inference or maximum likelihood estimation (e.g. for software tool ADMIXTURE), are employed to model population ancestry components (Alexander, D. H., Novembre, J., & Lange, K., 2009). Methods based on f-statistics, including qpAdm and qpWave, provide further resolution in testing specific admixture scenarios and evaluating relationships among ancient and modern populations (Patterson, N., et al., 2012). Principal component analysis (PCA) is another commonly used approach for summarizing genetic variation and visualizing population structure. It allows researchers to assess how ancient individuals relate to one another and to present-day populations in multidimensional genetic space(Poullet, M., & Orlando, L. , 2020).

In addition to these major analyses, several other downstream applications provide valuable insights. Kinship analysis reveals familial relationships within burial sites or communities, shedding light on social structures and inheritance patterns. Molecular sex determination enables the identification of biological sex from skeletal remains, which is essential for both demographic reconstructions and contextual archaeological interpretations. Microbial profiling of aDNA datasets can also offer information about the ancient microbiome, pathogens present at the time of death, and overall health and environmental exposures (Poullet, M., & Orlando, L. , 2020). Recently it has been shown that shared genomic segments, known as identity-by-descent (IBD) segments, can bto infer fine-scale genealogical connections of past human populations (Ringbauer H. et al., 2024).

Altogether, these downstream analytical steps transform ancient DNA sequences into meaningful narratives about human evolution, health, and population history, making them central to aDNA research.

1.4 Application of Pangenome Graphs in Ancient DNA Analyses

Ancient genomes are typically reconstructed using reference-based methods, in which sequencing reads are aligned to a known reference genome. The reference genome thus serves as a ground truth, guiding the reconstruction of the ancient genome. This approach uses a linear haploid reference genome, which is an assembled sequence derived either from a single individual or a mosaic of several individuals. As such, each genomic position in this reference represents only one allele. In reality, many genomic sites harbor multiple possible alleles, which can differ between individuals, populations, or even be associated with certain phenotypes. However, the linear reference captures only one of these allelic possibilities, often arbitrarily or based on the population used to generate the reference. As a result, when reads are mapped to the reference genome, any non-reference allele is treated as a mismatch, leading to a lower mapping score for those reads. This effect becomes more pronounced with increasing genetic distance between the sample and the reference, particularly in the case of ancient DNA, and even more acute when these sequences are from non-European individuals. Consequently, there is an inherent bias toward the allele represented in the reference. This phenomenon is known as reference bias (Günther, T., & Nettelblad, C., 2019).

Currently, there is no fully effective method to overcome reference bias in paleogenomics. As discussed in chapter *1.3.2 Reconstructing Genomic Sequences from Ancient DNA* mapping of aDNA is typically performed using adjusted alignment parameters. While these modifications improve alignment quality and facilitate the recovery of higher number of reads, they do not resolve the problem of reference bias. As a result, non-reference alleles remain systematically underrepresented in ancient DNA datasets (Rubin, J., et al., 2024).

However, one promising, yet still underexplored, approach is the use of pangenome graphs, which have shown potential in improving aDNA mapping and reducing reference bias (Martiniano, R., et al., 2020). Unlike linear haploid references that represent only a single allele per genomic position, pangenome graphs are designed to incorporate multiple alternative alleles at a given site by encoding genetic variation directly within the graph structure (Eizenga, J. M., et al., 2020). This allows for more inclusive and representative mapping of ancient sequences, particularly in cases where the sequenced individual diverges substantially from the reference. As such, pangenome-based methods offer a compelling avenue for addressing key challenges in the reconstruction of ancient genomes, and their potential was explored in this thesis.

1.5 The Rising Importance of Pangenome Approaches

The pangenome represents the complete set of DNA sequences found within a species. While the first pangenomes were developed for relatively small and easy-to-sequence bacterial genomes, the concept is much newer in eukaryotic and especially human genomics. Until recently, nearly all genomic research has relied on a single linear reference genome, serving as the mapping backbone and basis for most genomic analyses today. However, the rise of high-throughput sequencing technologies has led to an exponential increase in genomic data. The increasing availability of sequenced genomes has made it clear that a single reference genome cannot fully represent the genetic diversity within a species.To better capture this variation, the concept of a pangenome has emerged as a powerful alternative to the traditional reference genome (Sherman, R. M., & Salzberg, S. L. , 2020).

1.5.1 Limitations of the Linear Reference Genome

The current human reference genome, GRCh38.p13, is a linear reference, meaning it is organized as a continuous string of nucleotides across chromosomes. Although it incorporates genomic data from more than 20 individuals, approximately 70% of the sequence originates from a single individual. Despite its foundational role in genomics, this reference still contains errors, includes rare structural variants not representative of the general population, and has gaps in regions that are highly repetitive or polymorphic and therefore difficult to assemble (Wang, T., et al., 2022).

The limitations of the current human reference genome became increasingly evident with the rise of large-scale population sequencing projects, such as the 1000 Genomes Project (1kGP). Findings from 1kGP highlighted the extent of population-specific genetic variation, showing that 86% of the discovered variants were exclusive to a single continental group across five continents and 26 populations (1000 Genomes Project Consortium, 2015). This underscores the fact that the existing reference genome fails to adequately represent global human diversity. Furthermore, another study analyzing genomic sequences from 19 diverse populations estimated that up to 10% of the human genome, approximately 296.5 million base pairs, is missing from the current reference genome (Sherman, R. M., et al., 2019). These findings collectively reveal not only a lack of allelic diversity but also the absence of substantial and biologically relevant sequences in the current reference framework.

Another major limitation of the linear reference genome is its inability to accurately detect structural variants (SVs). Structural variants are variants longer than 50 base pairs, and they include deletions, insertions, duplications, and inversions. SVs are considered among the most variable forms of genomic variation. SVs can be highly individual-specific and exhibit considerable variation in size across different human genomes (Sudmant, P. H., et al., 2015). Accurate detection of structural variants (SVs) is essential, as many of them carry clinical significance and have important implications

for both genomic and clinical research. However, current sequencing practices, primarily based on short-read mapping to a linear reference genome, face major challenges in identifying SVs. Short reads often fail to span long or repetitive genomic regions, leading to mapping ambiguities, especially where multiple repetitive elements are identical or overlap. This can make it impossible to determine the correct genomic location of the reads. This is further supported by studies showing that up to 70% of structural variants may be missed in traditional whole-genome sequencing approaches that rely on short-read mapping to a linear reference (Wang, T., et al., 2022).

Known genomic variation has long been cataloged, leading to the development of large public databases, as for example dbSNP and ClinVar (Sherry, S. T., et al., 2001; Landrum, M. J., et al., 2014). Some of this variation has been incorporated into GRCh38 as alternative loci, but this inclusion is largely limited to single nucleotide polymorphisms (SNPs) and does not adequately account for structural variants (Sherman, R. M., et al., 2019). Moreover, this approach remains a suboptimal solution, especially as sequencing across diverse global populations continues to reveal novel sequences missing from GRCh38. These findings point to significant gaps in the current reference genome and suggest that a linear model is no longer sufficient for representing human genomic diversity. As with any foundational technology, the reference genome must evolve. This is where the pangenome emerges as a promising alternative, able to capture the full spectrum of human genetic diversity (Wang, T., et al., 2022).

1.5.2 Graph-based Pangenome Models

A pangenome model is a data structure that represents the complete set of genomic sequences within a defined group, such as a population, species, clade, or even a metagenome. This structure enables the representation of relationships among sequences, allowing for a more comprehensive description of genomic diversity. While "pangenome model" is an umbrella term encompassing various forms, this thesis specifically focuses on graph-based pangenome models (Eizenga, J. M., et al., 2020).

Graph-based pangenome models are often implemented as sequence graphs, a well-established structure for representing large sets of genomic sequences in a compact and efficient manner (Hein, J., 1989). Like other graph-based data structures, sequence graphs consist of nodes and edges. In this context, nodes are labeled with DNA sequences, while edges define the possible connections between nodes, representing how sequences can be combined or varied. This structure captures genomic variation by allowing for multiple alternative paths, known as walks, through the graph. Each walk corresponds to one of the original sequences used to construct the graph. Typically, sequence graphs are bidirectional, allowing for representation of both strands of DNA. Notable examples of sequence graphs in bioinformatics include de Bruijn graphs, widely used in *de novo* genome assembly, and variation graphs, which represent all observed variants in a population s (Eizenga, J. M., et al., 2020).

Variation graphs are a type of sequence graph that include not only nodes and edges but also a set of paths, each representing a possible sequence from a population (Garrison, E., et al., 2018). A key advantage of variation graphs is their ability to represent genetic variants by adding them as additional nodes between shared sequences. This makes it possible to represent complex variations that linear references cannot capture. For example, nested variants, such as SNPs inside insertions, or multiple versions of an insertion that overlap partially can be represented accurately. As a result, variation graphs offer a more complete representation of complex variants, highly variable regions, and divergent repeat sequences than any current linear reference can(Sherman, R. M., & Salzberg, S. L. , 2020). Because of these benefits, variation graphs were chosen as the foundation for building the human pangenome (more on this in *1.7 Draft Human Pangenome*)(Liao, Wen-Wei, et al., 2023). However, the increased complexity of variation graphs introduces several technical challenges, including the development of efficient graph-based alignment algorithms, scalable data storage solutions, and standardized formats and tools for graph construction, annotation, and analysis.

1.6 Technical Solutions for Pangenome Graphs

1.6.1 Constructing Pangenome Graphs

The pioneering construction of pangenome graphs began with the vg toolkit, which introduced the idea of building variant graphs directly from VCF files (Eizenga, J. M., et al., 2020). However, this method depends on pre-called variants derived from genome assemblies. With the rise of high-throughput whole-genome assemblies, a new idea emerged; to construct pangenome graphs directly from genome assemblies, entirely bypassing the variant calling step. This process is essentially equivalent to performing a whole-genome multiple alignment. Multiple genome alignment is a well-known and computationally challenging problem in bioinformatics. Due to its complexity, it typically requires heuristic algorithms. These algorithmic solutions scale with both the number of input genomes and their combined sequence length. To manage this, the multiple alignments are usually decomposed into smaller sub-alignments, which are later combined. Additionally, seed-and-extend heuristics are applied to efficiently align sequences across large and complex regions (Hickey, G., et al., 2024).

There are several strategies currently under development for constructing pangenome graphs, and no clear consensus yet exists on which is best. However, each method offers a unique perspective on the variations within a single pangenome graph. Minigraph is a tool that uses an iterative sequence-to-graph mapping approach. For the human genome, it begins with the GRCh38 reference and progressively adds new assemblies, recording only SVs of $\geq$50 base pairs (Li, H., Feng, X., & Chu, C., 2020). This makes it effective at capturing complex variants such as duplications, insertions, and inversions. An extension of this method, called Minigraph-Cactus, enhances the original approach by integrating homology relationships between assemblies using the Progressive Cactus genome aligner, resulting in a richer and more accurate graph (Hickey, G., et al., 2024;

Armstrong, J., et al., 2020). Another method, the PanGenome Graph Builder (PGGB), constructs pangenomes via all-to-all alignment of assemblies. This generates reference-free and unbiased pangenome graphs (Garrison, E., et al., 2024). The Human Pangenome Reference Consortium (HPRC) has used all three pipelines—Minigraph, Minigraph-Cactus, and PGGB—to create complementary views of the human pangenome (Liao, Wen-Wei, et al., 2023). In this thesis, I focus on the Minigraph-Cactus version of the human pangenome graph, as it provides a well-balanced trade-off between accuracy and computational efficiency for my research objectives.

1.6.2 Storing and Interpreting Pangenome Graphs

Although pangenome graphs can be constructed using various pipelines and alignment strategies, a standardized data model for their storage has been established; the Graphical Fragment Assembly (GFA) format. This standard simplifies work in the field of graphical pangenomics by providing a consistent way to represent graph structures. In this format, segments correspond to nodes in the graph, while lines represent the edges that connect them (Li, H., 2016). An extended version of this format, known as reference GFA (rGFA), enables the translation between linear and graph-based coordinate systems. It introduces three additional tags: SN, which indicates the name of the stable sequence (such as chr1); SO, which specifies the position or offset on the stable sequence; and SR, which defines the rank (rank 0 for reference sequences; rank > 0 for non-reference paths). The rGFA format also includes paths, or walks, which describe valid combinations of segments and represent possible genomic sequences within the graph (Li, H., Feng, X., & Chu, C., 2020). In parallel with these developments, the format to standardize how alignments are stored in graph-based contexts has been introduced; Graphical Fragment Assembly (GFA) (Li, H., Feng, X., & Chu, C., 2020). By enabling coordinate translation between graph and linear references, this standardization makes it possible to apply existing bioinformatics tools such as SAMtools, BAMtools, and GATK to graph-based alignments. Together, these formats provide a practical and interoperable framework for working with pangenome graphs.

1.6.3 Aligning to Pangenome Graphs

The shift from linear to graph-based reference genomes in pangenomics has created a need for new, efficient alignment algorithms, as traditional methods like Smith–Waterman are not suited for genome graphs (Eizenga, J. M., et al., 2020). The main challenge in read mapping to pangenome graphs is the expanded alignment search space. The complexity arises from the presence of numerous paths in the graph, representing many possible sequence combinations that need to be considered during alignment. As a result, mapping becomes computationally intensive and often acts as a bottleneck in genomic analysis (Sirén, J. et al., 2021). To address this, different mappers have been developed, each with its own limitations. Some tools are fast but restricted to directed acyclic graphs and low-variation settings, where variants are added to a linear backbone (Kim, D., et al., 2019). Others can handle more complex graph structures but

are limited to long-read sequencing data (Rautiainen, M., & Marschall, T., 2020). Currently, Giraffe is an optimal solution for aligning short reads to pangenome graphs, given the limitations and trade-offs of existing mapping tools (Sirén, J. et al., 2021).

Giraffe is a short-read mapper designed to efficiently address the large alignment search space posed by pangenome graphs. To achieve this, Giraffe limits the search to paths observed in actual individuals' genomes, known as reference haplotypes. It traces two haplotypes, one from each parent, through the reference graph. This approach significantly speeds up the mapping process and allows Giraffe to achieve performance comparable to traditional linear mappers. Giraffe follows a seed-and-extend strategy. It first identifies short exact matches between the sequencing reads and the reference (called *seeds*), then extends high-quality seeds into full-read alignments. The algorithm uses several heuristics to prioritize and refine alignments effectively. Beyond its speed, Giraffe reduces reference bias compared to linear mappers. This effect is especially pronounced for longer insertions. Its accuracy was demonstrated by genotyping 167,000 structural variants, previously discovered in long-read studies, across 5,202 diverse human genomes sequenced with short reads (Sirén, J. et al., 2021). Given its characteristics and advantages, Giraffe was used to demonstrate the range of applications enabled by the newly developed human pangenome reference (Liao, Wen-Wei, et al., 2023). It also served as the primary mapping tool throughout this thesis.

1.7 Draf Human Pangenome

The Human Pangenome Reference Consortium has created the first draft of the human pangenome reference, which includes 47 phased, diploid assemblies from a genetically diverse cohort. These individuals represent populations from nearly every continent, including Africa, Asia, the Americas, and Europe, aiming to capture the full spectrum of human genetic diversity. However there is a particular focus on populations underrepresented in the current reference genome, GRCh38. The draft pangenome was shown to capture novel sequences absent from GRCh38. These novel sequences include 119 million base pairs of euchromatic polymorphic regions and 1,115 gene duplications. Additionally, about 90 million of extra base pairs are derived from structural variations.
Beyond adding novel sequences, the draft pangenome significantly improves variant detection. The errors in discovering small variants from short-read data were reduced by 34%, while structural variant detection increased by 104% per haplotype compared to GRCh38-based workflows (Liao, Wen-Wei, et al., 2023). These improvements make the human draft pangenome a promising candidate to replace the traditional linear reference genome in genomics research.

The terms "human draft pangenome" and "human pangenome" are used interchangeably throughout this thesis.

1.8 The children of Llullaillaco

The Inca Empire (called *Tawantinsuyu*, eng. The FourParts Together) was the largest and most powerful prehispanic empire in the Americas with its territory encompassing large parts of present-day Colombia, Ecuador, Peru, Bolivia, Chile, Argentina. At its height in the 15th and early 16th centuries, the empire was renowned for its highly organized social structure, efficient administrative system, and remarkable architectural feats. Among their most extraordinary accomplishments was the construction of stone structures on nearly 100 Andean mountains, each over 5,000 meters high and within a span of just 60 years (*ca.* 1470–1532 CE)(Besom, T., 2009; D'Altroy, 2015).

Such high mountains held a special place in Inca culture. They were sacred sites where important ceremonies took place, including human sacrifices. As a result, several ice mummies have been discovered at high altitudes on Andean peaks, including El Plomo in Chile, Mount El Toro on the Chile–Argentina border, and the lower slopes of Aconcagua in Argentina(Reinhard & Ceruti, 2010; González Holguín, 1993).

Andean mummies exhibit distinctive features. They typically became frozen at the time of death, rather than later, after decomposition had begun. Several environmental factors contributed to their remarkable preservation. The consistently below-freezing temperatures at the high altitudes of the Andes, along with volcanic ash in some areas, that inhibited bacterial growth and helped retain moisture (Reinhard & Ceruti, 2010). This unique combination of conditions makes Andean summits ideal for preserving organic material. This is exemplified by the archaeological site at the summit of Llullaillaco (6,715 meters), where the remains of three individuals were discovered. These mummies are among the best-preserved in the world, and the Llullaillaco site holds the distinction of being the highest archaeological site on Earth (Reinhard, & Ceruti, 2010) .

The individuals found near the summit of the Llullaillaco volcano, probably died in a sacrificial ritual known as *capacocha*. The capacocha ritual was a significant Inca ceremony performed to honor the gods of the Inca Empire (Reinhard & Ceruti, 2010; González Holguín, 1993). The ritual involved offering children and *acllas*, (*"chosen women"*) who were believed to be the purest and most perfect individuals. Selected children were brought to the Inca capital Cuzco, where they lived in seclusion under strict ceremonial conditions and were placed on a regulated diet to prepare them for upcoming sacrificial rites (Clinker, 2022). *Capacocha* ceremonies were primarily conducted at sacred shrines and on high mountain summits.

In this master's thesis, the Llullaillaco mummies serve as the underlying dataset. A detailed description of the burial context can be found in *2.2 Individuals included in the study* in Methods section.

1.9 Project overview

Ancient DNA is a unique form of genetic material characterized by specific damage patterns, including DNA fragmentation and post-mortem chemical modifications that often lead to nucleotide substitutions. These molecular alterations pose significant challenges for sequencing and analysis. One of the central issues in ancient DNA analysis is the increased divergence between the degraded DNA fragments and the modern reference genome. This divergence, driven both by molecular damage and the evolutionary distance between ancient and reference genomes, leads to reference bias; a phenomenon where alleles matching the linear reference are preferentially mapped, masking true genetic diversity present in the sample.

This bias distorts our understanding of human genetic variation, obscuring evolutionary patterns and hindering the detection of population-specific or rare variants that could be critical for reconstructing human history. To address this, an emerging and promising approach is the use of graph-based references, called pangenomes. Pangenomes can represent the full genetic diversity of a species within a single, non-linear structure.

Such pangenome references are already established in microbial genomics and are increasingly gaining traction in human genomics. A major milestone in this field is the Human Pangenome Reference Consortium release of the first draft human pangenome. This reference incorporates diploid assemblies from genetically diverse individuals across multiple continents, offering a more inclusive and accurate representation of human genetic variation.

This project seeks to assess human draft pangenome advantages over the linear reference for ancient DNA studies. The analysis will focus on identifying specific genomic loci,such as genes or regulatory regions, where mapping quality and variant discovery are significantly improved. In addition, the study will classify the genomic features (e.g., exons, introns, intergenic regions) where these improvements are most pronounced. Particular attention will be paid to the detection of clinically relevant pathogenic variants and the discovery of novel SNPs and insertions/deletions (indels) that may be overlooked when using the linear reference.

Ultimately, this work aims to highlight the value of using pangenome-based references in aDNA research, contributing to a more accurate reconstruction of human genetic diversity and enhancing our understanding of past populations and their evolutionary trajectories.

Although ancient DNA is generally characterized by high damage levels, it is important to note that the Llullaillaco mummies represent exceptionally well-preserved samples. Despite the relatively short read lengths typical for ancient DNA, these individuals show comparatively low levels of post-mortem damage, making them a particularly reliable dataset for exploring the benefits of pangenome-based analyses.

2 Methods

This Methods section is divided into three parts: Individuals included in the study, Laboratory Methods, and In Silico Analyses. While all parts are relevant to the thesis and provide important contextual and methodological background, my personal contribution focused on the In Silico Analyses. The inclusion of the other two sections ensures a comprehensive overview of the entire workflow.

2.1 Key resources table

2.1.1 Software and algorithms

Cutadapt 5.0	Martin, M. ,2011	https://cutadapt.readthedocs.io/en/stable/
Pear 0.9.1	Zhang,J. , et al. ,2014	https://github.com/xflouris/PEAR
FASTQC 0.12.1	Andrews, S. , 2010	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
BWA 0.7.19	Li, H., & Durbin, R. , 2009	http://bio-bwa.sourceforge/
KMC 3.2.4	Kokot, M., Długosz, M., & Deorowicz, S. , 2017	https://github.com/refresh-bio/KMC
vg toolkit 1.63.1	Garrison, E. et al., 2018	https://github.com/vgteam/vg
SAMtools 1.21	Li, H. & Durbin, R., 2009	http://www.htslib.org/doc/1.9/samtools.html
GATK 4.6.1.0/3.8.1.0	McKenna, A., et al., 2010	https://github.com/broadinstitute/gatk
BEDtools 2.31.1	Quinlan, A. R., and Ira M. Hall., 2010	https://github.com/arq5x/bedtools2
ABRA2 2.23	Mose, L. E., et al., 2014	https://github.com/mozack/abra2
Picard Tools 3.3.0	Picard-Tools	https://broadinstitute.github.io/picard/
BCFtools 1.21	Li, H. & Durbin, R., 2009	https://github.com/samtools/bcftools
R 4.4.3	R Core Team	https://cran.r-project.org/
OptiType 1.3.3	Szolek, A., et al., 2014	https://github.com/FRED-2/OptiType

2.1.2 Alignments

Genome Reference Consortium. GRCh38.p13 (GCA_000001405.28)	Schneider, V.A., et al. , 2017	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39
Human Pangenome (HPRC v1.1, Minigraph-Cactus graph)	Liao, Wen-Wei, et al., 2023	https://github.com/ComparativeGenomicsToolkit/cactus/blob/master/doc/mc-pangenomes/hprc-v1.1-mc.md

2.2 Individuals included in the study

In this study, the genomes of three individuals discovered near the summit of Llullaillaco volcano were examined. These mummies are among the best-preserved naturally mummified remains known. Two female individuals (approximately 14 and 6 years old) and one male individual (approximately 7 years old) were discovered in 1999. (Rowe, J. H., 1946). Their burial displayed characteristic features of capacocha rituals, similar to those observed at other sites such as El Plomo, Ampato, Misti and Aconcagua (Schobinger, J., 2001; Castro, M. et al., 2005).

La Doncella

La Doncella (*"The Maiden"*), the oldest female, had a simpler burial than the others. She wore traditional Inca clothing; a brown dress, a shawl with gold pins, leather moccasins, and a white feather headdress. Her hair was braided, her face painted red, and she wore bone and metal ornaments. Offerings included pottery, wooden items, food bags, wool belts, and figurines made of gold, silver, and Spondylus shell (Rowe, J. H. , 1946).

La Niña del Rayo

La Niña del Rayo (*"The lightning girl"*), the younger girl, was buried on the eastern side of the platform. Her body was partly burned by lightning, leaving a hole in her chest. She had an artificially shaped head (Previgliano, C. H., et al., 2003) and wore a dress, shawl with tupu pins, and leather moccasins. Two silver plaques were on her forehead. Offerings included four female figurines (silver, gold, and Spondylus), ceramics, wooden vessels, food and feather bags, sandals, and bags with hair.

El Niño

El Niño (*"The boy"*) , the boy sacrifice, had an artificially shaped head (Previgliano, C. H., et al., 2003). He wore a red mantle and tunic, leather moccasins, a silver bracelet, and fur anklets. A sling with white feathers was tied around his head. His grave included

slings, sandals, a black tunic, a bag with coca leaves, pottery, and two skin bags with hair. Two Spondylus figurines (a llama and a male) were also found.

All the victims were buried in sitting positions (Rowe, J. H. , 1946).

Since 2007 the mummies have been on exhibition in the Museum of High Altitude Archaeology (MAAM) in the Argentine city of Salta.

2.3 Laboratory Methods

2.3.1 Sampling

Sampling took place in a restricted, clean room at Museo de Arqueologia de alta Montana (MAAM) in Salta (Argentina) designed for mummy preservation. Access was controlled, and the setup mimicked a clean aDNA lab. Each mummy could only be outside its compartment for under 10 minutes. This influenced the sampling strategy, which prioritized the use of cotton swabs due to their time efficiency. However, skin and hair samples were also collected. All three mummies were sampled at multiple body locations using a combination of all three methods (see Supplementary Table 2 and 3). Sampling was done in one day under strict contamination controls. Samples were stored in a freezer at the MAAM and transported to Vienna under cold conditions.

The entire process was carried out with permission from the Government in Salta (Disposición interna N° 016/2021), and the Argentinian National Government (Disposición número DI-2022-44-APN-INAYPL#MC) and with approval from the ethics committee of the University of Vienna (Reference number 01004).

2.3.2 DNA extraction, library preparation, and sequencing

All laboratory work took place in the specialized aDNA laboratory of the University of Vienna, following stringent contamination prevention measures, which included wearing suitable protective clothing such as hair nets, masks, and a full-body suit.

Both skin and swab samples were obtained using aDNA-optimized methods as described by Dabney et al. (2013). Rather than using 50 mg of powdered material, the cotton swabs were placed in 1 mL of extraction buffer (0.5 M EDTA, 0.25 mg/mL Proteinase K, and dH2O) and allowed to incubate for 18 hours at 37°C while being continuously shaken at 1200 rpm. Hair samples were cut into 2-3 cm segments, treated with 10 mL of a 0.5% bleach solution, and rinsed three times with ddH2O. Afterward, they were dried and transferred to an active reaction buffer (H2O, 1M Tris, 5M NaCl, 1M

CaCl2, 0.5M EDTA, 20% SDS, 1M DTT, 15 mg/mL Proteinase K). These samples were then incubated overnight at 55°C and shaken at 1200 rpm.

A 25 µL portion of each extract was used to build double-stranded sequencing libraries following Meyer, M., & Kircher, M., 2010, without applying any methods that might modify DNA fragment length. For clean-up, the Qiagen MinElute PCR Purification kit was used instead of the originally recommended SPRI beads. The purified libraries were diluted in EBT buffer to a final volume of 40 µL. Real-time PCR was then performed to determine the optimal number of indexing PCR cycles and avoid over-amplification
Library preparation included double indexing and amplification with NebNext Q5U DNA Polymerase (Agilent) and IS5/IS6 primers. 1.2x NGS magnetic beads were used for clean-up, followed by elution of the samples in 25 µL of EBT buffer.

