JSBS Journal of Sistan and Baluchistan Studies
Volume 5, Issue 1 (2025) 37-47
Doi: 10.22034/jsbs.2025.501869.1068

STUDYING ANCIENT LIVESTOCK-ORIGINATED DNA AND BIOINFORMATICS

Hamidreza HOSSEINI[1], Ali MAGHSOUDI[2]✉, Parastou ERFANMANESH[3]

[1] Researcher and lecturer in the Department of Computer and Bioinformatics, University of Zabol, Zabol, Iran.

[2] Department of Animal Science, Faculty of Agriculture, Tarbiat Modares University, Tehran, Iran, (alimaghsouditmu@gmail.com)

[3] Biology Laboratory of the Research Institute for Protection and Restoration of Historical- Cultural Monuments Research.

Received: 24 January 2025 Accepted: 15 April 2025 Available online: 20 April 2025

Abstract: The study of ancient DNA (aDNA) is a rapidly evolving field within bioinformatics, offering valuable insights into the genetic diversity, migration, and evolution of past populations. Advances in high-throughput sequencing technologies have enabled the retrieval of genomic-scale data from archaeological and historical specimens, including subfossil remains. Bioinformatics tools are essential for processing this data, addressing challenges posed by the degradation of aDNA and recovering useful genetic and even epigenetic information. Key bioinformatics applications include sequence alignment, phylogenetic analysis, and identifying genetic relationships between extinct and extant species. These studies have broad interdisciplinary implications in fields such as archaeology, anthropology, human genetics, ecology, and evolutionary biology. aDNA research has contributed to understanding ancient diets, domestication processes, and microbiomes, with samples extracted from sediments, ice cores, and other environmental sources. However, challenges remain: aDNA is often fragmented and chemically altered, and a high proportion of sequenced DNA belongs to non-target species. Effective separation and identification of target DNA rely on tools like BLAST, Bowtie2, and BWA, and on microbial databases, despite their limitations. Furthermore, the preservation conditions, such as temperature, significantly affect DNA survival. Fossils like those of mammoths or Aurochs offer valuable material for genetic studies, though phylogenetic isolation, as seen in saber-toothed cats, can hinder comparative analysis. Nevertheless, ongoing technological progress continues to refine the understanding of ancient genomes.

Keywords: Bioinformatics, ancient DNA, Alignment, Computational Biology, aDNA Sequencing.

چکیده: مطالعه DNA باستانی (aDNA) یکی از حوزه‌های نوظهور و پویای زیست‌اطلاعاتی (بیوانفورماتیک) است که بینش‌های ارزشمندی درباره‌ی تنوع ژنتیکی، مهاجرت و فرگشت جمعیت‌های گذشته فراهم می‌کند. پیشرفت در فناوری‌های تعیین توالی پرظرفیت (high-throughput) امکان بازیابی داده‌های ژنومی در مقیاس وسیع از نمونه‌های باستان‌شناسی و تاریخی، از جمله بقایای نیمه‌فسیلی را فراهم کرده است. ابزارهای بیوانفورماتیکی نقش اساسی در پردازش این داده‌ها، رفع چالش‌های ناشی از تخریب DNA باستانی و بازیابی اطلاعات ژنتیکی و حتی اپی‌ژنتیکی ایفا می‌کنند. کاربردهای کلیدی بیوانفورماتیک شامل هم‌ترازی توالی‌ها، تحلیل‌های تبارشناختی و شناسایی روابط ژنتیکی میان گونه‌های منقرض‌شده و موجودات زنده امروزی است. این مطالعات پیامدهای بین‌رشته‌ای گسترده‌ای در حوزه‌هایی چون باستان‌شناسی، انسان‌شناسی، ژنتیک انسانی، بوم‌شناسی و زیست‌شناسی فرگشتی دارند. پژوهش بر روی aDNA به درک رژیم‌های غذایی باستانی، فرآیند اهلی‌سازی و میکروبیوم‌های کهن کمک کرده است و نمونه‌هایی از رسوبات، هسته‌های یخی و منابع محیطی دیگر استخراج می‌شوند. با این حال، چالش‌هایی نیز وجود دارد: aDNA معمولاً قطعه‌قطعه و دچار آسیب شیمیایی است و بخش زیادی از DNA تعیین توالی‌شده متعلق به گونه‌های غیرهدف است. شناسایی و جداسازی مؤثر DNA هدف نیازمند ابزارهایی چون BLAST،Bowtie2 و BWA و استفاده از پایگاه‌های داده میکروبی، با وجود محدودیت‌های آن‌هاست. همچنین، شرایط نگهداری مانند دما نقش مهمی در بقای DNA دارد. فسیل‌هایی مانند ماموت‌ها یا گونه منقرض‌شده اوروک (Aurochs) مواد ارزشمندی برای مطالعات ژنتیکی فراهم می‌کنند، هرچند که انزوای تبارشناختی در برخی گونه‌ها مانند ببر دندان‌خنجری، تحلیل تطبیقی را دشوار می‌سازد. با این وجود، پیشرفت‌های فناورانه به بهبود مستمر درک ما از ژنوم‌های باستانی کمک می‌کنند.

کلیدواژه: بیوانفورماتیک، DNA باستانی، هم‌ترازی، زیست‌شناسی محاسباتی، تعیین توالی aDNA.

I. Introduction

Bioinformatics represents an interdisciplinary domain that integrates principles from biology, computer science, and statistics to facilitate the collection, storage, processing, and analysis of biological data. A particularly dynamic and intriguing application within bioinformatics is the investigation of ancient DNA (aDNA). This area of study focuses on the extraction and analysis of DNA from archaeological and historical specimens, thereby enhancing our comprehension of genetic diversity, migration patterns, and the evolutionary trajectories of past populations. Traditionally, our insights into the relationships between extinct species and their extant counterparts have been derived from morphological analyses of fossil remains. However, the recovery and examination of DNA from these fossilized specimens—referred to as "ancient DNA"—offers a complementary approach to elucidating evolutionary processes. Analysis of ancient DNA to resolve genetic relationships between extinct and extant species (Cooper et al., 2001; Höss, Dilling, Currant, & Pääbo, 1996; Krause et al., 2006; Serre et al., 2006) and it has been used to infer the geographical range of extinct organisms (Krause et al., 2007) and their phenotypic characteristics (Lalueza-Fox et al., 2007). Methods The investigation of ancient DNA (aDNA) encompasses a series of intricate and demanding procedures, which include DNA extraction,

library preparation, sequencing, and subsequent data analysis. The high throughput capabilities of next-generation sequencing facilitate the efficient shotgun sequencing of DNA obtained from fossilized bones (Green et al., 2006; Noonan et al., 2006; Poinar et al., 2006). Even though most DNA recovered is from microbes that colonized the bone after death (Höss et al., 1996; Huson, Auch, Qi, & Schuster, 2007), the sheer volume of sequence generated means that a few percent of what is typically a species of interest still constitutes a large enough sequence dataset for genome-scale analysis. Furthermore, since ancient DNA molecules are often fragmented into very short fragments (Huson et al., 2007), ancient DNA sequencing is practically not limited by the short-read lengths of current sequencers. The average ancient DNA fragment length has varied between 60 and 150 bp in most recent large-scale sequencing studies (Green et al., 2008; Pääbo et al., 2004; Shapiro & Hofreiter, 2014) but can vary greatly from sample to sample.

