MASTERARBEIT | MASTER'S THESIS

Titel | Title

Assessing the effect of long-term storage on ancient DNA samples

verfasst von | submitted by

Alexandra Raimo 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 066 827
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Evolutionäre Anthropologie
Betreut von | Supervisor:	Univ.-Prof. Ron Pinhasi PhD
Mitbetreut von | Co-Supervisor:	Olivia Cheronet MSc MRes PhD

Table of Contents

Popular scientific summary:

Preserving the Past: How Storage Affects Ancient DNA

Ancient DNA (aDNA) has revolutionized our knowledge of human evolution and human migration patterns, by providing a lens looking in the past through genetic evidence. aDNA has provided us insight into many different research topics, like zoonotic diseases, domestication and among other the discoveries of new hominin species. Ancient DNA is characterized by unique properties that complicate the handling of these specific specimens. Some of these characteristics are small amounts, fragmentation, as well as chemical alteration of the DNA strand. Therefore, the sequencing and analysis of aDNA is a complex and delicate process and requires specifically developed preparation methods. One property of aDNA that remains largely unexplored is the effect of long-term storage on the quality of aDNA samples. In detail: does storage influence the integrity of bone powder, extracted DNA, or indexed DNA libraries over time? Our research addresses this question by investigating whether long-term storage influences DNA degradation and whether such degradation could affect future research.

Ancient DNA has a wide range of possible applications to different scientific fields. Not only has aDNA research a huge importance in human evolution, but as well in plant and animal evolution, (host-)pathogen (co)evolution, archaeology as well as animal domestication. It also holds applications and great importance to related fields such as forensic sciences.

To explore this, we reprocessed bone powder, DNA extracts, and indexed DNA libraries initially prepared eight years ago in 2015. We further proceeded to compare the DNA quality from the samples processed in 2015 and the results we obtained now, from the laboratory work in 2023. We utilized a standardized approach for all laboratory steps (extraction, library preparation, PCR). The final analysis was done by using bioinformatics tools and statistical approaches. While this research provides deeper understanding of aDNA degradation during long-term storage, the findings could have wider implications. By deepening our understanding of the damage that occurs during long-term storage we can improve storage conditions and employ standardized storage protocols, thereby improving the preservation of samples. Genetic material from our past encompasses our bio-cultural heritage and therefore is inherently valuable. Every archaeological sample is a unique witness of the past and therefore irreplaceable. Moreover, improving preservation of samples also amplifies the amount of information we can obtain from aDNA.

Here we report an increase in degradation over the 8-year long-term storage period for the samples produced from Extract. Cytosine to Thymine deamination at position 1 was statistically significantly higher in the samples produced from Extract compared to the 2015 samples. Additionally, the Guanine to Adenine deamination at position 1 also showed statistically significantly higher levels of deamination in the Extract group compared to 2015. Indicating in both cases an increase in degradation over the 8-year long-term storage period. In contrast, samples produced from Indexed Library and Powder exhibited a statistically significantly lower Cytosine to Thymine deamination, suggesting better preservation compared to 2015. This is probably related to the higher stability of bone and Indexed library samples. The results also show an increase in mean fragment length, fragment length distribution and the rate of unique reads across all categories between the 2015 to 2023 samples. These findings show, that despite the higher degradation in the Extract group, the recovery of reads improved for all samples, likely due to improved sequencing technology from the NovaSeq and NextSeq technology compared to MiSeq in 2015. This implies that, as sequencing technologies improve, the negative effect of degradation on aDNA can be mitigated.

Abstract: Assessing the effect of long-term storage on ancient DNA (aDNA) samples

Ancient DNA (aDNA) research has revolutionized our understanding of human evolution through direct genetic evidence. However, aDNA is an inherently degraded molecule, characterized by fragmented structures, low quantities, short read lengths (less than 100 base pairs), and numerous chemical alterations, which complicate its handling and analysis. Numerous studies have shown the importance of sample preservation in achieving high-quality results. Despite the critical importance of preservation, few studies have examined a novel aspect of preservation which is the impact of long-term storage conditions on the quality of aDNA samples.

This project has investigated the effect of long-term storage on aDNA by reprocessing bone powders, extracts, and indexed libraries that were initially processed eight years prior from a total of 12 individuals. Utilizing standardized protocols consistent with those utilized eight years ago, DNA degradation was assessed. Methods include DNA extraction, library preparation and indexing, PCR amplification, Next generation Illumina sequencing, bioinformatics and statistical analysis using R and SPSS. The majority of the samples have been sequenced twice, using NextSeq date and NovaSeq Illumina Platforms. Samples from both sequencing platforms were incorporated in the final analysis, although most of the samples selected for final analysis were sequenced using the NextSeq platform. For the final analysis only 11 individuals were incorporated (*n = 11*). The selection of samples included in the final analysis was based on bioinformatic analysis.

We found that Cytosine to Thymine deamination at position 1 was statistically significantly higher in samples produced from Extract compared to the 2015 samples, indicating increased degradation. Similarly, Guanine to Adenine deamination at position 1 showed a statistically significantly higher deamination for the Extract group, indicating higher degradation over the 8 years of storage. Conversely, samples produced from Indexed Library and Powder exhibited a statistically significantly lower Cytosine to Thymine deamination, suggesting better preservation compared to 2015. This is probably related to the higher stability of bone and Indexed library samples.

Moreover, we report a statistically significant increase in mean fragment length, fragment length distribution and the rate of unique reads across all categories between the samples from 2015 to 2023. These findings suggest, that despite the higher degradation in the Extract group, the recovery rate of unique reads has improved for all samples, probably attributed due to improved sequencing technologies from MiSeq to NovaSeq and NextSeq Illumina platform. This is the first time similar findings have been reported for NextSeq data, the sequencing platform used for the majority of the samples selected for final analysis, while similar findings have previously been reported for NovaSeq data.

This research aimed to provide insights into the mechanisms on aDNA degradation, as DNA quality is central in achieving high-quality results and the accuracy of analysis. Additionally, the study seeks to provide insight into the importance of establishing standardized storage protocols that minimize DNA degradation, potentially improving the quality of stored aDNA samples and minimizing maintenance costs for laboratories. Moreover, we want to highlight the importance of constant temperature in the storage facility. Ultimately, this work aims to improve storage conditions for aDNA research for laboratories.

Keywords: Ancient DNA, aDNA, DNA degradation, ancient DNA decay, long term DNA survival, aDNA preservation, long-term storage aDNA, Storage time, Storage effect, Anthropology, Paleogenomic.

Abstract: Untersuchung des Effekts von Langzeitlagerung auf antike DNA (aDNA) Proben

Die Erforschung antiker DNA hat unser Verständnis von menschlicher Evolution revolutioniert durch Verwendung genetischer Daten als Grundlage. Antike DNA-Moleküle sind durch spezifische Eigenschaften charakterisiert, welche sowohl die Analyse als auch die Sequenzierung kompliziert gestalten. Zu den Eigenschaften zählen: Degradierung, geringe Mengen endogener DNA, fragmentierte, kurze DNA-Stränge (mit einer typischen Basenpaarlänge unter 100 Bp) und zahlreichen chemischen Modifikation taphonomischen-Ursprungs. Obwohl, zahlreiche Studien gezeigt haben, wie wichtig der Erhaltungszustand von Proben für qualitativ hochwertige Forschungsergebnisse ist, wurde ein Aspekt der Proben Konservierung bisher kaum untersucht, nämlich den Einfluss von Langzeitlagerung auf die Qualität von aDNA-Proben.

Dieses Forschungsprojekt untersucht die Auswirkungen von Langzeitlagerung auf die Qualität von aDNA-Proben. Proben unterschiedlicher Herstellungsschritte, die ursprünglich vor 8 Jahren von 12 Individuen produziert und analysiert wurden, wurden unter Verwendung der gleichen standardisierten Protokolle erneut hergestellt. Diese unterschiedlichen Proben-Herstellungsschritte beinhalteten Knochenpulver, DNA-Extrakte und indexierte DNA-Bibliotheken. Verwendete Methoden beinhalteten DNA-Extraktion, Herstellung von DNA-Bibliotheken und Indexierung, PCR, Illumina Sequenzierung, bioinformatische Analysen und statistische Analysen mit SPSS und R. Der Großteil der Proben wurde zweimal sequenziert, sowohl mit der NextSeq als auch der NovaSeq Illumina Plattformen. Proben beider Sequenzierungsplattformen wurden in die finale Analyse einbezogen, jedoch stammten die meisten Proben von NextSeq Illumina-plattformen. Für die finale Analyse wurden nur 11 Individuen integriert (*n = 11*). Die Entscheidung welche Proben für die finale statistische Auswertung inkludiert wurden, erfolgte mittels bioinformatischer Analysen.

Die Ergebnisse zeigen, dass die Cytosin zu Thymin Deaminierung an Position 1 statistisch signifikant höher war in den Proben, die vom Extrakt hergestellt wurden, im Vergleich zu den in 2015 hergestellten Proben, dies deutet auf eine erhöhte Degradation hin. Die Guanin zu Adenin Deaminierung an Position 1 hat ergeben, dass die Proben der Gruppe Extrakt eine statistisch signifikant höhere Deaminierung aufweist, dies deutet auf eine höhere Degradierung aufgrund der Langzeit-Lagerung hin. Im Gegensatz dazu zeigt sich eine statistisch signifikante Abnahme der Cytosin zu Thymin Deaminierung an Position 1 für die Proben, die von Knochenpulver und indexierten DNA-Bibliotheken hergestellt wurden. Dies wiederum deutet auf eine Verbesserung der Proben-Erhaltung im Vergleich zu 2015 hin. Dies ist wahrscheinlich auf eine höhere Stabilität der Knochen und indexierten DNA Bibliothek Proben zurückzuführen. Weiters, wurde ein Anstieg der mittleren DNA-Fragmentlänge, der DNA Fragmentlängen-verteilung und der Anzahl von „unique reads“ für alle untersuchten Proben-Kategorien zwischen 2015 und 2023 festgestellt. Diese Ergebnisse deuten darauf hin, dass obwohl die Extraktion Gruppe eine höhere Degradation aufweisen, sich die Länge der DNA-Fragmente für alle Proben erhöht hat. Dies ist wahrscheinlich auf eine Verbesserung der Sequenzierungstechnologie, von der MiSeq Plattform in 2015 zu der NovaSeq und NextSeq Plattform in 2023, zurückzuführen. Das ist das erste Mal, dass solche Ergebnisse für NextSeq Daten gezeigt wurden, die Plattform, von der die meisten Proben für die finale Analyse stammten, jedoch wurden bereits von ähnlichen Ergebnissen für NovaSeq Sequenzierungsdaten berichtet.

Das Ziel dieses Forschungsprojekts war es Einblicke in die Mechanismen der DNA-Degradation zu gewinnen, da die Qualität der DNA-Proben entscheidend ist, um Resultate hoher Qualität und akkurate Analysen zu erhalten. Zusätzlich zielt dieses Projekt darauf ab Einblicke für den Entwurf standardisierter

Lagerungsprotokolle mit konstanten Temperaturen zu gewähren, welche darauf abzielen DNA-Degradierungsprozesse auf ein Minimum zu reduzieren. Standardisierten Lagerungsprotokolle könnten potenziell die Qualität von gelagerten aDNA-Proben verbessern, und wenn möglich zugleich Instandhaltungskosten für aDNA-Labore senken. Dieses Projekt strebt an einen Beitrag zu leisten die Lagerungsbedingungen in aDNA-Labors zu verbessern.

Schlüsselwörter: Antike DNA, aDNA, DNA-Degradierung, Langzeit DNA Stabilität, aDNA-Erhaltung, Langzeitlagerung aDNA, Lagerungszeit, Lagerungseffekt, Anthropologie, Paläogenomik

Danksagung

I would like to express my sincere gratitude to all those who have supported and mentored me throughout the course of my research, the completion of this thesis and my Master's.

First, I want to greatly thank my supervisor, Prof. Dr. Ron Pinhasi, for entrusting me with this project and for his support throughout this research.

I am especially grateful to my co-supervisor Dr. Olivia Cheronet for her supervision, support, and continuous help. Your guidance throughout the entire project, from the initial design, through the laboratory work, to the final writing stage, has been crucial to this research.

I would also like to thank Dr. Daniel Fernandes for teaching me bioinformatics and for always being open to offer help whenever I needed it. Your expertise has been indispensable.

A special thanks goes to Dr. Mario Novak for his ongoing support over the years. Without his project in 2015, I would not have been able to conduct this research project.

Lastly, I want to express my deep appreciation to the entire team at the Pinhasi aDNA Lab. Their support and the excellent work environment have greatly contributed to the success of this research project. I am truly grateful for everything that I have learned through this project and as a member of this team. My deep gratitude goes to everyone in the Pinhasi team. It has been a truly enriching experience working with such a dedicated and talented group.

Finally, I also want to express my deep gratitude to the entire department of Evolutionary Anthropology.

1 Introduction

DNA is the carrier of the genetic information (Franklin & Gosling, 1953; Watson & Crick, 1953). Many newly developed techniques in the field of genetics have laid the foundation for the biomolecular techniques that rely on today to analyse ancient DNA data. Among these groundbreaking techniques are Sanger sequencing (Sanger et al., 1977; Sanger & Coulson, 1975), PCR (Saiki et al., 1985) and Next-generation sequencing (Margulies et al., 2005).

The field of ancient genomics is a rapidly growing field. The field started in the 1980s with the molecular bacterial cloning of an ancient Egyptian mummy (Pääbo, 1985) and the sequencing of the mtDNA of the quagga, an extinct member of the horse family (Higuchi et al., 1984). Yet, the first whole ancient human genome was successfully sequenced only in 2011 (Rasmussen et al., 2010). For the past ten years ancient studies have produced thousands of samples, and this number is growing constantly. As of April 2023 the number of human ancient genomes has surpassed 10.000 (Callaway, 2023). Some researchers have referred to this rapid improvement of technologies and increase in data about the human past to the “ancient DNA revolution” (Pinhasi et al., 2019).

The origins of our species have long intrigued the interest of biologists and anthropologists alike. The question where we come from is one of the most fundamental questions for us as human beings. Ancient DNA has provided insights in our human evolution for example by identifying a new hominin species, who are the Denisovans (Krause, Fu, et al., 2010; Reich et al., 2010). Ancient DNA provides a crucial bridge to answer this question from an evolutionary and genetic point perspective. This is one of the reasons why aDNA research is gaining so much popularity and emphasizes the importance of this type of research.

Given the increasing number of genomes being sequenced and the growing quantities of bone powders, extracts, and libraries produced in ancient DNA laboratories, it would be crucial to develop standardized storage conditions to ensure good quality genetic data. As of now, storage conditions are not subject to formal guidelines, with each lab following different approaches.

1.1 Properties of aDNA

Ancient DNA (aDNA) refers to highly fragmented molecules from long-dead organism (Briggs et al., 2007) that have undergone taphonomic molecular alterations, thus resulting in unique properties that complicate handling and analysis. Ancient DNA is isolated from historical or pre-historical contexts (Llamas et al., 2017). Specifically, aDNA is characterized by: small amounts, short fragments (Pääbo, 1989) and chemical modifications.

Only a small fraction of the total DNA in an ancient sample derives from the individual itself (henceforth referred to as ‘endogenous DNA’). This endogenous DNA typically comprises less than 1-5% (Höss et al., 1996; Llamas et al., 2017). The rest of the non-endogenous DNA can be either of exogenous or environmental origin (Pedersen et al., 2015). As a result, a sample’s exogenous DNA can readily surpass the amount of endogenous DNA. Therefore, it is essential to maximize the amount of endogenous DNA by improving methods, such as optimizing the sample’s preservation or the long-term storage. The degree of DNA preservation is the limiting factor for aDNA studies (Bollongino et al., 2008).

The degree of pre-existing degradation of a sample prior to excavation, ergo the burial conditions, are the determining and uncontrollable factor for the quality of a sample. Numerous environmental factors influencing the burial environment are critical to the rate of biomolecular degradation or preservation of a sample. Some of these environmental factors include (Smith et al., 2001) humidity, mean temperature of the burial sedimentary layers, pH of the soil, phosphorus concentration, microbial-driven degradation.

Since these environmental factors contributing to a sample's preservation are uncontrollable, scientific method improvements should focus on preventing further degradation. This means, the conditions under which the sample is stored after the excavation process are crucial (Höss et al., 1996).

An example for an inherent property of the DNA molecule is its molecular half-life. Assuming stable conditions of -5°C and a pH of 7.5 the mitochondrial DNA's half-life is modelled to be 521 years, according to Allentoft et al. (2012). More specifically, this model predicts a half-life of 158,000 years for a mitochondrial DNA fragment of 30 base pairs in bone. Furthermore, the team also estimated the maximum mtDNA survival under assumed ideal conditions of -5°C. Their model suggests that no DNA will survive in intact bones after 6.8 Myr. This estimation indicates that it will never be possible to retrieve DNA from fossil specimens, or even fossil specimen such as Dinosaurs. Nonetheless the findings suggest that short DNA fragments could persist for significant timeframes. These calculations indicate at which sample age, under optimal storage conditions, one can still expect successful amplifications.

In a living organism, various processes regularly destabilize the chemical integrity of DNA molecules. However, during an individual's lifetime, these damaging processes are counteracted by DNA repair mechanisms. Thereby maintaining the genomic stability (Lindahl, 1993). In deceased individuals, these repair mechanisms stop functioning, leading to the accumulation of DNA damage characteristics in ancient DNA.

1.2 Ancient DNA damage

As mentioned above the onset of death is the starting point for taphonomic processes, that modify the composition of the DNA molecule (Dabney et al., 2013). Another hallmark characteristic of ancient DNA damage is low endogenous DNA content or even its absence in many samples (Briggs et al., 2007). Consequently, standardized methods (Pinhasi et al., 2019) have been developed for the retrieval of aDNA samples and aiming at maximizing the yield of the endogenous DNA. These taphonomic alterations explain the difficulties of working with aDNA. Furthermore as highlighted by (Dabney, Meyer, et al., 2013) ancient DNA damages are characterized as three primary destructive chemical alterations: diminished DNA fragment size, misincorporation of wrong nucleotides during DNA replication and further lesions which impede the polymerase's function during the elongation phase of the DNA replication in the lab, thereby interfering with multiple analytical techniques.