Upon library preparation, shallow sequencing was conducted at the Vienna BioCenter Core Facility. Sequencing was performed on an Illumina NovaSeq instrument using 100 cycles.

Additional libraries were prepared for samples exhibiting high endogenous DNA content and low levels of modern DNA contamination, including: NiñaDelRayo_Hair_SHALLOW_SEQ, Niño_Hair_SHALLOW_SEQ, Doncella_Hair_SHALLOW_SEQ, NiñaDelRayo_Arm_SHALLOW_SEQ, Doncella_Mouth_SHALLOW_SEQ, NiñaDelRayo_Mouth_SHALLOW_SEQ, NiñaDelRayo_Ear_SHALLOW_SEQ, NiñaDelRayo_Anus_SHALLOW_SEQ, Niño_Anus_SHALLOW_SEQ, and Doncella_Anus_SHALLOW_SEQ. These libraries were sequenced using 150-cycle runs on an Illumina NextSeq500 platform at the Polo GGB - Polo d'Innovazione di Genomica, Genetica e Biologia in Siena, Italy.

2.4 In Silico Analyses

2.4.1 Preprocessing and Quality Control of Sequencing Reads

We obtained both single- and paired-end sequencing data. Sequencing quality was assessed using *FastQC* (version 0.12.1) (Andrews, S., 2010.). Adapter trimming of raw reads was subsequently performed with *Cutadapt* (version 5.0) (Martin, M. ,2011). A minimum read length of 10 bases (*-m 10*) was enforced during trimming, and standard Illumina adapter sequences were specified. For single-end libraries, only the forward adapter (*-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA*) was used. For paired-end libraries, adapter sequences for both forward and reverse reads were provided (*-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT)*. Additionally, for paired-end libraries, overlapping read pairs were collapsed using Pear (version 0.9.11) (Zhang, J. , et al.

,2014). The criteria for collapsing reads included a minimum overlap of 11 base pairs (*-v 11*) and a maximum assembled read length of 30 base pairs (*-n 30*).

2.4.2 Alignment to the Linear and Pangenome References

Preprocessed reads were aligned in parallel to two human reference assemblies: the linear GRCh38 reference genome (Genome Reference Consortium, GCA_000001405.28) and the draft human pangenome (HPRC version 1.1), represented as a graph constructed using the Minigraph-Cactus pipeline (Liao, Wen-Wei, et al., 2023).

Alignment to the Linear Reference

Alignment to GRCh38 was performed with BWA aln (version 0.7.19) (Li, H., & Durbin, R., 2009), disabling seeding by setting a very large seed length (*-l 556565*). Setting the seed length to such a high value ensures that the full read is considered during alignment, which can increase sensitivity, particularly for highly degraded or divergent sequences. A gap open penalty of 2 (*-o 2*) was used to penalize the introduction of gaps in the alignment, and the maximum allowed edit distance was set to 0.01 mismatches per base (*-n 0.01*), ensuring stringent alignment quality.

Alignment to the Pangenome Reference

Alignment to HPRC version 1.1, Minigraph-Cactus graph was performed in multiple steps:

1) Sample-specific pangenome reference

To improve alignment accuracy, a sample specific pangenome reference graph was constructed, following the approach described by Sirén, J., et al. (2024). This method leverages k-mer profiles derived from sequencing reads to identify haplotypes within the pangenome graph that are most similar to the sample sequence.

To generate k-mer profiles, trimmed reads were processed using *KMC* (version 3.2.4) (Kokot, M., Długosz, M., & Deorowicz, S. , 2017) with a k-mer size of 29 (*-k29*), and results were output in KFF format for compatibility with the *vg toolkit* (version 1.63.1) (Garrison, E. et al., 2018). These k-mer counts served as input to *vg haplotypes*, which was run using parameters optimized for diploid genome inference (*--diploid-sampling, --include-reference*). For each sample, the four most likely haplotypes were selected (*--num-haplotypes 4*) based on a haplotype information file from the Human Pangenome (*version 1.1;* file name: *hprc-v1.1-mc-grch38.hapl;* downloaded from: https://github.com/human-pangenomics/hpp_pangenome_resources?tab=readme-ov-file). The final output was a personalized haplotype graph in GBZ format, which was subsequently

indexed using *vg index* to generate a distance index and *vg minimizer* to construct a minimizer index, which are both required for mapping.

```
# Count k-mers from sequencing reads using KMC
kmc -k29 \
    -m2 \
    -okff \
    @reads_list.txt \                       # Input: list of raw FASTQ files (one per
line, can be gzipped)
    kmc_output/${sample}_kmers \            # Output prefix: will produce
${sample}_kmers.kff
    kmc_output/temp_kmc_${sample}           # Temporary directory for intermediate KMC
files

# Generate haplotype paths from a variation graph using VG toolkit
vg haplotypes -v 2 \
--num-haplotypes 4 \
--present-discount 0.9 \
--het-adjustment 0.05 \
--absent-score 0.8 \
--include-reference \
--diploid-sampling \
-i hprc-v1.1-mc-grch38.hapl \
-k kmc_output//${sample}_kmers.kff \
-g ${sample}_haplotypes.gbz\
hprc-v1.1-mc-grch38.gbz

# Generate distance index for the haplotypes graph
vg index  -j  ${sample}_haplotypes.dist    ${sample}_haplotypes.gbz

# Generate minimizer index for the haplotypes graph
vg minimizer -d  ${sample}_haplotypes.dist -o ${sample}_haplotypes.min
${sample}_haplotypes.gbz
```

2) Alignment to pangenome reference

Alignment to the sample specific pangenome reference was performed with Giraffe (part of vgtoolkit version 1.63.1) (Sirén, J. et al., 2021) using default parameters. The resulting alignments were stored in a graphical alignment file (GAF).

```
#Note: This code snippet is intended for paired-end reads. For single-end reads, use only
one #-f option and supply a single FASTQ file.
vg giraffe --progress \
        --sample "${sample}" \
        --output-format gaf \
        -f  ${fastq_read1} \
        -f   ${fastq_read1} \
        -Z  {sample}_haplotypes.gbz\
        -d  {sample}_haplotypes.dist \
        -m  {sample}_haplotypes.min  | gzip > {sample}_alignments.gaf.gz
```

3) Projection onto a linear reference

To obtain BAM files aligned to a linear coordinate system, the GAF files were projected onto a linear reference using *vg surject*. Reference paths used for projection are displayed in Supplement Table 1. Sample metadata was incorporated using the *--read-group* option (*"ID:1 LB:lib1 SM:${sample} PL:illumina PU:unit1"*), and low-complexity alignments were excluded using the *--prune-low-cplx* parameter.

The surjected alignments were then passed through a renaming script (*rename_bam_stream.py*, available at: https://github.com/jmonlong/docker-vg-work) to ensure compatibility with downstream tools. Next, indels were left-aligned using bamleftalign, and the output was coordinate-sorted with *SAMtools* sort (version 1.21) (Li, H. · Durbin, R., 2009).

```
# Surject GAF alignments onto the linear GRCh38 reference using VG, then left-align and
sort the resulting BAM
vg surject \
    -F hprc-v1.1-mc-grch38.paths_list.txt \                  # List of reference paths
(see Supplement Table 1)
    -x ${sample}_haplotypes.gbz \
    --sam-output \
    --gaf-input \
    --sample ${sample} \
    --read-group "ID:1 LB:lib1 SM:${sample} PL:illumina PU:unit1" \
    --prune-low-cplx ${sample}_alignments.gaf.gz \
| python3 rename_bam_stream.py -f hg38.fa.fai -p "GRCh38#0#" \ # Rename reference names
to GRCh38 format
| bamleftalign --fasta-reference hg38.fa --compressed \        # Left-align indels
| samtools sort -O BAM > ${sample}_surjected.bam               # Final output: sorted
surjected BAM
```

4) *Realignment*

To refine read alignments around indels, local realignment was performed. First, realignment target regions were identified using *GATK* (version 3.8.1.0) (McKenna, A., et al., 2010) with the *RealignerTargetCreator* module. This step was executed with disabling duplicate read filtering (*-drf DuplicateRead*) and BAM indexing (*--disable_bam_indexing*). The resulting interval file was transformed into BED format using awk, adjusting for 0-based coordinate conventions, and subsequently extended by 160 base pairs in both directions using *BEDtools* (version 2.31.1) (Quinlan, A. R., and Ira M. Hall., 2010) slope with the corresponding reference index file.

```
#Identify regions around potential indels for local realignment
java -Xmx16G -jar GenomeAnalysisTK.jar \
    -T RealignerTargetCreator \
    --remove_program_records \
    -drf DuplicateRead \
    --disable_bam_indexing \
```

```
    -R hg38.fa \
    -I ${sample}_surjected.bam \
    --out ${sample}.IndelRealign.intervals

# Convert GATK interval format to BED format (0-based, half-open)
awk -F '[:-]' 'BEGIN { OFS = "\t" } { if ($3 == "") { print $1, $2-1, $2 } else { print
$1, $2-1, $3 }}'
    ${sample}.IndelRealign.intervals > ${sample}.IndelRealign.bed

#Expand BED intervals by ±160 bp using bedtools
bedtools slop   -i ${sample}.IndelRealign.bed  -g hg38.fa.fai   -b 160 >
${sample}.IndelRealign_targets.bed
```

The resulting target BED file was then used as input for *ABRA2* (version 2.23) (Mose, L. E., et al., 2014) to perform the actual indel realignment. *ABRA2* was run with the default parameters, producing BAM files with locally realigned reads.

```
# Perform indel realignment using ABRA2
java -Xmx26G -jar abra2-2.23.jar \
    --targets ${sample}.IndelRealign_targets.bed \
    --in  ${sample}_surjected.bam \
    --out ${sample}_surjected_realn.bam \
    --ref hg38.fa
```

The multi-step alignment process to the HPRC reference followed the structure of the snakemake workflow developed for the vg toolkit, available at https://github.com/vgteam/vg_snakemake. This process is specifically tailored for sequencing reads longer than 40 base pairs. For shorter reads, the k-mer size used in the first step *1) Sample-specific pangenome reference* be adjusted according to the read length.

2.4.3 Postprocessing of Alignments

Reads with mapping qualities below 30 were removed using *SAMtools view* (version 1.21). Duplicates were removed by using *Picard Tools* (version 3.3.0) module *MarkDuplicates* with default parameters. This process was repeated in the same manner for both sets of the files, one mapped to the pangenome and the other mapped to the linear genome. For each resulting BAM file, the total number of mapped reads was determined using *SAMtools flagstat*. This number was then divided by the total number of reads in the corresponding raw FASTQ file to calculate the mapping rate for each file. These mapping rates are summarized in Supplementary Table 2 and 3 and visualized as box plots in Figure 2B in the Results section.

Postprocessed BAM files originating from different tissue types were merged into a single BAM file per individual using *SAMtools*, as illustrated in Figure 1B in the Results

section. Unique read groups were then added using *Picard Tools* (version 3.1.1) module *AddOrReplaceReadGroups* to ensure consistency and proper annotation within each merged file. Subsequently, the whole genome coverage was calculated with *SAMtools depth* and displayed in Figure 1D in the Results section. Whole genome coverages can be found in Supplement Table 4.

Identification and Classification of Unique and Shared Reads Between Pangenome and Linear Genome Alignments

This method was used to generate the data shown in Figure 2C of the Results section.

To identify reads uniquely mapped to the pangenome or the linear genome, as well as those shared between the two alignments, all read names were first extracted from the respective BAM files. This was done by converting the BAM files to SAM format using *SAMtools view*, and then extracting the first column, which contains the read names, using *cut -f1*. The resulting read names were saved into separate text files for each alignment.

These text files were then sorted, and the *comm* command was used to compare them line by line. This allowed the classification of reads into three categories: unique to the pangenome, unique to the linear genome, and shared between both alignments. The read names from each category were subsequently used to filter the original BAM files with *FilterSamReads* module from *Picard Tools* (version 3.1.1), generating new BAM files containing only the relevant reads. Lastly, venn diagram was generated in RStudio (version 4.4.3) using the VennDiagram package to visualize the overlap and exclusivity of reads across the two alignment types. Total counts of unique and shared reads per individual are reported in Supplementary Table 8 and visually summarized in Figure 2C.

```
# Extract read names from pangenome/linear genome alignment BAM
samtools view pangenome/linear_aligned.bam | cut -f1 | sort > pangenome/linear_reads.txt

# Unique to pangenome
comm -23 pangenome_reads.txt linear_reads.txt > reads_unique_pangenome.txt

# Unique to linear genome (lines only in second file)
comm -13 pangenome_reads.txt linear_reads.txt > reads_unique_linear.txt

# Shared reads (lines common to both files)
comm -12 pangenome_reads.txt linear_reads.txt > reads_shared.txt

# Filter BAM files to only keep reads unique to pangenome/linear genome or shared reads
Note: This code snippet is intended for reads unique to pangenome alignment. For reads
uniq to linear alignment and shared reads, repeat this command with correct files.
picard FilterSamReads \
    I=pangenome_aligned.bam \
    O=pangenome_unique.bam \
    READ_LIST_FILE=reads_unique_pangenome.txt \
    FILTER=includeReadList
```

To assess the distribution of sequencing reads across different genomic features, we analyzed BAM files derived from both linear and pangenome alignments for each individual sample. Using *BEDtools intersect*, we identified overlaps between aligned reads and predefined genomic intervals corresponding to exonic, intronic, and intergenic regions(see *2.4.9 Supporting Procedures* for details on how these annotations were created).

For each genomic region type, the number of uniquely overlapping reads was determined. The relative proportion of reads per region was then calculated based on the total number of mapped reads for each sample. These proportions were visualized using the *create_pie* function in R and are depicted in Figure 2C of the Results section.

```
Note: This code snippet is intended for reads unique to pangenome alignment. For reads
uniq to linear alignment, repeat this command with correct files.
    bedtools intersect -abam $pangenome_unique.bam -b exons_protein_coding.bed -wa -wb
-bed > $pangenome_exon.bed
    bedtools intersect -abam $pangenome_unique.bam -b introns_protein_coding.bed -wa -wb
-bed > $pangenome_intron.bed
    bedtools intersect -abam $pangenome_unique.bam -b intergenic.bed -wa -wb -bed >
$pangenome_intergenic.bed
```

2.4.5 Variant Calling

Variant calling was performed using the GATK (version 4.6.1.0). Module *HaplotypeCaller* was run in *gvcf* mode (*-ERC GVCF*) with default parameters. Subsequently, *GenotypeGVCFs* was used to convert GVCF files into a *VCF* file. Furthermore, heterozygous variants were filtered, keeping only variants with an allele ratio between 0.4 and 0.6. An additional depth-based filtering step was applied to retain only variants supported by sufficient read coverage, adjusted according to the whole-genome sequencing depth of each individual (see Figure 1D in Results section). Specifically, variants were retained if their read depth exceeded individual-specific thresholds: ≥5 for La Nina del Rayo, ≥7 for El Nino, and ≥10 for La Doncella. This procedure was applied for both pangenome and linear genome alignment in the exact same manner.

Identification of Unique and Shared Variants Between from Pangenome and Linear Genome Alignments

This method was used to generate the data shown in Figure 3A,B,C of the Results section.

To assess the concordance and uniqueness of variant calls generated from linear and pangenome-based alignments, we performed a pairwise intersection of the two VCF files using *BCFtools* (version 1.21) (Li, H. & Durbin, R., 2009) module *isec*. This analysis produced separate VCF files containing variants unique to the linear reference, variants unique to the pangenome reference, and variants shared between both references.

Summary statistics for each variant set were then computed using *BCFtools stats*. These statistics were subsequently visualized using the *VennDiagram* package in R. The results are presented in the Results section (Figure 3, Panels A,B,C).

```
#Index VCF files
bcftools index -f $pangenome_filtered_variants.vcf.gz >
pangenome_filtered_variants.gz.csi
bcftools index -f $linear_filtered_variants.vcf.gz > linear_filtered_variants.gz.csi

# Perform intersection using BCFtools isec
bcftools isec pangenome_filtered_variants.gz.csi linear_filtered_variants.gz.csi -p
$output_dir

#Obtain summary statistics
bcftools stats $output_dir/pangenome_uniq_varinats.vcf > $stats_output.txt
bcftools stats $output_dir/linear_uniq_varinats.vcf >>  $stats_output.txt
bcftools stats $output_dir/shared_varinats.vcf >>  $stats_output.txt
```

2.4.6 HLA Class 1 Typing

This method was used to generate the data shown in Figure 5C and D of the Results section.

Genotyping of HLA class I loci was performed using OptiType (version 1.3.3) (Szolek, A., et al., 2014), specifying the DNA sequence filter option (*--dna*) with default settings. Each individual had two separate alignments: one against the linear GRCh38 genome and another against the draft human pangenome (HPRC version 1.1, Minigraph-Cactus graph). For both alignments, reads uniquely mapped to each reference were extracted using the BEDtools module *bamtofastq* and saved in FASTQ format. The resulting FASTQ files were used as input for OptiType. The resulting genotypes, along with their supporting objective values and read coverages, are shown in Figure 5C and 5D in the Results section.

```
#Extract uniquely mapped reads
bedtools bamtofastq  -i pangenome/linear_aligned.bam -fq  pangenome/linear_aligned.fastq

#Run OptiType (installation required)
python OptiTypePipeline.py -i pangenome/linear_aligned.fastq  --dna -v -o ./output_dir
```

2.4.7 Identification of Clinically Relevant Pathogenic Variants

This method was used to generate the data shown in Figure 6 of the Results section.

To evaluate the clinical relevance of identified variants, all filtered VCF files (for more details refer to *2.4.5 Variant calling* section above) were annotated using the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/; Landrum, M. J., et al., 2014). *BCFtools* was used to perform the annotation with the most recent ClinVar VCF file (*downloaded*

date:2025_03_23). Following annotation, variants classified as "Pathogenic" (i.e., those containing the *CLNSIG=Pathogenic flag in the INFO field*) were extracted. The resulting pathogenic variant profiles are summarized in Figure 6.

```
#Annotation with BCFtools
bcftools annotate -a clinvar.vcf.gz -c INFO pangenome/linear_filtered_variants.vcf.gz -o
pangenome/linear_filtered_variants.annotated.vcf.gz

# Extract pathogenic variants
pathogenic_variantst=$(zcat pangenome/linear_filtered_variants.annotated.vcf.gz | grep
'CLNSIG=Pathogenic' )
```

2.4.8 Regression Analysis for Genome-Wide Comparison of Mapping Performance

This method was used to generate the data shown in Figure 4 A,B,C of the Results section.

To perform the regression line analysis, the genome was first divided into 1 Mb windows with a 10 kb overlap using *BEDtools makewindows* (*-w 1000000 -s 900000*). The chromosome size file for GRCh38 (available at the UCSC Genome Browser; https://hgdownload.soe.ucsc.edu/goldenpath/hg38/bigZips/) was used as reference. Subsequently, base coverage for each 1 Mb genomic window was calculated with *BEDtools coverage*. This process was performed for both genome alignments, resulting in two BED files with identical 1 Mb windows, but coverage values specific to each alignment. This pair of BED files was used to generate the regression plots shown in Figure 4A, B, and C. Plots were created in RStudio using the *ggplot2* package.

```
# Make 1 Mb windows with 100 kb overlap across GRCh38
bedtools makewindows -g hg38.chrom.sizes -w 1000000 -s 900000 >
1Mb_regions_with_100kb_overlap.bed

# Calculate coverage using the generated windows
bedtools coverage -sorted -a 1Mb_regions_with_100kb_overlap.bed -b
pangenome/linear_aligned.bam > coverage_pangenome/linear_1MbRegions_100kbOverlap.bed

# Pseudocode for plotting a regression line with ggplot2
library(ggplot2)
ggplot(df, aes(x = coverage_linear, y = coverage_pangenome)) +
  geom_point() +
  geom_smooth(method = "lm")
```

After fitting the regression line, the 1 Mb regions in the plot were classified into three categories according to their deviation from the fitted line: (i) regions within ±2 standard deviations (SD), (ii) regions below −2 SD, and (iii) regions above +2 SD. Each set of classified regions was saved as a separate BED file. For all regions in these BED files, the percentage of bases annotated as intronic, exonic, or intergenic was calculated using *BEDtools intersect.* This was done by intersecting these BED files with annotation

files containing exon, intron, and intergenic intervals (see *2.4.9 Supporting Procedures* for details on how these annotations were created). This process was performed separately for each category (within 2 SD, below −2 SD, above +2 SD) and for each individual. The results were stored into summary tables (see Supplementary Table 5,6), indicating for each 1 Mb region the percentage of exonic, intronic, and intergenic content. Additionally, the results were visualized using box plots in *RStudio* with the *ggplot2* package and presented in Figure 4 A2, B2, C2 .

```
Note: This procedure is repeated for each category and each individual

category="categories"        # e.g., within_2std, below_2std, above_2std
individual="mummies_individuals"    # e.g., La Niña del Rayo, El Niño, La Doncella

bedtools intersect -a "${category}.bed" -b exons_protein_coding.bed -wo
  > "exon_coverage_${category}_${individual}.txt"

bedtools intersect -a "${category}.bed" -b introns_of_protein_coding_genes.bed -wo
  > "intron_coverage_${category}_${individual}.txt"

bedtools intersect -a "${category}.bed" -b intergenic.bed -wo  >
"intergenic_coverage_${category}_${individual}.txt"
```

Mapping Performance Analysis for Challenging Medically Relevant Genes

To assess whether the pangenome performs better in challenging genomic regions, the benchmarking set for challenging medically relevant autosomal genes (CMRG) was used (Wagner, Justin et al., 2022). The BED file containing CMRG gene coordinates was downloaded from the following source: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/CMRG_v1.00/GRCh38/SupplementaryFiles/. Base coverage was calculated using BEDtools coverage for each gene in the benchmark set across both the pangenome and linear genome alignments. Next, linear regression was fitted using the calculated base coverages to identify outliers. Genes with coverage deviating by more than two standard deviations from the regression line were classified as outliers. These genes were considered to have significantly different coverage between the linear and pangenome reference alignments.
Using this method, we identified 22 outlier genes for *La Niña del Rayo*, 23 for *El Niño*, and 23 for *La Doncella*—corresponding to approximately 8.5% of the 273 genes in the benchmark set. These findings are summarized in Figure 4.1 A in the Results section. Additionally, box plots illustrating the base coverage of outlier genes are presented in Figure 4.1 B in the Results section. All visualizations were generated using RStudio and the *ggplot2* package.

```
Note: This procedure is repeated for each alignment type and each individual

alignment="alignments"        # e.g.,linear genome, pangenome
individual="mummies_individuals"    # e.g., La Niña del Rayo, El Niño, La Doncella
bedtools coverage -sorted -g hg38.fa.fai -a
GRCh38_CMRG_benchmark_gene_coordinates_sorted.bed  -b $${individual}_${alignment}.bam >
coverage_CMRG_${individual}_${alignment}.bed
```

2.4.9 Supporting Procedures

Visualization of Coverage Profiles on Chromosome 6

This method was used to generate the data shown in Figure 5 and Supplementary Figure 3 of the Results section.

Read coverage along chromosome 6 was visualized using the karyoploteR package in R. Coverage density was computed from BAM alignment files using the *kpPlotBAMDensity* function, with a window size of 10,000 bp throughput chromosome 6. Separate plots were generated for alignments to the linear and the pangenome reference, using light pink and light blue colors, respectively, to distinguish between references. The MHC region (chr6:29,602,228–33,410,226) was highlighted with a semi-transparent rectangle.

```
kp <- plotKaryotype(genome = genome_reference, chromosomes = target_chromosome)
kpPlotBAMDensity(kp, data = bam_file, window.size = window_size, r0 = 0.1, r1 = 0.9,
                 col = plot_color, lwd = 1, ymin = y_min, ymax = y_max)
kpRect(kp, data = highlight_region_data, y0 = 0, y1 = 1.1,
       border = adjustcolor(highlight_color, alpha.f = 0.5), lwd = 2)
```

Generation of Genomic Annotation Files

Genomic features were extracted from the GENCODE v44 comprehensive gene annotation file (gencode.v44.chr_patch_hapl_scaff.annotation.gtf.gz)(downloaded from: https://www.gencodegenes.org/human/release_44.html) to classify regions into exonic, intronic, and intergenic segments specific to protein-coding genes. First, all entries with *feature == "exon"* and *gene_type == "protein_coding"* were extracted using awk command and stored in BED format. Coordinates were adjusted to 0-based start indexing, and exons were merged using *BEDtools merge* (with parameters *-c 4 -o distinct*) to remove overlaps across transcripts. In parallel, all entries with *feature == "gene"* and *gene_type == "protein_coding"* were extracted and saved as gene intervals. Intronic regions were defined by subtracting the merged exon intervals from the corresponding gene spans using *BEDtools subtract*. To create a comprehensive set of genic regions, exon and intron BED files were concatenated, sorted (*sort -k1,1 -k2,2n*), and then followed overlapping intervals were merged with *BEDtools merge*. Intergenic regions were then computed using *BEDtools complement*, which identifies genome

intervals not covered by any genic region. This approach ensures reproducible classification of the genome into protein-coding exons, introns, and intergenic regions based on standardized GENCODE annotations.

```
# Extract exonic intervals from protein-coding genes
zcat <annotation_file>.gtf.gz | \
  awk -F"\t" 'BEGIN {OFS="\t"} $3 == "exon" && /gene_type "protein_coding"/ {print $1,
$4-1, $5, $9}' | \
  sort -k1,1 -k2,2n | \
  bedtools merge -c 4 -o distinct > exons_protein_coding.bed

# Extract full gene intervals for protein-coding genes
zcat <annotation_file>.gtf.gz | \
  awk -F"\t" 'BEGIN {OFS="\t"} $3 == "gene" && /gene_type "protein_coding"/ {print $1,
$4-1, $5, $9}' | \
  sort -k1,1 -k2,2n > genes_protein_coding.bed

# Derive intronic intervals by subtracting exons from genes
bedtools subtract -a genes_protein_coding.bed -b exons_protein_coding.bed >
introns_protein_coding.bed

# Concatenate introns and exons, then sort and merge intervals
cat introns_protein_coding.bed exons_protein_coding.bed | \
  sort -k1,1 -k2,2n > exons_introns_protein_coding.sorted.bed

bedtools merge -i exons_introns_protein_coding.sorted.bed -c 4 -o distinct >
merged_exons_introns_protein_coding.bed

sort -k1,1 -k2,2n merged_exons_introns_protein_coding.bed >
merged_exons_introns_protein_coding.sorted.bed

# Create intergenic intervals by complementing merged exons and introns with genome
coordinates
bedtools complement -i merged_exons_introns_protein_coding.sorted.bed -g
hg38.chrom.sizes.sorted > intergenic.bed
```

3 Results

3.1 Background Information and Overview of Sampling and Genomic Data of Llullaillaco Mummies

This study was based on three ancient individuals *La Doncella* (the maiden, approx. 14 years old), *La Niña del Rayo* (the lightning girl, approx. 6 years old) and *El Niño* (the boy, approx. 7 years old). In Figure 1C, exhibition photographs of these individuals are shown. These individuals were victims of a sacrificial ritual known as *capacocha* and were found near the summit of Llullaillaco at an altitude of 6,715 meters above sea level (see Figure 1A).