The analysis of ancient DNA (aDNA) presents several challenges, including issues of contamination, degradation, and low coverage. aDNA is frequently found to be significantly degraded and often contaminated with contemporary DNA, which complicates the acquisition of reliable results. This paper emphasizes the critical importance of sample preservation and storage in maintaining the integrity of aDNA. Subsequently, it outlines the various methodologies employed in aDNA extraction, library preparation, sequencing, and data analysis. Given the inherent limitations and conditions associated with aDNA, which often results in fragmentation and low yield, specialized amplification techniques are necessary. Different library preparation strategies, such as shotgun sequencing and targeted enrichment, are implemented to enhance the coverage and quality of aDNA datasets. Quality control measures are paramount in aDNA research, with one effective approach being the use of negative controls to detect and mitigate contamination. aDNA has significant applications in disciplines such as archaeology, evolutionary biology, and population genetics, facilitating investigations into the genetic history of ancient populations, the migratory patterns of early humans, and the evolution of human diseases. Research on ancient DNA (aDNA) has demonstrated its significant contribution to addressing enduring inquiries within the field of evolutionary biology, particularly regarding the origins of modern humans and the domestication processes of flora and fauna. The initial phase of aDNA research involves the extraction of DNA from biological samples, a process that presents considerable challenges due to the degradation and contamination that occur over time. To mitigate DNA damage and enhance yield, several extraction techniques have been developed, including silica-based extraction, phenol-chloroform extraction, and magnetic bead-based extraction. Following the extraction of DNA, the subsequent step is the preparation of a DNA library, which entails the generation of a collection of DNA fragments suitable for sequencing via high-throughput sequencing technologies. A primary obstacle in the preparation of aDNA libraries is the typically limited quantity of available DNA, necessitating amplification through methods such as polymerase chain reaction (PCR). However, it is important to note that PCR amplification may introduce biases and errors into the DNA sequences, thereby impacting the accuracy of subsequent analyses. The next phase in aDNA research involves sequencing the prepared DNA library using advanced high-throughput sequencing technologies, such as Illumina or PacBio platforms. The sequencing data obtained from aDNA is frequently characterized by low quality, short read lengths, low coverage, and elevated error rates, which complicates data analysis. To address these challenges, various methodologies have been developed, including read mapping, de novo assembly, and variant calling. The final stage of aDNA research encompasses the analysis of sequencing data to elucidate the genetic diversity, population structure, and evolutionary history of the samples under investigation.

The analysis of ancient DNA (aDNA) presents a range of challenges, including DNA degradation and contamination, low data coverage and quality, and the absence of reference genomes for numerous ancient populations. To mitigate these issues, researchers have developed various methodologies, such as statistical models, machine learning algorithms, and tools from population genetics. The study of aDNA is particularly fraught with difficulties that are specific to this domain, including the deterioration and contamination of DNA over time, the limited yield and quality of the samples, and the biases and inaccuracies that may arise from polymerase chain reaction (PCR) amplification and sequencing processes. These factors can significantly impact the accuracy and reliability of research findings, necessitating meticulous consideration and validation of the employed methodologies. Additionally, the lack of reference genomes for many ancient populations complicates the comparison of aDNA data with contemporary populations and hinders the inference of evolutionary relationships. This challenge can be addressed through the development of reference genomes for ancient populations, the application of comparative genomics techniques, and the integration of archaeological and historical data into the analyses. The applications of aDNA research are extensive, spanning fields such as archaeology, evolutionary biology, and population genetics. A primary application of aDNA studies is to elucidate the genetic diversity, migratory patterns, and evolutionary history of ancient

populations. Such insights can enhance our understanding of the origins and movements of human groups, their interactions, and the influence of environmental factors on these dynamics.

The first genome of an ancient human was sequenced in 2010 (Rasmussen et al., 2010), and soon after, the genomes of two extinct ancient humans, Neanderthal (Green et al., 2010) and Denisovan (Krause et al., 2010), were sequenced. Since then, hundreds of ancient genomes have been characterized in many phyla, including humans, horses, dogs, pigs, cattle, goats, and woolly mammoths, as well as many human pathogens and agricultural crops such as maize, sorghum, and barley. Ancient genome time series have made it possible to chart migration, admixture, and selection across space and time with unprecedented resolution. They have provided many opportunities to review evolutionary scenarios created from patterns of cultural diversity among archaeological sites (e.g., steppe-related ancestral expansions during the Stone Age and early Bronze Age (Damgaard et al., 2018; Narasimhan et al., 2018; Wang et al., 2019) and from the patterns of genetic variation in modern populations (such as the temporal and geographical increase of lactose tolerance in western Eurasia (Neolithic; Ségurel & Bon, 2017).

The diversity in ancient DNA sequences not only informs us about the genetic affinities of past individuals, populations, and species. It can also provide insights into ancient epigenetic landscapes, which play an important role in regulating gene expression (Malik et al., 2018) in response to infection (Smith et al., 2014) and as social and environmental cues can help predict individual phenotypes in the past. (See (Pedersen et al., 2014) and (Hanghøj et al., 2016) for ancient age predictions or (Gokhman et al., 2020) for morphological predictions).

While methodologies have been established to deduce nucleosome maps in ancient biological specimens, the majority of research in ancient epigenetics has predominantly concentrated on the identification of DNA methylation at CpG dinucleotide sites. Various molecular techniques, including bisulfite sequencing and immunoprecipitation, have been employed; however, comprehensive genome-wide DNA methylation maps have primarily been constructed through statistical inference. This process involves analyzing differential sequence footprints resulting from postmortem DNA damage at both methylated and unmethylated sites, with a particular emphasis on CpG sites. In instances where molecular techniques hinder the sequencing of unmethylated CpGs that have undergone degradation to UpG, previously methylated CpGs can be detected in the sequencing data via the CpG→TpG missense mutation. Recent advancements in methodology have been proposed to mitigate the effects of evolutionary divergence and/or sequence variation at CpG sites on the computation of DNA methylation scores (Hanghøj, Renaud, Albrechtsen, & Orlando, 2019).