1.2.1 Low endogenous DNA content

Ancient DNA samples often containing less than 1% endogenous DNA, although the DNA yield can be increased substantially by using specialized methods targeting regions of the bone with good preservation (Pinhasi et al., 2015) or by employing specialized protocols to process very short reads, meaning reads shorter than 50 bp . The majority of surviving DNA has a read length of less than 100 base pairs (Pinhasi, 2022).

1.2.2 Nucleotide misincorporation

Since DNA is composed of two DNA strands complementary with complementary base pairs, one would expect an equal representation of substitution, where each base change corresponds to its complementary base, unless nucleotide misincorporation has taken place (Briggs et al., 2007). The most common example of nucleotide misincorporations is Hydrolytic deamination of Cytosine to Thymine (Dabney, Meyer, et al., 2013). By analysing 12 different substitutions relative to their distance to the 3' and 5' fragment end Briggs et al., (2007) observed in accordance with previous studies that C to T and G to A were strongly elevated. Moreover, they found that the remaining substitutions did not exceed the estimated error rate of four per 10,000 bp. Therefore, they have a considerably low substitution rate. In

summary, for the purpose of this project, which is the assessment of the degradation in long-term storage, it will be sufficient to analyse substitutions derived from Cytosine to Thymine and Guanine to Adenine deamination.

1.2.3 Hydrolytic depurination and deamination

Hydrolytic depurination and deamination are among the most common damage patterns present in ancient DNA samples (Dabney, Meyer, et al., 2013). These two damage patterns are collectively referred to as hydrolytic damages (Lindahl, 1993).

Hydrolytic deamination in the context of ancient DNA typically refers to the apparent deamination of Cytosine to Thymine (CtoT). Although apparent Guanine to Adenine conversion on the complementary strand is also a common pattern of aDNA. It is referred to as apparent Cytosine to Thymine deamination since the actual product of Cytosine deamination is Uracil. Uracil will incorporate Adenine, and the complementary base of Adenine is Thymine. As a result, the data shows a high prevalence of apparent Cytosine to Thymine deamination, with a gradient towards the end of the ends of the molecule, where cytosines are more prone to be deamination, compared to more internally of the molecule. This can be identified by an overrepresentation of deaminated Cytosines at the 5'end (Briggs et al., 2007; Dabney, Meyer, et al., 2013; Sawyer et al., 2012). This higher prevalence of deamination rates at the end of the read is a key characteristic of aDNA. Hydrolytic deamination of Cytosine, shown as Thymine, and its counterpart the deamination of Guanine to Adenine is so prevalent in ancient DNA, its occurrence has become a common method to verify the authenticity of ancient DNA molecules ((Ginolhac, Rasmussen, et al., 2011; Jónsson et al., 2013); see chapter *Authentication of aDNA samples*).

Conversions of Guanine to Adenine (GtoA) bases are also common damage patterns in aDNA, even if not as prevalent as Cytosine to Thymine. This is because Guanine to Adenine deamination is the complimentary process to Cytosine to Thymine deamination and is caused by Cytosine deamination itself (Green et al., 2010). According to the model proposed by (Briggs et al., 2007) Cytosine to Thymine deamination itself is indirectly the cause for Guanine to Adenine deamination. As Cytosine to Thymine deamination causes single-stranded ends and nicks, causing the DNA molecule to be more decayed, Guanine to Adenine deamination increases in the 3'end and is deaminated to a similar extent as Cytosine to Thymine at the 5'end. Contrary to apparent Thymine accumulations at the end of a fragment, which is mainly caused by hydrolytic Cytosine to Thymine deamination, accumulation of Guanine and Adenine can be indicative of depurination (Briggs et al., 2007).

In this research project the deamination rates will be used as a comparative method to assess the degradation of the long-term stored samples (see chapter Assessing DNA degradation through Deamination of Cytosine to Thymine).

Fragmentation has been shown to preferentially arise at Purine bases (Briggs et al., 2007), suggesting that the surplus of Guanine and Adenine adjacent to the 5'end to strand breaks may be attributed to depurination. Hydrolytic depurination is the loss of purine bases, which are Adenine or Guanine, resulting in abasic sites in the DNA strand. Depurination happens when the N-glycosyl bond between a Purine base and the sugar is cleaved, leading to the formation of an abasic site (Dabney, Meyer, et al., 2013). Following the cleavage of the N-glycosylic bond the DNA is fragmented through β-elimination. Therefore, these molecular processes are the cause for many characteristic properties of aDNA.

Moreover, age-dependent differences in patterns of purine (adenine and guanine) representation adjacent to DNA strand breaks have been identified (Sawyer et al., 2012). In samples 100 years or younger, tend to exhibit a greater overrepresentation of Guanine than Adenine. In contrast, samples

older than 40,000 years, display a higher overrepresentation of Guanine than Adenine overrepresentation, with Guanine exhibiting a higher frequency of Guanine near strand breaks in these samples. These findings suggest that different patterns of depurination are caused at different time points. Specific patterns of damage seen in ancient DNA molecules can be used as an age estimator. These results (Sawyer et al., 2012) indicate that patterns of hydrolytic depurination increase specifically over time and can therefore could be used as a measure of time-dependent DNA damage pattern.

1.2.4 Patterns of strand breaks

Many exterior factors can cause strand breaks of the DNA molecule. Some of those are, according to (Shapiro et al., 2019), Nuclease activity, microorganism degradation, desiccation, heat, chemicals, direct cleavage (hydrolysis), as well as depurination causing abasic sites. The data shows that strand breaks are occur predominantly ahead of purine residues (Sawyer et al., 2012). Strand breaks, resulting from hydrolytic direct cleavage or as a consequence of depurination have been suggested as one of the major causes of post-mortem DNA fragmentation and loss (Lan & Lindqvist, 2018).

1.2.5 Fragmentation of aDNA samples

As previously mentioned, fragmentation in aDNA molecule is caused by depurination, which is shown by the formation of overhanging ends, so-called 5'-overhangs (Orlando et al., 2021). Previous research by Pääbo (1989) suggested that DNA fragmentation into small sizes primarily results from autolytic processes occurring shortly after an organism's death by enzymatic reactions, or by nonenzymatic reactions such as hydrolytic cleavage (Lindahl, 1993; Pääbo et al., 2004). Supporting this hypothesis, Sawyer et al. (2012) found no consistent difference in the degree of median fragmentation size among samples ranging from 18 to 16,000 years old. These findings support the idea that significant DNA degradation occurs soon after death rather than progressively over time. The general aim of the study was to examine on how typical DNA damage patterns, such as fragmentation, accumulate over time.

Thermal fluctuations and precipitation, which serve in this model as an indicator for humidity, have been identified as a strong predictor for the degree of fragmentation (Kistler et al., 2017). If the findings of the model were applied to storage conditions, samples that have been stored at constant temperatures, as well as low humidity, would be expected to show less fragmentation than samples that have undergone frequent thermal fluctuations and have been exposed to humidity. The samples used in this experiment have been stored by bone powders at constant 4 -6° degrees Celsius, samples that have been produced from extract and indexed library were stored at constant -20 degrees Celsius. Considering all literature provided above, the extent of fragmentation will be assessed and its possible effect on long-term stored samples evaluated. On another note this experiment could provide insight into the differences between the effect of age-related fragmentation as highlighted by (Sawyer et al., 2012) and the effect of long-term storage on fragmentation.

1.2.6 DNA oxidation: oxidative damage

Oxidative damage causes most commonly blocking lesions (Lan & Lindqvist, 2018). Furthermore, miscoding lesions can be caused by deamination or oxidative damage (Lan & Lindqvist, 2018).

1.2.7 Influence of Climate on Taphonomic Processes

Specifically, taphonomic processes are slowed down by absence of humidity and hot climate, the best conditions are presented in ice. This means, that DNA preservation is mainly influenced by geographic and climatic conditions, hence the burial conditions (Bollongino et al., 2008). As such, the oldest aDNA sample found till now (January 2025) are teeth from a 1.6 million year old mammoth and these have been found in the Permafrost (van der Valk et al., 2021). Permafrost conditions cannot be recreated in

labs and storage rooms, and the preservation of the archaeological samples before excavation cannot be influenced. Yet, setting up good storing conditions in the aDNA labs optimizes the conditions for the samples. In the Pinhasi aDNA Lab (University of Vienna) powders and bones are stored in the fridge (4°-6°C) and extracts, libraries etc. in the freezer (-20°C). When storing samples, the objective is to halt or minimize DNA degradation processes (Dabney, Meyer, et al., 2013) as efficiently as possible.

1.2.8 DNA processing in the laboratory

The DNA is processed in different steps before being sequenced. To prevent further degradation and optimize the DNA yield, specific steps in handling and analysis of aDNA have been introduced. During the grinding process (Pinhasi et al., 2015, 2019) the bones are carefully processed to avoid as much contamination as possible. During DNA extraction (Dabney, Knapp, et al., 2013) the DNA is being treated with extraction buffer to dissolve the bone matrix for releasing the DNA. The EDTA and Proteinase K in the extraction buffer releases the DNA and binds the DNA guanidine hydrochloride and isopropanol to a silica. For removal of contaminants the DNA in the sample will be washed with PE buffer. Library preparation (Meyer & Kircher, 2010) involves different steps to prepare the DNA for sequencing. During blunt-end repair, the overhanging 5'- and 3' of the fragment are being repaired using T4 DNA polymerase. The specific P7 adapter was added during adapter ligation. The Bst polymerase, known for its strand-displacement function, fills in the nicks in the DNA strand. Lastly, the sequencing library is being indexed (Meyer & Kircher, 2010). An amplification mastermix composed of the polymerase Accuprime Pfx Supermix and the Primer IS4 (10mM), for the ligation of the index, are mixed together with 1µl of the respective indexing adapter and 3µl of the DNA sample. The indexing PCR will now be run for 12 cycles in the thermocycler. Lastly, the samples are being washed with different buffers such as PB, EB and EBT. The obtained purified indexed library will now be pooled together with all samples going into the same sequencing run. Now the samples can further proceed to sequencing.

1.3 Storage of aDNA samples

While considerable knowledge exists about the types of damage that characterize ancient DNA (aDNA), much remains unknown about the processes that contribute to its preservation. As the field of aDNA is relatively new (Higuchi et al., 1984; Pääbo, 1985; Rasmussen et al., 2010), long-term storage has only recently emerged as a significant concern. This gap in understanding presents a significant challenge in working with aDNA, because it is not fully understood how its integrity is maintained over time. This is one of the challenges that render working with aDNA difficult. There have been only a few studies in regards to storage of modern DNA samples (Chen et al., 2018; Love et al., 2002; Rubio et al., 2009).

Two studies examining the storage of modern DNA samples have yielded different results, albeit by investigating different sample materials. One study focused on the storage of human blood samples (Chen et al., 2018) and one on the storage of horse semen (Love et al., 2002). An example for the modern DNA storage is provided by a study that examined the long-term storage of blood extracts (Chen et al., 2018). Samples were tested for 16 different storage periods ranging from 2 - 19 years of storage, all samples were long-term stored at -30°C. The blood samples were stored for an average of 12 years, concluding no significant difference or correlation between the storage duration and the quantity or quality of the extracted DNA.

These data suggest that under these conditions, storage at -30°C up to 19 years, no effect of storage on the samples could be observed and high-quality modern DNA was extracted. However, the extracts stored in the Pinhasi aDNA lab are stored at -20 °C, which is 10 degrees higher than the conditions used in the study. Another study on modern DNA storage (Love et al., 2002) assessed the effect of storage on stallion sperm. Exposing the samples to three different temperatures: 5°C, 20°C or 37 °C at 7, 20, 31 and

46h. The study concluded that temperature plays a crucial role in the preservation of semen. However, they determined, as did (Allentoft et al., 2012) that temperature is not the only important factor influencing DNA preservation.

Another study examining the storage of DNA samples from a forensic context is (Rubio et al., 2009). They examined the effect of long-term storage of teeth in an interval of 0 to 10 years. The background in research was a forensic one, as DNA degradation is relevant also in this context. The teeth were stored at room temperature. These are typical storing conditions for museum specimen. The evidence shows that there was a significant decline in DNA concentration during the first 2 years of storage, with no significant further degradation between the second and tenth years. Moreover, statistically significant differences were found between fresh teeth and teeth from the 2- and 5-year groups, but no such differences were observed from the 10-year group. Interestingly in this study the DNA damage occurred mainly in the first two years, for the teeth, that were stored at room-temperature.

A number of studies have aimed to determine how storage conditions affect ancient DNA, frequently investigating the relationship between storage duration and temperature. Yielding different results.

The following study, conducted in 1996, identified temperature as a key factor for ancient DNA preservation. From their sample pool all the samples, that were buried at constant low temperatures resulted in samples with low damage and further in samples where the DNA has been successfully amplified. Concluding that temperature is a key factor for DNA preservation (Höss et al., 1996). It is important to highlight that the studies were published in 2007 and 1996, when very old DNA recovery methods were used, based on PCR. Thus, the conclusions may not reflect the performance of more advanced techniques developed since then.

Since than another study (Allentoft et al., 2012) has examined the effect of storage on aDNA samples. The effect of storage on samples was not the main research objective, but, by trying to create a model estimating the half-life of DNA (Allentoft et al., 2012), storage time was taking into account as an influencing variable. This experiment used Illumina HiSeq data, which is not the same sequencing technology (MiSeq) as used for the samples in 2015. They estimated the half-life of DNA, as well as analysed the effect of storage on the samples from the extinct New Zealand moa (Allentoft et al., 2012), using two sets of samples. They concluded that the “Rosslea” samples, which had been stored for less time, compared to the “PV” samples showed more degradation. The Rosslea samples had been stored for a similar period as the samples in this experiment. The Rosslea samples being stored for 7 years and the samples in this experiment for 8 years. The PV samples had been stored for over 50 years. They concluded that the higher degradation in the samples, which had been stored for less time is probably attributed to the findings from their model that storage time is not the highest factor to determine sample decay. The decay kinetics model assumed -5°C and pH of 7,5. A multiple regression analysis revealed that both storage time and geological age significantly affected DNA preservation, together explaining 45.2% of the variation in DNA preservation. However, storage time alone accounted for only 7.7% of this variation.

Another example of a study examining the preservation of aDNA samples is (Pruvost et al., 2007), which examined the effect of storage conditions and bone treatments prior to storing on the yield of amplifiable DNA. Specifically, they compared freshly excavated samples with fossil samples stemming from museum collections. Freshly excavated samples were defined in this experiment as samples that were treated to minimize degradation of endogenous DNA and preventing contamination. This was achieved my measures such as not washing the bones after excavation. The goal for these samples was to replicate, as closely as possible, the natural sediment burial conditions, during storage. Although the

study primarily focused on examining how sample preservation is affected by post-excavation treatments, they also examined the effect of storage conditions on the samples. The authors concluded that the freshly excavated samples, defines as the samples that have not been washed or additionally treated, produced six times more DNA and twice as many authentic DNA sequences compared to bones that were handled with standardized procedures used in these museums.

The authors claim that the storing conditions in the museums negatively affected the DNA preservation. This research emphasizes the importance of standardized treatments on osseous samples, which has shown to influence the yield of amplifiable DNA, claiming that the storing conditions from the museum collections negatively affected the DNA preservation, emphasizing the importance of standardized storage conditions.

In a second experiment of the same study, (Pruvost et al., 2007) compared the success of amplified DNA between the a sample from the same site and individuum excavated and stored for approximately 50 years earlier (1947) with the a sample from the same site and individuum excavated and subjected to improved standardized storage procedures in 2004. In 2004, modern contamination was avoided, and the bones were immediately stored at -20°C. The 1947 samples showed no successful DNA amplification, in comparison all nine samples from 2004 were successfully amplified, obtaining a 153bp and 201 bp fragments using UQPCR.

It is important to note that the results from this project are not expected to be the same as the results from (Pruvost et al., 2007), since they were not subjected to the same standardized treatment procedures. Additionally, if the experiment were to be repeated the amount of amplifiable DNA may differ since DNA recovery methods are not PCR (UQPCR) based nowadays. In (Pruvost et al., 2007)'s second experiment the bones had been frozen and stored at -20°C. (Pruvost et al., 2007).

The storing conditions for samples from the Pinhasi aDNA laboratory are similar, but not identical, to the freshly excavated samples from the first experiment. In addition to not being washed, the bones in the Pinhasi aDNA lab are not brushed or treated with any chemicals, and contamination is at all times avoided. However, the exact temperature for the freshly excavated samples was not specified in the original study (Pruvost et al., 2007). The results of the first experiment emphasize the importance of maintaining good, standardised storage conditions.

The bone powders at the Pinhasi aDNA lab facility were stored at 4°-6°C in the fridge and are not being washed after excavation. Bone washing is not a post-excavation standard procedure anymore, since this can negatively affect the DNA preservation. Moreover, samples other than bone powders, like extracts and indexed libraries are being stored at -20°C.

1.4 Criteria for Sample selection

Samples were chosen from an existing pool of samples in the inventory stored at the Pinhasi ancient DNA laboratory facility. The samples were selected according to the following criteria: there should be a sufficient number of samples available at various stages of sample preparation. These stages of sample preparation are bone powder, extract and indexed libraries.

1.5 Aim of the research project

Considering the degraded properties of ancient DNA (aDNA) molecules, sequencing and analysing ancient data is very costly and time-consuming. Also, ancient DNA samples are very precious and unique and are considered by some as our bio-cultural heritage (Orlando et al., 2021), preserving these samples effectively is of central importance. Furthermore, preserving them enables future studies to re-use them

using potentially new methods that did not yet exist at the time of the initial sample processing. This project aims to address some of these issues.

Few studies (Allentoft et al., 2012; Pruvost et al., 2007) have been published about the storage of ancient DNA samples . It is well-established that the characteristics of aDNA are of a fragmented molecule, with low quantities of DNA, short reads (about 30-80 bp) and chemical modifications (Oliva et al., 2021). The question is if the sample in its state, as a bone powder, extract and indexed library further degrades and if so, are these degradation processes relevant for further research purposes. In this research the aim is to explore this question further. Bearing in mind, that aDNA analysis are destructive per se (Pinhasi et al., 2015), it is our aim to use the archaeological material as efficiently as possible. Every sample is precious and should therefore be handled accordingly.

Ultimately, this research seeks to optimize storage conditions for aDNA samples. By addressing the gaps in current knowledge regarding aDNA storage, this project contributes to the broader field of ancient genomics and supports ongoing efforts to unravel the genetic histories of extinct and historical populations.