The Llullaillaco mummies were sampled as part of this project and a related PhD project conducted in our lab (for more details about the sampling process refer to *2.3.1 Sampling*). In Figure 1B, the specific anatomical sites where samples were taken from each mummy are indicated. After sampling, the obtained samples were sequenced (for more details about sequencing procedure see *2.3.2 DNA extraction, library preparation, and sequencing*). Sequencing data from different body locations were merged per individual, and all downstream analyses were conducted on these merged datasets (for more details refer to *2.4.3 Postprocessing of Alignments* in Methods section). In Figure 1D, the whole genome coverage (WGC) of the merged datasets is shown. WGC ranges from approximately 5× to 25×, which represents exceptionally high coverage for ancient DNA and provides a solid foundation for confident downstream analyses.

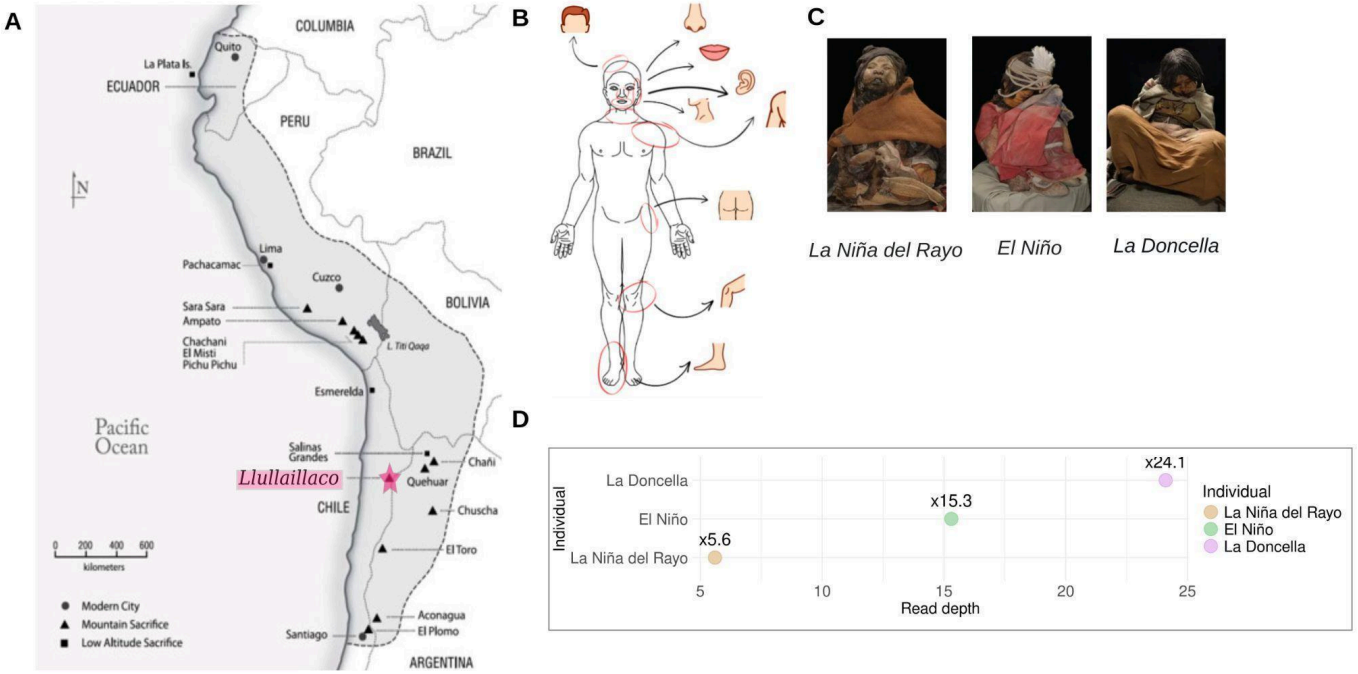


Figure 1 - Children of Llullaillaco

A) Map showing the locations of known capacocha sacrifice sites, with the Llullaillaco mountain site highlighted. B) Anatomical sampling locations from the Llullaillaco mummies. C) Photographs of the Llullaillaco mummies (Children of Llullaillaco; Image credit: Johan Reinhard). D) Whole genome sequencing (WGS) read depth per individual.

3.2 Assessing Genome Alignment of Llullaillaco Mummies to Human Pangenome and GRCh38 References

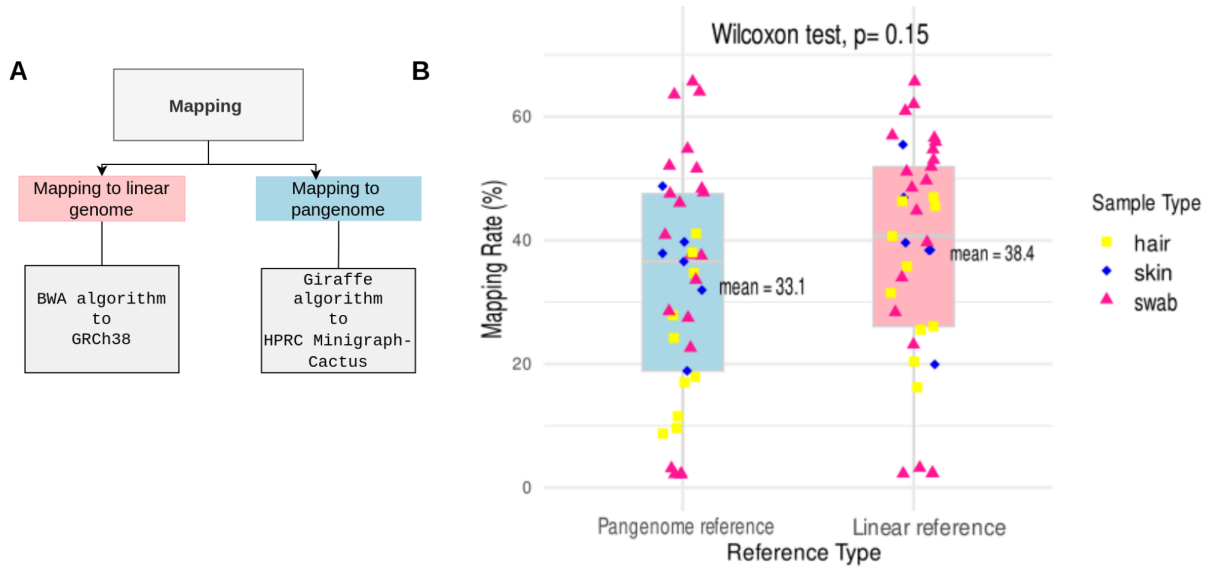
The Llullaillaco mummies were mapped to two human genome assemblies: the linear GRCh38 reference genome (Genome Reference Consortium, GCA_000001405.28) and the draft human pangenome (HPRC version 1.1), represented as a graph constructed using the Minigraph-Cactus pipeline (Liao, Wen-Wei, et al., 2023) (see Figure 2A). Mapping was performed using closely matched parameters to ensure comparability between the two references (for more details, refer to *2.4.2 Alignment to the Linear and Pangenome References* in Methods section).

Each tissue sample extracted from the mummies (see Figure 1B) was mapped to both the pangenome and the linear reference genome. Mapping rates were then calculated for each sample as the percentage of reads that successfully aligned relative to the total number of raw reads (see *Supplement Table 2 and 3*). These mapping rates are presented as box plots in Figure 2B, with one box plot summarizing all mapping rates to the pangenome and another summarizing those to the linear genome (mean values: 33.1 and 38.4, respectively).

To further investigate differences in mapping between the pangenome and the linear genome, we first merged the individual tissue sample files belonging to each Llullaillaco individual. We then extracted all reads from both alignments and generated Venn diagrams for each individual to visualize the overlap and unique read sets (see Figure 2C). The Venn diagrams show that most reads are shared between the pangenome and linear genome alignments, representing a core set of confidently mapped reads. In addition, there are smaller subsets of reads that map uniquely to either the pangenome or the linear genome. For La Niña del Rayo and El Niño, the proportion of uniquely mapped reads is slightly higher for the linear genome. In contrast, for La Doncella , a greater proportion of uniquely mapped reads aligns to the pangenome. This may be attributed to the overall higher sequencing coverage for La Doncella, which includes a greater number of total reads and, consequently, may increase the chance of mapping to additional regions present only in the pangenome.

Furthermore, we analyzed the distribution of unique reads across genomic features - specifically exons, introns, and intergenic regions (see Figure 3C, pie charts corresponding to each Venn diagram). In this analysis, exons refer exclusively to protein-coding exons, and introns to introns within protein-coding genes (for more details, refer to *2.4.9 Supporting Procedures* in Method section). As expected, the

distribution of unique reads aligns with the general patterns observed in human genome annotations; the smallest proportion of reads mapped to regions annotated as exons, followed by intronic regions, while the majority of unique reads were found in intergenic regions for both the pangenome and linear reference alignments (Francis, W. R., & Wörheide, G. , 2017). The proportion of unique reads mapped to exonic regions in the linear reference showed a mean value of 5,4% across all three Llullaillaco individuals. For intronic regions, the mean was 52,4%, and for intergenic regions, 46,5%. In comparison, the unique reads mapped to the pangenome reference had a mean value of 3,3% for exons, 63,3% for introns, and 36,7% for intergenic regions. These distributions are comparable; however, a noticeable difference emerges in the mean values for intergenic regions, which vary by about 10% between unique reads mapped to the pangenome and those mapped to the linear genome. The higher proportion of unique reads mapping to regions annotated as intergenic in the pangenome alignment likely reflects its improved representation of structurally complex and repetitive elements, such as Variable Number Tandem Repeats (VNTRs), centromeric/pericentromeric, and satellite sequences, which are often located in intergenic space but poorly captured by the linear reference (Liao, Wen-Wei, et al., 2023).



C

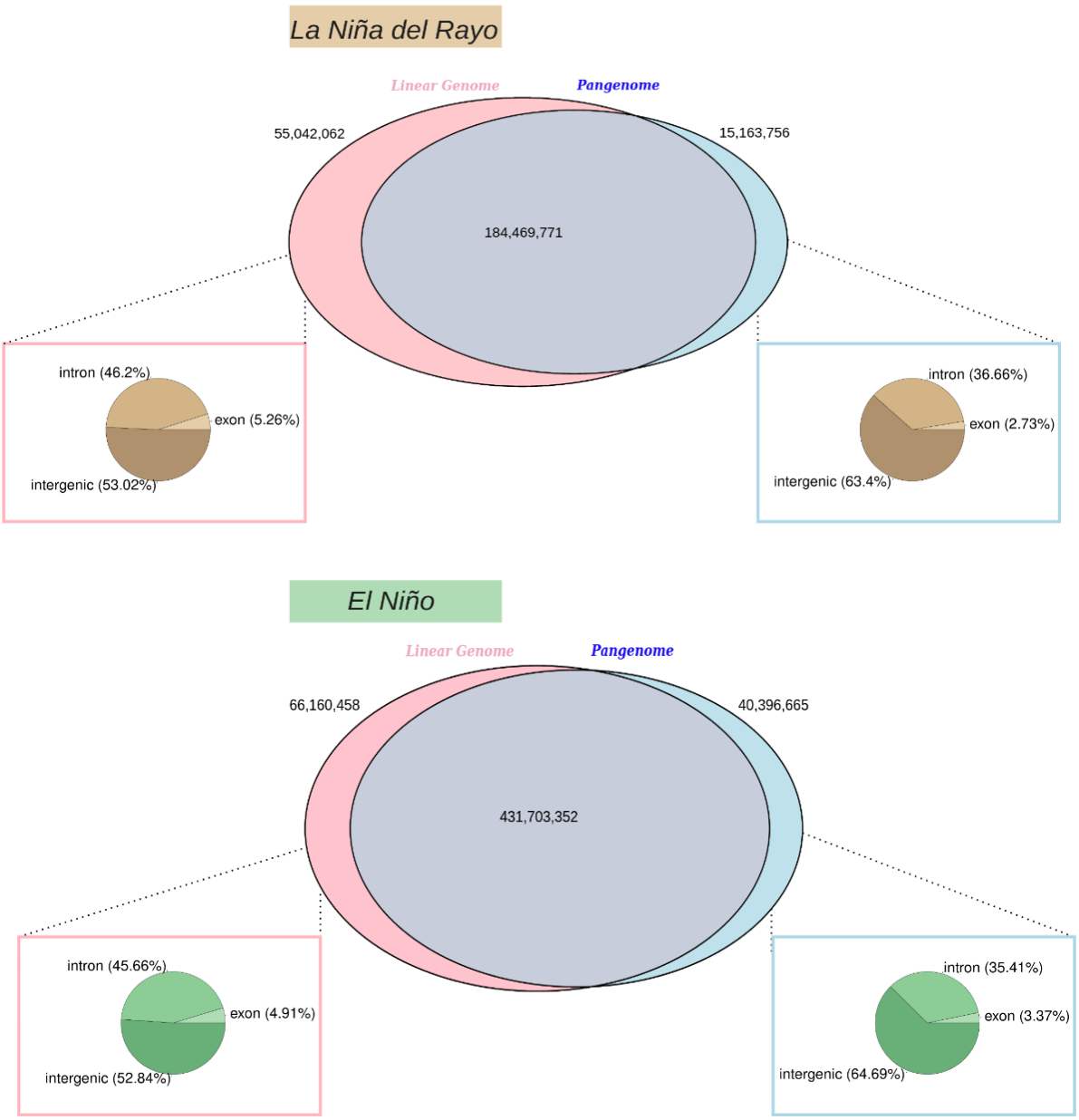

La Niña del Rayo
Linear Genome
Pangenome
55,042,062
184,469,771
15,163,756
intron (46.2%)
exon (5.26%)
intergenic (53.02%)
intron (36.66%)
exon (2.73%)
intergenic (63.4%)
El Niño
Linear Genome
Pangenome
66,160,458
431,703,352
40,396,665
intron (45.66%)
exon (4.91%)
intergenic (52.84%)
intron (35.41%)
exon (3.37%)
intergenic (64.69%)
La Doncella
Linear Genome
Pangenome
111,476,673
572,942,975
143,160,257
intron (47.63%)
exon (6.2%)
intergenic (51.44%)
intron (38.25%)
exon (3.78%)
intergenic (61.87%)

Figure 2 - Comparison of mapping performance and read distribution for Llullaillaco mummies

A) Schematic overview of the mapping strategy applied to Llullaillaco mummy samples B) Box plots displaying the mapping rates of all individual tissue samples, comparing alignment to the pangenome and to the linear reference genome. C) Venn diagrams showing the overlap between reads mapped to the pangenome and the linear genome. Blue represents reads unique to the pangenome, pink indicates reads unique to the linear genome, and grey shows reads shared between both references. Accompanying pie charts illustrate the distribution of uniquely mapped reads across genomic features annotated as protein-coding exons, introns of protein-coding genes, and intergenic regions. Pie charts with a pink border correspond to reads unique to the linear genome, while those with a blue border represent reads unique to the pangenome.

3.3 Variant Detection in Llullaillaco Mummies

After aligning reads to both GRCh38 and the pangenome reference, we performed variant calling using identical parameters for both alignments. This ensures that the results are directly comparable and that any differences arise solely from the mapping step, not the variant calling itself (see Figure 3A,B,C). Shown here are the filtered variants, based on allele ratio and read depth (for more details please refer to *2.4.5 Variant Calling* in Methods section).

The Venn diagrams represent the total set of variants identified by GATK HaplotypeCaller ("labeled as Total variants in figures below"), encompassing all detectable variant types: SNPs (single nucleotide polymorphisms), INDELs (insertions and deletions), MIXED variants (combinations of SNPs and indels at the same position), MNPs (multi-nucleotide polymorphisms, such as dinucleotide substitutions), and SYMBOLIC variants (https://gatk.broadinstitute.org/hc/en-us/articles/360035530752-What-types-of-variants-can-GATK-tools-detect-or-handle). In addition, we separately analyzed the two most commonly studied variant categories, SNPs and INDELs. One prominent pattern is that substantially more INDELs were detected in the pangenome alignment compared to the linear reference. This is consistent with findings from previous studies, which demonstrate that human pangenome reference enables better representation of structural variants and complex genomic loci, thereby significantly improving the accuracy and sensitivity of variant detection at these sites (Hickey et al., 2020; Sirén et al., 2021; Liao et al., 2023; Groza et al., 2024). SNP detection yielded comparable results between the two reference genomes. In the samples from La Doncella and El Niño, a slight increase in SNP calls was observed when using the pangenome alignment. In contrast, the La Niña del Rayo sample exhibited a marginally higher number of SNP calls with the linear genome alignment. Finally, the total set of variants detected was consistently higher across all three Llullaillaco individuals when using the pangenome

reference, further supporting previous evidence of its enhanced sensitivity for variant detection.

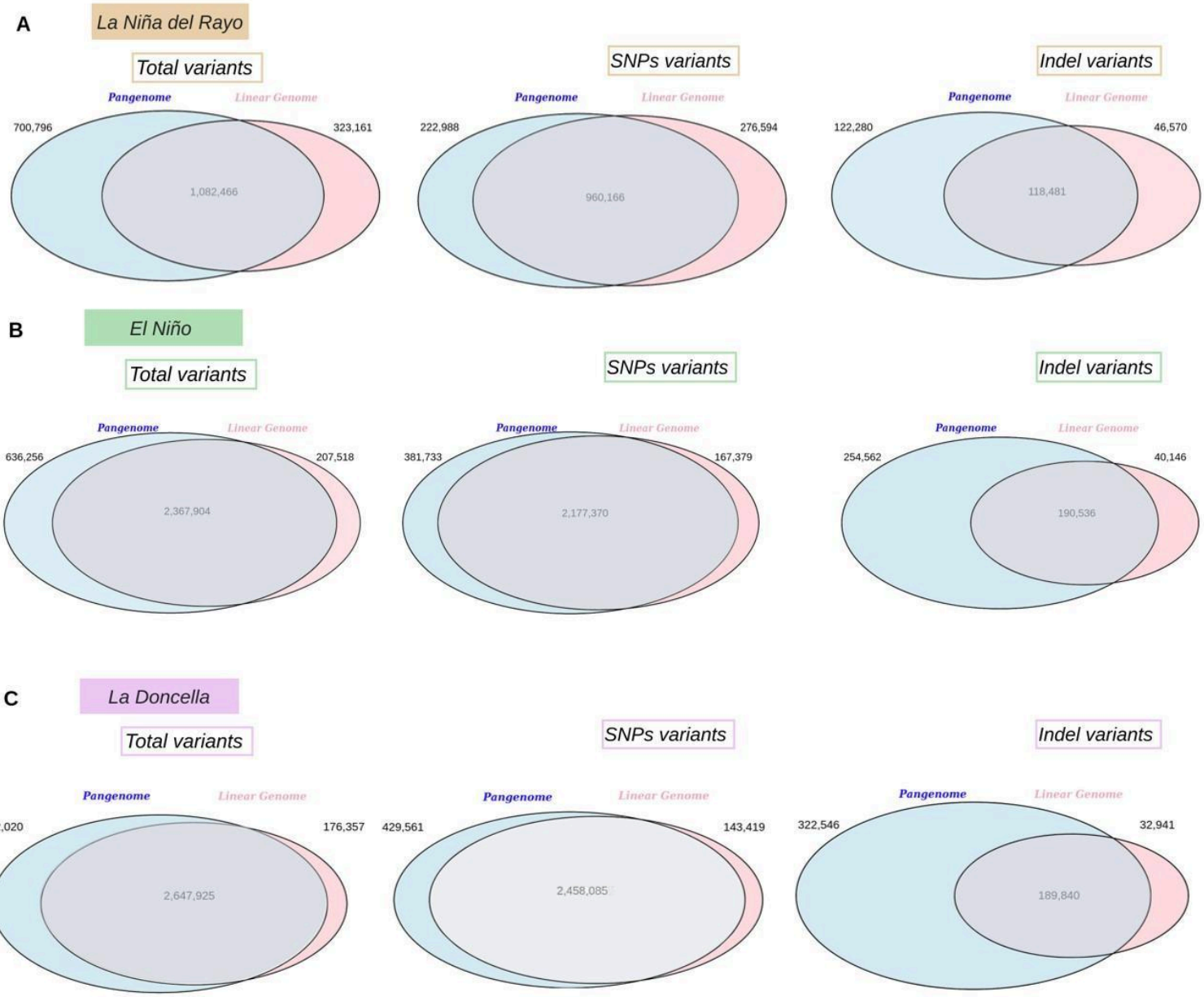


Figure 3 - Comparison of Variant Calls Between Pangenome and Linear Genome Alignment

Venn diagrams illustrate the overlap of variants called from alignments to the pangenome and the linear genome. Three variant categories are presented: *Total variants* (left), *SNPs* (middle), and *INDELs* (right). Blue areas represent variants uniquely identified through the pangenome alignment, pink areas correspond to variants unique to the linear genome alignment, and grey regions indicate variants detected by both approaches.

Venn Diagrams of Variant Calls in La Niña del Rayo

Venn Diagrams of Variant Calls in El Niño

Venn Diagrams of Variant Calls in La Doncella

3.4 Genome-Wide Comparison of Mapping Performance

To investigate the specific genomic regions where the pangenome provides an advantage, we divided the GRCh38 genome into 1 Mb windows with 10 kb overlaps between consecutive windows. For each window, we calculated read coverage separately for the pangenome and linear genome alignments. Read coverage was defined as the percentage of bases within each 1 Mb window that are covered by sequencing reads. This approach provides a genome-wide, coarse-grained measure of where the pangenome may offer improved read alignment compared to the linear reference.

Upon calculation of read coverage, we then performed a regression analysis (see Figures 4 A1,B1,C1). In these figures, three distinct categories of data points can be observed. The first category consists of data points within two standard deviations from the regression line (grey), indicating genomic regions where mapping performance is comparable between the linear and pangenome references. The second category includes points above two standard deviations (blue), corresponding to regions where the linear genome shows superior mapping performance. Conversely, the third category consists of points below two standard deviations (red), representing regions where the pangenome provides improved mapping.

To further characterize regions with differing mapping performance, we assessed their genomic composition (see Figures 4 A2,B2,C2). Specifically, for each 1 Mb window, we calculated the proportion of the region classified as intergenic, intronic, or exonic. This allowed us to identify which types of genomic regions are associated with improved or diminished mapping performance. One striking observation is, that regions where the linear genome outperformed the pangenome in mapping ("blue data points" in regression line) were almost exclusively intergenic (mean intergenic content: 99.2%). Upon closer inspection, we found that these regions correspond to areas containing alternative contigs in the linear reference. In the pangenome, these alternative sequences are not represented as separately labeled contigs but are instead integrated as graph paths (see Supplementary Table 1 for details). Thus, the apparent improved performance of the linear genome in these regions is a labeling artifact rather than a genuine biological difference.

In regions where the pangenome outperforms the linear genome (red data points in the regression plot), the majority are annotated as intergenic, followed by intronic, with exonic regions comprising the smallest fraction. Same distribution was observed in the regions where mapping performance was comparable between the pangenome and the linear genome ("grey data points"). To assess whether the composition of genomic features differs between these two categories, we performed a Wilcoxon rank-sum test comparing the distributions of exonic, intronic, and intergenic content. We found that the proportion of exonic content is significantly higher in regions where the pangenome performs better, compared to regions with similar performance across references. This trend was consistent across all Llullaillaco individuals, with p-values of 2.4×10^{-5} for La Niña del Rayo, 2.0×10^{-3} for El Niño, and 2.0×10^{-3} for La Doncella. In contrast, no

significant differences were observed for the proportions of intronic and intergenic regions. These results suggest that the improved performance of pangenome alignments is particularly pronounced in regions enriched for exonic sequences.

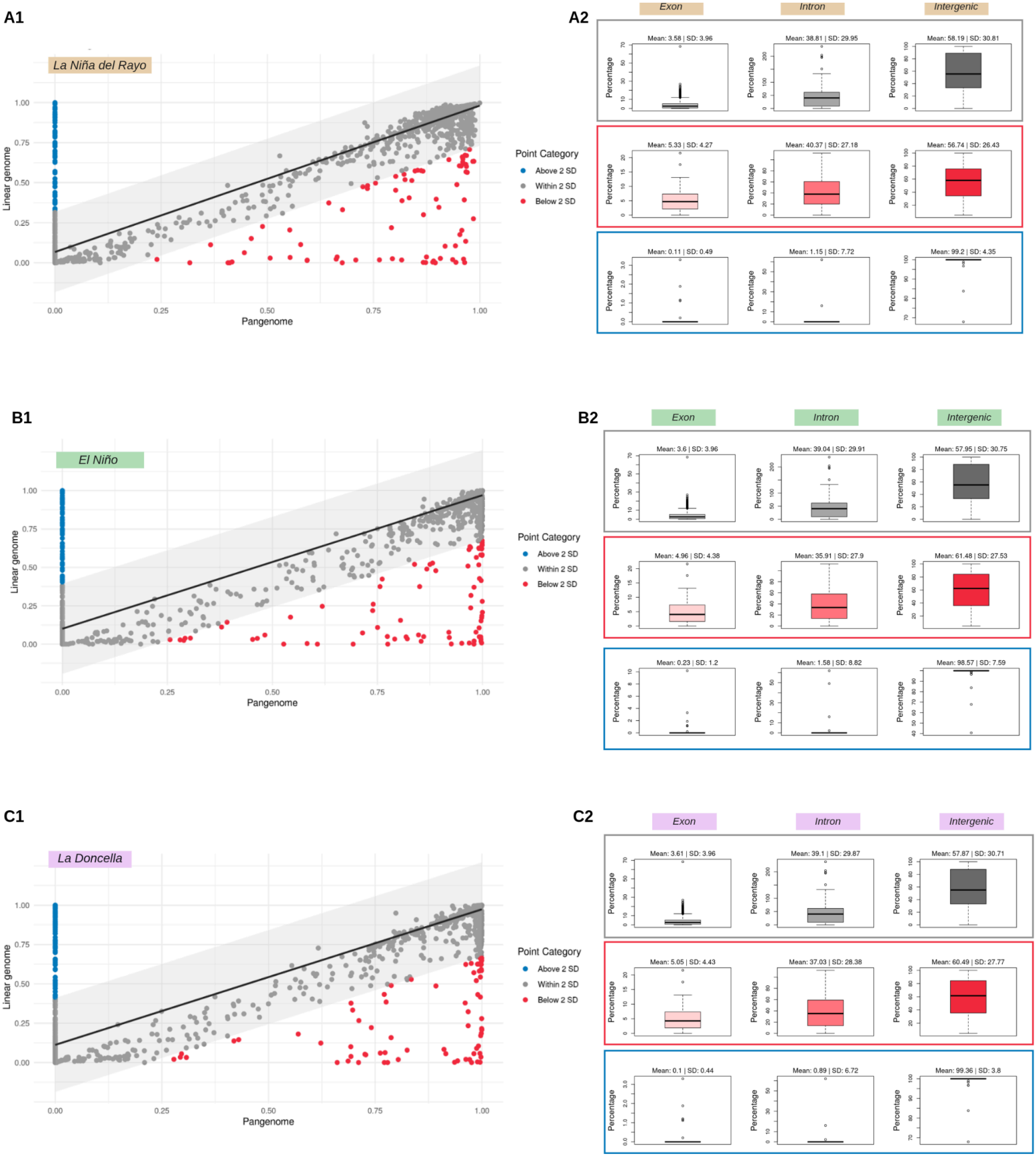


Figure 4 - Genome-Wide Comparison of Mapping Performance

Regression plots (left) illustrate the correlation of base coverage between alignments to the pangenome and the linear reference genome, calculated across overlapping 1 Mb genomic windows. Base coverage here is defined as the percentage of each 1 Mb region covered by sequencing reads. Data points are categorized based on their deviation from the regression line: grey (within ±2 standard deviations), blue (above +2 SD, favoring linear genome), and red (below −2 SD, favoring pangenome).