Accurate predictions of historical genetic and epigenetic alterations necessitate high-quality DNA sequence alignments against a reference genome. Nonetheless, the majority of ancient DNA research employs read aligner software that has been designed primarily for mapping short sequence reads derived from relatively intact DNA molecules obtained from contemporary tissues. Consequently, these tools are not optimized for the highly fragmented and degraded characteristics of ancient DNA templates. Numerous studies have evaluated various mapping conditions to ascertain the most specific and sensitive approaches or to mitigate reference bias. However, the sensitivity and specificity of alternative read aligners, such as Bowtie2, in the context of ancient DNA data, as well as the influence of different read strategies on the inference of ancient methylation, remain largely unexplored. Given that the inference of ancient methylation relies on the misincorporation patterns of CpG to TpG transitions introduced by postmortem DNA damage, it is anticipated that the distance of edited reads from the reference genome will increase at methylated sites. This phenomenon may compromise the sensitivity of alignment at these loci, thereby impacting the accuracy of DNA methylation assessments for ancient specimens. Therefore, it is imperative to investigate the sensitivity of read alignment methodologies at CpG dinucleotides to avoid underestimating genome-wide DNA methylation levels and to accurately delineate differentially methylated regions among individuals with varying degrees of postmortem DNA damage.

Shotgun sequencing of ancient DNA presents both notable challenges and distinct advantages. A significant complication arises from the high proportion of DNA derived from bacteria and other non-target organisms, necessitating the identification of relevant DNA molecules amidst this intricate background—a concern that is not applicable to PCR-based methodologies. Typically, this identification process involves comparative analysis against the genome of a closely related species and extensive databases of microbial sequences. However, this classification process may encounter failures for various reasons. One primary issue is that ancient DNA sequences frequently exhibit mismatches due to base damage (Briggs et al., 2007; Brotherton et al., 2007; Hofreiter, Jaenicke, Serre, Haeseler, & Pääbo, 2001). These errors can potentially lead to a false similarity or, more often, a failure to identify a similarity. Second, as mentioned above, ancient DNA fragments are usually very short (Poinar et al., 2006) and thus may not be similar enough for correct identification. Third, the microbial sequence

databases used to identify background sequences include only a small fraction of microbes in nature (Huson et al., 2007). Finally, the target genome used to detect fragments of interest may not be sufficiently similar to the genome of an extinct organism to allow unambiguous detection of all relevant sequences. This last problem can be exacerbated by using heuristics in fast database search programs such as BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990).

Several recent analyzes of ancient DNA shotgun data have largely employed case-by-case methods to deal with these issues (Green et al., 2006). While the necessity of using fast local alignment programs such as BLAST (Altschul et al., 1997), Mega BLAST (Zhang, Schwartz, Wagner, & Miller, 2000), or BLASTZ (Schwartz et al., 2003) when handling such large datasets, precise classification, and filtering organizations have not been standardized or even comprehensively investigated. In the simplest classification scheme, reads that match a specific target genome with sufficient similarity are classified as endogenous (i.e., from the target species). A simple extension of this method considers whether there are better alignments to other sequence databases and uses them to exclude potential microbial or other contamination (Green et al., 2006; Noonan et al., 2006). The pairwise divergence can then be calculated from the average similarity of all alignments for sequences considered endogenous (Green et al., 2006).

This study aims to identify and examine the biases that arise from the characteristics of ancient DNA when analyzing shotgun next-generation sequencing data. Given that a primary objective of numerous research projects is to elucidate the genetic relationships between extinct and extant species, our analysis specifically targets the classification of endogenous DNA fragments, defined as DNA that originates from the original organism of the bone, as opposed to microbial or other external DNA sources. We concentrate on calculating pairwise nucleotide differences and divergence. Through this analysis, we identify a selection of extinct species that may serve as valuable references for ancient DNA sequencing utilizing portable shotgun methods.

II. Materials and Methods

II1. aDNA Analysis Using Multiple Sequence Alignment Tools

Multiple Sequence Alignment (MSA) is an essential component of ancient DNA (aDNA) analysis, facilitating the elucidation of evolutionary relationships among sequences and the identification of conserved regions that may be significant for functional studies. When selecting an MSA tool for aDNA analysis, several critical factors must be considered, including accuracy, processing speed, scalability, and the capability to manage degraded and ancient DNA samples. ClustalW is one of the most widely utilized MSA tools in aDNA research, employing a progressive alignment methodology to align multiple sequences. Its rapid processing and accuracy, along with its capacity to handle extensive datasets, contribute to its popularity. However, ClustalW may not be optimal for aDNA analysis due to its potential limitations in aligning degraded and fragmented DNA sequences effectively.

Muscle is another prominent MSA tool that employs an iterative refinement approach for sequence alignment. It is recognized for its high accuracy and speed, as well as its ability to accommodate large datasets. Muscle's proficiency in managing gaps and insertions is particularly advantageous for aDNA analysis, where such features are common in degraded sequences. T-Coffee is frequently employed in aDNA studies as well, utilizing a consistency-based alignment approach. It is noted for its accuracy and its capability to handle incomplete and fragmented sequences, including gaps and insertions, which enhances its suitability for aDNA analysis. MAFFT is also a widely used MSA tool, known for its speed and accuracy, employing a fast Fourier transform method for sequence alignment. It is capable of managing large datasets and accommodating gaps and insertions, making it a favorable option for aDNA analysis.

Other notable MSA tools in the context of aDNA analysis include ProbCons, MUSCLE-Profile, and PAGAN. ProbCons utilizes a probabilistic consistency-based approach and is recognized for its accuracy and ability to manage gaps and insertions. MUSCLE-Profile employs a profile-based alignment strategy and is known for its effectiveness in handling incomplete and fragmented sequences. PAGAN, which adopts a phylogenetic approach, is also acknowledged for its accuracy and capacity to manage large datasets.

In summary, the selection of an appropriate MSA tool for aDNA analysis is contingent upon various factors, including accuracy, speed, scalability, and the ability to process degraded and ancient DNA sequences. ClustalW, Muscle, T-Coffee, MAFFT, ProbCons, MUSCLE-Profile, and PAGAN represent some of the most commonly employed MSA tools in aDNA research, each possessing distinct features and advantages. Researchers should carefully consider the specific requirements of their studies and assess the performance of different tools to identify the most suitable option for their analytical needs.