2 Materials and Methods

For the aim of this experiment samples originally processed 8 years ago, February 2015**,** were reprocessed and observed degradation of DNA in them was quantified. Previously produced bone powders, extracts, and amplified indexed libraries from 12 samples were selected. Identical methods to those originally used, and the quality and quantity of ancient DNA was assessed. To ensure reproducibility, the same protocols that were in use eight years ago were followed (Dabney & Meyer, 2019; Meyer & Kircher, 2010). The lab work was done between the end of 2022 and 2023.

2.1 Sample description

A total of 12 samples (*n= 12*) (see Table 1) were selected for the experiment, with all but one, OI57, derived from the temporal bone. The type of bone element for OI57 remains unknown. The samples originate from Medieval archaeological sites in Ireland and Croatia. Osteological analysis determined the sex for 5 individuals, since the remaining seven remain subadults for which the biological sex cannot definitely be determined through osteological analysis (see Table 1). All bone powders have been stored for constant 4°-6°C in the fridge of the storage room. Extracts and Indexed Libraries have been stored at constant -20°C in the freezers of the post-PCR (for Indexed Libraries) and pre-PCR lab (for Extracts).

Table 1: Overview of all analysed samples for the experiment, including Lab ID, archaeological site, country of origin, skeletal element, sex, age and type of bone element from which the DNA was extracted.

Lab ID	Site	Country	Skeletal element	Sex	Age	Bone element type
ARD13	Ardsallagh 1, Co. Meath	Ireland	SK 13	subadult (no assigned sex)	7-8 y	temporal bone
AUG175	Augherskea, Co. Meath	Ireland	SK 175	male	28-35 y	temporal bone
AUG182	Augherskea, Co. Meath	Ireland	SK 182	female	35-45 y	temporal bone
AUG87	Augherskea, Co. Meath	Ireland	SK 87	male	28-35 y	temporal bone
OI57	Omey Island, Co. Galway	Ireland	SK 57	subadult (no assigned sex)	11-12 y	unknown
OI117	Omey Island, Co. Galway	Ireland	SK 117	subadult (no assigned sex)	1-2 m	temporal bone
OI32	Omey Island, Co. Galway	Ireland	SK 32	male	40-50 y	temporal bone
OI89	Omey Island, Co. Galway	Ireland	SK 89	subadult (no assigned sex)	1-1.5 y	temporal bone
OI95	Omey Island, Co. Galway	Ireland	SK 95	subadult (no assigned sex)	0.5-1 y	temporal bone
S111	Donje polje-Sveti Lovre, Šibenik	Croatia	G 111	subadult (no assigned sex)	0-0.5 y	temporal bone

S176	Donje polje-Sveti Lovre, Šibenik	Croatia	G 176	subadult (no assigned sex)	13-14y	temporal bone
S202A	Donje polje-Sveti Lovre, Šibenik	Croatia	G 202A	female	28-35y	temporal bone

2.2 Original list of samples

Each sample has three ID's. One ID is the identifier for sequencing facility called "Sample ID". The ID called "Sequencing ID" encompasses the information about the preparation/production process, as of op1_nL1_nI1 for samples produced from powder, op1_oL1_nI1 for samples produced from extract, and op1_oL1_oI1 for samples produced from Indexed Library. The third ID called "Lab ID" identifies which samples belong to the same individual. The table also includes information about sample conditions, date of sequencing, as well as from which stage of sample the new sample has been produced. The following table (see Table 2) entails the original list of samples before data cleansing. For the Lab ID S176 there is by default only one version of ONN, being sample ID 222489. Since there was not enough indexed library left in the tube for a second pooling in June (see chapter 2.7.3 Indexing PCR and pooling).

Table 2: Original list of samples, corresponding to 13 samples 3 different versions of each individual, differing in preparation stage, have been created in the lab. These differing preparation stages are identifiable by the abbreviations at the end of the sample name by ONN, OON, OOO. If a sample is produced from the stage of old powder it is identified by ONN (old powder, new extract, new indexed library), if a sample is produced from the stage of old extract it is identified by OON (old powder, old extract, new indexed library), and lastly if a sample has been produced from the stage of old indexed library it is *identifiable by OOO(old powder, old extract, new indexed library). In the table are also included extraction, library and PCR blanks, excluded samples, as well as the 2015 samples.*

Sample ID	Sequencing ID	Lab ID	Sample Type by Preparation Stage	Sample Conditions °C	Sequencing date
222482	S176_op1_oL1_oI1	S176	Indexed Library	-20	February 2023
222483	S202A_op1_oL1_oI1	S202A	Indexed Library	-20	February 2023
222484	ARD13_op1_nL1_nI1	ARD13/RAD13	Powder	4 - 6	February 2023
222485	AUG175_op1_nL1_nI1	AUG175	Powder	4 - 6	February 2023
222486	OI117_op1_nL1_nI1	OI117	Powder	4 - 6	February 2023
222487	OI157_op1_nL1_nI1	OI157	Powder	4 - 6	February 2023
222488	AUG87_op1_nL1_nI1	AUG87	Powder	4 - 6	February 2023
222489	S176_op1_nL1_nI1	S176	Powder	4 - 6	February 2023
222490	S202A_op1_nL1_nI1	S202A	Powder	4 - 6	February 2023
222491	AUG175_op1_oL1_nI1	AUG175	Extract	-20	February 2023
222492	OI157_op1_oL1_nI1	OI157	Extract	-20	February 2023
222493	S176_op1_oL1_nI1	S176	Extract	-20	February 2023
222494	S202A_op1_oL1_nI1	S202A	Extract	-20	February 2023
222495	LB1_op1_nL1_nI1	LB1	Blank		February 2023
222496	EB2_op1_nL1_nI1	EB2	Blank		February 2023
222497	LB2_op1_nL1_nI1	LB2	Blank		February 2023
222498	LB3_old_op1_oL1_nI1	LB3	Blank		February 2023
222526	PCRB_1.1	PCRB_1.1	Blank		February 2023
222527	PCRB_1.2	PCRB_1.2	Blank		February 2023
225353	ARD13_op1_oL1_oI1	ARD13/RAD13	Indexed Library	-20	March 2023
225354	AUG175_op1_oL1_oI1	AUG175	Indexed Library	-20	March 2023
225355	OI117_op1_oL1_oI1	OI117	Indexed Library	-20	March 2023
225356	OI157_op1_oL1_oI1	OI157	Indexed Library	-20	March 2023

225357	AUG87_op1_oL1_oI1	AUG87	Indexed Library	-20	March 2023
225358	OI89_op1_oL1_oI1	OI89	Indexed Library	-20	March 2023
225359	OI89_op1_nL1_nI1	OI89	Powder	4 - 6	March 2023
225360	OI89_op1_oL1_nI1	OI89	Extract	-20	March 2023
225361	AUG182_op1_oL1_oI1	AUG182	Indexed Library	-20	March 2023
225362	AUG182_op1_nL1_nI1	AUG182	Powder	4 - 6	March 2023
225363	AUG182_op1_oL1_nI1	AUG182	Extract	-20	March 2023
225364	OI32_op1_oL1_oI1	OI32	Indexed Library	-20	March 2023
225365	OI32_op1_oL1_nI1	OI32	Extract	-20	March 2023
225366	OI32_op1_nL1_nI1	OI32	Powder	4 - 6	March 2023
225367	OI95_op1_oL1_oI1	OI95	Indexed Library	-20	March 2023
225368	OI95_op1_oL1_nI1	OI95	Extract	-20	March 2023
225369	OI95_op1_nL1_nI1	OI95	Powder	4 - 6	March 2023
225370	S111_op1_oL1_oI1	S111	Indexed Library	-20	March 2023
225371	S111_op1_oL1_nI1	S111	Extract	-20	March 2023
225372	S111_op1_nL1_nI1	S111	Powder	4 - 6	March 2023
225373	ARD13_op1_oL1_nI1	ARD13/RAD13	Extract	-20	March 2023
225374	AUG87_op1_oL1_nI1	AUG87	Extract	-20	March 2023
225375	OI117_op1_oL1_nI1	OI117	Extract	-20	March 2023
225376	EB1_op1_nL1_nI1	EB1	Blank		March 2023
233198	ARD13_op1_oL1_oI1	ARD13/RAD13	Indexed Library	-20	June 2023
233199	AUG175_op1_oL1_oI1	AUG175	Indexed Library	-20	June 2023
233200	OI117_op1_oL1_oI1	OI117	Indexed Library	-20	June 2023
233201	OI157_op1_oL1_oI1	OI157	Indexed Library	-20	June 2023
233202	AUG87_op1_oL1_oI1	AUG87	Indexed Library	-20	June 2023
233203	OI89_op1_oL1_oI1	OI89	Indexed Library	-20	June 2023
233204	OI89_op1_nL1_nI1	OI89	Powder	4 - 6	June 2023
233205	OI89_op1_oL1_nI1	OI89	Extract	-20	June 2023
233206	AUG182_op1_oL1_oI1	AUG182	Indexed Library	-20	June 2023
233207	AUG182_op1_nL1_nI1	AUG182	Powder	4 - 6	June 2023
233208	AUG182_op1_oL1_nI1	AUG182	Extract	-20	June 2023
233209	OI32_op1_oL1_oI1	OI32	Indexed Library	-20	June 2023
233210	OI32_op1_oL1_nI1	OI32	Extract	-20	June 2023
233211	OI32_op1_nL1_nI1	OI32	Powder	4 - 6	June 2023
233212	OI95_op1_oL1_oI1	OI95	Indexed Library	-20	June 2023
233213	OI95_op1_oL1_nI1	OI95	Extract	-20	June 2023
233214	OI95_op1_nL1_nI1	OI95	Powder	4 - 6	June 2023
233215	S111_op1_oL1_oI1	S111	Indexed Library	-20	June 2023
233216	S111_op1_oL1_nI1	S111	Extract	-20	June 2023
233217	S111_op1_nL1_nI1	S111	Powder	4 - 6	June 2023
233218	ARD13_op1_oL1_nI1	ARD13/RAD13	Extract	-20	June 2023
233219	AUG87_op1_oL1_nI1	AUG87	Extract	-20	June 2023
233220	OI117_op1_oL1_nI1	OI117	Extract	-20	June 2023
233221	EB1_op1_nL1_nI1	EB1	Blank		June 2023
233222	S176_op1_oL1_oI1	S176	Indexed Library	-20	June 2023
233223	S202A_op1_oL1_oI1	S202A	Indexed Library	-20	June 2023
233224	ARD13_op1_nL1_nI1	ARD13/RAD13	Powder	4 - 6	June 2023

233225	AUG175_op1_nL1_nI1	AUG175	Powder	4 - 6	June 2023
233226	OI117_op1_nL1_nI1	OI117	Powder	4 - 6	June 2023
233227	OI157_op1_nL1_nI1	OI157	Powder	4 - 6	June 2023
233228	AUG87_op1_nL1_nI1	AUG87	Powder	4 - 6	June 2023
233229	S202A_op1_nL1_nI1	S202A	Powder	4 - 6	June 2023
233230	AUG175_op1_oL1_nI1	AUG175	Extract	-20	June 2023
233231	OI157_op1_oL1_nI1	OI157	Extract	-20	June 2023
233232	S176_op1_oL1_nI1	S176	Extract	-20	June 2023
233233	S202A_op1_oL1_nI1	S202A	Extract	-20	June 2023
233234	LB1_op1_nL1_nI1	LB1	Blank		June 2023
233235	EB2_op1_nL1_nI1	EB2	Blank		June 2023
233236	LB2_op1_nL1_nI1	LB2	Blank		June 2023
233237	LB3_old_op1_oL1_nI1	LB3	Blank		June 2023
233238	PCRB_1.1	PCRB_1.1	Blank		June 2023
233239	PCRB_1.2	PCRB_1.2	Blank		June 2023
S202A.q30	S202A-2015	S202A	2015		Feb.15
S176.q30	S176-2015	S176	2015		Feb.15
S111.q30	S111-2015	S111	2015		Feb.15
RAD13.q30	RAD13-2015	ARD13-RAD13	2015		Feb.15
OI117.q30	OI117-2015	OI117	2015		Feb.15
OI95.q30	OI95-2015	OI95	2015		Feb.15
OI89.q30	OI89-2015	OI89	2015		Feb.15
OI57.q30	OI57-2015	OI57	2015		Feb.15
OI32.q30	OI32-2015	OI32	2015		Feb.15
AUG182.q30	AUG182-2015	AUG182	2015		Feb.15
AUG175.q30	AUG175-2015	AUG175	2015		Feb.15
AUG87.q30	AUG87-2015	AUG87	2015		Feb.15

2.3 Data cleansing

2.3.1 Excluded samples based on incompatibility: first stage of data cleansing

Based on various quality control assessments (chapter 3.2) the following samples 233201, 225375, 233220, 222487, 222492, 225356, 233198, 233203 were excluded. The sample 225374 (AUG87_op1_oL1_nI1) will be excluded because it matches as same individual or 1st degree with samples from LabID OI57 and because it failed the sex determination analysis. 225356 does not match with the original sample OI57-2015 and will be excluded, therefore. OI89 will be entirely excluded since there is no positive kinship between OI89-2015 and the corresponding samples from 2023 (see Tble 7). The Table below (see Table 3) entails a list of excluded samples, after the first stage of data cleansing (see chapter 3.2 Quality control assessment for final analysis).

Table 3: List of samples after data cleansing: excluded samples

Sample ID	Sequencing ID	Lab ID	Sample Type	Sample Conditions °C	Sequencing date
225375	OI117_op1_oL1_nI1	OI117	Extract	-20	March 2023
233220	OI117_op1_oL1_nI1	OI117	Extract	-20	June 2023
233201	OI157_op1_oL1_oI1	OI57	Indexed Library	-20	June 2023
222487	OI157_op1_nL1_nI1	OI57	Powder	4 - 6	February 2023
222492	OI157_op1_oL1_nI1	OI57	Extract	-20	February 2023

225356	OI157_op1_oL1_oI1	OI157	Indexed Library	-20	March 2023
233198	OI117	OI117	Indexed Library	-20	June 2023
225358	OI89_op1_oL1_oI1	OI89	Indexed Library	-20	March 2023
225359	OI89_op1_nL1_nI1	OI89	Powder	4 - 6	March 2023
225360	OI89_op1_oL1_nI1	OI89	Extract	-20	March 2023
233203	OI89_op1_oL1_oI1	OI89	Indexed Library	-20	June 2023
233204	OI89_op1_nL1_nI1	OI89	Powder	4 - 6	June 2023
233205	OI89_op1_oL1_nI1	OI89	Extract	-20	June 2023
225374	AUG87_op1_oL1_nI1	AUG87	Extract	-20	March 2023
Z00000	PCRB_1.2	PCRB_1.2	Blank		June 2023

2.3.2 Excluded samples: second stage of data cleansing for double presence of samples

In the dataset obtained from the first stage of data cleansing further samples had to be excluded, since the majority of samples had been sequenced twice. For these samples that had two identical versions, e.g. two versions of AUG182_op1_oL1_oI1, the second same one had to be excluded in order to compare the samples. Both versions had passed the data quality assessment, therefore uniform criteria were established to decide which samples to select for the final analysis. The samples for the final analysis should have the highest quality possible. Therefore, the chosen sample were chosen based on criteria, indicating good read quality. Specifically, samples were excluded based on not being related to a same individual degree, using kinship analysis with the program TKGWV2 (Fernandes et al., 2021). Another criterion is “Used_Snps” since it is used to calculate the HRC coefficient, which determines the degree of relatedness. Furthermore, the relatedness of the samples from 2023 to the samples from 2015 itself was used as a criterion for the inclusion of the final step, if the HRC was established by using less than 10 SNPS the sample was excluded. Lastly, the samples stemming from the June sequencing run were preferentially chosen since their sequencing quality is estimated higher compared to the March sequencing run. Below is a list of excluded samples based on the criteria above (see Table 4).

A detailed explanation on why which samples was excluded can be found in chapter 3.2.3

Table 4: Excluded samples, because of double presence.

Sample ID	Sequencing ID	Lab ID	Sample Type	Sample Conditions °C	Sequencing date
222491	ARD13_op1_oL1_nI1	ARD13-RAD13	Extract	-20	February 2023
233224	ARD13_op1_nL1_nI1	ARD13-RAD13	Powder	4 - 6	June 2023
222485	AUG175_op1_nL1_nI1	AUG175	Powder	4 - 6	February 2023
225373	AUG175_op1_oL1_nI1	AUG175	Extract	-20	March 2023
225354	AUG175_op1_oL1_oI1	AUG175	Indexed Library	-20	March 2023
225362	AUG182_op1_nL1_nI1	AUG182	Powder	4 - 6	March 2023
225363	AUG182_op1_oL1_nI1	AUG182	Extract	-20	March 2023
225361	AUG182_op1_oL1_oI1	AUG182	Indexed Library	-20	March 2023
222488	AUG87_op1_nL1_nI1	AUG87	Powder	4 - 6	February 2023
225357	AUG87_op1_oL1_oI1	AUG87	Indexed Library	-20	March 2023
222486	OI117_op1_nL1_nI1	OI117	Powder	4 - 6	February 2023
225355	OI117_op1_oL1_oI1	OI117	Indexed Library	-20	March 2023
225366	OI32_op1_nL1_nI1	OI32	Powder	4 - 6	March 2023
225365	OI32_op1_oL1_nI1	OI32	Extract	-20	March 2023
225364	OI32_op1_oL1_oI1	OI32	Indexed Library	-20	March 2023

225369	OI95_op1_nL1_nI1	OI95	Powder	4 - 6	March 2023
225368	OI95_op1_oL1_nI1	OI95	Extract	-20	March 2023
225367	OI95_op1_oL1_oI1	OI95	Indexed Library	-20	March 2023
225371	S111_op1_oL1_nI1	S111	Extract	-20	March 2023
225370	S111_op1_oL1_oI1	S111	Indexed Library	-20	March 2023
233217	S111_op1_nL1_nI1	S111	Powder	4 - 6	June 2023
222493	S176_op1_oL1_nI1	S176	Extract	-20	February 2023
222482	S176_op1_oL1_oI1	S176	Indexed Library	-20	February 2023
222490	S202A_op1_nL1_nI1	S202A	Powder	4 - 6	February 2023
222494	S202A_op1_oL1_nI1	S202A	Extract	-20	February 2023
222483	S202A_op1_oL1_oI1	S202A	Indexed Library	-20	February 2023
OI89	OI89-2015	OI89			2015

2.4 Final list of samples: after data cleansing

The following list (see Table 5) is the corrected list of samples meaning after quality control. This list encompasses all samples used for the final analysis:

Table 5: List of all samples for final data analysis. Blanks are not used for the analysis but encompass a quality control: Sample type refers to sample type by preparation stage.