Accompanying box plots (right) display the genomic composition (proportions of exonic, intronic, and intergenic regions) of the 1 Mb windows within each category.

Regression and box plot for La Niña del Rayo

Regression and box plot for El Niño

Regression and box plot for La Doncella

3.4.1 Enhanced Mapping in Medically Challenging Genes with a Pangenome Reference

As demonstrated in the previous figure, pangenome-based mapping shows improved performance in exonic regions compared to the linear genome. To further investigate which specific genes benefit from this improvement, we analyzed a set of challenging, medically relevant genes defined in the benchmark by Wagner et al., 2022. This benchmarking set comprises 273 autosomal genes known to be difficult for variant detection due to their complex genomic architecture. We evaluated how both the pangenome and linear reference alignments performed across this gene set. Our analysis revealed that for approximately 8.5% of these genes (22 in La Niña del Rayo, and 23 in both El Niño and La Doncella), the pangenome provides a more suitable reference for reliable variant detection, as evidenced by significantly higher coverage compared to the linear reference (see Supplementary Table 7 for details). These genes were largely consistent across Llullaillaco in and are highlighted in Figure 4.1A.

To illustrate this difference, we calculated base-level coverage (as defined in Figure 4) for all 273 genes using both reference alignments and visualized the data using boxplots (see Figures 4.1 B1,2,3). These plots clearly show that the subset of genes with significant differences had markedly improved coverage when aligned to the pangenome. This improvement was statistically supported by Wilcoxon rank-sum tests, yielding p-values of 5.2×10^{-8} for La Niña del Rayo, 1.5×10^{-8} for El Niño, and 1.1×10^{-8} for La Doncella.

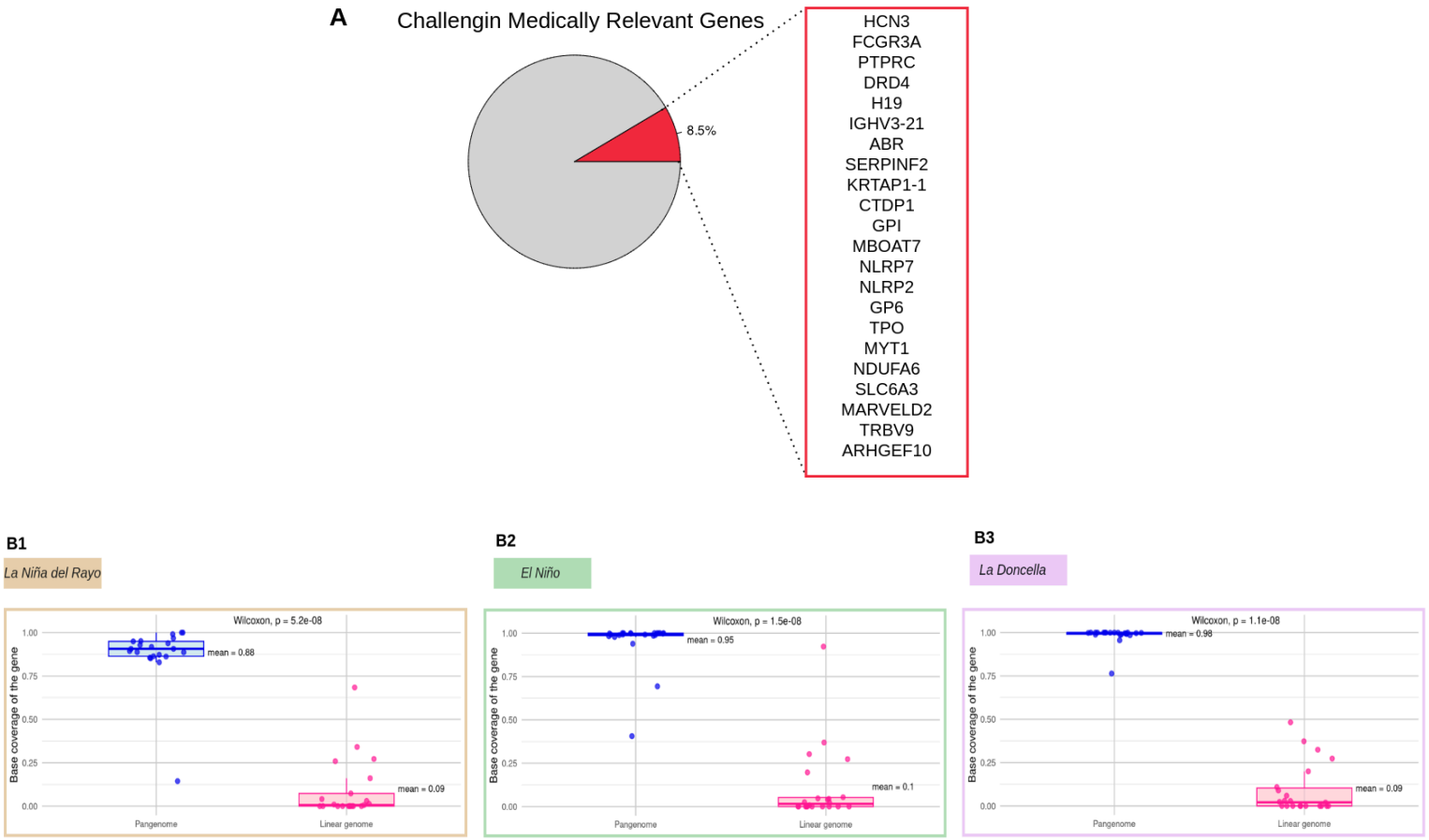


Figure 4.1 - Enhanced Mapping in Medically Challanging Genes with a Pangenome Refernce

A) Pie chart illustrating the proportion of medically relevant genes from the benchmark set that exhibit significantly different base coverage between pangenome and linear genome mappings. Gene names are displayed within the chart.

B) Box plots comparing read coverage between pangenome and linear genome alignments for the subset of genes identified in panel A for *La Niña del Rayo (B1), El Niño (B2), La Doncella(B3).*

3.5 Enhanced HLA Genotyping of Llullaillaco Mummies Using a Pangenome Reference

Building on the findings presented in Figure 4, where the pangenome reference showed improved mapping performance in specific genomic regions, including complex and medically relevant genes, we next investigated whether these advantages extend to one

of the most polymorphic region of the human genome: the human leukocyte antigen (HLA) cluster. Improved read alignment within this region is especially important for ancient DNA (aDNA) samples, where accurate genotyping is often hindered by degradation and low coverage.

The HLA region is located on the short arm of chromosome 6 (6p21.3) and it encodes the major histocompatibility complex (MHC), which plays a central role in immune function (Cruz-Tapias et al., 2013). It includes three major gene classes: HLA class I (HLA-A, HLA-B, HLA-C), class II (HLA-DP, HLA-DQ, HLA-DR), and class III (encoding components of the complement system, 21-hydroxylase, heat shock proteins, and tumor necrosis factors) (Medhasi & Chantratita, 2022). In this study, we focused our analysis on HLA class I genes (Figure 5A), as these have been shown to provide the most robust and reliable genotyping results in aDNA contexts (Plascencia et al., 2025).

At the beginning of our analysis, we visualized read depth across chromosome 6 using karyoplots, calculated in 10 kb windows (see Figure 5B). These plots clearly illustrate a striking discrepancy in read depth between the linear and pangenome alignments, particularly at the MHC class I loci. Further, we quantified these differences by calculating the exact read depth values for each Lullaico individual (see Figure 5C). For example,in the La Doncella sample, only 188 reads mapped to the MHC I region in the linear genome alignment, compared to 2,029 reads in the pangenome genome alignment. Similar trends are observed in La Niña del Rayo and El Niño samples, confirming that the discrepancy is not individual-specific but systematic. These results demonstrate that the pangenome can significantly enhance HLA genotyping in ancient DNA (aDNA) individuals, potentially enabling more accurate and comprehensive characterization of this highly polymorphic region.
To determine the HLA class I genotypes of the ancient individuals, we utilized OptiType (Szolek, A., et al., 2014)(for more details refer to *2.4.6 HLA Class 1 Typing* in Methods section). Notably, the genotyping results derived from pangenome-based alignments differed markedly from those obtained using the linear reference genome (see Figure 5D). However, in all cases, the pangenome-derived genotypes were more reliable, as evidenced by higher read depth and superior values of the “Objective” score provided by OptiType. This improvement can be attributed to the human pangenome graph's construction, which incorporates 90 haploid assemblies. For complex loci like the HLA-A region, the graph's structure allows different structural haplotypes to be represented as distinct paths, leading to enhanced genotyping accuracy. For more details on the graph's construction and representation, refer to Liao, Wen-Wei, et al., 2023, particularly Figure 5.

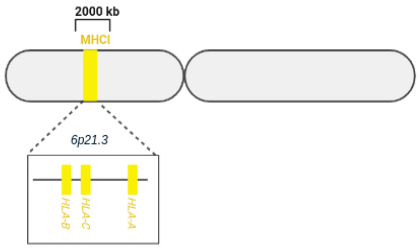
La Niña del Rayo was genotyped as *A02:01 / A31:01*, *B39:05 / B35:08*, and *C07:02 / C04:14*, supported by a total of 284 reads and score of 276 for Objective value. Notably, the haplotype *A02:01–B39:05–C07:02* is almost exclusively observed in Latin American populations, including Mexico (Mexico City), Colombia (Bogotá), and Panama, as well as among Hispanic American populations. This population-specific haplotype information

was obtained using the HLA Allele Frequency Net Database haplotype search tool (allelefrequencies.net). El Niño was genotyped as *A24:02 / A68:01*, *B40:02 / B48:01*, and *C03:04 / C08:01*, with 1208 supporting reads and a score of 1175 for Objective value. Both haplotypes are found in Latin American populations with high frequencies in Mexico (Chihuahua, Mexico City, Hidalgo Mezquital Valley) and Colombia (Bogotá), and are enriched among Hispanic American individuals. La Doncella was genotyped at as *A02:01 / A02:11*, along with *B15:01 / B51:01* and *C01:02 / C03:04*, supported by 2029 reads and a score of 1175 for Objective value. The haplotype *A02:01–B15:01–C01:02* is also reported at high frequency in Mexico and Nicaragua. This finding aligns with expectations that all three individuals share a common ancestral background or exhibit gene flow patterns indicative of Latin American origin.

Several of the identified alleles have documented associations with modern diseases. For instance, HLA-A*31:01 has been linked to hypersensitivity to carbamazepine treatment (Yip, V. L. M., & Pirmohamed, M., 2017; Dean, L., 2018; Amstutz, U., et al., 2013), while HLA-B*51:01 is associated with an increased susceptibility to Behçet's disease (Gül & Ohno, 2012; Wallace, G. R. , 2014 ;Khoshbakht, S., et al., 2023). However, these associations likely had limited relevance in ancient individuals, as many of the implicated conditions related to the detected alleles (Langtry, A., et al., 2024; Shi, Y. W., et al., 2017; Dolton, G., et al., 2024) are more prevalent in post-industrial, westernized populations. These connections are further explored in the ClinVar-based disease association analysis section below.

For quality control purposes, coverage plots of the predicted HLA genotypes were provided (see Supplementary Figure 2).

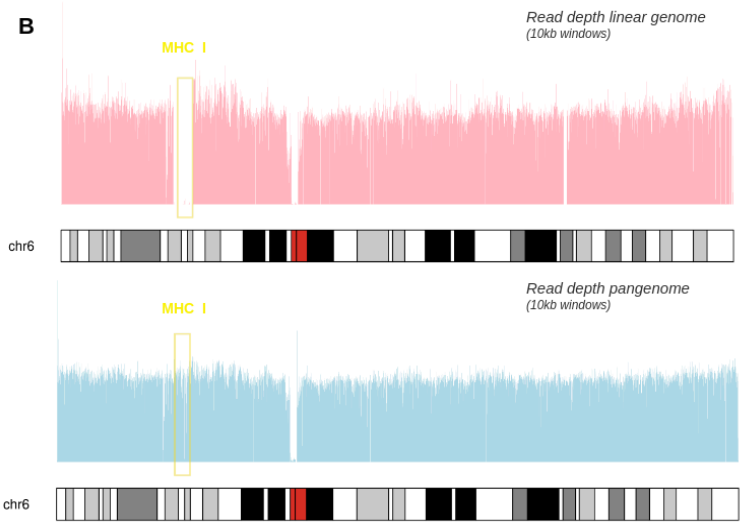
A



C

	La Niña del Rayo *linear genome*	La Niña del Rayo *pangenome*	El Niño *linear genome*	El Niño *pangenome*	La Doncella *linear genome*	La Doncella *pangenome*
A1	A*03:01	A*02:01	A*23:01	A*24:02	A*31:01	A*02:01
A2	A*03:01	A*31:01	A*68:01	A*68:01	A*68:02	A*02:11
B1	B*14:02	B*39:05	B*27:05	B*40:02	B*15:01	B*15:01
B2	B*14:02	B*35:08	B*48:01	B*48:01	B*51:01	B*51:01
C1	/	C*07:02	C*03:04	C*03:04	C*03:04	C*01:02
C2	/	C*04:14	C*04:01	C*08:01	C*03:04	C*03:04
Reads	**5.0**	**284.0**	**80.0**	**1208.0**	**188.0**	**2029.0**
Objective	**5.0**	**276.312**	**77.84**	**1175.374**	**184.606**	**1955.936**

B



D

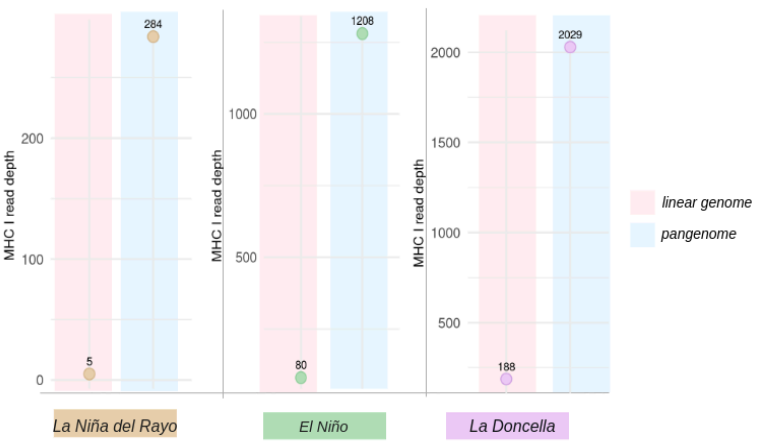


Figure 5 - HLA Genotyping of Llullaillaco Mummies

(A) Schematic overview of the MHC Class I loci on chromosome 6 (6p21.3), indicating the positions of HLA-A, HLA-B, and HLA-C genes. (B) Karyoplots of chromosome 6 showing read depth across 10 kb windows La Doncella aligned to the linear reference genome (pink) and the pangenome reference (blue) (C) Table summarizing HLA Class I genotypes of the Llullaillaco mummies, including allelic calls for HLA-A, -B, and -C loci, total supporting reads, and OptiType objective scores. (D) Read depth comparison plots of the MHC Class I loci for the Llullaillaco individuals.

3.6 Clinically Relevant Pathogenic Variants in the Llullaillaco Mummies

To assess the possible health-related genotype of the Llullaillaco individuals, we screened their genomes for clinically relevant pathogenic variants using the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/), following the approach outlined in Chapter *2.4.7 Identification of Clinically Relevant Pathogenic Variants* in the Methods section. Both linear and pangenome alignments were analyzed for this purpose and the results are shown in Figure 6. For La Niña del Rayo and El Niño, the linear genome alignment identified one additional variant compared to the pangenome alignment. However, this difference is likely due to post-variant-calling filtering criteria, which retain only sites that meet individual-specific read depth thresholds and an allele balance between 0.40 and 0.60 (for more details refer to *2.4.5 Variant Calling* in Methods section). These filters are applied to ensure that only high-confidence variants are retained. In these cases, the pangenome alignment provided fewer supporting reads for some loci, leading to their exclusion during filtering. This observation is consistent with our earlier findings (see Figure 2), where linear genome alignments generally yielded a higher total number of reads than pangenome alignments. Conversely, for La Doncella and La Niña del Rayo, one additional variant was detected in the pangenome alignment that was not observed with the linear reference. This result highlights that a pangenome provides a more comprehensive representation of genetic variation and enables the detection of variants that might be missed using a linear reference.

All three individuals carried four clinically relevant pathogenic variants (see Figure 6), which included homozygous variant associated with familial hypercholesterolemia type 1 and heterozygous variant linked to familial partial lipodystrophy type 8. Both of these variants are highly prevalent in populations with Native American ancestry. The alternative allele frequency for the familial hypercholesterolemia variant exceeds 76%, with nearly 60% of individuals being homozygous. For the lipodystrophy variant, the alternative allele frequency surpasses 70%, with nearly 50% of individuals being homozygous. Given their high frequency in the reference population, these variants are

unlikely to have had significant clinical impact in the context of the Lullaillaco individuals. Furthermore, all three individuals carried a variant associated with neutrophil inclusion bodies and renal tubular epithelial cell apoptosis. La Niña del Rayo and El Niño were homozygous for the alternative allele, while La Doncella was heterozygous. This variant is also highly prevalent in populations with Native American ancestry, with an alternative allele frequency of 81.1% and a homozygous alternative frequency of 65.9%.

Aside from the shared variants, several individual-specific variants were also identified. La Niña del Rayo carried a variant in the Xg blood group system; El Niño harbored a variant associated with increased susceptibility to asthma; and La Doncella carried a variant linked to malaria susceptibility. However, we suspect that the majority of these findings have limited relevance in the context of ancient DNA. Some conditions are strongly influenced by modern environmental factors. For instance, aspirin-induced asthma requires exposure to aspirin, which was not available in ancient populations. Similarly, susceptibility to malaria only manifests in individuals exposed to malaria-endemic regions, which may not apply to the mummies studied. Moreover, for some of the detected conditions, there were no phenotypic indications suggesting their expression. For example, familial partial lipodystrophy type 8 is characterized by abnormal fat distribution, specifically, fat loss in the limbs and accumulation in the face, neck, and trunk, which was not observed in the studied mummies (Bagias, C., et al., 2020). Likewise, familial hypercholesterolemia is generally associated with features such as tendon xanthomas (fatty deposits in tendons) and corneal arcus (a white or gray ring around the cornea), neither of which were present in the individuals analyzed (Rallidis, L. S., Iordanidis, D., & Iliodromitis, E., 2020).

In summary, all variants that passed our filtering criteria were either highly prevalent among individuals with Native American ancestry (as reflected in the “Latin America 2” dataset in dbSNP) or lacked sufficient clinical support to indicate meaningful disease relevance in these ancient individuals.

La Niña del Rayo

Disease association in ClinVar	Chr	Position	REF	ALT	SNP(rs)	DP	Genotype	ALT Allele Frequency	ALT HMOZ Frequency	Pangenome	Linear
Familial hypercholesterolemia (FHCL1)	chr1	161223893	G	A	rs5082	5	1|1	0.7685	0.5887	X	X
Renal tubular epithelial cell apoptosis Neutrophil inclusion bodies	chr1	976215	A	G	rs7417106	5	1|1	0.811	0.659236		X
Familial partial lipodystrophy type 8 (FPLD8)	chr10	111079821	A	G	rs553668	8	0|1	0.7026	0.490662		X
Xg blood group system	chrX	2748343	G	C	rs311103	6	0|1	0	0	X	

El Niño

Disease association in ClinVar	Chr	Position	REF	ALT	SNP(rs)	DP	Genotype	ALT Allele Frequency	ALT HMOZ Frequency	Pangenome	Linear
Familial hypercholesterolemia (FHCL1)	chr1	161223893	G	A	rs5082	21	1|1	0.7685	0.5887	X	X
Renal tubular epithelial cell apoptosis Neutrophil inclusion bodies	chr1	976215	A	G	rs7417106	17	1|1	0.811	0.659236	X	X
Familial partial lipodystrophy type 8 (FPLD8)	chr10	111079821	A	G	rs553668	24	0|1	0.7026	0.490662		X
Aspirin-induced susceptibility to asthma|Asthma and nasal polyps	chr17	47731462	T	C	rs4794067	16	0|1	0.354	0.12456	X	X

La Doncella

Disease association in ClinVar	Chr	Position	REF	ALT	SNP(rs)	DP	Genotype	ALT Allele Frequency	ALT HMOZ Frequency	Pangenome	Linear
Familial hypercholesterolemia (FHCL1)	chr1	161223893	G	A	rs5082	32	1|1	0.7685	0.5887	X	X
Renal tubular epithelial cell apoptosis Neutrophil inclusion bodies	chr1	976215	A	G	rs7417106	19	0|1	0.811	0.659236	X	X
Mild susceptibility to malaria |Malaria severe susceptibility	chr6	31593133	C	G	rs2736191	26	1|1	0.226	0.045902	X	X
Familial partial lipodystrophy type 8 (FPLD8)	chr10	111079821	A	G	rs553668	32	0|1	0.7026	0.490662	X	

Figure 6 - Tables of clinically relevant pathogenic variants identified in the Llullaillaco mummies

Genomic positions are based on the GRCh38 reference genome. The SNP(rs) column lists unique variant identifiers from the dbSNP database. DP indicates read depth from the linear genome alignment. ALT Allele Frequency and ALT Homozygous Frequency refer to the frequency of the alternative allele and the homozygous alternative genotype, respectively, in populations with Native American ancestry, as reported in the dbSNP "Latin America 2" dataset.

4 Disscussion

Aligning DNA sequences obtained from ancient biological materials, commonly referred to as aDNA, presents a unique set of challenges. These include the typically short fragment lengths, characteristic post-mortem damage patterns, and frequent contamination with modern DNA. One major obstacle is *reference bias*, where alignment algorithms favor the reference genome alleles, potentially leading to incorrect interpretations when the ancient sequences diverge significantly from the reference (Dolenz, S., et al. , 2024; Günther, T., & Nettelblad, C. , 2019; Poznyak, A. V., et al., 2024). Recent advances in genome sequencing and assembly technologies have paved the way for the construction of population-scale genomic references, most notably the Human Pangenome Reference developed by the Human Pangenome Reference Consortium (HPRC). This resource comprises 94 de novo haplotype assemblies derived from 47 individuals representing diverse ancestral backgrounds. By incorporating a broader range of human genetic diversity, the pangenome reference provides a more inclusive and representative framework for read alignment (Liao, Wen-Wei, et al., 2023; Taylor, D. J., et al., 2024; Abondio, P., Cilli, E., & Luiselli, D., 2023). These improvements are especially impactful in the field of ancient DNA research, where aligning degraded and divergent sequences to a more inclusive reference can lead to significantly enhanced mapping accuracy. Ultimately, this should enable more reliable detection of genetic variants and offer deeper insights into past populations, their health, adaptations, and evolutionary history. The focus of this thesis is to evaluate whether aligning ancient human DNA to the human pangenome reference improves mapping accuracy and variant detection, particularly within complex genomic regions.

In this thesis, I analyzed ancient DNA (aDNA) samples obtained from the Llullaillaco mummies, considered among the best-preserved human mummies in the world. Their exceptional state of preservation enabled the recovery of high-quality aDNA, with whole genome sequencing (WGS) coverage ranging from approximately 5x to 25x (see *Figure 1D* in Results section). This level of coverage is rare in the field of ancient genomics and provided a robust foundation for a comprehensive comparison between linear and pangenome-based alignments.

To compare alignment performance, tissue-specific samples were mapped to both the linear and pangenome references in parallel. Overall mapping rates were similar, with a slight advantage for the linear genome (see *Figure 2B* in Results section). Most reads were shared between alignments, while a small subset was unique to either the linear or pangenome alignment (see *Figure 2C* in Results section). Notably, pangenome-specific reads were more likely to map to intergenic regions, with an average increase of 10% compared to the linear alignment. This observation aligns with the work of Liao, Wen-Wei, et al., 2023, which shows that the human pangenome reference provides more complete representation of complex genomic areas like satellites, centromeric regions,

and structural variants, that are commonly annotated as intergenic. Upon mapping, we performed variant calling on both the linear and pangenome alignments. A distinct pattern emerged that pangenome alignment resulted in substantially more total and indel variants, while the number of SNPs was comparable between the two approaches (see *Figure 3A,B,C*). The substantial increase in variant detection provides a strong initial indication that the pangenome reference captures genetic diversity more comprehensively. From an applied perspective, this increase in variant detection holds particular promise for ancient DNA studies, where the ability to identify a broader range of genetic variation may enable deeper insights into population history, adaptation, and medically relevant loci that were previously underrepresented due to reference bias.

While we quantified and compared variants detected from both the linear and pangenome alignments, we did not perform formal benchmarking against a validated reference call set. Incorporating benchmarking would be a valuable next step in future studies, as it would enable a more precise assessment of sensitivity and specificity, and help distinguish true positives from potential false discoveries. Standard benchmarking resources, such as the Genome in a Bottle (GIAB) sample HG002, offer high-confidence small variant calls and could serve as a reference for evaluating variant detection accuracy. Additionally, the CMRG (Challenging Medically Relevant Genes) benchmark set (Wagner et al., 2022), provides both small and structural variant truth sets in clinically important regions, making it particularly relevant for testing medically significant variant recovery. Moreover, future work could also incorporate functional interpretation pipelines to assess the potential phenotypic relevance of detected variants. Our primary goal here was to explore whether pangenome alignment offers advantages over the linear reference in the context of aDNA variant discovery. As such, our approach focused on comparative detection rather than benchmarking accuracy, and also allowed for the inclusion of novel variants not present in current benchmark datasets. Nonetheless, combining benchmarking with functional annotation in future research would provide a more holistic evaluation of pangenome-based variant calling in ancient genomes.