II2. Data Simulation

Assessing the sensitivity and positive predictive value of DNA alignment software necessitates the identification of three categories of reads: (1) correctly mapped reads, referred to as true positives; (2) incorrectly mapped reads, known as false positives; and

(3) reads that are not mapped at all, termed false negatives. To conduct this evaluation of performance metrics, we generated simulated DNA sequence data utilizing the human reference genome in conjunction with Gargammel (Renaud, 2017). Gargammel is a suite of software applications designed for the cloning of ancient DNA fragments. This program is capable of detecting and simulating various degrees of microbial contamination in ancient hominin specimens. The software generates DNA sequences of specified lengths, which may include sequencing errors characteristic of Illumina DNA sequencing technologies, and may also incorporate DNA misfoldings that represent postmortem DNA degradation. Figure 1 illustrates the workflow of the Gargammel software:

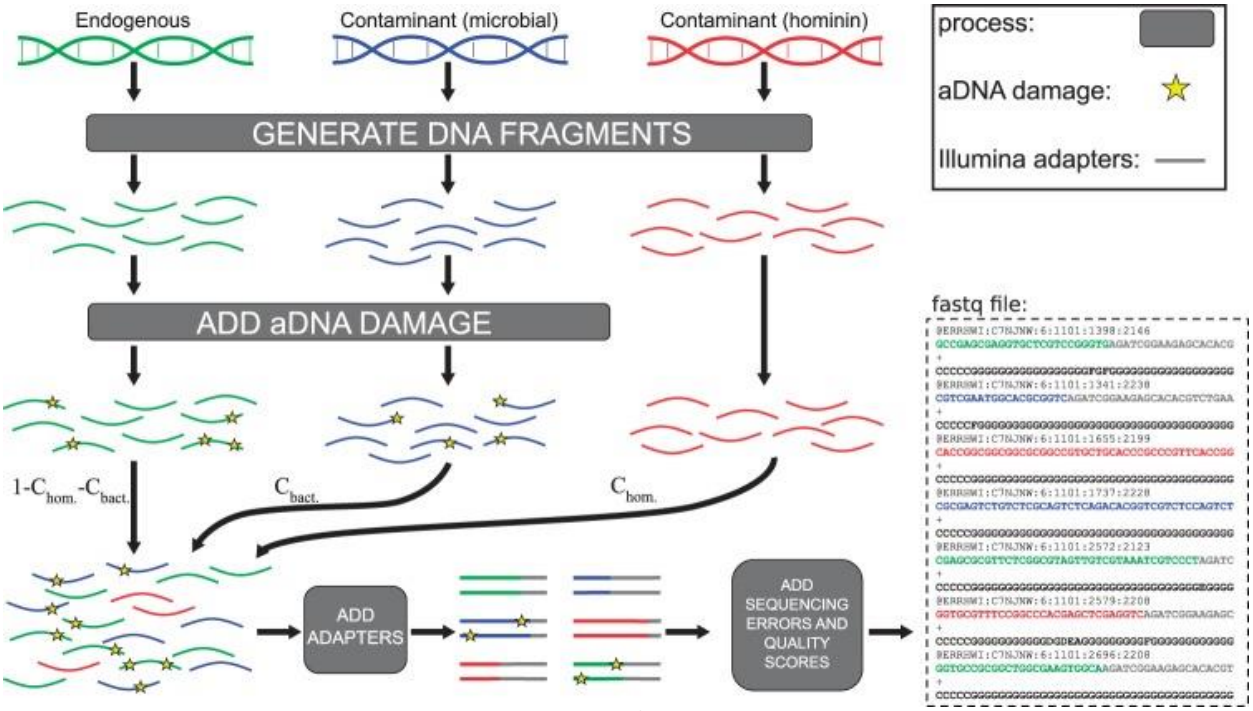

Figure 1: Flowchart of Gargammel software.

A comprehensive simulation involving 3.3 million read pairs was conducted, examining both the presence and absence of ancient DNA damage across a diverse array of DNA template sizes that correspond to the typical size distributions observed in ancient DNA. The simulation of DNA damage was achieved through the incorporation of SIII sample DNA misfolding, as generated by the mapDamage2 software (Purcell et al., 2007). The mapDamage2 software serves as a tool for tracking and quantifying patterns of DNA damage in ancient DNA sequences produced by next-generation sequencing technologies. The input alignment file for mapDamage2 is generated using PALEOMIX (Schubert et al., 2014), which is sensitive to the default global (end-to-end) alignment mode of Bowtie2. PALEOMIX comprises a suite of pipelines and tools designed to facilitate the efficient processing of high-throughput sequencing (HTS) data. The BAM pipeline processes demultiplexed reads from one or more samples through sequencing and alignment procedures, resulting in BAM alignment files that are valuable for subsequent analyses. The phylogenetic pipeline conducts genotyping and phylogenetic inference on BAM alignment files, which may be produced by the BAM pipeline or generated through other means. Additionally, the Zonkey pipeline performs a series of analyses on low-coverage alignments to identify the presence of F1 hybrids within archaeological assemblages. Although these pipelines were initially developed with ancient DNA (aDNA) in mind and incorporate several features tailored for the analysis of ancient samples, they are also applicable to modern samples. Bowtie2 is an ultra-fast and memory-efficient tool for aligning sequence reads to extensive reference sequences, demonstrating particular efficacy in aligning reads of approximately 50 to 1000 characters, especially in the context of relatively long genomes, such as those of mammals. Bowtie2 employs an FM Index for genome indexing, which minimizes its memory usage; consequently, the memory footprint for the human genome typically approximates 3.2 GB. Furthermore, Bowtie2 accommodates both gap and local pairwise alignment modes.

II3. Reading and Alignment Processing

The primary tools utilized for the alignment of ancient DNA sequences include (1) BWA-aln, (2) BWA-mem, (3) NovoAlign, and (4) Bowtie2. Both

simulated and authentic ancient DNA sequence data were processed through the Paleomix pipeline. This automated computational framework executes a series of read-processing steps, which encompass adapter trimming, pairwise assembly, mapping, quality and size filtering, deduplication, and local rearrangement. Mapping was conducted using BWA and Bowtie2, both of which are widely recognized software for read alignment in ancient DNA studies. In the present investigation, BWA was employed in conjunction with two primary alignment modes: backtrack and mem. In accordance with the recommendations of Schubert et al. for ancient DNA datasets, the backtracking algorithm was implemented either with seed or without seed, utilizing default parameters (n-0.04). The study also utilized Bowtie2 read mapper version 2.3.5.1, which incorporates both local and global alignment modes, offering four sensitivity options (very fast, fast, sensitive, and very sensitive). Collectively, this results in a total of 11 reading level conditions. Read pairs were automatically assembled into single reads when sufficient sequence overlap was detected, and base quality was recalibrated based on sequence matches at the overlapping positions, adhering to the default methodology established in the AdapterRemoval2 software (Hanghøj et al., 2016).