Sample ID	Sequencing ID	Lab ID	Sample Type by Preparation Stage	Sample Conditions °C	Sequencing date
225353	ARD13_op1_oL1_oI1	ARD13-RAD13	Indexed Library	-20	March 2023
233230	ARD13_op1_oL1_nI1	ARD13-RAD13	Extract	-20	June 2023
222484	ARD13_op1_nL1_nI1	ARD13-RAD13	Powder	4 - 6	February 2023
233199	AUG175_op1_oL1_oI1	AUG175	Indexed Library	-20	June 2023
233218	AUG175_op1_oL1_nI1	AUG175	Extract	-20	June 2023
233225	AUG175_op1_nL1_nI1	AUG175	Powder	4 - 6	June 2023
233206	AUG182_op1_oL1_oI1	AUG182	Indexed Library	-20	June 2023
233208	AUG182_op1_oL1_nI1	AUG182	Extract	-20	June 2023
233207	AUG182_op1_nL1_nI1	AUG182	Powder	4 - 6	June 2023
233202	AUG87_op1_oL1_oI1	AUG87	Indexed Library	-20	June 2023
233239	AUG87_op1_oL1_nI1	AUG87	Extract	-20	June 2023
233228	AUG87_op1_nL1_nI1	AUG87	Powder	4 - 6	June 2023
233200	OI117_op1_oL1_oI1	OI117	Indexed Library	-20	June 2023
233226	OI117_op1_nL1_nI1	OI117	Powder	4 - 6	June 2023
233209	OI32_op1_oL1_oI1	OI32	Indexed Library	-20	June 2023
233210	OI32_op1_oL1_nI1	OI32	Extract	-20	June 2023
233211	OI32_op1_nL1_nI1	OI32	Powder	4 - 6	June 2023
233219	OI57_op1_oL1_oI1	OI57	Indexed Library	-20	June 2023
233231	OI57_op1_oL1_nI1	OI57	Extract	-20	June 2023
233227	OI57_op1_nL1_nI1	OI57	Powder	4 - 6	June 2023
233212	OI95_op1_oL1_oI1	OI95	Indexed Library	-20	June 2023
233213	OI95_op1_oL1_nI1	OI95	Extract	-20	June 2023
233214	OI95_op1_nL1_nI1	OI95	Powder	4 - 6	June 2023
233215	S111_op1_oL1_oI1	S111	Indexed Library	-20	June 2023
233216	S111_op1_oL1_nI1	S111	Extract	-20	June 2023

225372	S111_op1_nL1_nI1	S111	Powder	4 - 6	March 2023
233222	S176_op1_oL1_oI1	S176	Indexed Library	-20	June 2023
233232	S176_op1_oL1_nI1	S176	Extract	-20	June 2023
222489	S176_op1_nL1_nI1	S176	Powder	4 - 6	February 2023
233223	S202A_op1_oL1_oI1	S202A	Indexed Library	-20	June 2023
233233	S202A_op1_oL1_nI1	S202A	Extract	-20	June 2023
233229	S202A_op1_nL1_nI1	S202A	Powder	4 - 6	June 2023

Samples from 2015			
ARD13-RAD13	RAD13-2015	ARD13-RAD13	2015
AUG175	AUG175-2015	AUG175	2015
AUG182	AUG182-2015	AUG182	2015
AUG87	AUG87-2015	AUG87	2015
OI117	OI117-2015	OI117	2015
OI32	OI32-2015	OI32	2015
OI57	OI57-2015	OI57	2015
OI89	OI89-2015	OI89	2015
OI95	OI95-2015	OI95	2015
S111	S111-2015	S111	2015
S176	S176-2015	S176	2015
S202A	S202A-2015	S202A	2015

Blanks				
222495	LB1_op1_nL1_nI1	LB1	Blank	February 2023
222496	EB2_op1_nL1_nI1	EB2	Blank	February 2023
222497	LB2_op1_nL1_nI1	LB2	Blank	February 2023
222498	LB3_old_op1_oL1_nI1	LB3	Blank	February 2023
222526	PCRB_1.1	PCRB_1.1	Blank	February 2023
222527	PCRB_1.2	PCRB_1.2	Blank	February 2023
225376	EB1_op1_nL1_nI1	EB1	Blank	March 2023
233221	EB1_op1_nL1_nI1	EB1	Blank	June 2023
233234	LB1_op1_nL1_nI1	LB1	Blank	June 2023
233235	EB2_op1_nL1_nI1	EB2	Blank	June 2023
233236	LB2_op1_nL1_nI1	LB2	Blank	June 2023
233237	LB3_old_op1_oL1_nI1	LB3	Blank	June 2023
233238	PCRB_1.1	PCRB_1.1	Blank	June 2023

2.5 Pretreatment of samples in the grinding lab facility

Pretreatment measurements for the samples include UV-radiation and sodium hypochlorite (bleach) to remove all possible sources of exogenous contamination. Bones are typically stored in bags. These bags are bleached prior entering of the lab facility and upon inside the grinding lab further decontaminated by UV and bleached on both bag sides.

2.6 Grinding process in the ancient/grinding Lab

The scientific experiment started with 12 already pre-ground bone powders. The grinding process was done in accordance to the protocols described by (Pinhasi et al., 2015, 2019). The inner part of the

osseous human ear the cochlea, referred to as part C, has been identified as the best anatomical specimen for DNA preservation and therefore DNA retrieval (Pinhasi et al., 2015). All bones were samples from the temporal bone, except for individual OI57 for which the bone of origin is unknown (see Table 1). The original samples were ground in 2015 at the University College Dublin.

2.7 DNA extraction, library preparation and Indexing PCR

In accordance with general laboratory framework, the following two protocols as described by Dabney, Knapp, et al., (2013) for DNA extraction and Meyer and Kircher (2010) for library preparation were utilized. This protocol was generally followed, with a few modifications to maintain consistency with the workflow used eight years ago (February of 2015). The specific adaptations will be discussed in detail when addressing the workflow. Detailed steps and specific reagent concentrations are provided in Appendix B: B.1 Detailed DNA Extraction Protocol and B.2 Library preparation protocol.

2.7.1 Ancient DNA Extraction

Ancient DNA was extracted (Dabney, Knapp, et al., 2013) from bone powder using a standardized protocol optimized for highly fragmented DNA molecules. Extraction starts with incubation of bone or teeth samples (~50 mg) with extraction buffer overnight (16 - 24 hours, at 37°C). The whole extraction process takes two days, because of the overnight incubation. The extraction buffer, containing EDTA and Proteinase K, dissolves the bone matrix, releasing the DNA. Following lysis, the released DNA is bound to a silica through guanidine hydrochloride and isopropanol. The sample is then washed with PE buffer to remove contaminants such as cellular debris. Finally, the DNA can be eluted (released) from the silica using a low-salt buffer, yielding the purified DNA extract. In the protocol used in 2015 the MinElute column was attached to the extension reservoir for binding, but this procedure was adapted to the use of the Roche High Volume columns opted not to follow for safety and practicality reasons.

2.7.2 Ancient DNA Library Preparation

Ancient DNA (aDNA) libraries (Meyer & Kircher, 2010) were constructed by following established protocols optimized for short, highly degraded DNA fragments. These protocols have been specifically developed to handle samples with minimal and highly fragmented DNA (Henneberger et al., 2019; Margulies et al., 2005), because it is necessary to employ specific adaptations tailored to the difficulties of handling aDNA samples. Unlike modern DNA, aDNA does not require fragmentation prior to sequencing (Sanger et al., 1977), since one of its characteristic is of fragment reads.

Library preparation for aDNA samples (Meyer & Kircher, 2010) consists of four main steps: blunt-end repair, adapter ligation, adapter fill-inn, and in the end indexing PCR (see chapter Indexing PCR and pooling). Blunt-end repair fills in or removes overhanging 5’- and 3’ DNA-ends using T4 DNA polymerase. During adapter ligation typically, both P5 and P7 adapters would be ligated to the fragment ends by the T4 DNA ligase. In this experiment, only the P7 adapter was used to adhere to the old protocol (see protocol in Appendix B: B.2 Library preparation protocol for Illumina sequencing). Ligation joins only single strands. Nicks in the DNA are filled in by Bst polymerase, an enzyme known for its strand-displacement function.

2.7.3 Indexing PCR and pooling

The final two step of the library preparation are indexing PCR (Meyer & Kircher, 2010) and pooling. Amplification with 5'-tailed primers normally adds both indexes, and complete adapter sequences. Single (P7) indexing was added, since this resembles the experimental set up of 2015, instead of double (P7 and P5) indexing, which is routinely done nowadays. Indexing ensures the identification of each individual sample after pooling of all samples together. The polymerase Accuprime Pfx Supermix was

used, and the PCR run for 12 cycles. The Primer Primer IS4 (10mM) was added for the ligation of the index. The Primer IS4 (10mM) and Accuprime Pfx Supermix are referred together as Amplification Mastermix. With a total reaction volume of 25 µl, which encompasses 21µl of mastermix, 1µl of appropriate indexing adapter (5µM) & 3µl of DNA sample, the sample will be pipette-mixed and is ready for the indexing PCR amplification step. After completion of indexing PCR, the final step is sample clean-up. This encompasses different steps of washing the sample with different buffers, such as PB, EB and EBT, and different steps in between of centrifugation to obtain the purified indexed library ready for pooling and subsequently for sequencing. Pooling of indexed libraries for sequencing is done in equimolar ratio. Pooling is done in a post-PCR facility. When the pooling is finalised, the samples can proceed to sequencing. In-between Indexing PCR and pooling, which is the final step before sequencing, two more laboratory steps are performed, which are tape station and qubit fluorometric measurements.

2.8 Tape station and Qubit fluorometric measurement

Both the Tape station and Qubit measurements were assessed as a qualitative measurement to decide if the sample can further proceed to sequencing. The higher the Qubit value the more DNA is present in the sample. The results from the Qubit measurements and confirmation that the samples have passed the tape station can be found in Appendix A: A.1 Qubit results.

2.9 qPCR

Quantitative PCR (qPCR) was done as an additional measurement in this experiment to assess the quantitative amount of the DNA present in the indexed libraries. The Cq value has an inverse relationship to the amount of DNA. A low Cq value means high amount of DNA, since the Cq-value indicates the number of cycles that have to be gone through to pass the amplification threshold. The qPCR is routinely done from sequencing libraries, that have not been indexed previously. Normally the qPCR is used to assess the amount of DNA to determine how many cycles will be used for the Indexing PCR. Since in this experimental setup the number of cycles was always the same and not corrected, the Cq cycles were not corrected, and therefore only the amount of DNA assessed. The qPCR values for the samples can be found in Appendix A: A.2 qPCR results.

2.10 DNA Illumina Sequencing

In 2023, three sequencing runs using Illumina sequencing technology were performed, taking place in February, March and June. The February and March sequencing run utilized NovaSeq technology, while the June sequencing run employed NextSeq technology to resequence all samples. All March and February samples were resequenced in June, except for one sample 222489, which could not be resequenced due to insufficient DNA left in the indexed library for pooling. Resequencing with “NextSeq” was required due to problems encountered during data evaluation for the March run. These issues have since been solved; all samples have been initially incorporated into the analysis. However final incorporation for the final analysis will depend on the successful passing of the quality control analysis (see chapter Quality control assessment).

2.11 Bioinformatics analysis

The data analysis was performed on a cluster shared by the Life Sciences (LISC) at the University of Vienna. This chapter provides a summary of all performed bioinformatics steps. The analysis began from already demultiplexed files.

The first step was to trim the sequence with the tool cutadapt (Version 5.0), using .fastq.gz files as input data (see Script Step1). The sequences were trimmed, and the minimum length was set as 30bp (option -m 30).

Next, the sequences were aligned (see script Step 2) with the tool Burrows-Wheeler Aligner (BWA) (Version 0.7.17) to the reference genome, which in this case was human_g1k_v37.fasta. The .rmdup function in SAMtools was applied to remove duplicates from the aligned sequences.

The software SAMtools (Version 1.21) is a suite of bioinformatic programs. SAMtools was used with following commands: view, sort, rmdup, index, flagstat, mpileup, in many of the scripts provided in Appendix C (see Step31 till Step43).

Contamination was calculated using the software calico (Version 0.2) (Krause, Adrian, et al., 2010) (available on https://github.com/pontussk/calico).

Kinship analysis was done by modifying a Python script from the software TKGWV2.py (Version 1.0) (Fernandes et al., 2021) (available on https://github.com/danimfernandes/tkgwv2) in a new customised bash shell script written for this project and named PKin.sh. PKin.sh (see Script PKin.sh in Appendix C) mirrors the Python script TKGWV2.py and launches the R script plinktotkrelated.R (also written by Daniel Fernandes) that calculates the relatedness between the chosen samples. Kinship analysis uses plink text files produced with Plink 1.9.

Sex-determination was assessed through the software ry_compute (Version 0.4) (Skoglund, 2013; Skoglund et al., 2013) (available on https://github.com/pontussk/ry_compute). A customized bash shell script was produced to run it in the bash environment on the cluster.

Ancient DNA was authenticated through the software mapDamage (Version 2.0) (available on https://ginolhac.github.io/mapDamage/) (Ginolhac, Rasmussen, et al., 2011; Jónsson et al., 2013). The output of the same software was used to obtain the deamination rates for subsequent statistical analysis with SPSS.

The software AMBER (Dolenz et al., 2024) (available on https://github.com/tvandervalk/AMBER) used to calculate the sample's fragment lengths and fragment length distribution. Normally, AMBER requires updating the sample name and path after each sample in the textfile "BamList.tsv". For this project, a customized script (see Appendix C4), was employed to run multiple BAM files automatically sequentially, if they were located in the same directory. This means this script automatically updates the list "BamList.tsv" with each new sample name, BAM-file and path.

The Microsoft software Excel (MS Office 365), the software R (Version 4.4.2) (included the toolchain bundle RTools Version 4.4) and RStudio (Version 2024.12.0) were used on a Personal Computer to complete the analysis.

A list of the most important customized bioinformatic bash scripts can be found in Appendix C1 – C6.

2.12 Statistical Analysis via SPSS and R

All statistical analysis will be computed with the Software IBM SPSS Statistics (Version 29. 0. 0. 0). The results of these analyses will be found in the chapter 4 Results. Customized scripts have also been created for the statistical analysis in SPSS, see chapter 9.4. Customized Scripts for analysis in SPSS.

- Mann Whitney Independent Samples 3 Tests.sps: This SPSS Syntax was realized to conduct an independent-samples Mann-Whitney U test to assess whether the distribution of Cytosine to Thymine deamination at position 1 (pos1) differs between the baseline group of 2015 across each of the sample preparation stages: Extract, Indexed Library and Powder. The same SPSS Syntax was then applied to examine the relationship for Guanine to Adenine at position1 between the two groups.

- CreateTransposeForRelatedTest_CtoT_MAX.sps: This SPSS Syntax was developed to transposes SPSS tables, so that the table originally used for the Mann-Whitney Independent samples analysis can be transformed into a table that can be used for a related test Wilcoxon statistic. Again, the same SPSS Syntax was then applied to examine the relationship for Guanine to Adenine at position1 between the two groups.
- WilcoxonRelatedRankTransposed.sps: This SPSS Syntax was realized to conduct a related-samples Wilcoxon test to assess whether the distribution of Cytosine to Thymine deamination at position 1 (pos1) differs between the 2015 group across each of the sample preparation stages: Extract, Indexed Library and Powder. The same SPSS Syntax was then adapted for two additional datasets, by replacing the variables pos1_2015, pos1_Extract, pos1_IndexedLibrary and pos1_Powder with the corresponding variables fragment_length_2015, fragment_length_Extract, fragment_length_IndexedLibrary, fragment_length_Powder for the mean fragment length analysis. For the related Wilcoxon analysis on the rate of unique reads the same SPSS Syntax was used, with variables adjuste to UniqueReads_2015, UniqueReads_Extract, Reads_IndexedLibrary and UniqueReads_Powder.
- Mann Whitney Independent Samples 15 Tests.sps is not added in the appendix but is a modification of the SPSS Syntax Mann Whitney Independent Samples 3 Tests.sps to calculate the fragment length distribution.

The program R was also used to create tables, which were further on used for the statistical analysis. The tables grouped using R are the following: Table 12 and Table 24.

A list of the most important customized SPSS Syntax and R scripts can be found in Appendix C7 and C8.

3 Results

In this section of the thesis, the results of the experiment studying the influence of long-term storage on aDNA samples from different laboratory preparation stages are being presented. In total 11 of the original 12 individuals were further used for the analysis. Bone powders, extracts, and indexed libraries processed in 2023 were compared with the same individuals *(n = 11*) produced 8 years ago in 2015, using the same protocols to ensure reproducibility.

In general, mostly two types of tests were conducted. For independent sample comparisons the Mann-Whitney U test *(n = 11*) was used. For the related sample comparison Wilcoxon test was used. Specifically, for the Wilcoxon related test comparison 11 different individuals *(n = 11*) were included for two test comparisons (Powder, Indexed Library) and only 10 individuals *(n = 10*) for the third group Extract. Missing values are marked with -999 in the SPSS dataset.

3.1 Main findings

Here we report, a statistically significant difference in Cytosine to Thymine deamination at position 1 for the independent samples Mann-Whitney U-test comparison between the groups Powder and the group 2015. The group Powder showed less deamination at position 1, indicating therefore better preservation compared to the group 2015. The samples processed from Powder in 2023 are statistically significantly less deaminated at position 1 than the samples in 2015.

Moreover, the Wilcoxon related sample comparison between samples from all preparation methods (Extract, Indexed Library, Powder) and the baseline group from 2015 show a statistically significant difference in Cytosine to Thymine deamination at position 1. However, the samples produced from

Extract are statistically significantly higher deaminated, hence suggesting higher degradation of the extracts during the 8-year storage time. On the other hand, the samples produced from Powder and Indexed Library are statistically significantly less deaminated, suggesting better preservation during the same 8-year long-term storage time.