Furthermore, we identified genome-wide regions where pangenome alignment provides improved read coverage compared to the linear genome (see *Figures 4A1–C1* in Results section). These regions were defined as 1 Mb windows, and we found 77 such regions in La Niña del Rayo, 80 in El Niño, and 72 in La Doncella (see *Supplementary Table 6a,b,c* in Supplements). Notably, 66 of these regions were consistent across all three individuals, suggesting a systematic advantage of the pangenome reference rather than a sample-specific effect. These regions also exhibited significantly higher exonic content compared to regions where no difference in read coverage was observed between the pangenome and linear genome (see *Figures 4A2–C2* in Results section). In summary, this finding indicates that the pangenome reference improves read coverage in specific genomic regions that are systematically observed across individuals and are particularly rich in exons. In addition to identifying broader 1 Mb regions, we also pinpointed specific medically relevant genes where the pangenome alignment demonstrated

improved coverage compared to the linear reference. Using a curated benchmarking gene set (Wagner et al., 2022), we found that 8.5% of these genes exhibited significantly better coverage when aligned to the pangenome (see *Figure 4.1* in Results section). This finding demonstrates that pangenome analysis enhances the recovery of medically relevant genomic regions even in degraded aDNA samples, enabling more accurate reconstruction of health-related genetic variation in ancient populations. This can deepen our understanding of the evolution of disease-associated alleles, population health, and adaptation over time.

In paleogenomics, HLA genotyping is currently performed almost exclusively through targeted enrichment of the HLA region followed by sequencing (Immel et al., 2021; Krause-Kyora et al., 2018; Barquera et al., 2020). However, there is limited evidence on whether computational HLA typing methods can be reliably applied to Whole Genome Sequencing (WGS) data without HLA-specific enrichment (Plascencia et al., 2025). To explore whether pangenome mapping can help address this limitation, we genotyped HLA class I genes using reads from both linear and pangenome alignments and compared the results (see *Figure 5* in the Results section). Our findings show a significant improvement in genotype prediction when using pangenome-aligned reads, as indicated by prediction metrics such as objective value and read depth. However, our samples had unusually high coverage for aDNA (see *Supplementary Table 4*). We found the genotyping approach to be reliable at 15× and 25× coverage, based on prediction metrics (objective value and read depth) and quality control in the form of coverage plots. At 5×, results were still significantly improved compared to linear alignment, although with reduced overall reliability (see *Supplementary Figure 2*). Since our evaluation was limited to samples with at least 5× coverage, future studies should investigate performance at lower and more variable coverage levels. Taken together, these results position pangenome alignment as a promising approach for improving HLA genotyping in ancient DNA (aDNA) research. Notably, recent work by Rubin, J., et al. ,2024 has also explored this direction, advancing the field further by developing aDNA damage-aware version of the pangenome aligner *vg giraffe*. This tool represents a promising advancement, and future research should explore its potential to further improve HLA genotyping from ancient DNA.

Beyond HLA genotyping, we also explored the broader potential of pangenome alignment for variant detection, focusing specifically on pathogenic variants identified in our ancient individuals. This analysis suggested that there is no evidence for a genetic predisposition to disease in all Llullaillaco individuals, as the pathogenic variants identified were either common among individuals with Native American ancestry or lacked strong clinical evidence for a clear disease association. This aligns with previous reports that the Inca selected healthy, high-status children for sacrifice in the capacocha ritual (Faux, J. L., 2012). The pangenome detected two variants that were completely missed by the linear genome, highlighting its potential to identify variants that linear mapping fails to capture. However, overall, more variants were detected using the linear

genome. This discrepancy is not due to the pangenome's inability to detect these variants, but rather because they did not pass read depth and allele balance filters.

Finally, I would like to acknowledge several important considerations and limitations of this study. The analysis was conducted on a small number of exceptionally well-preserved individuals with relatively high sequencing coverage (5x–25x). While this allowed for a controlled and informative comparison, it limits the generalizability of the findings. Many aDNA samples exhibit lower coverage and higher levels of degradation, which could affect the relative performance of linear and pangenome-based methods. Expanding future analyses to include a broader range of individuals, populations, and time periods will be crucial for assessing the robustness and broader applicability of the observed results. The pangenome alignment pipeline used in this study is still in its early stages of development. Improvements such as support for shorter read lengths, beyond the current 40 bp minimum, would make it more suitable for highly fragmented aDNA. Furthermore, the computational requirements are substantial. Typical whole-genome runs required 64–100 GB of memory, which may pose practical challenges. It is important to note that while the pangenome offers a more inclusive representation of human genetic diversity, reference bias can still persist. This is partly because European ancestry in pangenome reference is represented only by the GRCh38 assembly, whereas most of the added diversity stems from African, Asian, and South American populations. As a result, the benefits of pangenome alignment may be less pronounced for individuals of European descent. We observed this limitation during HLA class I genotyping of ancient DNA samples with European ancestry, where the use of the pangenome did not yield substantial improvements over the linear reference.

Ultimately, while challenges remain, the observed improvements underscore the transformative potential of pangenome references to reshape analytical standards in ancient DNA research, enabling more comprehensive, accurate, and inclusive reconstructions of human history and evolution.

5 Conclusion

This thesis highlights the advantages of using a graph-based pangenome reference for ancient DNA analysis, particularly in improving alignment accuracy across highly variable genomic regions. Through comparative analyses of high-altitude Inca mummy genomes, I demonstrate that pangenome-based alignments enable broader variant detection and improved coverage across multiple genomic regions, including those that are medically relevant and technically challenging. Most notably, we observed a significant improvement in HLA class I genotyping accuracy when using a pangenome reference.

These findings underscore the practical value of pangenome references in enhancing the resolution, sensitivity, and reliability of ancient genomic studies. As the field progresses, integrating diverse and representative reference models such as the pangenome will be essential for deepening our understanding of human population history, disease evolution, and genetic adaptation.

While certain limitations and methodological considerations remain, this work provides a compelling foundation for the continued development and broader application of pangenome-based strategies in ancient DNA research.

References:

1. Sawchuk, E. A., Potts, D. T., Harkness, E., Neelis, J., & McIntosh, R. J. (2021). Ancient DNA. The Encyclopedia of ancient history: Asia and Africa, 1-9.
2. Pääbo, S. (1985). Molecular cloning of ancient Egyptian mummy DNA. nature, 314(6012), 644-645.
3. Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., & Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. Nature, 312(5991), 282-284.
4. Lan, T., & Lindqvist, C. (2019). Paleogenomics: genome-scale analysis of ancient DNA and population and evolutionary genomic inferences. Population Genomics: concepts, approaches and applications, 323-360.
5. Xiang, Q. Y., Zuo, K. L., & Liu, H. (2024). Historical retrospects and evolutionary characteristics of paleogenomic science and technology. Hist Philos Med, 6(2), 8.
6. Meyer, M., Arsuaga, J. L., De Filippo, C., Nagel, S., Aximu-Petri, A., Nickel, B., ... & Pääbo, S. (2016). Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature, 531(7595), 504-507.
7. Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., ... & Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature, 468(7327), 1053-1060.
8. Kerner, G., Choin, J., & Quintana-Murci, L. (2023). Ancient DNA as a tool for medical research. Nature medicine, 29(5), 1048-1051.
9. Choin, J., & Quintana-Murci, L. (2022). Paleogenomics: The demographic past of prehistoric Europeans. Current Biology, 32(11), R535-R538.
10. Ottoni, C., Bekaert, B., & Decorte, R. (2017). DNA degradation: current knowledge and progress in DNA analysis. Taphonomy of Human Remains: Forensic Analysis of the Dead and the Depositional Environment: Forensic Analysis of the Dead and the Depositional Environment, 65-80.
11. Dolenz, S., van der Valk, T., Jin, C., Oppenheimer, J., Sharif, M. B., Orlando, L., ... & Heintzman, P. D. (2024). Unravelling reference bias in ancient DNA datasets. Bioinformatics, 40(7), btae436.
12. Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C. Y., Popejoy, A. B., Periyasamy, S., ... & Duncan, L. E. (2019). Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. Cell, 179(3), 589-603.
13. Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., ... & Hofreiter, M. (2004). Genetic analyses from ancient DNA. Annu. Rev. Genet., 38(1), 645-679.
14. Dabney, J., Meyer, M., & Pääbo, S. (2013). Ancient DNA damage. Cold Spring Harbor perspectives in biology, 5(7), a012567.
15. Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., ... & Warinner, C. (2021). Ancient DNA analysis. Nature reviews methods primers, 1(1), 14.

16. Lindahl, T. (1993). Instability and decay of the primary structure of DNA. nature, 362(6422), 709-715.
17. Briggs, A. W., Stenzel, U., Johnson, P. L., Green, R. E., Kelso, J., Prüfer, K., ... & Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. Proceedings of the National Academy of Sciences, 104(37), 14616-14621.
18. Raffone, C., Baeta, M., Lambacher, N., Granizo-Rodríguez, E., Etxeberria, F., & de Pancorbo, M. M. (2021). Intrinsic and extrinsic factors that may influence DNA preservation in skeletal remains: A review. Forensic Science International, 325, 110859.
19. Eriksen, A. M. H., Rodríguez, J. A., Seersholm, F., Hollund, H. I., Gotfredsen, A. B., Collins, M. J., ... & Matthiesen, H. (2025). Exploring DNA degradation in situ and in museum storage through genomics and metagenomics. Communications Biology, 8(1), 210.
20. Bollongino, R., Tresset, A., & Vigne, J. D. (2008). Environment and excavation: Pre-lab impacts on ancient DNA analyses. Comptes Rendus Palevol, 7(2-3), 91-98.
21. Adler, C. J., Haak, W., Donlon, D., Cooper, A., & Genographic Consortium. (2011). Survival and recovery of DNA from ancient teeth and bones. Journal of Archaeological Science, 38(5), 956-964.
22. Pinhasi, R., Fernandes, D., Sirak, K., Novak, M., Connell, S., Alpaslan-Roodenberg, S., ... & Hofreiter, M. (2015). Optimal ancient DNA yields from the inner ear part of the human petrous bone. PloS one, 10(6), e0129102.
23. Lan, T., & Lindqvist, C. (2019). Technical advances and challenges in genome-scale analysis of ancient DNA. Paleogenomics: genome-scale analysis of ancient DNA, 3-29.
24. Hofreiter, M., Paijmans, J. L., Goodchild, H., Speller, C. F., Barlow, A., Fortes, G. G., ... & Collins, M. J. (2015). The future of ancient DNA: Technical advances and conceptual shifts. BioEssays, 37(3), 284-293.
25. Gansauge, M. T., & Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nature protocols, 8(4), 737-748.
26. Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols, 2010(6), pdb-prot5448.
27. Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., ... & Willerslev, E. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature, 463(7282), 757-762.
28. Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., ... & Willerslev, E. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature, 499(7456), 74-78.
29. Raghavan, M., Skoglund, P., Graf, K. E., Metspalu, M., Albrechtsen, A., Moltke, I., ... & Willerslev, E. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature, 505(7481), 87-91.
30. Mascher, M., Schuenemann, V. J., Davidovich, U., Marom, N., Himmelbach, A., Hübner, S., ... & Stein, N. (2016). Genomic analysis of 6,000-year-old cultivated

grain illuminates the domestication history of barley. Nature Genetics, 48(9), 1089-1093.
31. Enk, J. M., Devault, A. M., Kuch, M., Murgha, Y. E., Rouillard, J. M., & Poinar, H. N. (2014). Ancient whole genome enrichment using baits built from modern DNA. Molecular biology and evolution, 31(5), 1292-1294.
32. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., & Orlando, L. (2013). mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics, 29(13), 1682-1684.
33. Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H. A., Kelso, J., & Pääbo, S. (2013). DNA analysis of an early modern human from Tianyuan Cave, China. Proceedings of the National Academy of Sciences, 110(6), 2223-2227.
34. Renaud, G., Slon, V., Duggan, A. T., & Kelso, J. (2015). Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome biology, 16, 1-18.
35. Poullet, M., & Orlando, L. (2020). Assessing DNA sequence alignment methods for characterizing ancient genomes and methylomes. Front Ecol Evol. 2020: 8: 105.
36. Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., Al-Rasheid, K. A., Willerslev, E., ... & Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. BMC genomics, 13, 1-15.
37. Oliva, A., Tobler, R., Cooper, A., Llamas, B., & Souilmi, Y. (2021). Systematic benchmark of ancient DNA read mapping. Briefings in Bioinformatics, 22(5), bbab076.
38. Childebayeva, A., & Zavala, E. I. (2023). Computational analysis of human skeletal remains in ancient DNA and forensic genetics. Iscience, 26(11).
39. Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome research, 19(9), 1655-1664.
40. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., ... & Reich, D. (2012). Ancient admixture in human history. Genetics, 192(3), 1065-1093.
41. Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS genetics, 15(7), e1008302.
42. Martiniano, R., Garrison, E., Jones, E. R., Manica, A., & Durbin, R. (2020). Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. Genome biology, 21, 1-18.
43. Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., ... & Garrison, E. (2020). Pangenome graphs. Annual review of genomics and human genetics, 21(1), 139-162.
44. Sherman, R. M., & Salzberg, S. L. (2020). Pan-genomics in the human genome era. Nature Reviews Genetics, 21(4), 243-254.
45. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., ... & Human Pangenome Reference Consortium. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature, 604(7906), 437-446.

46. 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526(7571), 68.
47. Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., ... & Salzberg, S. L. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nature genetics, 51(1), 30-35.
48. Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., ... & Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. Nature, 526(7571), 75-81.
49. Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research, 42(D1), D980-D985.
50. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research, 29(1), 308-311.
51. Hein, J. (1989). A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. Molecular Biology and Evolution, 6(6), 649-668.
52. Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., ... & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature biotechnology, 36(9), 875-879.
53. Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., ... & Paten, B. (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus. Nature biotechnology, 42(4), 663-673.
54. Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. Genome biology, 21, 1-19.
55. Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., ... & Paten, B. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature, 587(7833), 246-251.
56. Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., ... & Prins, P. (2024). Building pangenome graphs. Nature Methods, 1-5.
57. Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics, 32(14), 2103-2110.
58. Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature biotechnology, 37(8), 907-915.
59. Rautiainen, M., & Marschall, T. (2020). GraphAligner: rapid and versatile sequence-to-graph alignment. Genome biology, 21(1), 253.
60. Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., ... & Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312-324.
61. "Frozen Mummies of the Andes." *Expedition Magazine* 58, no. 2 (September, 2016): -. Accessed June 25, 2025. https://www.penn.museum/sites/expedition/frozen-mummies-of-the-andes/

62. Besom, T. (2009). *Of mummies, mountains, and immolations: An ethnohistoric study of human sacrifice and mountain worship among the Inkas*. University of Texas Press.
63. D'Altroy, T. N. (2015). Chapter 7. Funding the Inka Empire. In *The Inka Empire: a multidisciplinary approach* (pp. 97-118). University of Texas Press.
64. Reinhard, J., & Ceruti, M. C. (2010). *Inca rituals and sacred mountains: a study of the world's highest archaeological sites*.
65. González Holguín, D. (1993). Vocabulario de la lengua general de todo el Perú, llamada lengua quichua, o del Inca.
66. Clinker, S. (2022). The Inca Capacocha Ritual: Sacred, Social and Political Ties to Landscape. *Fields| Terrains*, *12*, 9.
67. Reinhard, J., & Ceruti, M. C. (2010). *Inca rituals and sacred mountains: a study of the world's highest archaeological sites*.
68. Rowe, J. H. (1946). Inca culture at the time of the Spanish conquest. US Government Printing Office.
69. Schobinger, J. (2001). El santuario incaico del cerro Aconcagua (No. 25). EDIUNC.
70. Castro, M., Busel, D., Camargo, C., Moraga, M., Llop, E., & Durán, E. (2005). The prince of el Plomo: new studies 50 years after its discovery. Journal of Biological Research-Bollettino della Società Italiana di Biologia Sperimentale, 80(1).
71. Previgliano, C. H., Ceruti, C., Reinhard, J., Araoz, F. A., & Diez, J. G. (2003). Radiologic evaluation of the Llullaillaco mummies. American Journal of Roentgenology, 181(6), 1473-1479.
72. Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., ... & Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences*, *110*(39), 15758-15763.
73. Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb-prot5448.
74. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), 10-12.
75. Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593
76. Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
77. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, *25*(14), 1754-1760.
78. Kokot, M., Długosz, M., & Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17), 2759–2761. https://doi.org/10.1093/bioinformatics/btx304
79. Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., ... & Paten, B. (2018). Variation graph toolkit improves read mapping by representing

genetic variation in the reference. *Nature Biotechnology*, 36(9), 875–879. https://doi.org/10.1038/nbt.4227

80. Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., ... & Paten, B. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, 374(6574), abg8871. https://doi.org/10.1126/science.abg8871
81. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
82. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297-1303.
83. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033
84. Mose, L. E., Wilkerson, M. D., & Hayes, D. N. (2019). ABRA2: Improved coding indel detection via assembly-based realignment. *Bioinformatics*, 35(16), 2966–2973. https://doi.org/10.1093/bioinformatics/btz035
85. Picard-Tools (2021) Github repository. http://broadinstitute.github.io/picard/.
86. Core Team, R.

    R: A Language and Environment for Statistical Computing

    R Foundation for Statistical Computing, 2013

87. Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., & Kohlbacher, O. (2014). OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*, 30(23), 3310–3316. https://doi.org/10.1093/bioinformatics/btu548
88. Wu, Z., Li, T., Jiang, Z., Zheng, J., Gu, Y., Liu, Y., ... & Xie, Z. (2024). Human pangenome analysis of sequences missing from the reference genome reveals their widespread evolutionary, phenotypic, and functional roles. Nucleic Acids Research, 52(5), 2212-2230.
89. Hickey, Glenn, et al. "Genotyping structural variants in pangenome graphs using the vg toolkit." *Genome biology* 21 (2020): 1-17.
90. Sirén, Jouni, et al. "Pangenomics enables genotyping of known structural variants in 5202 diverse genomes." *Science* 374.6574 (2021): abg8871.
91. Groza, C., Schwendinger-Schreck, C., Cheung, W. A., Farrow, E. G., Thiffault, I., Lake, J., ... & Pastinen, T. (2024). Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nature communications*, *15*(1), 657.
92. Wagner, J., Olson, N. D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., ... & Sedlazeck, F. J. (2022). Curated variation benchmarks for challenging medically relevant autosomal genes. *Nature biotechnology*, *40*(5), 672-680.

93. Cruz-Tapias, P., Castiblanco, J., & Anaya, J. M. (2013). HLA association with autoimmune diseases. In *Autoimmunity: From Bench to Bedside [Internet]*. El Rosario University Press.
94. Medhasi, S., & Chantratita, N. (2022). Human leukocyte antigen (HLA) system: genetics and association with bacterial and viral infections. *Journal of Immunology Research*, *2022*(1), 9710376.
95. Plascencia, A. G., Jakobsson, M., & Sánchez-Quinto, F. (2025). Ancient DNA HLA typing reveals significant shifts in frequency in Europe since the Neolithic. *Scientific Reports*, *15*(1), 6161.
96. Yip, V. L. M., & Pirmohamed, M. (2017). The HLA-A* 31: 01 allele: influence on carbamazepine treatment. *Pharmacogenomics and personalized medicine*, 29-38.
97. Dean, L. (2018). Carbamazepine therapy and HLA genotype.
98. Amstutz, U., Ross, C. J., Castro-Pastrana, L. I., Rieder, M. J., Shear, N. H., Hayden, M. R., ... & CPNDS Consortium. (2013). HLA-A* 31: 01 and HLA-B* 15: 02 as genetic markers for carbamazepine hypersensitivity in children. Clinical Pharmacology & Therapeutics, 94(1), 142-149.
99. Gul, A., & Ohno, S. (2012). HLA-B* 51 and Behçet disease. *Ocular immunology and inflammation*, *20*(1), 37-43.
100. Khoshbakht, S., Başkurt, D., Vural, A., & Vural, S. (2023). Behçet's disease: A comprehensive review on the role of HLA-B* 51, antigen presentation, and inflammatory cascade. International Journal of Molecular Sciences, 24(22), 16382.
101. Wallace, G. R. (2014). HLA-B* 51 the primary risk in Behcet disease. Proceedings of the National Academy of Sciences, 111(24), 8706-8707.
102. Langtry, A., Rabadan, R., Alonso, L., van Eijck, C., Macarulla, T., Lawlor, R. T., ... & López de Maturana, E. (2024). HLA-A* 02: 01 allele is associated with decreased risk and a longer survival in pancreatic cancer: Results from an exhaustive analysis of the HLA variation in PDAC. *medRxiv*, 2024-08.
103. Shi, Y. W., Min, F. L., Zhou, D., Qin, B., Wang, J., Hu, F. Y., ... & Liao, W. P. (2017). HLA-A* 24: 02 as a common risk factor for antiepileptic drug–induced cutaneous adverse reactions. *Neurology*, *88*(23), 2183-2191.
104. Dolton, G., Bulek, A., Wall, A., Thomas, H., Hopkins, J. R., Rius, C., ... & Sewell, A. K. (2024). HLA A* 24: 02–restricted T cell receptors cross-recognize bacterial and preproinsulin peptides in type 1 diabetes. *Journal of Clinical Investigation*, *134*(18), e164535.
105. Bagias, C., Xiarchou, A., Bargiota, A., & Tigas, S. (2020). Familial partial lipodystrophy (FPLD): recent insights. Diabetes, Metabolic Syndrome and Obesity, 1531-1544.
106. Rallidis, L. S., Iordanidis, D., & Iliodromitis, E. (2020). The value of physical signs in identifying patients with familial hypercholesterolemia in the era of genetic testing. Journal of cardiology, 76(6), 568-572.

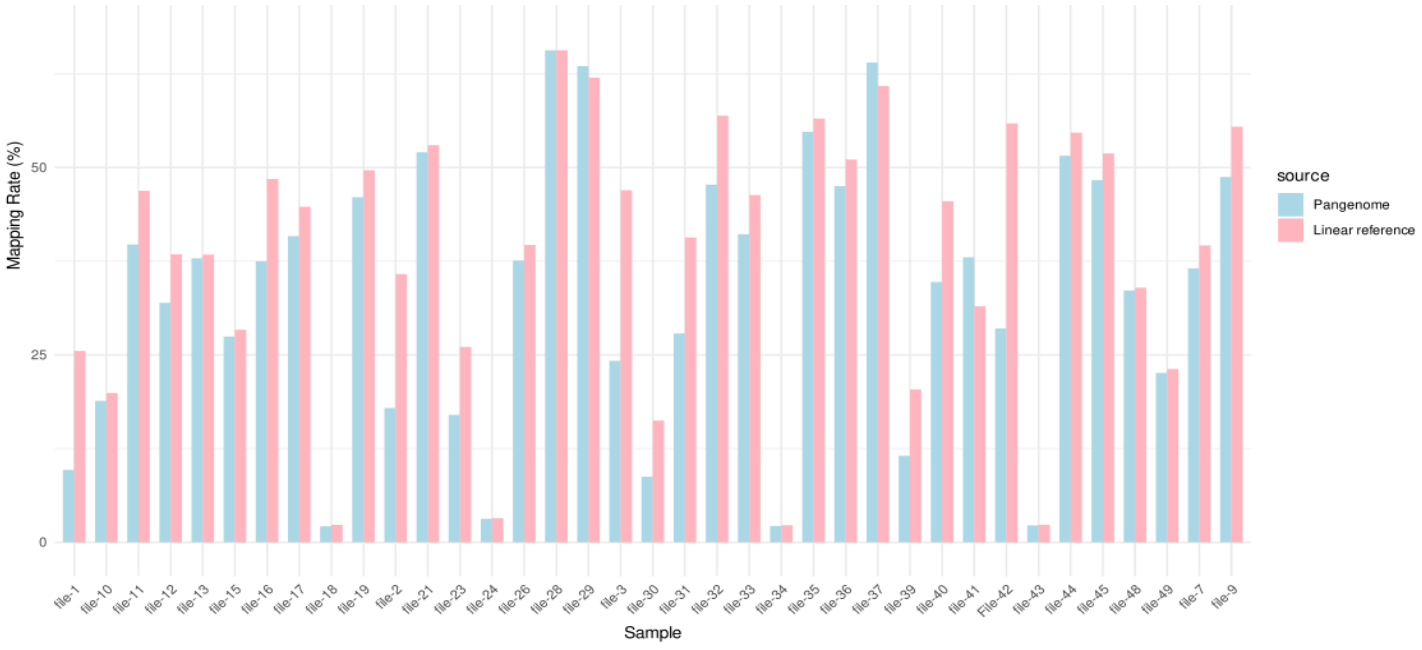
107. Dolenz, S., van der Valk, T., Jin, C., Oppenheimer, J., Sharif, M. B., Orlando, L., ... & Heintzman, P. D. (2024). Unravelling reference bias in ancient DNA datasets. Bioinformatics, 40(7), btae436.
108. Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS genetics, 15(7), e1008302.
109. Poznyak, A. V., Orekhov, N. A., Kovyanova, T. I., Starodubtseva, I. A., Elizova, N. V., Sukhorukov, V. N., & Orekhov, A. N. (2024). Ancient DNA studies: Common limitations and Genotyping. Journal of Angiotherapy, 8(6), 1-8.
110. Taylor, D. J., Eizenga, J. M., Li, Q., Das, A., Jenike, K. M., Kenny, E. E., ... & Schatz, M. C. (2024). Beyond the human genome project: the age of complete human genome sequences and pangenome references. Annual review of genomics and human genetics, 25.
111. Abondio, P., Cilli, E., & Luiselli, D. (2023). Human pangenomics: promises and challenges of a distributed genomic reference. Life, 13(6), 1360.
112. Plascencia, A. G., Jakobsson, M., & Sánchez-Quinto, F. (2025). Ancient DNA HLA typing reveals significant shifts in frequency in Europe since the Neolithic. Scientific Reports, 15(1), 6161.
113. Immel, A., Key, F. M., Szolek, A., Barquera, R., Robinson, M. K., Harrison, G. F., ... & Krause, J. (2021). Analysis of genomic DNA from medieval plague victims suggests long-term effect of Yersinia pestis on human immunity genes. Molecular biology and evolution, 38(10), 4059-4076.
114. Krause-Kyora, B., Nutsua, M., Boehme, L., Pierini, F., Pedersen, D. D., Kornell, S. C., ... & Nebel, A. (2018). Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. Nature communications, 9(1), 1569.
115. Barquera, R., Lamnidis, T. C., Lankapalli, A. K., Kocher, A., Hernández-Zaragoza, D. I., Nelson, E. A., ... & Krause, J. (2020). Origin and health status of first-generation Africans from early colonial Mexico. Current Biology, 30(11), 2078-2091.
116. Rubin, J., van Waaij, J., Kraft, L., Sirén, J., Sackett, P. W., & Renaud, G. (2024). SAFARI: Pangenome Alignment of Ancient DNA Using Purine/Pyrimidine Encodings. bioRxiv.
117. Faux, J. L. (2012). Hail the conquering gods: Ritual sacrifice of children in Inca society. Journal of contemporary anthropology, 3(1), 1.

6 Supplements

6.1 Supplementary Figures

6.1.1 Mapping Rates for Tissue - Derived Samples

This supplementary figure shows the mapping rates from both pangenome and linear alignments for samples coming from multiple tissue types (see Figure 1B). The data shown in this supplementary figure correspond to the values summarized in Supplementary Table 1 (linear alignment) and Supplementary Table 2 (pangenome alignment), which were also used to generate the box plot in Figure 2B.