AdapterRemoval is a tool designed to identify and eliminate adapter sequences from high-throughput sequencing (HTS) data. Additionally, it offers the option to remove low-quality bases from the 3-prime terminus of reads following the removal of adapters. This software is capable of processing both single-end and paired-end sequencing data, and it can also facilitate the merging of overlapping paired-end reads into longer consensus sequences.

II4. DNA Methylation Coverage and Calculations

Binary Alignment Map (BAM) files, along with summary files generated by Paleomix, were subjected to various analyses. Initially, the average depth of coverage was computed without regard to alignments, revealing quality scores consistently below 30. This finding aligns with the estimated endogenous coverage detailed in the Paleomix summary file. Subsequently, average coverage depth estimates for the dinucleotides CpG, CpA, CpC, and CpT were derived utilizing the Bedtools coverage function (Ruepp et al., 2010), based on the bed coordinates of each dinucleotide type identified in the human reference genome. The coordinates were obtained through Seqkit version 0.3.1.1 (Shen, Le, Li, & Hu, 2016). In the third step, we conducted a repetition of the prior calculations while applying a masking procedure to the truncated soft bases found in the read alignments, utilizing the Jvarkit tool Biostar84452 (Lindenbaum, 2015). All preceding analyses were executed on both the read and cloned DNA sequence datasets. Additionally, we assessed the sensitivity and positive predictive value for each alignment condition using the simulated data. The sensitivity of the alignments was quantified by calculating the ratio of true positive alignments to the total number of false negative alignments, expressed as true positive/(true positive + false negative).

The predicted value of positive alignment was calculated as the ratio of correctly mapped simulated reads, defined as true positives divided by the sum of true positives and false positives (Hanghøj et al., 2016). A read was classified as a true positive if it exhibited a minimum overlap of 80% of its length with known genomic coordinates utilized for cloning. Conversely, reads that did not map were categorized as false negatives, while those that mapped incorrectly were deemed false positives. These classifications were executed using Python version 2.7.5 in conjunction with the pysam library (H. Li, 2009). Additionally, DNA methylation analysis was conducted employing the recently developed DamMet package, which allows for the estimation of the DNA methylation fraction, denoted as f, for a specified genome. The analysis focused on a predetermined number of CpG dinucleotides. In this investigation, we selected 22,845 regions encompassing a total of 100 CpG dinucleotides within the human reference genome, as the volume of sequence data available was inadequate for obtaining reliable estimates in smaller regions. The relevant genomic coordinates were supplied to DamMet in the format of a BED coordinate file. DNA methylation values for each genomic window were directly extracted from the output of DamMet. Coverage estimates for each window were computed using the aforementioned method to ensure consistency. All visual representations were created utilizing RStudio version 1.1.463 (Team, 2016) and the ggplot2 library (Wickham, 2016).

II5. Identification of Endogenous Parts

The initial phase of analyzing ancient shotgun DNA data involves the identification of fragments belonging to the target (endogenous) species. The primary objective of this phase is to accurately discern endogenous components while minimizing the introduction of significant biases that could distort subsequent analyses. In theory, if only microbial contamination is present, there are two methodologies for detecting endogenous fragments. The first approach entails identifying microbial sequences and subsequently subtracting them, thereby allowing any remaining non-microbial sequences to be attributed to the target species. Alternatively, endogenous fragments may be recognized through their similarity to genomic sequences of closely related species. The former method is advantageous as it can uncover novel

sequences and highly divergent regions between the target species and comparative genomes. However, recent research indicates that the existing microbial sequence databases are inadequate for capturing the full diversity inherent in microbial communities. Consequently, the most viable method for identifying DNA fragments of a target species currently relies on their similarity to sequences from closely related species. For instance, Neanderthal sequences are identified through their resemblance to human or chimpanzee genomes, while mammoth sequences are recognized based on their similarity to elephant genomes. The specificity of this approach can be enhanced by stipulating that the similarity must be closer to that of a related genome than to any known microbial sequence (Venter et al., 2004).

Due to the limited proportion of endogenous fragments, particularly in less well-preserved and unfrozen specimens such as Neanderthal bones, extensive sequencing efforts are required to obtain a sufficient number of fragments for subsequent analyses. This necessity, in turn, demands considerable computational resources to conduct similarity searches across multiple genomic databases. Numerous commonly utilized local alignment programs facilitate the rapid comparison of sequences to extensive databases by initiating the alignment process with a short exact match sequence, referred to as a seed. This heuristic approach enhances search efficiency by confining the computationally intensive alignment process to sequences that share at least one short seed (Altschul et al., 1997; Zhang et al., 2000). However, as evolutionary distances increase, the occurrence of exact match seeds that facilitate alignments diminishes, thereby hindering the identification of certain similarities. This reduction in sensitivity is further compounded in ancient DNA shotgun data, as both genomic divergence in the reference genome and chemical degradation of the molecules contribute to shorter read lengths and the presence of erroneous bases (Gotea, Veeramachaneni, & Makalowski, 2003).

II6. Pairwise Differences

Upon the identification of endogenous reads, it is possible to assess their alignments in order to compute the average number of differences per site. However, the analysis of ancient DNA presents several unique challenges. Firstly, there is a risk of misclassifying unrelated microbial sequences as endogenous. Secondly, true endogenous reads that exhibit significant divergence may not be recognized as such. Thirdly, while endogenous reads may be accurately identified, they may be misaligned, potentially occurring within the same genomic region. Lastly, postmortem DNA damage results in miscoding lesions. Each of these factors can introduce bias into the calculation of pairwise differences: the failure to recognize highly divergent reads tends to produce a downward bias in the estimation of pairwise differences, whereas other factors may contribute to an upward bias.

II7. Evaluation of Potential Sequencing Targets

In this research, we conducted an analysis of complete shotgun genomes from various extinct species. To achieve this, multiple criteria were taken into account. The initial phase involved identifying a sample that contained endogenous DNA. Findings from several decades of fossil discoveries indicate that the presence of endogenous DNA is influenced primarily by two factors: the age of the specimen and the conditions under which it was preserved.

The oldest ancient DNA sequences obtained to date come from the silty portion of an ice core from Greenland (Willerslev et al., 2007) and are approximately 500,000 years old. However, in warmer environments, DNA may degrade much faster (Bollongino, Tresset, & Vigne, 2008). Given these limitations, several potentially interesting sequencing targets are probably currently out of reach for ancient DNA research. These include Homo fluorescens fossils, which were found in a warm environment and probably prevented the preservation of endogenous DNA. Other ancient humans, such as Australopithecus, whose extinction is associated with the oldest fossils that have produced endogenous DNA, are probably intolerant of ancient DNA work. On the other hand, endogenous DNA has been obtained from several younger or better-preserved fossils from a wide range of species, such as cave bears, mammoths, mastodons, or saber-toothed cats.