When repeating the same tests, for the Guanine to Adenine deamination, only the Extract group showed a statistically significant increase in Guanine to Adenine deamination at position 1 compared to 2015, suggesting increased degradation over the 8 years of long-term storage. No statistically significant difference in the Guanine to Adenine deamination at position 1 was found for the Powder and Indexed Library groups compared to 2015, suggesting the 8-year storage time did not cause a significant difference for these groups over the same period.

Despite, the higher rate of deamination for the Extract group, the mean fragment length and the fragment length distribution have statistically significantly increased across all categories. Moreover, the rate of unique reads has increased statistically significantly as well.

3.2 Quality control assessment for final analysis

Listed below are various analysis that have been done to screen, which samples to include in the final analysis to calculate the effect of long-term storage on aDNA samples.

3.2.1 Kinship analysis

Kinship analysis was performed as a measure of quality control. Plink-files were used as input data and the analysis was conducted using the software TKGWV2 (Fernandes et al., 2021).

The script "plink2tkrelated" from Daniel Fernandes program TKGWV2 (https://github.com/danimfernandes/tkgwv2) was used (see description below). The results include the following 7 columns:

- *"Sample1 - First individual of tested pair, Sample2 - Second individual of tested pair"*
- *"Used_SNPs - Used SNPs to calculate relatedness"*
- *"HRC - Halved Relatedness Coefficient. (Unrelated < 0.0625. 2nd Degree between 0.0625 and 0.1875. 1st Degree > 0.1875)"*
- *"counts0 - Number of non-shared alleles"*
- *"counts4 - Number of shared alleles"*
- *"Relationship - Descriptive relationship based on HRC value"*

(See (D. Fernandes, 2022) for further details https://github.com/danimfernandes/tkgwv2)

The kinship relationship is determined by the so-called HRC - Halved Relatedness Coefficient, which calculates the relationship based on the shared SNPs. The HRC produces a range of values, which determine the relationship between two samples. The values are the following: unrelated for values <= 0.0625, 2nd degree between 0.0625 > 0.1875, 1st degree between 0.1875=> 0.3125 and for same individual/twin >=0.3125. The kinship analysis was done in three steps. First, kinship relationships among the 2023 samples were compared. Secondly, the relatedness of the cleansed 2023 dataset was compared with the samples analysed 8 years ago in 2015. Samples that did not match the original sample from 2015 were excluded. Third, if duplicates of the same sequencing ID were present they had to be identified and removed to obtain the final dataset, for results see Table 6, Table 7, Table 8.

3.2.2 First and second stage of kinship analysis

The first stage of kinship analysis aimed to determine the relationships among all samples from the original sample pool (see chapter 2.2 Original list of samples). Each sample was assigned to a kinshipID

"KID". This assignment groups samples with the same KID as belonging to the same individual. Samples within the same KID can be related with each other as same individual/twins, 1st degree or 2nd degree. The higher the degree of relatedness, the more confidently the relationship can be established. Samples that match according to the label but result as unrelated with their original KID will be automatically excluded from that KID. Same individual or 1st degree relatedness was deemed sufficient to determine a sample's assignment to a KID. Samples showing only 2nd degree relatedness were subjected to additional testing to determine if the sample could potentially belong to another KID. If a sample could not be confidently assigned to a certain individual, then this specific samples was excluded (see chapter 2.3.1).

Kinship analysis identified discrepancies during sample preparation, evidenced by mismatches between sequencing IDs and their assigned KIDs. These included potential sample swaps and the misidentification of one blank as a sample. The following samples were reassigned through kinship analysis:

- The samples 222491 (originally AUG175_op1_oL1_nI1) and 233230 (originally AUG175_op1_oL1_nI1) which originally belonged to the Lab ID AUG175 were both correctly assigned through kinship analysis to the sequencing ID ARD13_op1_oL1_nI1 and hence the Lab ID ARD13-RAD13.
- The samples 225373 (originally ARD13_op1_oL1_nI1) and 233218 (originally ARD13_op1_oL1_nI1) were both correctly assigned through kinship analysis to the sequencing ID AUG175_op1_oL1_nI1 and Lab ID AUG175. ARD13 and RAD13 are the same individual.
- All samples from LabID OI89 were excluded, because none of the samples processed in 2023 matched OI89-2015.

Moreover, all samples with the Lab ID OI157 were reassigned to the Lab ID OI57: meaning the samples 233227, 233231, 233219 (see Table 6).

Furthermore, kinship analysis confirmed the following reassignments:

- The sample 233219 (originally AUG87_op1_oL1_nI1) was reassigned through kinship analysis to the sequencing ID OI57_op1_oL1_oI1.
- The sample 233239 (originally PCRB_1.2) was reassigned through kinship analysis to the sequencing ID AUG87_op1_oL1_nI1.

Table 6: Results of kinship analysis for 2023 samples, showing the relatedness only between the 2023 samples: (1) dark green indicate very good results, meaning that this samples is related with the other samples of the same individual as "twin/same individual" (2) light green indicates first degree relatedness and (3) amber indicates more than one 2nd degree relatedness. Light yellow (4) indicates bad results.

ID	KID	ONN		OON		OOO	
ARD 13 - RAD13	1	222484	233224	222491	233230	225353	
AUG175	2	222485	233225	225373	233218	225354	233199
AUG182	3	225362	233207	225363	233208	225361	233206
AUG87	4	222488	233228	225374	233239	225357	233202
OI57	5	233227		233231		225356	233219
OI117	6	222486	233226			225355	233200
OI32	7	225366	233211	225365	233210	225364	233209
OI95	8	225369	233214	225368	233213	225367	233212
S111	9	225372	233217	225371	233216	225370	233215
S176	10	222489		222493	233232	222482	233222

S202A	11	222490	233229	222494	233233	222483	233223
OI89	12	225359	233204	225360	233205	225358	

The second stage of kinship analysis compared the 2023 samples with the samples in 2015 (see Table 7 for results). Samples that did not match the 2015 original samples, highlighted as red in Table 7, were excluded.

Table 7: Kinship results of the comparison between the samples of 2023 with the 2015 samples: The colour indicates if they are related to the original samples from 2015: (1) dark green indicate very good results, meaning that this samples is related with the other samples of the same individual as "twin/same individual" (2) light green indicates first degree relatedness and (3) Red indicated not related to the 2015 samples.

ID	KID	ONN		OON		OOO	
ARD 13 - RAD13	1	222484	233224	222491	233230	225353	
AUG175	2	222485	233225	225373	233218	225354	233199
AUG182	3	225362	233207	225363	233208	225361	233206
AUG87	4	222488	233228	225374	233239	225357	233202
OI57	5	233227		225374	233231	225356	233219
OI117	6	222486	233226			225355	233200
OI32	7	225366	233211	225365	233210	225364	233209
OI95	8	225369	233214	225368	233213	225367	233212
S111	9	225372	233217	225371	233216	225370	233215
S176	10	222489		222493	233232	222482	233222
S202A	11	222490	233229	222494	233233	222483	233223
OI89	12	225359	233204	225360	233205	225358	

3.2.3 Third stage of kinship analysis: for further sample exclusion

Since both versions passed the data quality assessment further uniform criteria had to be established to decide which samples to select for the final analysis. These criteria are the following in the following order:

1. If two versions of the same sample are present, the sample which is not related as same individual/twin to the 2015 sample will be chosen.
2. For two samples that are both related as same individual to the 2015 sample the sample for which more than 2,3 times more Used_Snps were used for the calculation of the HRC coefficient were redeemed as of better quality and chosen.
3. If two samples are both related as same individual to the 2015 sample, the sample with less than 10 USED_Snps to calculate the relationship was excluded.
4. If two samples passed all the above criteria the remaining duplicates were excluded based on the sequencing run. The samples from the June run were favoured since their sequencing quality is redeemed higher.

Based on the above criteria, the following samples were excluded from the final dataset (see Table 8).

- Based on criterium 1 the following samples were excluded: 233224, 222486, 222483, 222494
- Based on criterium 2 the following samples were excluded: 222485, 222488, 222490, 222491, 225354, 225355, 225362, 225364, 225365, 225366, 225368, 225369, 225370, 225371, 225373
- Based on criterium 3 the following samples were excluded: 233217, 222482, 222493

- Based on criterium 4 the following samples were excluded: 225357, 225367

Table 8: Final list of samples after the third step of kinship analysis.

ID	KID	ONN	OON	OOO
ARD 13 - RAD13	1	222484	233230	225353
AUG175	2	233225	233218	233199
AUG182	3	233207	233208	233206
AUG87	4	233228	233239	233202
OI57	5	233227	233231	233219
OI117	6	233226		233200
OI32	7	233211	233210	233209
OI95	8	233214	233213	233212
S111	9	225372	233216	233215
S176	10	222489	233232	233222
S202A	11	233229	233233	233223

3.2.4 Sex determination

Sex determination was performed as another measure of quality control. For the sex determination the script "ry_compute" (Skoglund, 2013; Skoglund et al., 2013) was used. Additionally, a customized script was written to run the analysis in a bash environment and to process multiple samples at the same time. Also, the original python script "ry_compute" was edited so it works with python3.

The female limit is R_y < 0.016 and the male limit is R_y> 0.075, if there is a match with these criteria, the sex can be established.

Table 9: Sex determination for all individuals. The table includes the initial sex determination from the archeological record and the bioinformatical sex assessment done for this experiment. The table also contains information on which samples succesfully passed the sex determination analysis.

Lab ID	Assignment from the script	original Sex Assignment	Age	Samples,which passed the sex assignment
ARD13	XX	subadult (no assigned sex)	7-8 y	225353, 222484, 222491, 233224, 233230, RAD13-2015
AUG175	XY	male	28-35 y	222485, 225354, 225373, 233199, 233218, 233225, AUG175-2015
AUG182	XX	female	35-45 y	225361, 225362, 225363, 233206, 233207, 233208, AUG182-2015
AUG87	XY	male	28-35 y	222488, 225357, 233202, 233228, 233239, AUG87-2015
OI57	XX	subadult (no assigned sex)	11-12 y	233219, 233227, 233231, OI57-2015
OI117	XX	subadult (no assigned sex)	1-2 m	222486, 225355, 233200, 233226, OI117-2015
OI32	XY	male	40-50 y	225364, 225365, 225366, 233209, 233210, 233211, OI32-2015
OI89	XX	subadult (no assigned sex)	1-1.5 y	225358, 225359, 225360, 233204, 233205, OI89-2015
OI95	XX	subadult (no assigned sex)	0.5-1 y	225367, 225368, 225369, 233212, 233213, 233214, OI95-2015
S111	XX	subadult (no assigned sex)	0-0.5 y	225370, 225371, 225372, 233215, 233216, 233217, S111-2015
S176	XX	subadult (no assigned sex)	13-14y	222482, 222489, 222493, 233222, 233232, S176-2015
S202A	XX	female	28-35y	222483, 222490, 222494, 233223, 233229, 233233, S202A-2015

The specific values for each sample can be found in Appendix D: Supplementary Data Tables, as well as the values for the already previously excluded samples and the blanks. If it was not possible to

determine the sex, or if the determined sex did not match the correct one the samples was excluded (see chapter 2.3.1).

The only sample that did not pass the sex determination is 225374. It had already been excluded previously through kinship analysis (see chapter 3.2.2). The individual AUG87 is male (assigned XY), but the sample 225374 could not be assigned neither as male of female through the software and resulted in a R value of 0.0249.

3.2.5 Authentication of aDNA samples through mapDamage pattern

As a measure of the authenticity of the ancient DNA samples, the degree of deamination of each sample was assessed bioinformatically by using a so-called mapDamage pattern script (Ginolhac, Rasmussen, et al., 2011; Jónsson et al., 2013). The mapDamage, script utilizes mapped SAM files to a reference genome, to generate summary tables that are essential for assessing the authenticity of ancient DNA (aDNA) based on patterns of DNA damage. This mapDamage pattern clearly distinguishes between modern and ancient DNA samples. The authentication will be done visually.

According to the website of the developers of the mapDamage program the graph should be interpreted in the following way (Ginolhac, Rasmussen, et al., 2011; Jónsson et al., 2013):

- *"Red: C to T substitutions"*
- *"Blue: G to A substitutions"*
- *"Grey: All other substitutions"*
- *"Orange: Soft-clipped bases"*
- *"Green: deletions relative to the reference"*
- *"Purple: insertions relative to the reference"*

*(*See (Ginolhac, Schubert, et al., 2011) for further information https://ginolhac.github.io/mapDamage/)

The main damage patterns that the mapDamage software detects is cytosine deamination, and the guanine to adenine deamination on the complementary DNA strand.

Soft-clipped bases are defined as DNA fragments that have not been aligned to the reference genome, resulting in a partially mapped read (Suzuki et al., 2011). The mapDamage reports shows that there are some samples that show soft-clipped bases and a possible improvement of this work could be to perform a detailed analyse of the impact of this factor on the sample quality.

For all the samples that were selected for the final analysis (see chapter 0) and the samples from 2015 the authenticity has been verified, as an example for the individual ARD13-RAD13 see Figure 1 and Figure 2. Figure 1 is the sample 233230 sequenced and analysed in 2023 and Figure 2 is the sample ARD13-2015, hence sequenced and analysed in 2015.

Figure 1: graph produced by the software mapDamage for the sample 233230: showing the apparent Cytosine to Thymine deamination and the apparent Guanine to Adenine deamination between position 1 and position 25.

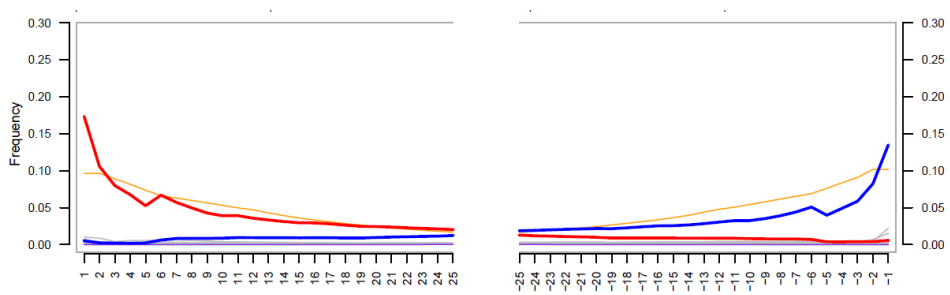
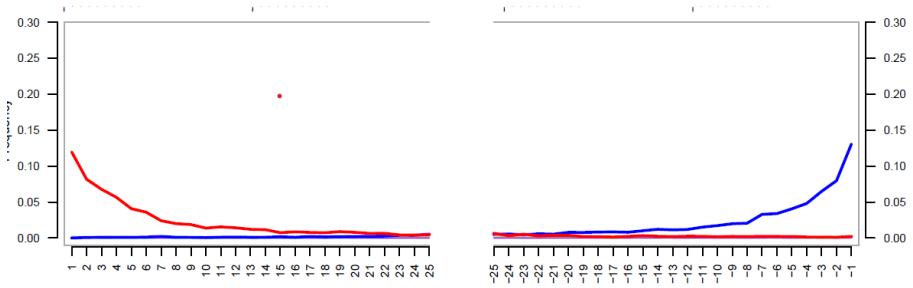


Figure 2: graph produced by the software mapDamage for the sample 233230: showing the apparent Cytosine to Thymine deamination and the apparent Guanine to Adenine deamination between position 1 and position 25.



3.2.6 Screening for contamination: degree of contamination

Below are the results for the samples, that were screened bioinformatically for contamination with the software "calico" (https://github.com/pontussk/calico) (Krause, Adrian, et al., 2010), see chapter 2.12 Statistical Analysis via SPSS and R. The tables below are colour coded. Green indicates contamination levels below 5%, yellow indicates contamination levels over 5% and red indicates contamination levels higher than 10%, see Table 10, Table 11.

Table 10: Colour coded contamination results for each Sample ID, obtained through the software calico.

ID	KID	ONN		OON		OOO	
ARD 13	1	222484	233224	222491	233230	225353	
AUG175	2	222485	233225	225373	233218	225354	233199
AUG182	3	225362	233207	225363	233208	225361	233206
AUG87	4	222488	233228	225374	233239	225357	233202
OI157	5	233227		233231		225356	233219
OI117	6	222486	233226			225355	233200
OI32	7	225366	233211	225365	233210	225364	233209
OI95	8	225369	233214	225368	233213	225367	233212
S111	9	225372	233217	225371	233216	225370	233215
S176	10	222489		222493	233232	222482	233222
S202A	11	222490	233229	222494	233233	222483	233223
OI89	12	225359*	233204	225360	233205	225358	

Table 11:Colour coded contamination results for each Sample ID, obtained through the software calico. It serves as a supplementary detail to Table 10, providing the exact contamination rates for those Sample IDs.

ID	KID	ONN		OON		OOO	
ARD 13	1	7.50%	8.70%	7.40%	13.50%	0.00%	
AUG175	2	6.50%	10.70%	6.10%	7.10%	8.60%	6.20%
AUG182	3	6.00%	10.50%	3.50%	9.70%	9.30%	15.00%
AUG87	4	6.80%	12.60%	10.70%	NA	7.90%	7.50%
OI157	5	13.80%		11.20%		NA	14.60%
OI117	6	7.70%	10.80%			7.70%	13.30%
OI32	7	7.30%	12.00%	9.00%	9.90%	10.70%	15.50%
OI95	8	5.10%	12.00%	4.20%	12.60%	10.20%	10.20%
S111	9	8.10%	0.00%	9.20%	10.00%	5.00%	11.40%
S176	10	7.00%		15.60%	13.60%	33.30%	5.40%
S202A	11	7.00%	11.10%	12.00%	9.40%	NA	6.90%
OI89	12	3.90%	NA	7.80%	33.30%	11.60%	

3.2.7 Pairwise comparison of contamination between the samples from 2015 and the samples from 2023

The degree of contamination of the 2023 and the 2015 samples was collected and assessed. By comparing the contamination values between the samples from 2015 and the samples from 2023 it becomes apparent that there are much less available values for the samples from 2015. The sample size between samples from different sample preparation stages in 2023 varies: *n = 4* for the group 2015, *n = 11* for the group Powder and Indexed Library, and lastly *n = 9* for the group Extract, see Table 12.

Table 12: Contamination values for all samples available.