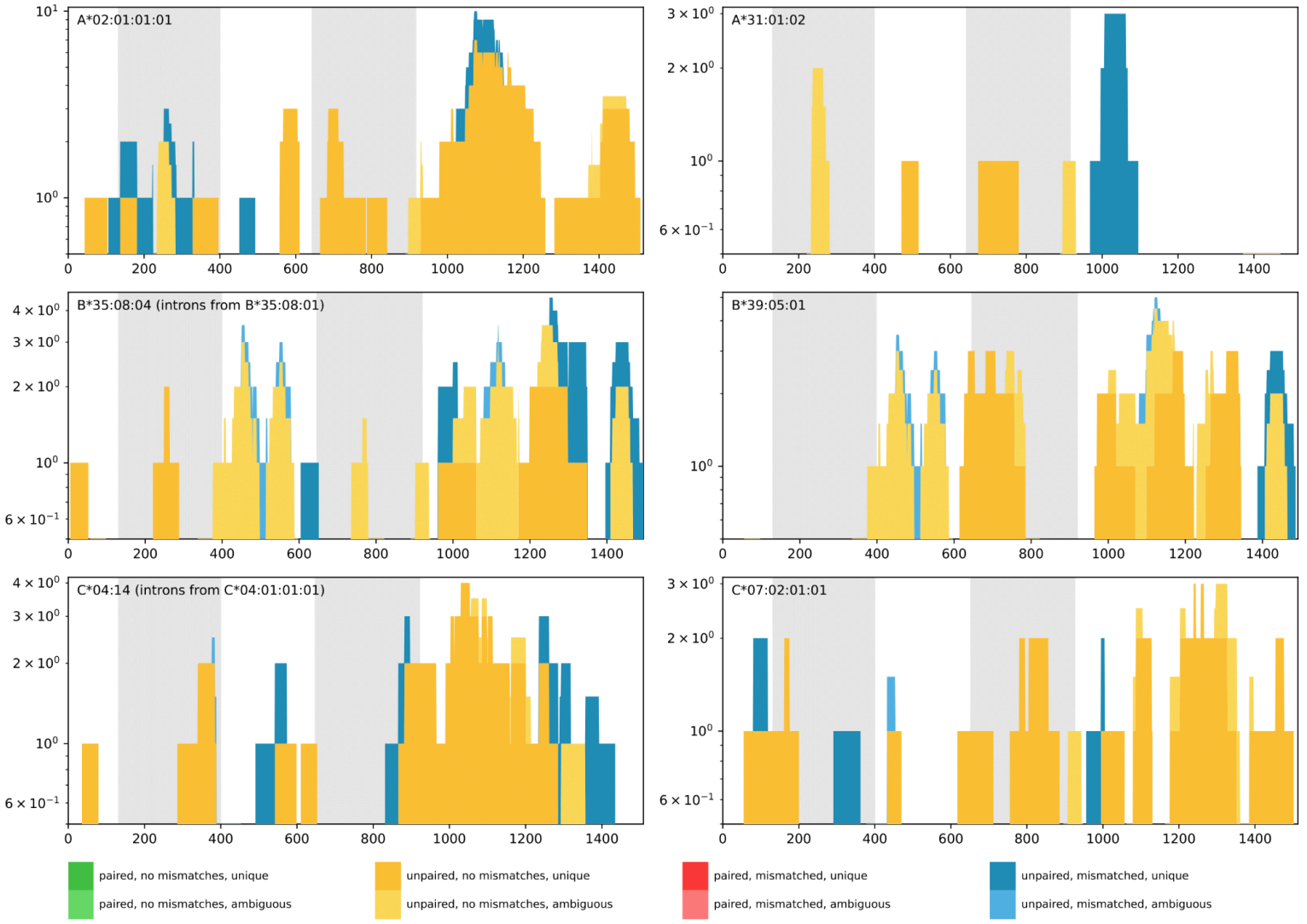
Supplementary Figure 1 - Mapping Rate Comparison for Tissue-Derived Samples

Bar plot showing the mapping rates for individual files corresponding to single tissue-derived samples. Each file (x-axis) represents one sample that was later merged into the final alignment per individual. Mapping rates are shown for both the pangenome and the linear reference genome. The y-axis indicates the mapping rate, calculated as the percentage of reads that successfully aligned out of the total number of raw reads.

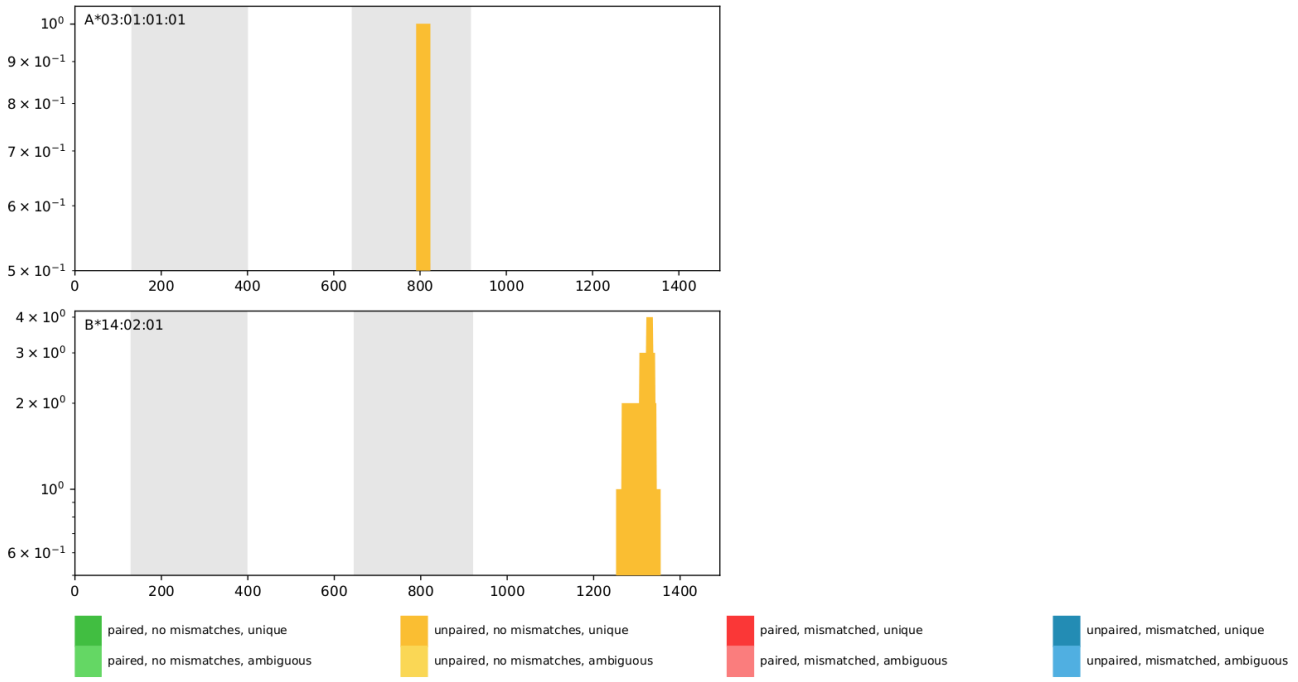
6.1.2 Coverage Plots of the Predicted HLA genotypes

The coverage plots shown illustrate the number of reads aligning to each position of the predicted HLA genotypes, presented in Figure 5C. These plots were generated alongside the genotype predictions produced by the OptiType software.

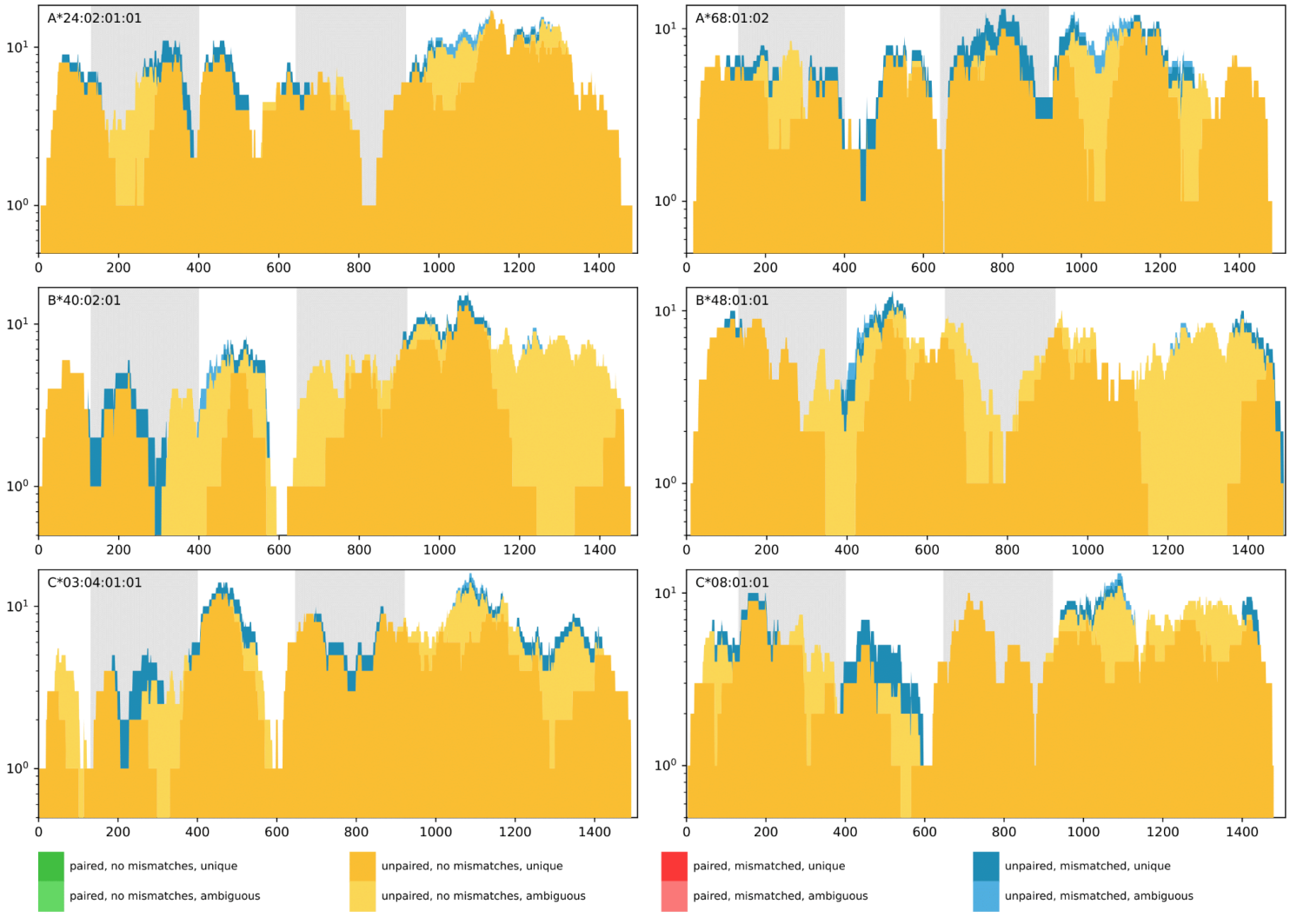
A1



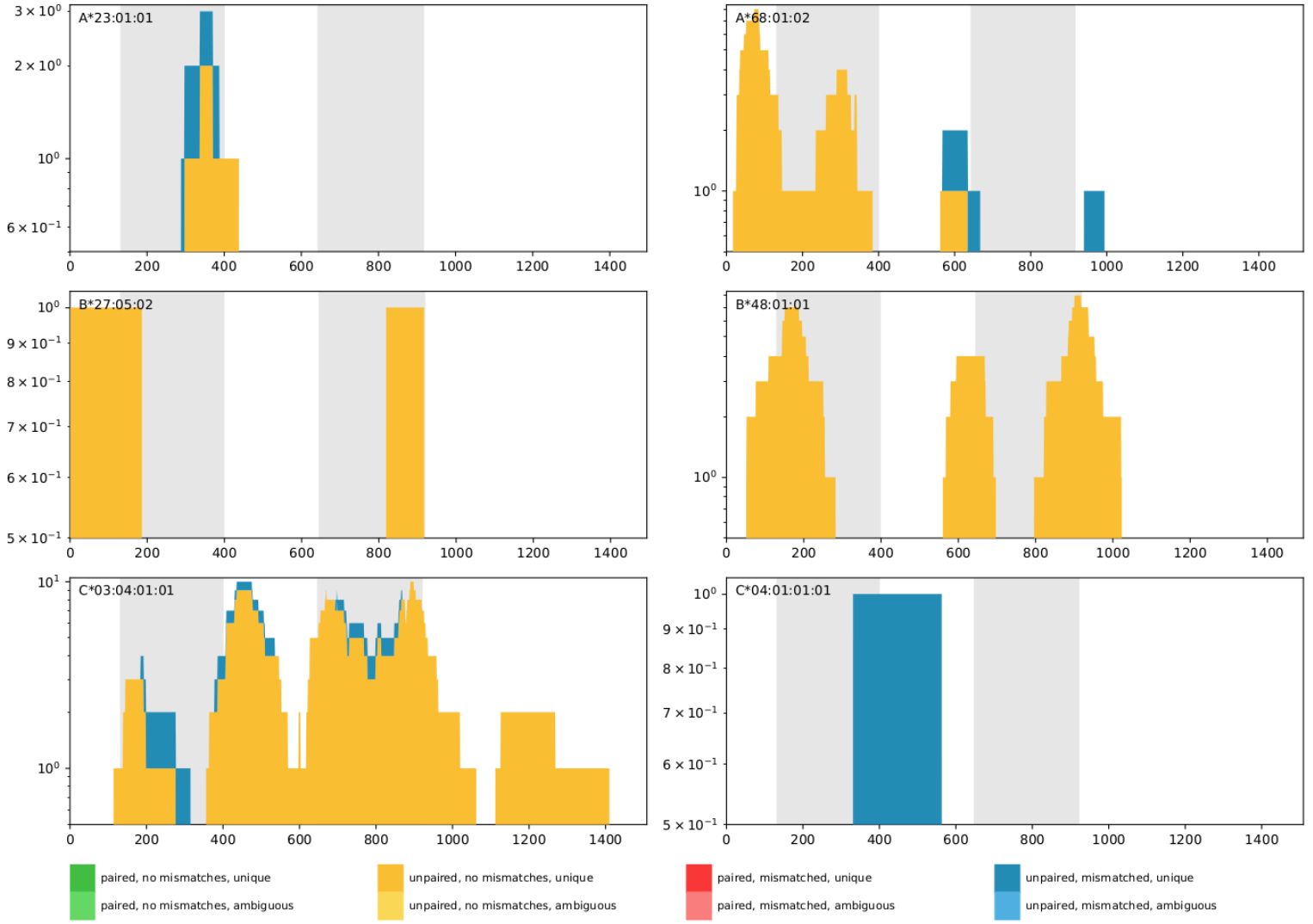
A2


A*03:01:01:01
B*14:02:01
paired, no mismatches, unique
paired, no mismatches, ambiguous
unpaired, no mismatches, unique
unpaired, no mismatches, ambiguous
paired, mismatched, unique
paired, mismatched, ambiguous
unpaired, mismatched, unique
unpaired, mismatched, ambiguous

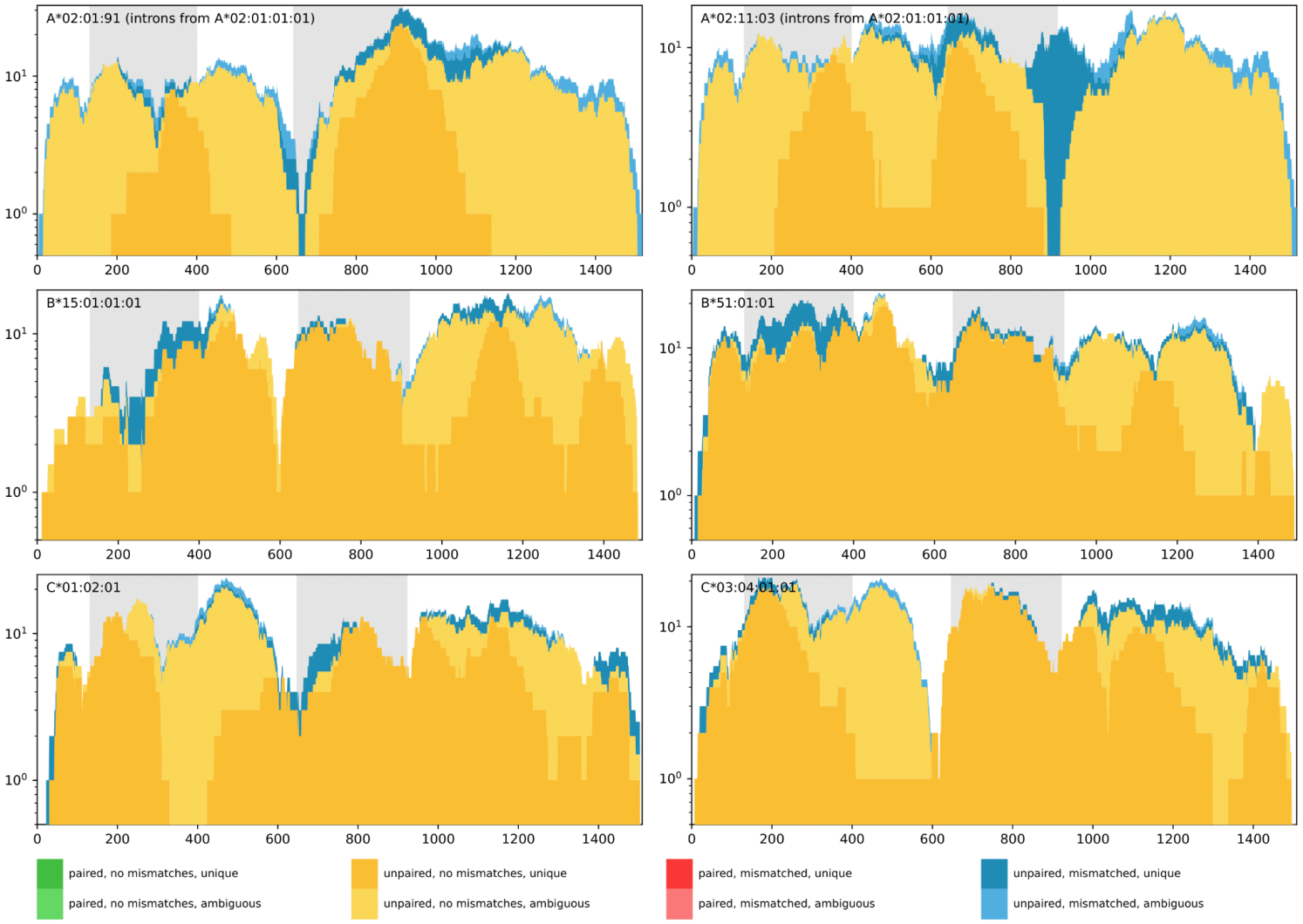
B1


A*24:02:01:01
A*68:01:02
B*40:02:01
B*48:01:01
C*03:04:01:01
C*08:01:01
paired, no mismatches, unique
paired, no mismatches, ambiguous
unpaired, no mismatches, unique
unpaired, no mismatches, ambiguous
paired, mismatched, unique
paired, mismatched, ambiguous
unpaired, mismatched, unique
unpaired, mismatched, ambiguous

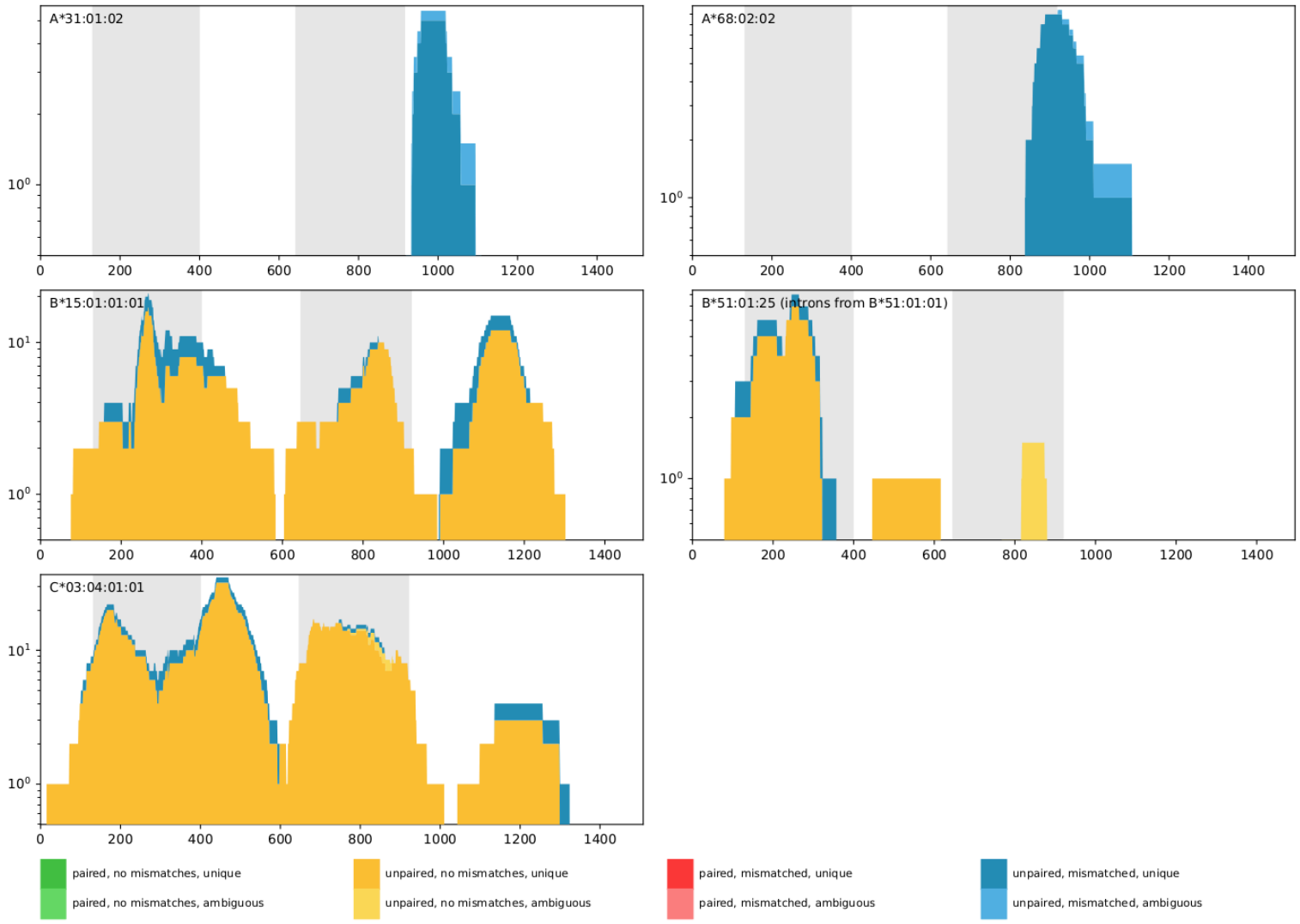
B2


A*23:01:01
A*68:01:02
B*27:05:02
B*48:01:01
C*03:04:01:01
C*04:01:01:01
paired, no mismatches, unique
paired, no mismatches, ambiguous
unpaired, no mismatches, unique
unpaired, no mismatches, ambiguous
paired, mismatched, unique
paired, mismatched, ambiguous
unpaired, mismatched, unique
unpaired, mismatched, ambiguous

C1


A*02:01:91 (introns from A*02:01:01:01)
A*02:11:03 (introns from A*02:01:01:01)
B*15:01:01:01
B*51:01:01
C*01:02:01
C*03:04:01:01
paired, no mismatches, unique
paired, no mismatches, ambiguous
unpaired, no mismatches, unique
unpaired, no mismatches, ambiguous
paired, mismatched, unique
paired, mismatched, ambiguous
unpaired, mismatched, unique
unpaired, mismatched, ambiguous

C2



Supplementary Figure 2 – Coverage Plots of Predicted HLA Genotypes

The plots display read coverage across the predicted HLA gene sequences, with color-coding indicating read type (paired/unpaired), presence of mismatches, and whether alignments are unique or ambiguous.

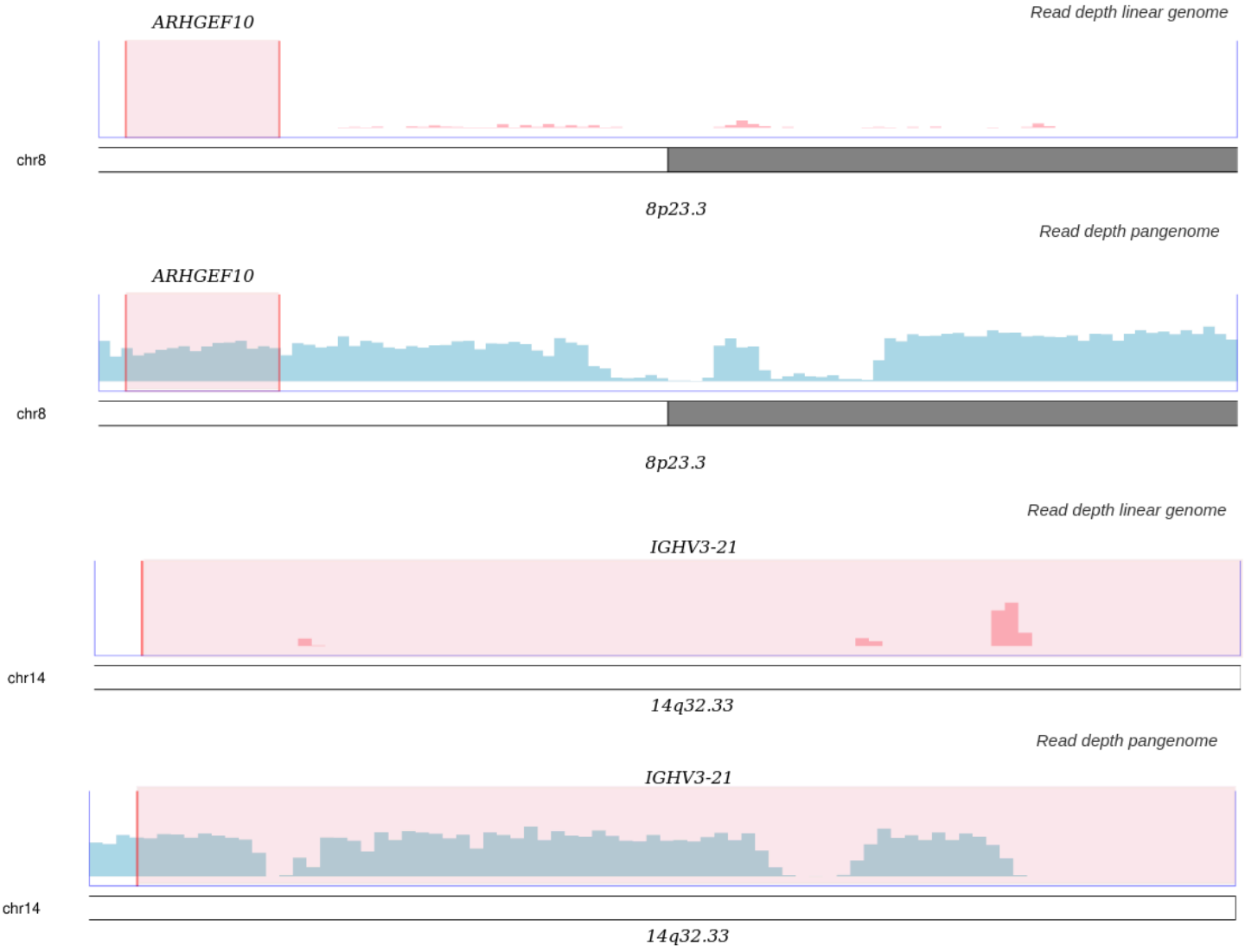
A) Coverage plot for La Niña del Rayo *(A1 pangenome, A2 lineare genome alignment)*

B) Coverage plot for El Niño *(B1 pangenome, B2 lineare genome alignment)*

C) Coverage plot for La Doncella *(C1 pangenome, C2 lineare genome alignment)*

6.1.3 Visualization of Coverage for Medically Relevant Genes

This figure illustrates the read coverage for two selected medically relevant genes (see Figure 4.1), where coverage differs significantly between the pangenome and linear reference alignments. The plots visually highlight the improved mapping performance achieved using the pangenome reference compared to the linear genome.