The identification and sequencing of well-preserved fossils necessitate the acquisition of an associated genome sequence to facilitate the identification of endogenous fragments and the elimination of contaminating sequences. Our analysis demonstrates that the number of fragments identifiable as endogenous is contingent upon the phylogenetic proximity of the comparative genomic sequences. In addition to enhancing the recovery of sequences for further analysis, a closely related genome sequence offers a more comprehensive representation of the ancient genome by mitigating biases associated with highly divergent regions. Conversely, the absence of a closely related extant species diminishes the utility of a genomic project for an extinct species, as any comparative sequence analysis is constrained to genomic regions exhibiting sufficient conservation to reliably identify ancient DNA sequences. The saber-toothed tiger serves as a pertinent example of this phenomenon; despite its intriguing morphological characteristics, this species occupies a relatively isolated position within the phylogenetic tree, thereby limiting

the potential value of a genomic project focused on it. In contrast, several other extinct species have closely related genomes available for study. For instance, the ongoing Neanderthal Genome Project utilizes genome sequences from humans and chimpanzees to identify endogenous fragments of Neanderthal DNA, while recently published sequences from a mammoth have been analyzed in conjunction with the draft genome sequence of the African elephant. Table 1 presents a compilation of several extinct species alongside their closest living relatives, whose genome sequences hold significant biological interest.

Table 1: Evaluation of potential ancient DNA shotgun sequencing targets

Species	DNA preservation	Biological relevance	Closely related genomes available
Neandertal	Yes, reasonable	Recent human evolution	Human, chimpanzee
Mammoth	Yes, very good. Draft genome published in 2008	Limited; possibly adaptation to arctic environments	Elephant
Mastodon	Yes, good	Limited; in combination with mammoth parallel adaptation to arctic environments	No close living relatives
Dwarf elephant	Maybe possible; young enough, but poor preservation conditions	Rapid decrease in body size due to island adaptation	Elephant
Cave bear	Yes, good	Limited; probably interesting in combination with genomes from modern bear species; ong hibernation without muscle atrophy may be medically interesting	Bear (not sequenced)
Ground sloth	Yes, reasonable	Size difference to modern species; parallel evolution in different lineages	Tree sloth (sequencing in progress)
Saber-tooth cat	Probably possible	Limited; unique morphological adaptations	No close living relatives
Aurochs (Bos primigenius)	Marginal; young enough, but poor preservation conditions in the region of domestication	Understanding of the domestication process	Cattle
Homo floresiensis	No, young enough, but too poor preservation conditions	Relationship to modern humans; recent human evolution; island adaptation in a hominid	Human, chimpanzee
Australopithecus	No, too old	Human evolution: potentially medical insights	Human, chimpanzee
Dinosaurs	No, far too old	Unique evolutionary lineage	No close living relatives

III. Results and Discussion

The investigation of ancient DNA (aDNA) represents a rapidly advancing and dynamic domain with the potential to significantly enhance our comprehension of the evolutionary trajectories of human populations. The methodologies, challenges, and applications associated with bioinformatics in aDNA research are intricate and diverse, necessitating interdisciplinary collaboration among biologists, computer scientists, and statisticians. Despite the inherent challenges and limitations present in aDNA research, the knowledge acquired from this field holds the promise of fundamentally altering our understanding of human history and evolution.

BWA (Burrows-Wheeler Aligner) (M. Li et al., 2012) is widely recognized as one of the predominant software tools for aligning ancient DNA sequences to reference genomes. Research has demonstrated that deactivating the seed function in BWA enhances the sensitivity of mapping for ancient DNA datasets. In a separate study conducted in 2018, Cahill et al. examined the specificity and sensitivity of Bowtie2 in the context of ancient DNA data. However, a comparative assessment of the performance of both alignment tools specifically focused on ancient DNA data, with the objective of evaluating their potential influence on the inference of ancient DNA methylation, has yet to be conducted. Consequently, we undertook a comparative analysis of the overall alignment performance of BWA and Bowtie2 using ancient DNA data derived from four previously published ancient human samples. This analysis encompassed a total of 11 mapping conditions, comprising 3 for BWA and 8 for Bowtie2. The alignment performance was quantified by normalizing

the genome coverage achieved under one mapping condition to that obtained when the seed function was disabled in BWA, following quality filtering and duplicate removal (Cahill et al., 2018) (Fig. 2).