	Contamination in %			
LabID	2015	Powder	Extract	Indexed Library
ARD13-RAD13	NA	7.50%	13.50%	6.30%
AUG175	5.90%	10.70%	7.10%	6.20%
AUG182	11.50%	10.50%	9.70%	15.00%
AUG87	0.00%	12.60%	NA	7.50%
OI117	NA	10.80%	NA	13.30%
OI32	0.00%	12.00%	9.90%	15.50%
OI57	NA	13.80%	11.20%	14.60%
OI95	0.00%	12.00%	12.60%	10.20%
S111	NA	8.10%	10.00%	11.40%
S176	NA	7.00%	13.60%	5.40%
S202A	NA	11.10%	9.40%	6.90%

Normal distribution was not tested, since it is not applicable in this context. A related-samples Wilcoxon test was conducted to determine whether there is a difference in the rate of contamination between the baseline group from 2015 and the three sample preparation stages: Extract, Indexed Library and Powder.

The results show a significant result for the groupwise comparison between the contamination of the samples from 2015 and the samples produced from Indexed Library in 2023 ($p<0.43$). The median of differences between the contamination of the group 2015 and the categories Powder ($p<0.273$) and Extract ($p<0.78$) did not yield a significant result.

Specifically, the median difference between the 2015 and Extract groups did not show a statistically significant increase in the rate of contamination (*Z = 1.095, p = .273, n = 4;* exact significance, two-tailed. Similarly, the median difference between the 2015 and Powder groups showed a statistically significant increase in the rate of contamination (*Z = 1.761, p = .078, n = 5;* exact significance, two-tailed). On the other hand, the median difference between the 2015 and Indexed Library groups did show a statistically significant increase in the rate of contamination (*Z = 2.023, p = .043, n = 5;* exact significance, two-tailed).

The null hypothesis, which states that there is no difference in the rate of contamination between the 2015 and the groups Extract and Powder were both rejected. The null hypothesis stating that there is no difference in the rate of contamination between the 2015 and the Indexed Library group, was retained. These results indicates that samples from the categories Powder and Extract have not experienced a difference in contamination rate over the course of the 8-year storage period. The group of Indexed Library has exhibited a higher rate of contamination compared to the sample from 2015. These results indicate that the long-term storage time of 8 years has not led to a reduction in rate of contamination for all groups (Extract, Indexed Library, Powder). This finding suggests that the 8-year long-term storage has

resulted in an increase of contamination rate compared to the samples from 2015, for results see Table 13.

Table 13: Hypothesis test summary of related Sample Wilcoxon test of median of differences between the contamination of the samples from the group 2015 against the other groups.

Variable: Contamination				
Groups	Null Hypothesis	Test	Sig.[a,b]	Decision
2015 vs. Extract	The median of differences between 2015 and Extract equals 0.	Related-Samples Wilcoxon Signed Rank Test	.273	Retain the null hypothesis.
2015 vs. Indexed Library	The median of differences between 2015 and Indexed Library equals 0.	Related-Samples Wilcoxon Signed Rank Test	.043	Reject the null hypothesis.
2015 vs. Powder	The median of differences between 2015 and Powder equals 0.	Related-Samples Wilcoxon Signed Rank Test	.078	Retain the null hypothesis.
a. The significance level is .050	b. Asymptotic significance is displayed.			

3.2.8 Degree of contamination of blanks

Blanks are checked for contamination as a measure of quality control (see Table 14).

Table 14: Degree of contamination in Blanks.

Sample ID	LabID	Sample	Month	SITES	MAJ	MIN	CONT	CI_low	CI_up
222495	LB1	LB1_op1_nL1_nI1	Feb	1	2	1	33.30%	0	0.866
222496	EB2	EB2_op1_nL1_nI1	Feb	0	0	0	NA	NA	NA
222497	LB2	LB2_op1_nL1_nI1	Feb	0	0	0	NA	NA	NA
222498	LB3	LB3_old_op1_oL1_nI1	Feb	1	2	1	33.30%	0	0.866
222526	PCRB_1.1	PCRB_1.1	Feb	0	0	0	NA	NA	NA
222527	PCRB_1.2	PCRB_1.2	Feb	1	2	1	33.30%	0	0.866
225376	EB1	EB1_op1_nL1_nI1	Mar	4	9	4	30.80%	0.057	0.559
233221	OI89	EB1_op1_nL1_nI1	Jun	0	0	0	NA	NA	NA
233234	S111	LB1_op1_nL1_nI1	Jun	0	0	0	NA	NA	NA
233235	S111	EB2_op1_nL1_nI1	Jun	0	0	0	NA	NA	NA
233236	ARD13	LB2_op1_nL1_nI1	Jun	0	0	0	NA	NA	NA
233237	AUG87	LB3_old_op1_oL1_nI1	Jun	0	0	0	NA	NA	NA
233238	OI117	PCRB_1.1	Jun	0	0	0	NA	NA	NA

3.3 Assessing DNA degradation through Deamination values

For all the tests assessing the effect of apparent Cytosine to Thymine deamination, and the apparent Guanine to Adenine deamination a dataset was created, which excluded not the previously excluded samples, but also the blanks for further analysis.

Position 1 (=pos1) was chosen for all tests since Cytosine tends to show the highest deamination at 5'end and Guanine tends to show the highest deamination at the 3'end end of a fragment, resulting in an increased rate of Thymine at the 5'end and of Adenine at the 5'end (Briggs et al., 2007).

3.4 Assessing DNA degradation through Deamination of Cytosine to Thymine

A key characteristic for the assessment of DNA degradation is the deamination of Cytosine to Thymine. The values were obtained through a software called mapDamage (Ginolhac, Rasmussen, et al., 2011; Jónsson et al., 2013), the software entails information about Cytosine to Thymine deamination between position 1 and 25. For the statistical analysis the percentage of Cytosine to Thymine deamination was compared only specifically at positions 1, across 4 groups. The groups were defined as 2015, Powder, Extract and Indexed Library. The group names encompass the information about the sample preparation (Powder, Extract and Indexed Library) as well as the baseline comparison group (2015).

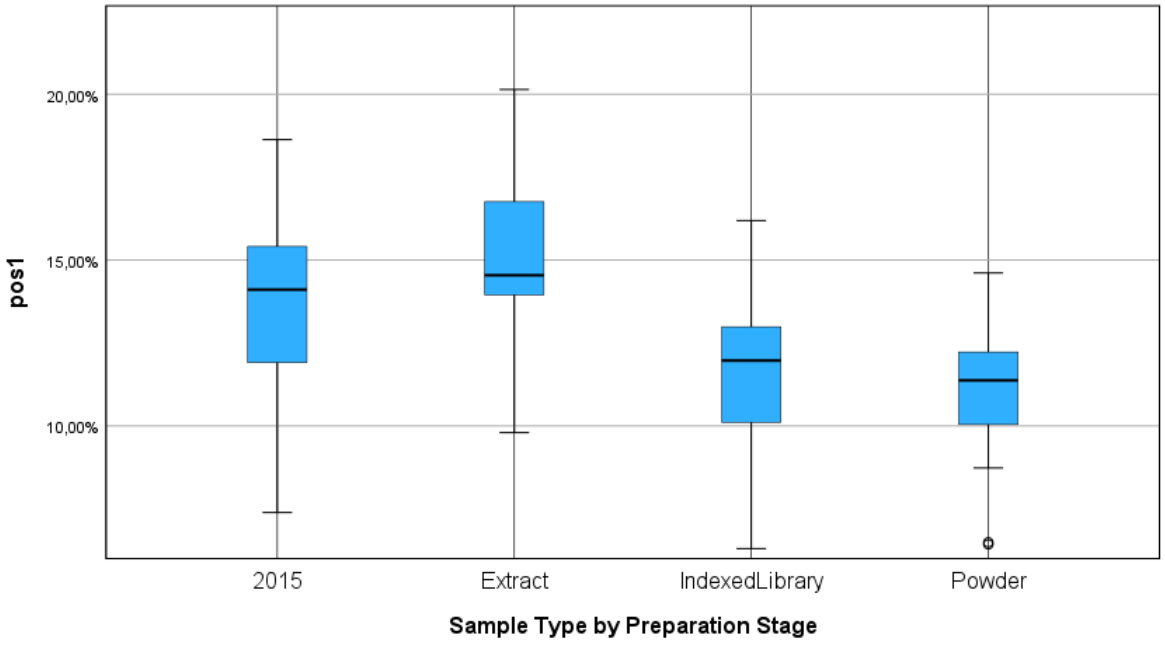
3.4.1 Descriptive Statistics

See table below (Table 15) for the descriptive statistics and Figure 3 for the boxplot comparison for the Cytosine to Thymine deamination (pos1) between the four groups.

Table 15: Median and interquartile range (IQR) for Cytosine to Thymine deamination (pos1) in the four Groups (2015, Extract, Indexed Library, Powder)

Sample Type by Preparation Stage			Descriptive Statistic - CtoT deamination pos1			
			Q1	Median	Q3	IQR
Weighted Average	pos1	2015	11.84%	14.00%	15.38%	3.54%
		Extract	14.09%	14.82%	16.91%	2.82%
		IndexedLibrary	9.87%	11.62%	12.97%	3.10%
		Powder	10.04%	11.28%	12.35%	2.31%

Figure 3: Boxplot comparison for the variable Cytosine to Thymine deamination (pos1) in the four groups (2015, Extract, Indexed Library, Powder)



3.4.2 Independent sample comparison of the samples through Mann-Whitney test

The data is not normally distributed. Independent-samples Mann-Whitney U test was conducted to determine whether the distribution of Cytosine to Thymine deamination at position 1 (pos1) differs

between the baseline group of 2015 across each of the sample preparation stages, being Extract, Indexed Library and Powder.

The Mann-Whitney-U test did not result in a statistically significant difference in the distribution of Cytosine to Thymine deamination at position 1 (pos1) between the 2015 group (*Mdn = 14.00%, IQR = 3.54%*) and the Extract group (*Mdn = 14.82%, IQR = 2.82%*) (Mann-Whitney's U-Test: *U = 73.00, Z = 1.268, p = .223, n = 21;* exact significance, two-tailed). The null hypothesis stating that the distribution of Cytosine to Thymine deamination at position 1 is the same between the groups of 2015 and the group of Extract, was retained.

The Mann-Whitney-U test did not result in a statistically significant difference in the distribution of Cytosine to Thymine deamination at position 1 (pos1) between the 2015 group (*Mdn = 14.00%, IQR = 3.54%*) and the Indexed Library group (*Mdn = 11.62% , IQR = 3.10%*) (Mann-Whitney's U-Test: *U = 35.00, Z = - 1.674, p = .101, n = 22;* exact significance, two-tailed). The null hypothesis stating that the distribution of Cytosine to Thymine deamination at position 1 is the same between the groups of 2015 and the group of Indexed Library, was retained.

However, the Mann-Whitney-U test did result in a statistically significant difference in the distribution of Cytosine to Thymine deamination at position 1 (pos1) between the 2015 group (*Mdn = 14.00%, IQR = 3.54%*) and the Powder group (*Mdn = 11.28%, IQR = 2.31%*) (Mann-Whitney's U-Test: *U = 26.00, Z = - 2.265, p = .023, n = 22;* exact significance, two-tailed). The null hypothesis stating that the distribution of Cytosine to Thymine deamination at position 1 is the same between the groups of 2015 and the group of Powder, was rejected.

This finding suggests that the long-term storage has not negatively impacted the quality of the samples in the Powder group, instead the findings indicate an improvement in the preservation of the samples from the Powder group. These results indicate a significant improvement in the deamination rate between the 2015 group (*Mdn = 14.00%, IQR = 3.54%*) and the Powder group (*Mdn = 11.28%, IQR = 2.31%*) exhibiting statistically significantly less deamination.

In summary, these results show a significant difference in Cytosine to Thymine deamination at position 1 for the independent sample comparison was observed only in the comparison between the samples processed from Powder and the samples processed in 2015, for results see Table 16.

Table 16: Hypothesis test summary of independent Sample Mann-Whitney test at pos1 across Sample Type by preparation stage for apparent Cytosine to Thymine deamination: comparing position 1 between the groups 2015 and the other groups.

Groups	Null Hypothesis	Test	Sig.[a,b]	Decision
2015 vs. Extract	The distribution of pos1 is the same across categories of Sample Type by Preparation Stage.	Independent-Samples Mann-Whitney U Test	.223[c]	Retain the null hypothesis.
2015 vs. Indexed Library	The distribution of pos1 is the same across categories of Sample Type by Preparation Stage.	Independent-Samples Mann-Whitney U Test	.101[c]	Retain the null hypothesis.
2015 vs. Powder	The distribution of pos1 is the same across categories of Sample Type by Preparation Stage.	Independent-Samples Mann-Whitney U Test	.023[c]	Reject the null hypothesis.
a. The significance level is .050	b. Asymptotic significance is displayed.	c. Exact significance is displayed for this test.		

3.4.3 Related sample comparison through Wilcoxon test

Normal distribution was not tested, since it is not applicable in this context. A related-samples Wilcoxon test was conducted to determine whether the distribution of Cytosine to Thymine deamination at

position 1 (pos1) differs between the baseline group of 2015 across each of the three sample preparation stages, being Extract, Indexed Library and Powder. The results showed significant results for all groupwise comparisons.

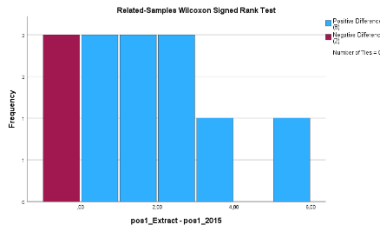


Figure 4: Column chart showing the frequency of positive and negative differences in Cytosine to Thymine deamination at position 1 (pos1) between the Extract group and the 2015 group (n = 10).

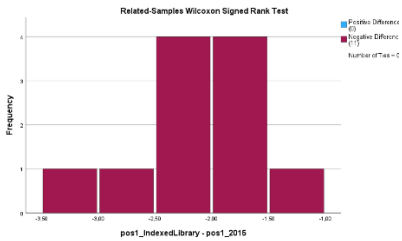


Figure 5: Column chart showing the frequency of positive and negative differences in Cytosine to Thymine deamination at position 1 (pos1) between the Indexed Library group and the 2015 group (n = 11).

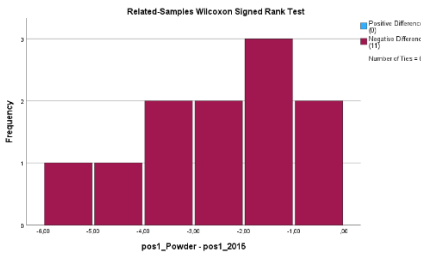


Figure 6: Column chart showing the frequency of positive and negative differences in Cytosine to Thymine deamination at position 1 (pos1) between the Powder group and the 2015 group (n = 11).

The median difference between the 2015 and Extract groups showed a statistically significant result in the distribution at position 1 (*Z = 2.369, p = .017, n = 10;* exact significance, two-tailed). The null hypothesis stating that the distribution of Cytosine to Thymine deamination at position 1 is the same between the 2015 and the Extract group, was rejected.

The median difference between the 2015 and Indexed Library group showed a statistically significant result in the distribution at position 1 (*Z = -2.934, p = .003, n = 11;* exact significance, two-tailed). The null hypothesis stating that the distribution of Cytosine to Thymine deamination at position 1 is the same between the 2015 and the Indexed Library group, was rejected.

The median difference between 2015 and Powder group showed a statistically significant result in the distribution at position 1 (*Z = -2.934, p = .003, n = 11;* exact significance, two-tailed). The null hypothesis stating that the distribution of Cytosine to Thymine deamination at position 1 is the same between the 2015 and the Powder group, was rejected.

These findings show that samples from all preparation methods (Extract, Indexed Library, Powder) show statistically significant differences in Cytosine to Thymine deamination at position 1 compared to the baseline group from 2015. However, these results show that only the samples from the Extract group show a significantly higher degree of deamination, hence indicating significantly higher degradation over the course of the 8 years long-term storage.

The samples from the groups Powder and Indexed Library show statistically significantly less deamination, hence indicating significantly less degradation and better preservation over the course of the 8 years long-term storage, for results see Table 17, as well as Figure 4, Figure 5, Figure 6.

Table 17: Hypothesis test summary of related Sample Wilcoxon test of pos1 across Sample Type by preparation stage for apparent Cytosine to Thymine deamination: comparing the median of differences of the Cytosine to Thymine rate at position 1 between the groups 2015 and the other groups.

Groups	Null Hypothesis	Test	Sig.[a,b]	Decision
2015 vs. Extract	The median of differences between pos1_2015 and pos1_Extract equals 0.	Related-Samples Wilcoxon Signed Rank Test	.017	Reject the null hypothesis.
2015 vs. Indexed Library	The median of differences between pos1_2015 and pos1_IndexedLibrary equals 0.	Related-Samples Wilcoxon Signed Rank Test	.003	Reject the null hypothesis.

2015 vs. Powder	The median of differences between pos1_2015 and pos1_Powder equals 0.	Related-Samples Wilcoxon Signed Rank Test	.003	Reject the null hypothesis.

a. The significance level is .050

b. Asymptotic significance is displayed.

3.5 Assessing DNA degradation through the Deamination of Guanine to Adenine

Guanine to Adenine deamination, on the complimentary strand to Cytosine to Thymine deamination, is an important property to assess the decay of aDNA. Similarly, the values were obtained through a software called "mapDamage" (Ginolhac et al., 2011), the software entails information about Guanine to Adenine deamination between position 1 and 25. For the statistical analysis the percentage of Adenine to Guanine deamination was compared only specifically at positions 1, across 4 groups. The groups were defined as 2015, Powder, Extract and Indexed Library. The group names encompass the information about the sample preparation (Powder, Extract and Indexed Library) as well as the baseline comparison group (2015).