Supplementary Figure 3 - Visualization of Coverage for Medically Relevant Genes

Karyoplots of gene ARHGEF10 and IGHV3-21 showing read depth across 10 kb windows La Doncella aligned to the linear reference genome (pink) and the pangenome reference (blue)

6.2 Supplementary Tables

6.2.1 Supplementary Table 1

GRCh38#0#chr1
GRCh38#0#chr2
GRCh38#0#chr3
GRCh38#0#chr4
GRCh38#0#chr5
GRCh38#0#chr6
GRCh38#0#chr7
GRCh38#0#chr8
GRCh38#0#chr9
GRCh38#0#chr10
GRCh38#0#chr11
GRCh38#0#chr12
GRCh38#0#chr13
GRCh38#0#chr14
GRCh38#0#chr15
GRCh38#0#chr16
GRCh38#0#chr17
GRCh38#0#chr18
GRCh38#0#chr19
GRCh38#0#chr20
GRCh38#0#chr21
GRCh38#0#chr22
GRCh38#0#chrX
GRCh38#0#chrY
GRCh38#0#chrM

Supplementary Table 1 - Reference Paths Used for Linear Projection

6.2.2 Supplementary Table 2

LINEAR MAPPING												
File name		ENA_Submission	Extract-ID	Library-ID	Samp	Sample	Raw seq data	Mapped read	Unique reads	Mapped read	Mapping rate	Mapping rate (postfiltering)
file-1	Ind_2361	NinaDelRayo_Hair_SHALLOW_SEQ	2853	2361	hair	1E_hair_HAI	17722387	8287275	4525570	6918473	46.76161851	25.5358942336605
file-2	Ind_2362	Nino_Hair_SHALLOW_SEQ	2854	2362	hair	2A_hair_HAI	14912619	8159847	5336755	7143344	54.71773268	35.786839320444
file-3	Ind_2363	Doncella_Hair_SHALLOW_SEQ	2855	2363	hair	3E_hair_HAI	17475803	12127031	8210334	10978764	69.39326908	46.9811544568224
file-7	Ind_2367	Nino_Foot_SHALLOW_SEQ	2859	2367	skin	2F_SKIN_FO	15239895	9067599	6037201	8647023	59.49909104	39.6144527242478
file-9	Ind_2369	Doncella_Mouth_SHALLOW_SEQ	2861	2369	skin	3G_SKIN_MO	14910102	11844063	8270233	11226218	79.43649882	55.4673133691507
file-10	Ind_2370	Doncella_Toe_SHALLOW_SEQ	2862	2370	skin	3h_Skineppi	11413156	4393415	2274498	4021931	38.49430429	19.9287383787622
file-11	Ind_2371	Nino_Arm_SHALLOW_SEQ	2863	2371	skin	2d_skinbag_A	15514003	11221243	7278034	10589449	72.32977201	46.9126762448093
file-12	Ind_2372	Doncella_Knee_SHALLOW_SEQ	2864	2372	skin	3d_skinbag_H	11937041	6581032	4587644	6191962	55.13118368	38.4320033750408
file-13	Ind_2373	NinaDelRayo_Mouth_SHALLOW_SEQ	2865	2373	skin	1d_swab_MO	8175091	4698243	3138147	4294567	57.47022266	38.3866919646521
file-15	Ind_2375	NinaDelRayo_Nose_SHALLOW_SEQ	2867	2375	swab	1g_swab_NO	12628638	5709077	3582367	5100199	45.20738499	28.3670099657619
file-16	Ind_2376	NinaDelRayo_Anus_SHALLOW_SEQ	2868	2376	swab	1h_swab_AN	19595842	13659369	9501347	12923950	69.70544568	48.4865462785421
file-17	Ind_2377	Nino_Ear_1_SHALLOW_SEQ	2869	2377	swab	2b_swab_EA	13740199	8773769	6152496	8335646	63.85474475	44.7773427444537
file-18	Ind_2378	Nino_Anus_SHALLOW_SEQ	2870	2378	swab	2c_swab_AN	13930025	472295	325472	442763	3.39048207	2.33647821881152
file-19	Ind_2379	Nino_Ear_2_SHALLOW_SEQ	2871	2379	swab	2e_swab_EA	13785501	9793020	6845539	9367876	71.03854985	49.6575278620632
file-21	Ind_2381	Doncella_Mouth_SHALLOW_SEQ	2873	2381	swab	3b_swab_MO	20149173	16069189	10677855	15419279	79.75110939	52.9940112182272
file-23		Nino_Hair_1_DEEP_SEQ	2854	2537	hair	2A_hair_HAI	694	214	181	181	30.83573487	26.0806916426513
file-24		Nino_Anus_1_DEEP_SEQ	3054	2538	swab	2c_swab_AN	255118274	9245982	8206050	8826849	3.624194322	3.2165669167235
file-26		Nino_Ear_1_DEEP_SEQ	2869	2540	swab	2b_swab_EA	267503149	132035330	106146497	124993124	49.35842082	39.6804663409775
file-28		Doncella_Mouth_DEEP_SEQ	2873	2542	swab	3b_swab_MO	104972395	79069638	68920045	75644889	75.32422024	65.6553992123358
file-29		Doncella_Anus_1_DEEP_SEQ	2874	2543	swab	3c_swab_AN	310685420	223590717	192660212	215866817	71.96691657	62.011346396622
file-30		NinaDelRayo_Hair_1_DEEP_SEQ	2853	2544	hair	1E_hair_HAI	229135545	63992120	37216377	53297804	27.92762686	16.2420793334356
file-31		Doncella_Hair_1_DEEP_SEQ	2855	2545	hair	3E_hair_HAI	237831063	118976468	96750799	107662010	50.02562176	40.6804720037769
file-32		NinaDelRayo_Anus_2_DEEP_SEQ	2868	2546	swab	1h_swab_AN	95305946	64251877	54265678	60584372	67.41644115	56.9383971069339
file-33		Nino_Hair_2_DEEP_SEQ	2854	2547	hair	2A_hair_HAI	33237462	17832560	15403154	16759040	53.65199064	46.3427502376686
file-34		Nino_Anus_2_DEEP_SEQ	2870	2548	swab	2c_swab_AN	550244063	14251443	12528438	13576837	2.59002213	2.27688744730718
file-35		Nino_Ear_3_DEEP_SEQ	2871	2549	swab	2e_swab_EA	78941872	51768196	44644796	49434263	65.57761387	56.5540122990749
file-36		Nino_Ear_4_DEEP_SEQ	2869	2550	swab	2b_swab_EA	173431588	103443749	88609418	97937337	59.64527581	51.0918564615807
file-37		Doncella_Mouth_DEEP_SEQ	2873	2382	swab	3b_swab_MO	183524472	129052687	111784691	124992509	70.31906186	60.9099646394842
file-39		NinaDelRayo_Hair_2_DEEP_SEQ	2853	2361	hair	1E_hair_HAI	11857016	3703633	2417932	2959364	31.23579322	20.392415764641
file-40		Doncella_Hair_2_DEEP_SEQ	2855	2363	hair	3E_hair_HAI	396066420	228839417	180283337	212545306	57.77804061	45.5184605147793
file-42		NinaDelRayo_Anus_3_DEEP_SEQ	3057	2376	swab	1h_swab_AN	77774201	53513367	43481241	50641105	68.80606462	55.9070237185722
file-41		Nino_Hair_3_DEEP_SEQ	3053	2362	hair	2A_hair_HAI	97438976	39644795	30686651	35942880	40.68679355	31.4931993948705
file-43		Nino_Anus_3_DEEP_SEQ	2870	2378	swab	2c_swab_AN	334554182	8814115	7811749	8399357	2.63458521	2.33497275487652
file-44		Nino_Ear_5_DEEP_SEQ	2871	2379	swab	2e_swab_EA	141013514	92758539	77086656	88328238	65.77989327	54.6661478133223
file-45		Nino_Ear_6_DEEP_SEQ	3056	2377	swab	2b_swab_EA	145992525	88453737	75764724	83541993	60.58785338	51.8963035949957
file-48		NinaDelRayo_Mouth_DEEP_SEQ	2865	2373	swab	1d_swab_MO	124054994	73916553	42144538	67508339	59.58369802	33.9724638574405
file-49		NinaDelRayo_Nose_DEEP_SEQ	2867	2375	swab	1g_swab_NO	169668815	83864411	39238636	75020511	49.42830007	23.1266046149966
											52.18825871	38.3960314168248

Supplementary Table 2 - Mapping Statistics for Linear Genome Alignment per Sample

6.2.3 Supplementary Table 3

PANGENOME MAPPING												
File name		ENA_Submission	Extract-ID	Library-ID	Sample	Sample	Raw seq data	Mapped read	Unique reads	Mapped read	Mapping rate	Mapping rate (postfiltering)
file-1	Ind_2361	Ni□aDelRayo_Hair_SHALLOW_SEQ	2853	2361	hair	1E_hair_HAI	17722387	3653083	1709150	2767853	20.61281587	9.64401691487721
file-2	Ind_2362	Ni□o_Hair_SHALLOW_SEQ	2854	2362	hair	2A_hair_HAI	14912619	4408667	2672853	3588265	29.56333157	17.9234311558553
file-3	Ind_2363	Doncella_Hair_SHALLOW_SEQ	2855	2363	hair	3E_hair_HAI	17475803	6806009	4231075	5701274	38.94532915	24.2110476983518
file-7	Ind_2367	Ni□o_Foot_SHALLOW_SEQ	2859	2367	skin	2F_SKIN_FO	15239895	9090940	5572504	7970157	59.65224826	36.5652388024983
file-9	Ind_2369	Doncella_Mouth_SHALLOW_SEQ	2861	2369	skin	3G_SKIN_MO	14910102	11426343	7270179	9862096	76.63490833	48.7600889651862
file-10	Ind_2370	Doncella_Toe_SHALLOW_SEQ	2862	2370	skin	3h_Skineppi	11413156	4762119	2154800	3790664	41.72482178	18.8799662424662
file-11	Ind_2371	Ni□o_Arm_SHALLOW_SEQ	2863	2371	skin	2d_skinbag_A	15514003	10507276	6165324	8942340	67.72769091	39.7403816410246
file-12	Ind_2372	Doncella_Knee_SHALLOW_SEQ	2864	2372	skin	3d_skinbag_H	11937041	6053642	3814409	5134239	50.7130871	31.9543930526837
file-13	Ind_2373	Ni□aDelRayo_Mouth_SHALLOW_SEQ	2865	2373	skin	1d_swab_MO	8175091	5241245	3097407	4221493	64.11237502	37.8883488881041
file-15	Ind_2375	Ni□aDelRayo_Nose_SHALLOW_SEQ	2867	2375	swab	1g_swab_NO	12628638	6290507	3467664	4911486	49.81144443	27.4587330795292
file-16	Ind_2376	Ni□aDelRayo_Anus_SHALLOW_SEQ	2868	2376	swab	1h_swab_AN	19595842	12171027	7351458	9873164	62.11025278	37.5153973991013
file-17	Ind_2377	Ni□o_Ear_1_SHALLOW_SEQ	2869	2377	swab	2b_swab_EA	13740199	8819541	5614122	7580735	64.18786948	40.8591025501159
file-18	Ind_2378	Ni□o_Anus_SHALLOW_SEQ	2870	2378	swab	2c_swab_AN	13930025	457643	297019	403694	3.285299201	2.13222158610627
file-19	Ind_2379	Ni□o_Ear_2_SHALLOW_SEQ	2871	2379	swab	2e_swab_EA	13785501	9871446	6347187	8665447	71.60745192	46.0424833308561
file-21	Ind_2381	Doncella_Mouth_SHALLOW_SEQ	2873	2381	swab	3b_swab_MO	20149173	17030634	10487842	15087945	84.52274443	52.0509799583338
file-23		Ni□o_Hair_1_DEEP_SEQ	2854	2537	hair	2A_hair_HAI	694	143	118	118	20.60518732	17.0028818443804
file-24		Ni□o_Anus_1_DEEP_SEQ	3054	2538	swab	2c_swab_AN	255118274	9426853	7975613	8597309	3.695091242	3.12624136050717
file-26		Ni□o_Ear_1_DEEP_SEQ	2869	2540	swab	2b_swab_EA	267503149	131589797	100601290	118589251	49.19186839	37.6075161642303
file-28		Doncella_Mouth_DEEP_SEQ	2873	2542	swab	3b_swab_MO	104972395	82952260	68918258	75869146	79.02292788	65.6536968600173
file-29		Doncella_Anus_1_DEEP_SEQ	2874	2543	swab	3c_swab_AN	310685420	237670758	197466292	221867135	76.49884504	63.5582744758348
file-30		Ni□aDelRayo_Hair_1_DEEP_SEQ	2853	2544	hair	1E_hair_HAI	229135545	35985285	20052479	28774096	15.70480259	8.75136112120885
file-31		Doncella_Hair_1_DEEP_SEQ	2855	2545	hair	3E_hair_HAI	237831063	82440327	66287968	72693855	34.66339761	27.8718713879692
file-32		Ni□aDelRayo_Anus_2_DEEP_SEQ	2868	2546	swab	1h_swab_AN	95305946	56959893	45487833	50715130	59.76530887	47.7282214899792
file-33		Ni□o_Hair_2_DEEP_SEQ	2854	2547	hair	2A_hair_HAI	33237462	16700226	13665060	14885324	50.2451902	41.1134279747353
file-34		Ni□o_Anus_2_DEEP_SEQ	2870	2548	swab	2c_swab_AN	550244063	14237093	11938339	12970972	2.587414196	2.16964430927445
file-35		Ni□o_Ear_3_DEEP_SEQ	2871	2549	swab	2e_swab_EA	78941872	52526601	43242507	48003878	66.53832709	54.7776558934402
file-36		Ni□o_Ear_4_DEEP_SEQ	2869	2550	swab	2b_swab_EA	173431588	101185783	82429028	91260009	58.34334112	47.5282668806561
file-37		Doncella_Mouth_DEEP_SEQ	2873	2382	swab	3b_swab_MO	183524472	141129196	117493022	131956523	76.89938811	64.0203569145809
file-39		Ni□aDelRayo_Hair_2_DEEP_SEQ	2853	2361	hair	1E_hair_HAI	11857016	2052617	1369107	1698187	17.31141292	11.546809079114
file-40		Doncella_Hair_2_DEEP_SEQ	2855	2363	hair	3E_hair_HAI	396066420	177194255	137601133	159909601	44.73852012	34.741933688799
File-42		Ni□aDelRayo_Anus_3_DEEP_SEQ	3057	2376	swab	1h_swab_AN	77774201	29628854	22189838	25931069	38.09599278	28.531103778231
file-41		Ni□o_Hair_3_DEEP_SEQ	3053	2362	hair	2A_hair_HAI	97438976	48170507	37062476	43127997	49.43658993	38.036602519304
file-43		Ni□o_Anus_3_DEEP_SEQ	2870	2378	swab	2c_swab_AN	334554182	8877219	7504435	8084265	2.653447327	2.24311498817253
file-44		Ni□o_Ear_5_DEEP_SEQ	2871	2379	swab	2e_swab_EA	141013514	91421249	72750707	83318837	64.83155153	51.591303840645
file-45		Ni□o_Ear_6_DEEP_SEQ	3056	2377	swab	2b_swab_EA	145992525	86712096	70579919	77772004	59.39488751	48.3448854658826
file-48		Ni□aDelRayo_Mouth_DEEP_SEQ	2865	2373	swab	1d_swab_MO	124054994	82478663	41683323	66325423	66.48556446	33.6006811624206
file-49		Ni□aDelRayo_Nose_DEEP_SEQ	2867	2375	swab	1g_swab_NO	169668815	92669955	38352630	72569411	54.61814241	22.604407297829
											48.01456402	33.0750293866411

Supplementary Table 3 - Mapping Statistics for Pangenome Alignment per Sample

6.2.4 Supplementary Table 4

PANGENOME	WGC (whole genome coverage)
NinaDelRayo	5.43443
Nino	15.0492
Doncella	23.9613
LINEAR GENOME	WGC (whole genome coverage)
NinaDelRayo	5.68427
Nino	15.363
Doncella	24.1896

Supplementary Table 4 - Whole genome coverage for Pangenome and Linear Alignment

6.2.5 Supplementary tables 5a-c

Supplementary Table 5a

chr	start	end	region_size	exon_pct	intron_pct	intergenic_pct
GL000009.2	0	201709	201709	1.10902339508896	0	98.890976604911
GL000194.1	0	191469	191469	1.1573675111898	61.8899142942199	67.8976753417002
GL000205.2	0	185591	185591	0	0	100
GL000208.1	0	92689	92689	0	0	100
GL000216.2	0	176608	176608	0	0	100
GL000218.1	0	161147	161147	1.87840915437458	0	98.1215908456254
GL000219.1	0	179198	179198	0.20759160258485	16.0247324188886	83.7676759785265
GL000225.1	0	211173	211173	0	0	100
KI270303.1	0	1942	1942	0	0	100
KI270304.1	0	2165	2165	0	0	100
KI270310.1	0	1201	1201	0	0	100
KI270320.1	0	4416	4416	0	0	100
KI270330.1	0	1652	1652	0	0	100
KI270333.1	0	2699	2699	0	0	100
KI270336.1	0	1026	1026	0	0	100
KI270337.1	0	1121	1121	0	0	100
KI270363.1	0	1803	1803	0	0	100
KI270392.1	0	971	971	0	0	100
KI270417.1	0	2043	2043	0	0	100
KI270420.1	0	2321	2321	0	0	100
KI270429.1	0	1361	1361	0	0	100
KI270435.1	0	92983	92983	0	0	100
KI270438.1	0	112505	112505	0	0	100
KI270442.1	0	392061	392061	0	0	100
KI270448.1	0	7992	7992	0	0	100
KI270466.1	0	1233	1233	0	0	100
KI270467.1	0	3920	3920	0	0	100
KI270468.1	0	4055	4055	0	0	100
KI270507.1	0	5353	5353	0	0	100
KI270508.1	0	1951	1951	0	0	100
KI270509.1	0	2318	2318	0	0	100
KI270510.1	0	2415	2415	0	0	100
KI270511.1	0	8127	8127	0	0	100
KI270512.1	0	22689	22689	0	0	100
KI270515.1	0	6361	6361	0	0	100
KI270516.1	0	1300	1300	0	0	100
KI270517.1	0	3253	3253	0	0	100
KI270518.1	0	2186	2186	0	0	100
KI270519.1	0	138126	138126	0	0	100
KI270521.1	0	7642	7642	0	0	100
KI270522.1	0	5674	5674	0	0	100
KI270538.1	0	91309	91309	0	0	100
KI270580.1	0	1553	1553	0	0	100
KI270583.1	0	1400	1400	0	0	100
KI270584.1	0	4513	4513	0	0	100
KI270587.1	0	2969	2969	0	0	100
KI270588.1	0	6158	6158	0	0	100
KI270590.1	0	4685	4685	0	0	100
KI270591.1	0	5796	5796	0	0	100
KI270593.1	0	3041	3041	0	0	100
KI270706.1	0	175055	175055	0	0	100
KI270707.1	0	32032	32032	0	0	100

KI270708.1	0	127682	127682	0	0	100
KI270709.1	0	66860	66860	0	0	100
KI270710.1	0	40176	40176	0	0	100
KI270713.1	0	40745	40745	3.29120137440177	0	96.7087986255982
KI270720.1	0	39050	39050	0	0	100
KI270722.1	0	194050	194050	0	0	100
KI270723.1	0	38115	38115	0	0	100
KI270724.1	0	39555	39555	0	0	100
KI270729.1	0	280839	280839	0	0	100
KI270732.1	0	41543	41543	0	0	100
KI270735.1	0	42811	42811	0	0	100
KI270736.1	0	181920	181920	0	0	100
KI270738.1	0	99375	99375	0	0	100
KI270746.1	0	66486	66486	0	0	100
KI270749.1	0	158759	158759	0	0	100
KI270757.1	0	71251	71251	0	0	100

Supplementary Table 5a – Genomic Regions with Better Coverage in Linear Alignment (La Niña del Rayo)

This table lists all genomic regions where the linear alignment outperformed the pangenome alignment, defined as regions falling above +2 standard deviations from the regression line (Figure 4A). For each region, the table also reports the percentage of exonic, intronic, and intergenic content.

Supplementary Table 5b

chr	start	end	region_size	exon_pct	intron_pct	intergenic_pct
GL000008.2	0	209709	209709	0	0	100
GL000009.2	0	201709	201709	1.10902339508896	0	98.890976604911
GL000194.1	0	191469	191469	1.1573675111898	61.8899142942199	67.8976753417002
GL000195.1	0	182896	182896	1.20013559618581	2.20398477823463	96.5958796255796
GL000205.2	0	185591	185591	0	0	100
GL000208.1	0	92689	92689	0	0	100
GL000216.2	0	176608	176608	0	0	100
GL000218.1	0	161147	161147	1.87840915437458	0	98.1215908456254
GL000219.1	0	179198	179198	0.20759160258485	16.0247324188886	83.7676759785265
GL000225.1	0	211173	211173	0	0	100
KI270330.1	0	1652	1652	0	0	100
KI270333.1	0	2699	2699	0	0	100
KI270336.1	0	1026	1026	0	0	100
KI270337.1	0	1121	1121	0	0	100
KI270362.1	0	3530	3530	0	0	100
KI270363.1	0	1803	1803	0	0	100
KI270378.1	0	1048	1048	0	0	100
KI270389.1	0	1298	1298	0	0	100
KI270392.1	0	971	971	0	0	100
KI270417.1	0	2043	2043	0	0	100
KI270420.1	0	2321	2321	0	0	100
KI270422.1	0	1445	1445	0	0	100
KI270423.1	0	981	981	0	0	100
KI270429.1	0	1361	1361	0	0	100
KI270435.1	0	92983	92983	0	0	100
KI270438.1	0	112505	112505	0	0	100
KI270442.1	0	392061	392061	0	0	100

KI270423.1	0	981	981	0	0	100
KI270429.1	0	1361	1361	0	0	100
KI270435.1	0	92983	92983	0	0	100
KI270438.1	0	112505	112505	0	0	100
KI270442.1	0	392061	392061	0	0	100
KI270448.1	0	7992	7992	0	0	100
KI270465.1	0	1774	1774	0	0	100
KI270466.1	0	1233	1233	0	0	100
KI270467.1	0	3920	3920	0	0	100
KI270468.1	0	4055	4055	0	0	100
KI270507.1	0	5353	5353	0	0	100
KI270508.1	0	1951	1951	0	0	100
KI270509.1	0	2318	2318	0	0	100
KI270510.1	0	2415	2415	0	0	100
KI270511.1	0	8127	8127	0	0	100
KI270512.1	0	22689	22689	0	0	100
KI270515.1	0	6361	6361	0	0	100
KI270516.1	0	1300	1300	0	0	100
KI270517.1	0	3253	3253	0	0	100
KI270518.1	0	2186	2186	0	0	100
KI270519.1	0	138126	138126	0	0	100
KI270521.1	0	7642	7642	0	0	100
KI270522.1	0	5674	5674	0	0	100
KI270538.1	0	91309	91309	0	0	100
KI270539.1	0	993	993	0	0	100
KI270580.1	0	1553	1553	0	0	100
KI270583.1	0	1400	1400	0	0	100
KI270584.1	0	4513	4513	0	0	100
KI270587.1	0	2969	2969	0	0	100
KI270588.1	0	6158	6158	0	0	100
KI270590.1	0	4685	4685	0	0	100
KI270591.1	0	5796	5796	0	0	100
KI270593.1	0	3041	3041	0	0	100
KI270706.1	0	175055	175055	0	0	100
KI270707.1	0	32032	32032	0	0	100
KI270708.1	0	127682	127682	0	0	100
KI270709.1	0	66860	66860	0	0	100
KI270710.1	0	40176	40176	0	0	100
KI270711.1	0	42210	42210	10.168206586117	49.0950011845534	40.7367922293295
KI270713.1	0	40745	40745	3.29120137440177	0	96.7087986255982
KI270714.1	0	41717	41717	0	0	100
KI270719.1	0	176845	176845	0	0	100
KI270720.1	0	39050	39050	0	0	100
KI270723.1	0	38115	38115	0	0	100
KI270724.1	0	39555	39555	0	0	100
KI270725.1	0	172810	172810	0	0	100
KI270729.1	0	280839	280839	0	0	100
KI270732.1	0	41543	41543	0	0	100
KI270735.1	0	42811	42811	0	0	100
KI270736.1	0	181920	181920	0	0	100
KI270738.1	0	99375	99375	0	0	100
KI270740.1	0	37240	37240	0	0	100
KI270742.1	0	186739	186739	0	0	100
KI270743.1	0	210658	210658	0	0	100
KI270746.1	0	66486	66486	0	0	100
KI270749.1	0	158759	158759	0	0	100
KI270750.1	0	148850	148850	0	0	100
KI270754.1	0	40191	40191	0	0	100
KI270756.1	0	79590	79590	0	0	100
KI270757.1	0	71251	71251	0	0	100

Supplementary Table 5b – Genomic Regions with Better Coverage in Linear Alignment (El Niño)

This table lists all genomic regions where the linear alignment outperformed the pangenome alignment, defined as regions falling above +2 standard deviations from the regression line (Figure 4B). For each region, the table also reports the percentage of exonic, intronic, and intergenic content.