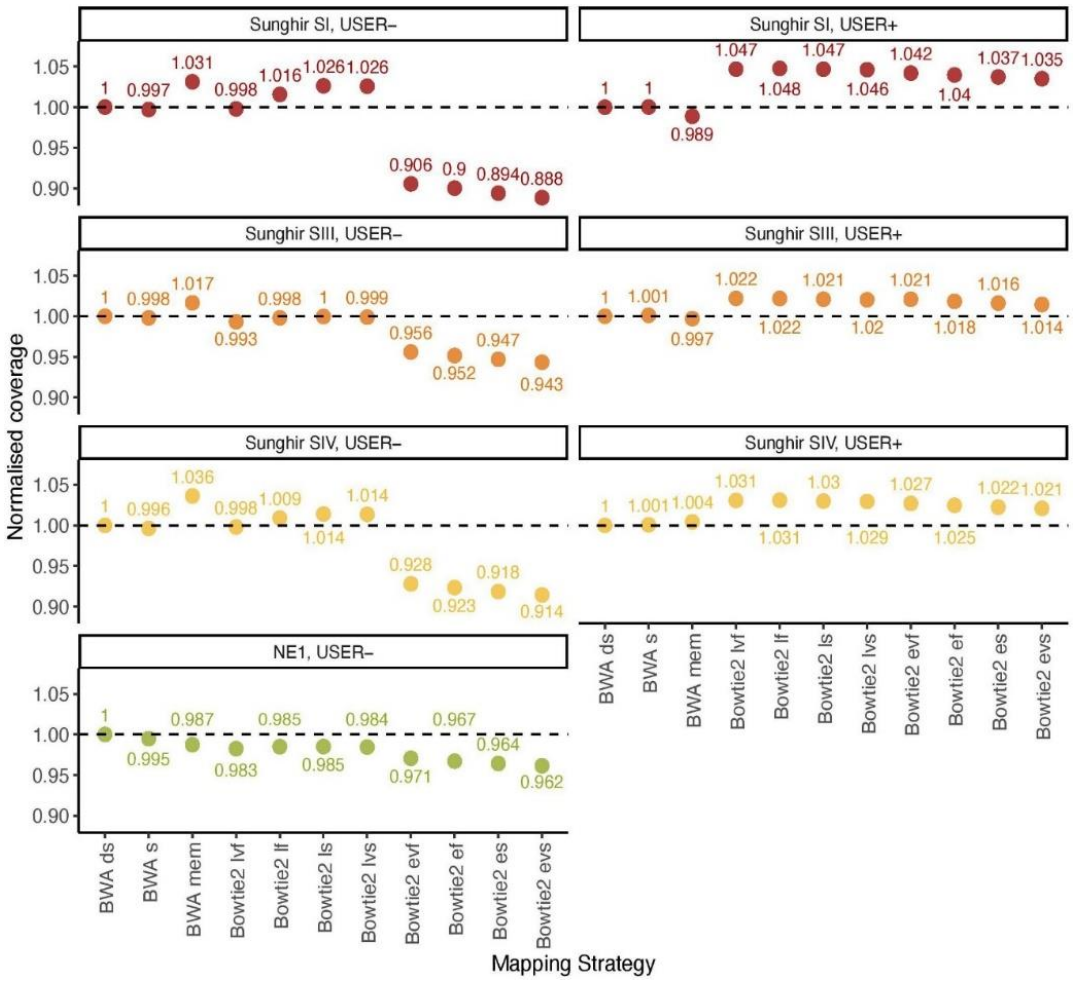

Figure 2: Normal coverage in all 11 mapping conditions examined (real data).

The individual characteristics of ancient DNA datasets, which are indicative of varying conditions of postmortem DNA preservation, can exert both beneficial and detrimental effects on the efficacy of the BWA alignment method. A similar pattern was observed with the four sensitivity options available in Bowtie2's local alignment mode (very fast, fast, sensitive, and very sensitive), where coverage could reach up to 2.63%, while losses could be as high as 1.75%, contingent upon the alignment method employed. Notably, the very fast sensitivity option was the sole choice associated with performance degradation across all four samples analyzed, with losses ranging from 0.20% to 1.75%. In contrast to the local alignment mode, the end-to-end mode in Bowtie2 consistently exhibited performance decline, resulting in coverage losses between 2.93% and 11.15%. Generally, the sensitivity of the local alignment mode was diminished for DNA templates shorter than 38 base pairs, yet it remained superior to that observed with BWA. The quality scores generated by BWA are conservative within the short size range, which leads to the omission of a substantial proportion of true positives (9.4–10.0%) when stringent quality thresholds are applied. Furthermore, the inference of local DNA methylation based on Bowtie2 alignments can be derived from a larger dataset compared to that based on BWA. This is particularly significant, as prior research has demonstrated that the accuracy of inferring ancient DNA methylation levels improves with increased sequencing depth (Hanghøj et al., 2019).

The conditions under which postmortem DNA is preserved have a substantial impact on the efficacy of alignment methodologies. Notably, the majority of DNA fragments retrieved from archaeological and paleontological specimens are of limited size. The mapping parameters evaluated using Bowtie2 are anticipated to enhance genome coverage estimates and, consequently, the quality of the data. Our analysis of the resultant sequence data indicated that Bowtie2 could

increase the average genome coverage depth by 1.62-3.72%. While this improvement may appear modest at first, it constitutes a significant advancement in the field of ancient DNA research, particularly given the constraints associated with the extraction of non-renewable and fragile DNA materials. Additionally, the inherent limitations in the molecular complexity of the DNA libraries available for sequencing often preclude further enhancements in sequencing depth. Crucially, the conditions for read alignment were found to influence both the depth of coverage and the assessment of regional ancient DNA methylation levels.

IV. Conclusion

The investigation of ancient DNA (aDNA) through bioinformatics constitutes a significant advancement in the endeavor to elucidate the intricate narrative of Earth's evolutionary history. The incorporation of high-throughput sequencing technologies has markedly enhanced our capacity to extract and analyze genetic material from archaeological and historical specimens, thereby providing valuable insights into genetic diversity, migration patterns, and the domestication processes of ancient livestock. This study emphasizes the critical role of bioinformatics tools in addressing the challenges inherent in aDNA analysis, such as base damage and the short fragment lengths characteristic of ancient samples. By employing sophisticated computational techniques for sequence alignment and similarity searches, researchers are able to identify relationships between extinct species and their contemporary relatives, thereby enriching our comprehension of evolutionary mechanisms.

Furthermore, research into ancient microbiomes and pathogens has shed light on the lifestyles and dietary practices of historical societies, offering a broader perspective on the ecological and evolutionary dynamics at play. The potential to extract new information from diverse environmental matrices, including ice cores and sediments, further highlights the necessity of interdisciplinary approaches within this domain. Looking forward, the ongoing development and application of bioinformatics in the study of ancient DNA promise to significantly enhance our understanding of biodiversity and the history of life on Earth. As methodologies and databases continue to be refined, we can expect to gain deeper insights into the relationships between extinct and extant organisms, thereby illuminating the evolutionary pathways that have shaped contemporary ecosystems.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research, 25*(17), 3389-3402.

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.

Briggs, A. W., Stenzel, U., Johnson, P. L., Green, R. E., Kelso, J., Prüfer, K., . . . Lachmann, M. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences, 104*(37), 14616-14621.

Brotherton, P., Endicott, P., Sanchez, J. J., Beaumont, M., Barnett, R., Austin, J., & Cooper, A. (2007). Novel high-resolution characterization of ancient DNA reveals C> U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Research, 35*(17), 5717-5728.

Cahill, J. A., Heintzman, P. D., Harris, K., Teasdale, M. D., Kapp, J., Soares, A. E., . . . Graim, K. (2018). Genomic evidence of widespread admixture from polar bears into brown bears during the last ice age. *Molecular biology evolution, 35*(5), 1120-1129.

Cooper, A., Lalueza-Fox, C., Anderson, S., Rambaut, A., Austin, J., & Ward, R. J. N. (2001). Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *409*(6821), 704-707.

Damgaard, P. d. B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., . . . Usmanova, E. (2018). 137 ancient human genomes from across the Eurasian steppes. *Nature, 557*(7705), 369-374.

Gokhman, D., Mishol, N., de Manuel, M., de Juan, D., Shuqrun, J., Meshorer, E., … Carmel, L. (2020). Reconstructing denisovan anatomy using DNA methylation maps. *Cell, 180*(3), 601.