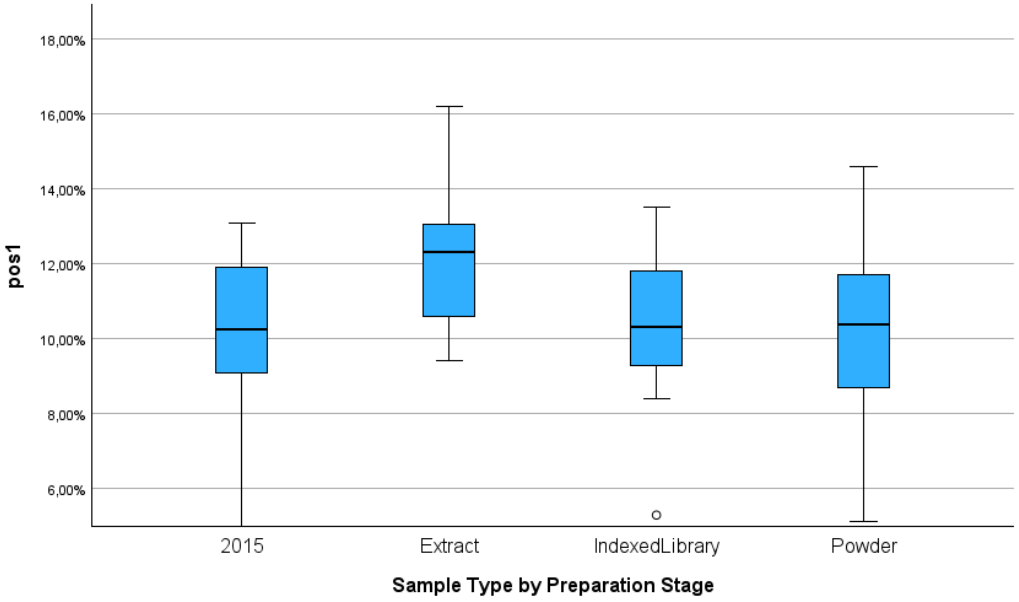
3.5.1 Descriptive Statistics

See table below (Table 18) for the descriptive statistics and Figure 7 for the boxplot comparison of the Guanine to Adenine deamination (pos1) between the four groups.

Table 18: Median and interquartile range (IQR) for Guanine to Adenine deamination (pos1) in the four Groups (2015, Extract, Indexed Library, Powder)

Sample Type by Preparation Stage			Descriptive Statistic - GtoA deamination pos1			
			Q1	Median	Q3	IQR
Weighted Average	pos1	2015	8.69%	10.25%	12.73%	4.04%
		Extract	10.49%	12.32%	13.16%	2.67%
		IndexedL	8.68%	10.33%	12.77%	4.09%
		Powder	8.41%	10.39%	11.97%	3.56%

Figure 7: Boxplot comparison for the variable Guanine to Adenine deamination (pos1) in the four Groups (2015, Extract, Indexed Library, Powder)



3.5.2 Independent sample comparison of the samples through Mann-Whitney test

The data is not normally distributed. An independent-samples Mann-Whitney U test was conducted to determine whether the distribution of Guanine to Adenine deamination at position 1 (pos1) differs between the baseline group of 2015 across each of the sample preparation stages, being Extract, Indexed Library and Powder.

The Mann-Whitney-U test did not result in a statistically significant difference in the distribution of Guanine to Adenine deamination at position 1 (pos1) between the 2015 group (*Mdn = 10.25%, IQR = 4.04%*) and the Extract group (*Mdn = 12.32%, IQR = 2.67%*) (Mann-Whitney's U-Test: *U = 78.00, Z = 1.62, p = .114, n = 21;* exact significance, two-tailed). The null hypothesis stating that the distribution of Guanine to Adenine deamination at position 1 is the same between the groups of 2015 and the group of Extract, was retained.

The Mann-Whitney-U test did not result in a statistically significant difference in the distribution of Guanine to Adenine deamination at position 1 (pos1) between the 2015 group (*Mdn = 10.25%, IQR = 4.04%*) and the Indexed Library (*Mdn = 10.33%, IQR = 4.09%*) (Mann-Whitney's U-Test: *U = 61.00, Z = .033, p = 1.000, n = 22;* exact significance, two-tailed). The null hypothesis stating that the distribution of Guanine to Adenine deamination at position 1 is the same between the groups of 2015 and the group of Indexed Library, was retained.

The Mann-Whitney-U test did not result in a statistically significant difference in the distribution of Guanine to Adenine deamination at position 1 (pos1) between the 2015 group (*Mdn = 10.25%, IQR = 4.04%*) and the Powder (*Mdn = 10.39%, IQR = 3.56%*) (Mann-Whitney's U-Test: *U = 62.00, Z = .0098, p = .949, n = 22;* exact significance, two-tailed). The null hypothesis stating that the distribution of Guanine to Adenine deamination at position 1 is the same between the groups of 2015 and the group of Powder was retained.

The results did not show for the independent Mann-Whitney comparison any significant difference in Guanine to Adenine deamination at position 1 for any of the three groups Extract, Powder and Indexed Library compared to the samples from 2015. These results indicate that for this comparison no effect of long-term storage degradation has been observed (for results see Table 19).

Table 19: Results of Hypothesis test summary of independent Sample Mann-Whitney test of pos1 across Sample Type by preparation stage for apparent Guanine to Adenine deamination: comparing position 1 between the groups 2015 and the other groups.

Groups	Null Hypothesis	Test	Sig.[a,b]	Decision
2015 vs. Extract	The distribution of pos1 is the same across categories of Sample Type by Preparation Stage.	Independent-Samples Mann-Whitney U Test	.114[c]	Retain the null hypothesis.
2015 vs. Indexed Library	The distribution of pos1 is the same across categories of Sample Type by Preparation Stage.	Independent-Samples Mann-Whitney U Test	1.000[c]	Retain the null hypothesis.
2015 vs. Powder	The distribution of pos1 is the same across categories of Sample Type by Preparation Stage.	Independent-Samples Mann-Whitney U Test	.949[c]	Retain the null hypothesis.
a. The significance level is .050	b. Asymptotic significance is displayed.	c. Exact significance is displayed for this test.		

3.5.3 Related sample comparison through Wilcoxon test

Normal distribution was not tested, since it is not applicable in this context. A related-samples Wilcoxon test was conducted to determine whether the distribution of Guanine to Adenine deamination at

position 1 (pos1) differed between the baseline group from 2015 across each of the three sample preparation stages, being Extract, Indexed Library and Powder. The results indicate significant results for one groupwise comparisons, being Extract.

Figure 8: Column chart showing the frequency of positive and negative differences in Guanine to Adenine deamination at position 1 (pos1) between the Extract group and the 2015 group (n = 10).

Figure 9 :Column chart showing the frequency of positive and negative differences in Guanine to Adenine deamination at position 1 (pos1) between the Indexed Library group and the 2015 group (n = 11).

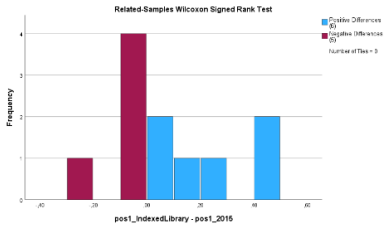
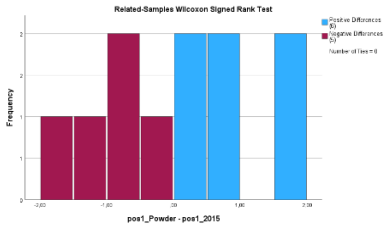


Figure 10: Column chart showing the frequency of positive and negative differences in Guanine to Adenine deamination at position 1 (pos1) between the Powder group and the 2015 group (n = 11).



The median difference between the 2015 and Extract groups showed a statistically significant result in the distribution at position 1 (*Z = 2.701, p = .007, n = 10;* exact significance, two-tailed). The null hypothesis stating that the distribution of Guanine to Adenine deamination at position 1 is the same between the 2015 and the Extract group, was rejected.

However, the remaining two groupwise comparisons did not show statistically significant differences. The median difference between the 2015 and Indexed Library group did not show a statistically significant result in the distribution at position 1 *(Z = 1.334, p = .182, n = 11; exact significance, two-tailed).* The null hypothesis stating that Guanine to Adenine deamination at position 1 is the same between the 2015 and the Indexed Library group, was retained. Similarly, the median difference between 2015 and Powder group did not show a statistically significant result in the distribution at position 1 (*Z = .089, p = .929, n = 11;* exact significance, two-tailed). The null hypothesis stating that the distribution of Guanine to Adenine deamination at position 1 is the same between the 2015 and the Powder group, as well as the null hypothesis stating that the distribution of Guanine to Adenine deamination at position 1 is the same between the 2015 and the Indexed Library, were both retained.

These findings suggest that only the group Extract showed a significant increase in Guanine to Adenine deamination at position 1 compared to the baseline group from 2015, indicating increased degradation over the 8 years of long-term storage.

However, no significant differences were observed for the Powder and Indexed Library groups, suggesting that these groups did not experience any statistically relevant difference for the Guanine to Adenine deamination at position 1 over the same period. This also suggests that these samples have not experienced a significant increase or decrease in degradation over the 8-year storage period, for results see Table 20 as well as Figure 8, Figure 9, Figure 10.

Table 20: Hypothesis test summary of related Sample Wilcoxon test pos1 across Sample Type by preparation stage for apparent Guanine to Adenine deamination: comparing position 1 in the samples from the group 2015 against the other groups

Groups	Null Hypothesis	Test	Sig.[a,b]	Decision
2015 vs. Extract	The median of differences between pos1_2015 and pos1_Extract equals 0.	Related-Samples Wilcoxon Signed Rank Test	.007	Reject the null hypothesis.
2015 vs. Indexed Library	The median of differences between pos1_2015 and pos1_IndexedLibrary equals 0.	Related-Samples Wilcoxon Signed Rank Test	.182	Retain the null hypothesis.
2015 vs. Powder	The median of differences between pos1_2015 and pos1_Powder equals 0.	Related-Samples Wilcoxon Signed Rank Test	.929	Retain the null hypothesis.

a. The significance level is .050

b. Asymptotic significance is displayed.

3.6 Assessing DNA degradation through the software "AMBER"

To asses DNA degradation the program AMBER (https://github.com/tvandervalk/AMBER) (Dolenz et al., 2024) was used. To perform multiple samples at the same time customized scripts were employed (see Appendix C)

3.7 Screening for patterns of DNA fragmentation

Patterns of DNA fragmentation were screened, to assess if there is an effect of long-term storage on fragmentation. Through the software AMBER the mean fragment length and fragment length distribution was assessed. The samples analysed in 2015 contained bam-files with fragments smaller than 30bp, entailing reads ranging from 17 bp to 29 bp. Since the samples produced in 2023 comprise a limit of 30bp to accurately compare the dataset and not affect the results the dataset was corrected accordingly and reads below 30bp were removed from the 2015 dataset.

3.7.1 Comparing mean fragment length

The mean fragment length was assessed for all samples and compared between the groups Extract, Indexed Library, Powder to the group of 2015 for all 11 Individuals (*n = 11*). In the table below a list of the mean fragmentation values of all samples can be found, see Table 21.

Figure 11: Mean fragment length: on the x-axis is the mean fragment length and on the y-axis the individuals (shown as Lab ID). The plot visualizes the difference in mean fragment length for each individual, depending on preparation stage.

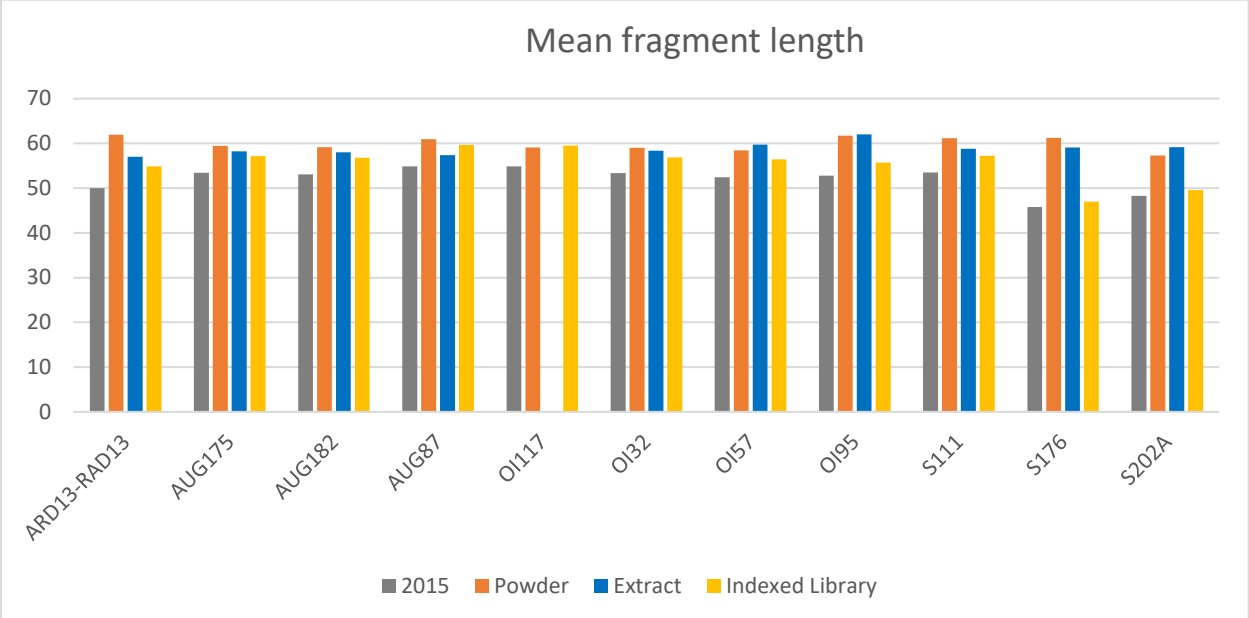


Table 21: Mean fragment length for all samples. Each sample is identifiable by LabID and preparation stage.

	Mean Fragment Length			
LabID	2015	Powder	Extract	Indexed Library
ARD13-RAD13	49.9969	61.932	57.0133	54.8506
AUG175	53.4684	59.429	58.2256	57.1749
AUG182	53.0997	59.1224	58.0101	56.8164
AUG87	54.8806	60.9665	57.38	59.6863
OI117	54.8781	59.0957	NA	59.539
OI32	53.3454	59.0441	58.3564	56.8409
OI57	52.4482	58.4279	59.7525	56.4242
OI95	52.8133	61.7363	61.9888	55.7585
S111	53.5362	61.1258	58.8013	57.2272
S176	45.7665	61.2157	59.0982	47.0046
S202A	48.2858	57.275	59.1827	49.6025

3.7.2 Differences in mean fragmentation length through Wilcoxon related test analysis

Normal distribution was not tested, since it is not applicable in this context. A related-samples Wilcoxon test was conducted to determine whether there is a difference in the mean fragment length between the baseline group from 2015 and the three sample preparation stages: Extract, Indexed Library and Powder. The results showed significant results for all groupwise comparisons. For all samples produced in 2023, for all categories being produced from Powder, Extract and Indexed Library the related groupwise comparison with the samples produces in 2015 yielded significant results.

The median difference between the 2015 and Extract groups showed a statistically significant result in the mean fragment length (*Z = 2.803, p = .005, n = 10;* exact significance, two-tailed). The median difference between the 2015 and Indexed Library groups showed a statistically significant result in the distribution of the mean fragment length (*Z = 2.934, p = .003, n = 11;* exact significance, two-tailed). The median difference between the 2015 and Powder groups showed a statistically significant result in the mean fragment length (*Z = 2.934, p = .007, n = 11;* exact significance, two-tailed).

The null hypothesis stating that the distribution of the mean fragment length is the same between the 2015 and the three respective sample preparation methods, were all rejected. This result indicates that samples from all categories (Powder, Extract, Indexed Library) exhibit less fragmentation compared to the mean fragment length of 2015. These results indicate that the long-term storage time of 8 years has not led to an increase in fragmentation. The results showed that the mean fragment length has increased between the samples produced in 2023 and the samples produced in 2015, for results see Table 22.

Table 22: Hypothesis test summary of related Sample Wilcoxon test of the median of difference between the mean fragment length of the samples from the group 2015 against the other groups.

Variable: Mean Fragment Length				
Groups	Null Hypothesis	Test	Sig.[a,b]	Decision
2015 vs. Extract	The median of differences between 2015 and Extract equals 0.	Related-Samples Wilcoxon Signed Rank Test	.005	Reject the null hypothesis.
2015 vs. Indexed Library	The median of differences between 2015 and Indexed Library equals 0.	Related-Samples Wilcoxon Signed Rank Test	.003	Reject the null hypothesis.
2015 vs. Powder	The median of differences between 2015 and Powder equals 0.	Related-Samples Wilcoxon Signed Rank Test	.003	Reject the null hypothesis.

a. The significance level is .050

b. Asymptotic significance is displayed.

3.7.3 Fragment Length Distribution: Descriptive Statistic

The table below (Table 23) entails the descriptive statistics for the fragment length distribution of the samples.

Table 23: Median and interquartile range (IQR) for aggregated Fragment Length distribution in the four Groups (2015, Extract, Indexed Library, Powder) and the 5 categories (L30_39_sum, L40_49_sum, L50_59_sum, L60_69_sum, L70_79_sum).