Supplementary Table 5c

chr	start	end	region_size	exon_pct	intron_pct	intergenic_pct
GL000008.2	0	209709	209709	0	0	100
GL000009.2	0	201709	201709	1.10902339508896	0	98.890976604911
GL000194.1	0	191469	191469	1.1573675111898	61.8899142942199	67.8976753417002
GL000195.1	0	182896	182896	1.20013559618581	2.20398477823463	96.5958796255796
GL000205.2	0	185591	185591	0	0	100
GL000208.1	0	92689	92689	0	0	100
GL000216.2	0	176608	176608	0	0	100
GL000218.1	0	161147	161147	1.87840915437458	0	98.1215908456254
GL000219.1	0	179198	179198	0.20759160258485	16.0247324188886	83.7676759785265
GL000224.1	0	179693	179693	0	0	100
GL000225.1	0	211173	211173	0	0	100
KI270303.1	0	1942	1942	0	0	100
KI270304.1	0	2165	2165	0	0	100
KI270310.1	0	1201	1201	0	0	100
KI270320.1	0	4416	4416	0	0	100
KI270330.1	0	1652	1652	0	0	100
KI270333.1	0	2699	2699	0	0	100
KI270336.1	0	1026	1026	0	0	100
KI270337.1	0	1121	1121	0	0	100
KI270362.1	0	3530	3530	0	0	100
KI270363.1	0	1803	1803	0	0	100
KI270378.1	0	1048	1048	0	0	100
KI270386.1	0	1788	1788	0	0	100
KI270389.1	0	1298	1298	0	0	100
KI270392.1	0	971	971	0	0	100
KI270411.1	0	2646	2646	0	0	100
KI270414.1	0	2489	2489	0	0	100
KI270417.1	0	2043	2043	0	0	100
KI270420.1	0	2321	2321	0	0	100
KI270422.1	0	1445	1445	0	0	100
KI270423.1	0	981	981	0	0	100
KI270429.1	0	1361	1361	0	0	100
KI270435.1	0	92983	92983	0	0	100
KI270438.1	0	112505	112505	0	0	100
KI270442.1	0	392061	392061	0	0	100
KI270448.1	0	7992	7992	0	0	100
KI270465.1	0	1774	1774	0	0	100
KI270466.1	0	1233	1233	0	0	100
KI270467.1	0	3920	3920	0	0	100
KI270468.1	0	4055	4055	0	0	100
KI270507.1	0	5353	5353	0	0	100
KI270508.1	0	1951	1951	0	0	100
KI270509.1	0	2318	2318	0	0	100
KI270510.1	0	2415	2415	0	0	100
KI270511.1	0	8127	8127	0	0	100
KI270512.1	0	22689	22689	0	0	100
KI270515.1	0	6361	6361	0	0	100
KI270516.1	0	1300	1300	0	0	100
KI270517.1	0	3253	3253	0	0	100
KI270518.1	0	2186	2186	0	0	100
KI270519.1	0	138126	138126	0	0	100
KI270521.1	0	7642	7642	0	0	100

KI270522.1	0	5674	5674	0	0	100
KI270529.1	0	1899	1899	0	0	100
KI270530.1	0	2168	2168	0	0	100
KI270538.1	0	91309	91309	0	0	100
KI270539.1	0	993	993	0	0	100
KI270580.1	0	1553	1553	0	0	100
KI270583.1	0	1400	1400	0	0	100
KI270584.1	0	4513	4513	0	0	100
KI270588.1	0	6158	6158	0	0	100
KI270590.1	0	4685	4685	0	0	100
KI270591.1	0	5796	5796	0	0	100
KI270593.1	0	3041	3041	0	0	100
KI270706.1	0	175055	175055	0	0	100
KI270707.1	0	32032	32032	0	0	100
KI270708.1	0	127682	127682	0	0	100
KI270709.1	0	66860	66860	0	0	100
KI270710.1	0	40176	40176	0	0	100
KI270713.1	0	40745	40745	3.29120137440177	0	96.7087986255982
KI270714.1	0	41717	41717	0	0	100
KI270719.1	0	176845	176845	0	0	100
KI270720.1	0	39050	39050	0	0	100
KI270722.1	0	194050	194050	0	0	100
KI270723.1	0	38115	38115	0	0	100
KI270724.1	0	39555	39555	0	0	100
KI270725.1	0	172810	172810	0	0	100
KI270729.1	0	280839	280839	0	0	100
KI270732.1	0	41543	41543	0	0	100
KI270735.1	0	42811	42811	0	0	100
KI270736.1	0	181920	181920	0	0	100
KI270738.1	0	99375	99375	0	0	100
KI270742.1	0	186739	186739	0	0	100
KI270743.1	0	210658	210658	0	0	100
KI270746.1	0	66486	66486	0	0	100
KI270749.1	0	158759	158759	0	0	100
KI270750.1	0	148850	148850	0	0	100
KI270754.1	0	40191	40191	0	0	100
KI270756.1	0	79590	79590	0	0	100
KI270757.1	0	71251	71251	0	0	100

Supplementary Table 5c – Genomic Regions with Better Coverage in Linear Alignment (La Doncella)

This table lists all genomic regions where the linear alignment outperformed the pangenome alignment, defined as regions falling above +2 standard deviations from the regression line (Figure 4C). For each region, the table also reports the percentage of exonic, intronic, and intergenic content.

6.2.6 Supplementary tables 6a-c

Supplementary Table 6a

chr	start	end	region_size	exon_pct	intron_pct	intergenic_pct
chr1	122400000	123400000	1000000	0	0	100
chr1	123300000	124300000	1000000	0	0	100
chr1	124200000	125200000	1000000	0	0	100
chr1	198000000	199000000	1000000	2.1568	28.0603	69.7829
chr1	243000000	244000000	1000000	3.2567	70.4541	27.4033
chr1	244800000	245800000	1000000	4.7126	74.4409	20.8465
chr10	45000000	46000000	1000000	3.1839	37.6982	59.1179
chr11	900000	1900000	1000000	13.0549	37.9618	50.636
chr11	2700000	3700000	1000000	6.1771	40.6571	53.1658
chr11	55800000	56800000	1000000	6.8834	4.7003	88.6842
chr12	10800000	11800000	1000000	3.6711	112.1077	37.3839
chr13	16200000	17200000	1000000	0	0	100
chr13	17100000	18100000	1000000	0	0	100
chr13	111600000	112600000	1000000	1.2884	15.3204	83.3912
chr14	90900000	91900000	1000000	4.6965	81.7061	13.5974
chr14	92700000	93700000	1000000	4.9941	75.6805	20.8626
chr14	93600000	94600000	1000000	7.3259	55.629	37.6936
chr14	105300000	106300000	1000000	3.9743	19.8312	80.5585
chr14	106200000	107043718	843718	0	0	100
chr15	27900000	28900000	1000000	5.2614	46.7256	48.013

chr15	28800000	29800000	1000000	3.0415	76.1624	20.7961
chr15	29700000	30700000	1000000	4.9388	41.7207	55.3801
chr15	30600000	31600000	1000000	7.0083	60.7147	34.3166
chr15	31500000	32500000	1000000	3.7845	60.994	35.2215
chr15	32400000	33400000	1000000	5.9736	67.1247	34.058
chr15	65700000	66700000	1000000	4.9612	85.2713	17.9167
chr16	14400000	15400000	1000000	5.4993	57.2457	42.093
chr16	15300000	16300000	1000000	7.6086	89.5443	16.2937
chr16	16200000	17200000	1000000	3.1977	20.6058	76.1965
chr17	0	1000000	1000000	8.3756	60.9216	30.7028
chr17	900000	1900000	1000000	10.8339	70.9015	18.3172
chr17	35100000	36100000	1000000	11.9662	51.9435	39.3796
chr17	36000000	37000000	1000000	6.9223	18.846	74.6392
chr17	36900000	37900000	1000000	6.9631	66.6247	27.8926
chr17	37800000	38800000	1000000	8.2296	34.4666	57.3038
chr17	40500000	41500000	1000000	9.5268	14.4661	77.2642
chr17	45000000	46000000	1000000	9.463	64.5087	33.6839
chr17	45900000	46900000	1000000	10.8457	73.111	21.8103
chr18	78300000	79300000	1000000	1.2907	24.0669	74.6424
chr19	20700000	21700000	1000000	7.4167	35.8263	58.0617
chr19	21600000	22600000	1000000	4.413	37.3818	58.2052
chr19	54000000	55000000	1000000	12.9872	45.9259	43.1163
chr2	241200000	242193529	993529	8.7184168756	48.05586953	43.26144480936
chr2	242100000	242193529	93529	0	0	100
chr20	63900000	64444167	544167	10.605567776	57.49062328	31.95967414415
chr21	17100000	18100000	1000000	1.4996	31.3985	67.1019
chr22	23400000	24400000	1000000	8.2363	58.5304	47.4441
chr3	198000000	198295559	295559	1.9962173373	12.79609147	85.20769118856
chr4	34200000	35200000	1000000	0	0	100
chr4	68400000	69400000	1000000	1.9136	22.4674	75.619
chr4	189000000	190000000	1000000	0.1623	2.071	97.7667
chr4	189900000	190214555	314555	1.826707571	10.06087966	88.1124127736
chr5	900000	1900000	1000000	4.7685	32.4668	62.7647
chr5	69300000	70300000	1000000	4.8308	22.5715	72.6515
chr5	71100000	72100000	1000000	3.4282	15.8838	80.688
chr6	27900000	28900000	1000000	7.0848	21.4643	72.9601
chr6	28800000	29800000	1000000	7.247	23.3105	70.514
chr6	29700000	30700000	1000000	8.6113	19.9573	72.6646
chr6	30600000	31600000	1000000	13.1265	28.1847	61.0306
chr6	31500000	32500000	1000000	21.5843	51.1514	32.8487
chr6	32400000	33400000	1000000	12.5667	24.3452	66.7029
chr6	127800000	128800000	1000000	1.1164	65.9299	32.9537
chr6	170100000	170805979	705979	2.9024942668	23.49078372	75.77590834855
chr7	141300000	142300000	1000000	7.346	66.9078	27.4964
chr7	142200000	143200000	1000000	3.596	9.9601	86.4439
chr8	0	1000000	1000000	3.3002	51.3278	45.372
chr8	900000	1900000	1000000	2.7721	102.5737	4.7301
chr8	1800000	2800000	1000000	2.5872	41.1058	68.7592
chr8	38700000	39700000	1000000	5.6383	56.5652	37.7965
chr8	39600000	40600000	1000000	1.7911	38.6182	59.5907
chr8	143100000	144100000	1000000	17.5734	48.6483	35.7728
chr9	100800000	101800000	1000000	4.0928	59.7757	36.1315
chrX	0	1000000	1000000	2.7196	10.2198	87.0606
chrX	900000	1900000	1000000	2.3205	25.6816	71.9979
chrX	1800000	2800000	1000000	1.371	37.7378	61.8616
chrX	91800000	92800000	1000000	1.1375	81.1855	17.677
chrX	155700000	156040895	340895	5.8727760747	47.49174966	52.21666495548

Supplementary Table 6a – Genomic Regions with Better Coverage in Pangenome (La Niña del Rayo)

This table lists all genomic regions where the pangenome alignment outperformed the linear genome alignment, defined as regions falling below -2 standard deviations from the regression line (Figure 4A). For each region, the table also reports the percentage of exonic, intronic, and intergenic content.

Supplementary Table 6b

chr	start	end	region_size	exon_pct	intron_pct	intergenic_pct
chr1	122400000	123400000	1000000	0	0	100
chr1	123300000	124300000	1000000	0	0	100
chr1	124200000	125200000	1000000	0	0	100
chr1	198000000	199000000	1000000	2.1568	28.0603	69.7829
chr1	243000000	244000000	1000000	3.2567	70.4541	27.4033
chr1	244800000	245800000	1000000	4.7126	74.4409	20.8465
chr10	45000000	46000000	1000000	3.1839	37.6982	59.1179
chr11	900000	1900000	1000000	13.0549	37.9618	50.636
chr12	10800000	11800000	1000000	3.6711	112.1077	37.3839
chr13	16200000	17200000	1000000	0	0	100
chr13	17100000	18100000	1000000	0	0	100
chr13	111600000	112600000	1000000	1.2884	15.3204	83.3912
chr14	92700000	93700000	1000000	4.9941	75.6805	20.8626
chr14	93600000	94600000	1000000	7.3259	55.629	37.6936
chr14	105300000	106300000	1000000	3.9743	19.8312	80.5585
chr14	106200000	107043718	843718	0	0	100
chr15	20700000	21700000	1000000	1.1913	15.4611	83.3476
chr15	27900000	28900000	1000000	5.2614	46.7256	48.013
chr15	28800000	29800000	1000000	3.0415	76.1624	20.7961
chr15	29700000	30700000	1000000	4.9388	41.7207	55.3801
chr15	30600000	31600000	1000000	7.0083	60.7147	34.3166
chr15	31500000	32500000	1000000	3.7845	60.994	35.2215
chr15	32400000	33400000	1000000	5.9736	67.1247	34.058
chr15	65700000	66700000	1000000	4.9612	85.2713	17.9167
chr16	14400000	15400000	1000000	5.4993	57.2457	42.093
chr16	15300000	16300000	1000000	7.6086	89.5443	16.2937
chr16	16200000	17200000	1000000	3.1977	20.6058	76.1965
chr17	0	1000000	1000000	8.3756	60.9216	30.7028
chr17	900000	1900000	1000000	10.8339	70.9015	18.3172
chr17	25200000	26200000	1000000	0	0	100
chr17	35100000	36100000	1000000	11.9662	51.9435	39.3796
chr17	36000000	37000000	1000000	6.9223	18.846	74.6392
chr17	36900000	37900000	1000000	6.9631	66.6247	27.8926
chr17	37800000	38800000	1000000	8.2296	34.4666	57.3038
chr17	40500000	41500000	1000000	9.5268	14.4661	77.2642
chr17	45000000	46000000	1000000	9.463	64.5087	33.6839
chr17	45900000	46900000	1000000	10.8457	73.111	21.8103
chr18	18000000	19000000	1000000	0	0	100
chr18	18900000	19900000	1000000	0	0	100
chr18	19800000	20800000	1000000	0	0	100
chr19	20700000	21700000	1000000	7.4167	35.8263	58.0617
chr19	21600000	22600000	1000000	4.413	37.3818	58.2052
chr19	54000000	55000000	1000000	12.9872	45.9259	43.1163
chr2	241200000	242193529	993529	8.71841688	48.0558695	43.2614448094
chr2	242100000	242193529	93529	0	0	100
chr20	63900000	64444167	544167	10.6055678	57.4906233	31.9596741442
chr22	23400000	24400000	1000000	8.2363	58.5304	47.4441
chr3	198000000	198295559	295559	1.99621734	12.7960915	85.2076911886
chr4	34200000	35200000	1000000	0	0	100
chr4	68400000	69400000	1000000	1.9136	22.4674	75.619
chr4	189000000	190000000	1000000	0.1623	2.071	97.7667
chr4	189900000	190214555	314555	1.82670757	10.0608797	88.1124127736
chr5	900000	1900000	1000000	4.7685	32.4668	62.7647
chr5	49500000	50500000	1000000	0.524	4.1817	95.2943
chr5	69300000	70300000	1000000	4.8308	22.5715	72.6515
chr5	70200000	71200000	1000000	2.2503	11.412	86.3377
chr5	71100000	72100000	1000000	3.4282	15.8838	80.688
chr6	27900000	28900000	1000000	7.0848	21.4643	72.9601
chr6	28800000	29800000	1000000	7.247	23.3105	70.514
chr6	29700000	30700000	1000000	8.6113	19.9573	72.6646
chr6	30600000	31600000	1000000	13.1265	28.1847	61.0306
chr6	31500000	32500000	1000000	21.5843	51.1514	32.8487
chr6	32400000	33400000	1000000	12.5667	24.3452	66.7029
chr6	127800000	128800000	1000000	1.1164	65.9299	32.9537
chr6	170100000	170805979	705979	2.90249427	23.4907837	75.7759083485
chr7	141300000	142300000	1000000	7.346	66.9078	27.4964
chr7	142200000	143200000	1000000	3.596	9.9601	86.4439
chr8	0	1000000	1000000	3.3002	51.3278	45.372
chr8	900000	1900000	1000000	2.7721	102.5737	4.7301
chr8	1800000	2800000	1000000	2.5872	41.1058	68.7592
chr8	7200000	8200000	1000000	5.33	17.8071	77.2764
chr8	38700000	39700000	1000000	5.6383	56.5652	37.7965
chr8	143100000	144100000	1000000	17.5734	48.6483	35.7728
chr9	44100000	45100000	1000000	0	0	100
chr9	100800000	101800000	1000000	4.0928	59.7757	36.1315

chrX	0	1000000	1000000	2.7196	10.2198	87.0606
chrX	900000	1900000	1000000	2.3205	25.6816	71.9979
chrX	1800000	2800000	1000000	1.371	37.7378	61.8616
chrX	155700000	156040895	340895	5.87277607	47.4917497	52.2166649555
chrY	22500000	23500000	1000000	1.1289	17.6561	81.215

Supplementary Table 6b – Genomic Regions with Better Coverage in Pangenome (El Niño)

This table lists all genomic regions where the pangenome alignment outperformed the linear genome alignment, defined as regions falling below -2 standard deviations from the regression line (Figure 4B). For each region, the table also reports the percentage of exonic, intronic, and intergenic content.

Supplementary Table 6c

chr	start	end	region_size	exon_pct	intron_pct	intergenic_pct
chr1	122400000	123400000	1000000	0	0	100
chr1	123300000	124300000	1000000	0	0	100
chr1	124200000	125200000	1000000	0	0	100
chr1	198000000	199000000	1000000	2.1568	28.0603	69.7829
chr1	243000000	244000000	1000000	3.2567	70.4541	27.4033
chr1	244800000	245800000	1000000	4.7126	74.4409	20.8465
chr11	900000	1900000	1000000	13.0549	37.9618	50.636
chr11	52200000	53200000	1000000	0	0	100
chr11	53100000	54100000	1000000	0	0	100
chr12	10800000	11800000	1000000	3.6711	112.1077	37.3839
chr13	16200000	17200000	1000000	0	0	100
chr13	17100000	18100000	1000000	0	0	100
chr14	92700000	93700000	1000000	4.9941	75.6805	20.8626
chr14	93600000	94600000	1000000	7.3259	55.629	37.6936
chr14	105300000	106300000	1000000	3.9743	19.8312	80.5585
chr14	106200000	107043718	843718	0	0	100
chr15	20700000	21700000	1000000	1.1913	15.4611	83.3476
chr15	27900000	28900000	1000000	5.2614	46.7256	48.013
chr15	28800000	29800000	1000000	3.0415	76.1624	20.7961
chr15	29700000	30700000	1000000	4.9388	41.7207	55.3801
chr15	30600000	31600000	1000000	7.0083	60.7147	34.3166
chr15	31500000	32500000	1000000	3.7845	60.994	35.2215
chr15	32400000	33400000	1000000	5.9736	67.1247	34.058
chr15	65700000	66700000	1000000	4.9612	85.2713	17.9167
chr16	14400000	15400000	1000000	5.4993	57.2457	42.093
chr16	15300000	16300000	1000000	7.6086	89.5443	16.2937
chr16	16200000	17200000	1000000	3.1977	20.6058	76.1965
chr17	900000	1900000	1000000	10.8339	70.9015	18.3172
chr17	36000000	37000000	1000000	6.9223	18.846	74.6392
chr17	36900000	37900000	1000000	6.9631	66.6247	27.8926
chr17	37800000	38800000	1000000	8.2296	34.4666	57.3038
chr17	40500000	41500000	1000000	9.5268	14.4661	77.2642
chr17	45000000	46000000	1000000	9.463	64.5087	33.6839
chr17	45900000	46900000	1000000	10.8457	73.111	21.8103
chr19	20700000	21700000	1000000	7.4167	35.8263	58.0617
chr19	21600000	22600000	1000000	4.413	37.3818	58.2052
chr19	54000000	55000000	1000000	12.9872	45.9259	43.1163
chr2	241200000	242193529	993529	8.7184169	48.05587	43.2614448094
chr2	242100000	242193529	93529	0	0	100
chr20	63900000	64444167	544167	10.605568	57.490623	31.9596741442
chr22	23400000	24400000	1000000	8.2363	58.5304	47.4441
chr3	198000000	198295559	295559	1.9962173	12.796091	85.2076911886
chr4	34200000	35200000	1000000	0	0	100
chr4	68400000	69400000	1000000	1.9136	22.4674	75.619
chr4	189000000	190000000	1000000	0.1623	2.071	97.7667
chr4	189900000	190214555	314555	1.8267076	10.06088	88.1124127736
chr5	900000	1900000	1000000	4.7685	32.4668	62.7647
chr5	49500000	50500000	1000000	0.524	4.1817	95.2943
chr5	69300000	70300000	1000000	4.8308	22.5715	72.6515
chr5	71100000	72100000	1000000	3.4282	15.8838	80.688
chr6	27900000	28900000	1000000	7.0848	21.4643	72.9601
chr6	28800000	29800000	1000000	7.247	23.3105	70.514

chr6	29700000	30700000	1000000	8.6113	19.9573	72.6646
chr6	30600000	31600000	1000000	13.1265	28.1847	61.0306
chr6	31500000	32500000	1000000	21.5843	51.1514	32.8487
chr6	32400000	33400000	1000000	12.5667	24.3452	66.7029
chr6	127800000	128800000	1000000	1.1164	65.9299	32.9537
chr6	170100000	170805979	705979	2.9024943	23.490784	75.7759083485
chr7	141300000	142300000	1000000	7.346	66.9078	27.4964
chr7	142200000	143200000	1000000	3.596	9.9601	86.4439
chr8	0	1000000	1000000	3.3002	51.3278	45.372
chr8	900000	1900000	1000000	2.7721	102.5737	4.7301
chr8	1800000	2800000	1000000	2.5872	41.1058	68.7592
chr8	38700000	39700000	1000000	5.6383	56.5652	37.7965
chr8	143100000	144100000	1000000	17.5734	48.6483	35.7728
chr9	39600000	40600000	1000000	0.1888	5.6076	94.2036
chr9	44100000	45100000	1000000	0	0	100
chr9	100800000	101800000	1000000	4.0928	59.7757	36.1315
chrX	0	1000000	1000000	2.7196	10.2198	87.0606
chrX	900000	1900000	1000000	2.3205	25.6816	71.9979
chrX	1800000	2800000	1000000	1.371	37.7378	61.8616
chrX	155700000	156040895	340895	5.8727761	47.49175	52.2166649555

Supplementary Table 6c – Genomic Regions with Better Coverage in Pangenome (La Doncella)

This table lists all genomic regions where the pangenome alignment outperformed the linear genome alignment, defined as regions falling below -2 standard deviations from the regression line (Figure 4C). For each region, the table also reports the percentage of exonic, intronic, and intergenic content.

6.2.7 Supplementary Table 7

Nino

coverage pan	coverage line	gene_file1	gene_file2
0.9979007	0	HCN3	HCN3
0.4061038	0.9230341	FCGR3A	FCGR3A
1	0.2732925	PTPRC	PTPRC
1	0.0398603	DRD4	DRD4
0.6933588	0	H19	H19
1	0	IGHV3-21	IGHV3-21
0.9887853	0.1968173	ABR	ABR
0.9975868	0.0536519	SERPINF2	SERPINF2
1	0	KRTAP1-1	KRTAP1-1
0.9829199	0.3691756	CTDP1	CTDP1
0.9999534	0.3024946	GPI	GPI
0.9981785	0.0224651	MBOAT7	MBOAT7
0.9925005	0	NLRP7	NLRP7
0.9894609	0	NLRP2	NLRP2
0.9928743	0	GP6	GP6
0.9393321	0.0441039	TPO	TPO
0.9783711	0.0071715	MYT1	MYT1
1	0	NDUFA6	NDUFA6
0.9998481	0.0468981	SLC6A3	SLC6A3
0.9850333	0.024842	MARVELD2	MARVELD2
1	0	TRBV9	TRBV9
0.992087	0	ARHGEF10	ARHGEF10

NinaDelRayo

coverage pan	coverage line	gene_file1	gene_file2
0.8867986	0	HCN3	HCN3
0.1435732	0.6840053	FCGR3A	FCGR3A
0.9913039	0.2716373	PTPRC	PTPRC
0.9511201	0	DRD4	DRD4
1	0	IGHV3-21	IGHV3-21
0.8551808	0.1607108	ABR	ABR
0.8616474	0.0092503	SERPINF2	SERPINF2
0.8723872	0	KRTAP1-1	KRTAP1-1
0.8522896	0.3409185	CTDP1	CTDP1
0.9388117	0.258589	GPI	GPI
0.8937462	0.0407407	MBOAT7	MBOAT7
0.9176226	0	NLRP7	NLRP7
0.9062526	0	NLRP2	NLRP2
0.9284987	0	GP6	GP6
0.8635327	0.0298624	TPO	TPO
0.8289205	0.0065066	MYT1	MYT1
0.9497947	0	NDUFA6	NDUFA6
0.8876458	0.0135813	SLC6A3	SLC6A3
0.9678797	0.072954	MARVELD2	MARVELD2
1	0	TRBV9	TRBV9
0.9068329	0	ARHGEF10	ARHGEF10

Doncella

coverage pan	coverage line	gene_file1	gene_file2
0.9966088	0	HCN3	HCN3
1	0.2733765	PTPRC	PTPRC
0.9991271	0.0587722	DRD4	DRD4
0.9901732	0.4823629	DEAF1	DEAF1
0.7639021	0	H19	H19
1	0	IGHV3-21	IGHV3-21
0.9848378	0.1994077	ABR	ABR
1	0.1085907	SERPINF2	SERPINF2
1	0	KRTAP1-1	KRTAP1-1
0.9859719	0.3728165	CTDP1	CTDP1
1	0.3242494	GPI	GPI
0.9926533	0.0295082	MBOAT7	MBOAT7
0.9979674	0	NLRP7	NLRP7
0.9981671	0	NLRP2	NLRP2
0.9964168	0	GP6	GP6
0.9556603	0.0889387	TPO	TPO
0.9885742	0.0189834	MYT1	MYT1
1	0	NDUFA6	NDUFA6
0.999962	0.0232306	SLC6A3	SLC6A3
0.9992824	0.0283274	MARVELD2	MARVELD2
1	0	TRBV9	TRBV9
0.9944475	0	ARHGEF10	ARHGEF10

Supplementary Table 7- Medically Relevant Genes with Significantly Different Coverage between Pangenome and Linear Genome Alignment Across All Three Individuals

6.2.8 Supplementary Table 8

NinaDelRayo	pangenome_uniq_reads	15163756
Nino	pangenome_uniq_reads	40396665
Doncella	pangenome_uniq_reads	143160257
NinaDelRayo	linear_uniq_reads	55042062
Nino	linear_uniq_reads	66160459
Doncella	linear_uniq_reads	111476673
NinaDelRayo	shared_reads	184469771
Nino	shared_reads	432703352
Doncella	shared_reads	572942975

Supplementary Table 8- Total Counts of Unique and Shared Reads per Individual in Pangenome and Linear Genome Alignments

6.3 Data and code availability

All code and data are stored at:

```
/lisc/project/anthropology/SocialGenomics/Nina
```

Pangenome mapping was performed using the script located at:

```
/lisc/project/anthropology/SocialGenomics/Nina/mapping_to_pangenome/mapping_to_pangenome.sh
```

To repeat this process, you can refer to the prototype script provided in the same directory. Descriptions of all other directories and their contents can be found in the accompanying README file.