Gotea, V., Veeramachaneni, V., & Makałowski, W. (2003). Mastering seeds for genomic size nucleotide BLAST searches. *Nucleic Acids Research, 31*(23), 6935-6941.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., . . . Fritz, M. H.-Y. (2010). A draft sequence of the Neandertal genome. *Science, 328*(5979), 710-722.

Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., Simons, J. F., . . . Paunovic, M. (2006). Analysis of one million base pairs of Neanderthal DNA. *Nature, 444*(7117), 330-336.

Green, R. E., Malaspinas, A.-S., Krause, J., Briggs, A. W., Johnson, P. L., Uhler, C., . . . Stenzel, U. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell, 134*(3), 416-426.

Hanghøj, K., Renaud, G., Albrechtsen, A., & Orlando, L. (2019). DamMet: ancient methylome mapping accounting for errors, true variants, and post-mortem DNA damage. *GigaScience, 8*(4), giz025.

Hanghøj, K., Seguin-Orlando, A., Schubert, M., Madsen, T., Pedersen, J. S., Willerslev, E., & Orlando, L. (2016). Fast, accurate and automatic ancient nucleosome and methylation maps with epiPALEOMIX. *Molecular biology evolution, 33*(12), 3284-3298.

Hofreiter, M., Jaenicke, V., Serre, D., Haeseler, A. v., & Pääbo, S. (2001). DNA sequences from multiple amplifications reveal artifacts induced by cytosine

deamination in ancient DNA. *Nucleic Acids Research, 29*(23), 4793-4799.

Höss, M., Dilling, A., Currant, A., & Pääbo, S. (1996). Molecular phylogeny of the extinct ground sloth Mylodon darwinii. *Proceedings of the National Academy of Sciences, 93*(1), 181-185.

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome research, 17*(3), 377-386.

Krause, J., Dear, P. H., Pollack, J. L., Slatkin, M., Spriggs, H., Barnes, I., . . . Hofreiter, M. (2006). Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature, 439*(7077), 724-727.

Krause, J., Fu, Q., Good, J. M., Viola, B., Shunkov, M. V., Derevianko, A. P., & Pääbo, S. (2010). The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature, 464*(7290), 894-897.

Krause, J., Orlando, L., Serre, D., Viola, B., Prüfer, K., Richards, M. P., . . . Pääbo, S. (2007). Neanderthals in central Asia and Siberia. *Nature, 449*(7164), 902-904.

Lalueza-Fox, C., Rompler, H., Caramelli, D., Staubert, C., Catalano, G., Hughes, D., . . . Condemi, S. (2007). A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. *Science, 318*(5855), 1453-1455.

Li, H. (2009). The sequence alignment/Map format and SAMtools. Bioinformatics. *Wysoker T. Fennell J. Ruan N. Homer G. Marth G. Abecasis R. Durbin, 1000*, 2078-2079.

Li, M., Du, X., Villaruz, A. E., Diep, B. A., Wang, D., Song, Y., . . . Lu, Y. (2012). MRSA epidemic linked to a quickly spreading colonization and virulence determinant. *Nature medicine, 18*(5), 816-819.

Lindenbaum, P. (2015). JVarkit: java-based utilities for Bioinformatics. *figshare, 10*, m9.

Malik, A. M., Miguez, R. A., Li, X., Ho, Y.-S., Feldman, E. L., & Barmada, S. J. (2018). Matrin 3-dependent neurotoxicity is modified by nucleic acid binding and nucleocytoplasmic localization. *elife, 7*, e35977.

Narasimhan, V. M., Patterson, N., Moorjani, P., Lazaridis, I., Lipson, M., Mallick, S., ... Nakatsuka, N. (2018). The genomic formation of South and Central Asia. *BioRxiv*, 292581.

Neolithic, A. Archaeological context for 83 newly reported ancient samples.

Noonan, J. P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., … Pritchard, J. K. (2006). Sequencing and analysis of Neanderthal genomic DNA. *Science, 314*(5802), 1113-1118.

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*, 645-679.

Pedersen, J. S., Valen, E., Velazquez, A. M. V., Parker, B. J., Rasmussen, M., Lindgreen, S., . . . Vang, S. (2014). Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome research, 24*(3), 454-466.

Poinar, H. N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R. D., Buigues, B., . . . Auch, A. (2006). Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science, 311*(5759), 392-394.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics, 81*(3), 559-575.

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., . . . Gupta, R. (2010). Bjarne Grønnow. *Morten Meldgaard, Claus Andreasen, Sardana A Fedorova, Ludmila P Osipova, Thomas FG Higham, Christopher Bronk Ramsey, Thomas VO Hansen, Finn C Nielsen, Michael H Crawford, Soren Brunak, Thomas Sicheritz-Pontén, Richard Villems, Rasmus Nielsen, Anders Krogh, Jun Wang, 190*, 757-762.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., . . . Mewes, H.-W. (2010). CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Research, 38*(suppl_1), D497-D501.

Schubert, M., Ermini, L., Sarkissian, C. D., Jónsson, H., Ginolhac, A., Schaefer, R., … McCue, M. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature protocols, 9*(5), 1056-1082.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., . . . Miller, W. (2003). Human–mouse alignments with BLASTZ. *Genome research, 13*(1), 103-107.

Ségurel, L., & Bon, C. (2017). On the evolution of lactase persistence in humans. *Annual review of genomics human genetics, 18*, 297-319.

Serre, D., Langaney, A., Chech, M., Teschler-Nicola, M., Paunovic, M., Mennecier, P., . . . Pääbo, S. (2006). No evidence of Neandertal mtDNA contribution to early modern humans. *Early Modern Humans at the Moravian Gate: The Mladeč Caves their Remains*, 491-503.

Shapiro, B., & Hofreiter, M. (2014). A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science, 343*(6169), 1236573.

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one, 11*(10), e0163962.

Smith, O., Clapham, A., Rose, P., Liu, Y., Wang, J., & Allaby, R. (2014). Genomic methylation patterns in archaeological barley show biotic stress-associated siRNA activity and time-dependent demethylation as a diagenetic process. *Sci. Rep, 4*(5559), 10.1038.

Team, R. (2016). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., … Nelson, W. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science, 304*(5667), 66-74.

Wang, C.-C., Reinhold, S., Kalmykov, A., Wissgott, A., Brandt, G., Jeong, C., . . . Keating, D. (2019). Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. *Nature communications, 10*(1), 590.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York, NY, USA. In.

Willerslev, E., Cappellini, E., Boomsma, W., Nielsen, R., Hebsgaard, M. B., Brand, T. B., . . . Dahl-Jensen, D. (2007). Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science, 317*(5834), 111-114.

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology, 7*(1-2), 203-214.