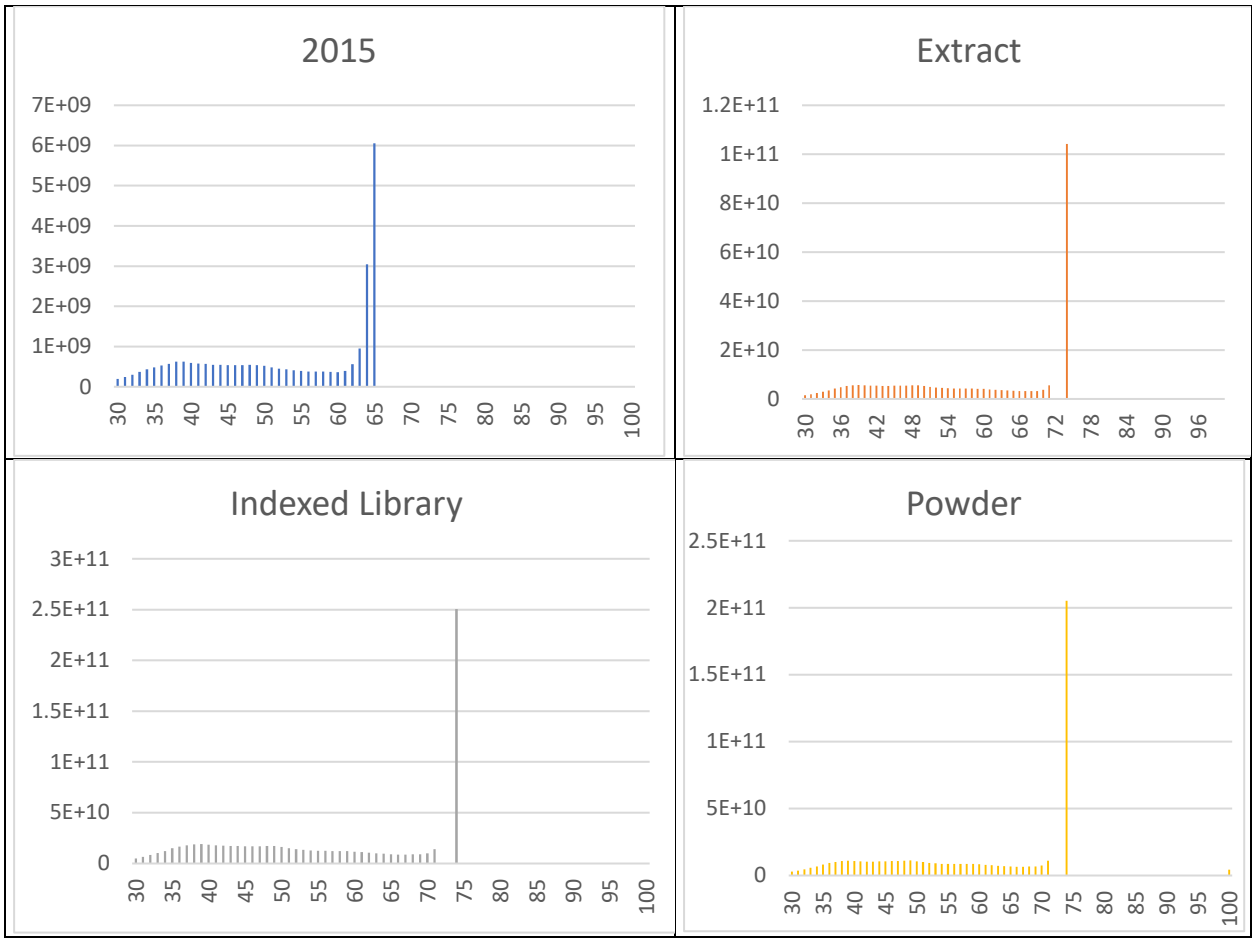
Sample Type by Preparation Stage			Descriptive Statistic - Fragment Length distribution			
			Q1	Median	Q3	IQR
Weighted Average	L30_39_sum	2015	1.29E+08	2.72E+08	5.65E+08	4.36E+08
		Extract	2.27E+09	3.61E+09	5.60E+09	3.34E+09
		Indexed Library	2.28E+09	7.84E+09	1.82E+10	1.59E+10
		Powder	3.01E+09	6.65E+09	8.83E+09	5.82E+09
Weighted Average	L40_49_sum	2015	1.15E+08	4.01E+08	7.41E+08	6.26E+08
		Extract	3.97E+09	5.47E+09	7.03E+09	3.05E+09
		Indexed Library	2.13E+09	1.24E+10	2.40E+10	2.18E+10
		Powder	4.06E+09	9.72E+09	1.40E+10	9.91E+09
Weighted Average	L50_59_sum	2015	6.19E+07	3.00E+08	5.61E+08	4.99E+08
		Extract	3.17E+09	4.79E+09	5.48E+09	2.31E+09
		Indexed Library	1.19E+09	9.52E+09	1.78E+10	1.66E+10
		Powder	3.22E+09	8.50E+09	1.13E+10	8.12E+09
Weighted Average	L60_69_sum	2015	1.10E+08	7.84E+08	1.58E+09	1.47E+09
		Extract	2.45E+09	3.76E+09	4.20E+09	1.75E+09
		Indexed Library	6.24E+08	6.89E+09	1.28E+10	1.22E+10
		Powder	2.38E+09	6.98E+09	8.67E+09	6.28E+09
Weighted Average	L70_79_sum	2015	0.00	0.00	0.00	0.00
		Extract	7.76E+09	1.20E+10	1.39E+10	6.16E+09
		Indexed Library	6.25E+08	1.61E+10	3.54E+10	3.48E+10
		Powder	1.71E+09	2.28E+10	2.46E+10	2.29E+10

3.7.4 Increased Fragment Length Distribution Across Sample Preparation Methods Following Long-Term Storage analysed through Mann-Whitney U analysis

Another test was conducted to assess another point of view of fragmentation. An independent-samples Mann-Whitney U test was conducted to determine whether there is a difference in the fragmentation distribution between the categories 2015 and the samples of 2023, with were categorized based on their respective preparation stages: Extract, Indexed Library, Powder and 2015. For all samples produced in 2023, for all categories being produced from Powder, Extract and Indexed Library the fragment length distribution compared to the samples produces in 2015 yielded significant results.

For this purpose, five variables were created assigning the fragment lengths into 5 categories: L30_39_sum for reads ranging from 30-39bp, L40_49_sum for reads ranging from 40-49bp, L50_59_sum for reads ranging from 50-59bp, L60_69_sum for reads ranging from 60-69bp and L70_79_sum for reads ranging from 70-79bp. This means 15 tests were conducted.

Figure 12: Fragment length distribution in the four groups 2015, Extract, Indexed Library and Powder.



For the comparison between 2015 and Extract, the test did result in a statistically significant difference for the fragment lengths across all categories. Specifically, the comparison between the groups 2015 and Extract showed statistically significant differences in the distribution of 30 – 39 bp (L30_39_sum; Mann-Whitney's U-Test: *U = 103.00, Z = 3.380, p < .001, n = 21;* exact significance, two-tailed), 40 – 49 bp (L40_49_sum; Mann-Whitney's U-Test: *U = 102.00, Z = 3.310, p < .001, n = 21;* exact significance, two-tailed), 50 – 59 bp (L50_59_sum; Mann-Whitney's U-Test: *U = 103.00, Z = 3.380, p < .001, n = 21;* exact significance, two-tailed), 60-69 bp (L60_69_sum; Mann-Whitney's U-Test: *U = 99.00, Z = 3.098, p < .001, n = 21;* exact significance, two-tailed), and 70-79 bp (L70_79_sum; Mann-Whitney's U-Test: *U = 110.00, Z = 4.183, p < .001, n = 21;* exact significance, two-tailed).

The null hypothesis stating that the distribution of fragment lengths across all categories (L30_39_sum, L40_49_sum, L50_59_sum, L60_69_sum, L70_79_sum) is the same between the 2015 and Extract group, were rejected.

Figure 13: Boxplot showing the difference between the fragment distribution, in the range of 30-39 reads, of the samples from 2015 and samples produced from Powder.

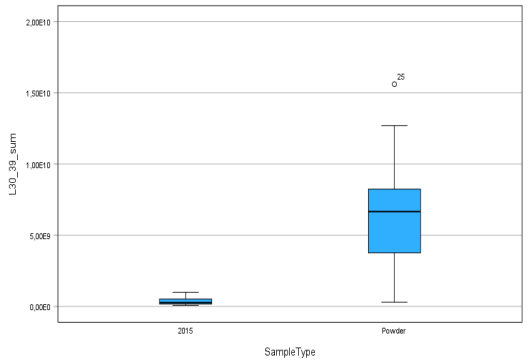
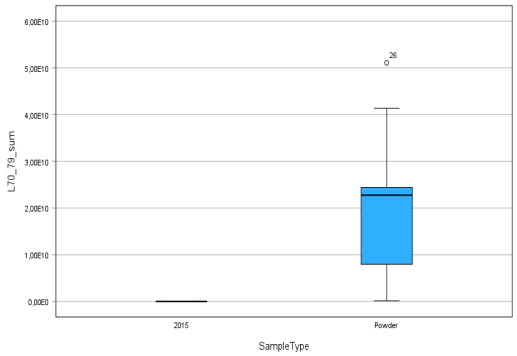


Figure 14: Boxplot showing the difference between the fragment distribution, in the range of 70-79 reads, of the samples from 2015 and samples produced from Powder.



For the comparison between 2015 and Indexed Library, the test did result in a statistically significant difference for the fragment lengths across all categories. Specifically, the comparison between the groups 2015 and Indexed Library showed statistically significant differences in the distribution of 30 – 39 bp (L30_39_sum; Mann-Whitney's U-Test: *U = 121.00, Z = 3.973, p < .001, n = 22;* exact significance, two-tailed), 40 – 49 bp (L40_49_sum; Mann-Whitney's U-Test: *U = 120.00, Z = 3.907, p < .001, n = 22;* exact significance, two-tailed), 50 – 59 bp (L50_59_sum; Mann-Whitney's U-Test: *U = 119.00, Z = 3.841, p < .001, n = 22;* exact significance, two-tailed), 60-69 bp (L60_69_sum; Mann-Whitney's U-Test: *U = 97.00, Z = 3.098, p = .016, n = 22;* exact significance, two-tailed), and 70-79 bp (L70_79_sum; Mann-Whitney's U-Test: *U = 121.00, Z = 4.245, p < .001, n = 22;* exact significance, two-tailed).

The null hypothesis stating that the distribution of fragment lengths across all categories (L30_39_sum, L40_49_sum, L50_59_sum, L60_69_sum, L70_79_sum) is the same between the 2015 and Indexed Library group, were rejected.

For the comparison between 2015 and Powder, the test did result in a statistically significant difference for the fragment lengths across all categories. Specifically, the comparison between the groups 2015 and Powder showed statistically significant differences in the distribution of 30 – 39 bp (L30_39_sum; Mann-Whitney's U-Test: *U = 116.00, Z = 3.644, p < .001, n = 22;* exact significance, two-tailed), 40 – 49 bp (L40_49_sum; Mann-Whitney's U-Test: *U = 115.00, Z = 3.579, p < .001, n = 22;* exact significance, two-tailed), 50 – 59 bp (L50_59_sum; Mann-Whitney's U-Test: *U = 115.00, Z = 3.579, p < .001, n = 22;* exact significance, two-tailed), 60-69 bp (L60_69_sum; Mann-Whitney's U-Test: *U = 110.00, Z = 3.250, p < .001, n = 22;* exact significance, two-tailed), and 70-79 bp (L70_79_sum; Mann-Whitney's U-Test: *U = 121.00, Z = 4.245, p < .001, n = 22;* exact significance, two-tailed).

The null hypothesis stating that the distribution of fragment lengths across all categories (L30_39_sum, L40_49_sum, L50_59_sum, L60_69_sum, L70_79_sum) is the same between the 2015 and Powder group, were rejected. In total, all 15 Mann-Whitney tests examining fragment length distribution across the respective categories yielded statistically significant results. These findings suggest that all samples prepared in 2023 have experienced an increase in fragment length distribution compared to the samples from 2015. These results indicate that for this comparison no effect of the 8-year term storage degradation has been observed, for results see Table 24, Figure 13 and Figure 14.

The results specifically indicate that for all categories being L30_39_sum, L40_49_sum, L50_59_sum, L60_69_sum and L70_79_sum there is an at least 10-fold increase in the fragment length between the samples from 2015 to the samples from 2023. Notably, for reads ranging from 70-79 bp the values for 2015 are even 0.

Table 24: Hypothesis test summary of independent Sample Mann-Whitney test of the distribution different categories across samples from the group 2015 against the other groups.

Category	Groups	Null Hypothesis	Test	Sig.[a,b]	Decision
L30_39	2015 vs. Extract	The distribution of L30_39_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L30_39	2015 vs. Indexed Library	The distribution of L30_39_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L30_39	2015 vs. Powder	The distribution of L30_39_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L40_49	2015 vs. Extract	The distribution of L40_49_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L40_49	2015 vs. Indexed Library	The distribution of L40_49_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L40_49	2015 vs. Powder	The distribution of L40_49_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L50_59	2015 vs. Extract	The distribution of L50_59_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L50_59	2015 vs. Indexed Library	The distribution of L50_59_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L50_59	2015 vs. Powder	The distribution of L50_59_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L60_69	2015 vs. Extract	The distribution of L60_69_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L60_69	2015 vs. Indexed Library	The distribution of L60_69_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	.016[c]	Reject the null hypothesis.
L60_69	2015 vs. Powder	The distribution of L60_69_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L70_79	2015 vs. Extract	The distribution of L70_79_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L70_79	2015 vs. Indexed Library	The distribution of L70_79_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.
L70_79	2015 vs. Powder	The distribution of L70_79_sum is the same across categories of SampleType.	Independent-Samples Mann-Whitney U Test	<.001[c]	Reject the null hypothesis.

a. The significance level is .050

b. Asymptotic significance is displayed.

c. Exact significance is displayed for this test.

3.8 Increase in rate of unique reads

The rate of all aligned/unique reads was assessed for all samples, for the comparison between the groups Extract, Indexed Library, Powder to the group of 2015 for all 11 Individuals (*n = 11*). In the table below the rate of unique reads in percent for each sample can be found, see Table 25.

Table 25: Rate of unique reads in % of the total reads per sample, sorted by LabID and the four groups 2015, Powder, Extract Indexed Library.

	Unique Reads in Percent of total Reads			
LabID	2015	Powder	Extract	Indexed Library
ARD13-RAD13	25.09%	25.73%	34.46%	13.98%
AUG175	51.41%	60.51%	60.28%	63.89%
AUG182	61.42%	76.71%	73.41%	73.22%
AUG87	38.27%	48.95%	48.06%	46.38%
OI117	36.69%	48.42%	NA	44.85%
OI32	62.77%	79.78%	76.09%	75.67%
OI57	33.18%	46.16%	42.60%	41.68%
OI95	67.15%	84.92%	81.34%	80.14%
S111	40.90%	45.65%	54.71%	51.25%
S176	28.87%	37.32%	48.53%	36.17%
S202A	26.04%	45.67%	35.98%	34.30%

3.8.1 Differences in increase of the rate of unique reads through Wilcoxon related samples test

Normal distribution was not tested, since it is not applicable in this context. A related-samples Wilcoxon test was conducted to determine whether there is a difference in the rate of unique reads between the baseline group from 2015 and the three sample preparation stages: Extract, Indexed Library and Powder. The results show statistically significant results for all groupwise comparisons. For all samples produced in 2023, for all categories being produced from Powder, Extract and Indexed Library the related groupwise comparison with the samples produced in 2015 yielded statistically significant results. For all samples the rate of unique reads increased.

Specifically, the median difference between the 2015 and Extract groups showed a statistically significant increase in the percentage of unique reads (*Z = 2.803, p = .0.0051, n = 10;* exact significance, two-tailed). Likewise, the median difference between the 2015 and Indexed Library groups showed a statistically significant increase in the percentage of unique reads (*Z = 2.312, p = 0.021, n = 11;* exact significance, two-tailed). At last, the median difference between the 2015 and Powder groups showed a statistically significant increase in the percentage of unique reads (*Z = 2.934, p = 0.0033, n = 11;* exact significance, two-tailed).

The null hypothesis, which states that there is no difference in the rate of unique reads between the 2015 and the three respective sample preparation methods, was rejected in all cases. These results indicates that samples from all categories (Powder, Extract, Indexed Library) exhibit a higher rate of unique reads compared to the sample from 2015. These results indicate that the long-term storage time of 8 years has not led to a decrease in unique reads. On the contrary the rate of unique reads has increased between the samples produced in 2023 and those samples produced in 2015, for results see Table 26.

Table 26: Hypothesis test summary of related Sample Wilcoxon test of the median of difference between the Unique reads in % of total reads in the samples from the group 2015 against the other groups.

Variable: Unique reads in % of total reads				
Groups	Null Hypothesis	Test	Sig.[a,b]	Decision
2015 vs. Extract	The median of differences between Unique Reads (%) in 2015 and in Extract equals 0.	Related-Samples Wilcoxon Signed Rank Test	.0051	Reject the null hypothesis.
2015 vs. Indexed Library	The median of differences between Unique Reads (%) in 2015 and in Indexed Library equals 0.	Related-Samples Wilcoxon Signed Rank Test	.021	Reject the null hypothesis.
2015 vs. Powder	The median of differences between Unique Reads (%) in 2015 and in Powder equals 0.	Related-Samples Wilcoxon Signed Rank Test	.0033	Reject the null hypothesis.
a. The significance level is .050	b. Asymptotic significance is displayed.			

4 Discussion

The study here reports a significantly higher degree of Cytosine to Thymine Guanine to Adenine deamination in the long-term stored samples prepared from Extract compared to the samples produced in 2015. The sample's deamination rates varied depending on the preparation stage. The samples produced from powders and indexed libraries showed less Cytosine to Thymine deamination compared to 2015, indicating overall less degradation and better preservation compared to the 2015 samples. In contrast, samples from the extract group showed higher degradation during the 8 years of storage.

The findings regarding the increase of Cytosine to Thymine deamination in the Extract group are consistent with previous studies, which identified a clear relationship between age and increased deamination (Kistler et al., 2017). Cytosine to Thymine deamination is commonly used to authenticate ancient DNA samples (Ginolhac, Rasmussen, et al., 2011; Jónsson et al., 2013), as it is a known example of DNA degradation (Krause, Adrian, et al., 2010) and is expected to increase with passing time. Further, according to the model from Kistler et al. (2017) both sample age and temperature strongly impact cytosine deamination. Although the model did not examine long-term storage, but sample age.

The cause for the observed increased deamination in the Extract group, compared to Indexed Library, could probably be attributed to the differences in the chemical composition of the extracts, compared to Powder and Indexed Library groups. Both the repair of fragment ends during the blunt-end repair and the addition of adapters during library preparation (Meyer & Kircher, 2010), make sequencing libraries more stable than extracts (see chapter 2.7.2 Ancient DNA library preparation). Additionally, the better preservation of the DNA in the bone powder might be attributed to sample not being modified, other than physical grinding, and therefore being more stable. During DNA extraction (Dabney, Knapp, et al., 2013) the DNA in the sample is being separated from the bone matrix in order to obtain an extract of the endogenous DNA that was originally present in the bone. The results indicate that the DNA in the bone was better preserved through the 8-year storage time compared to the DNA stored in the extract, this better preservation could be caused by the DNA in the bone powder still being bound to the bone matrix during the storage period. Bone is mainly composed of three components: Osteocytes, extracellular matrix and interstitial liquid. The extracellular matrix, or short bone matrix, is mainly composed of organic collagen fibers and anorganic salts, composed mainly of hydroxyapatite (Campos et al., 2012; Faller & Schünke, 2016). Previous research (Brundin et al., 2013) demonstrated that when DNA is

incubated with hydroxyapatite, the retrieved DNA remains amplifiable up to 3 months, while DNA incubated in the same media without hydroxyapatite becomes non-amplifiable DNA after only 3 weeks. During DNA extraction (Dabney, Knapp, et al., 2013), the extraction buffer dissolves the DNA from the bone matrix, composed of hydroxyapatite, potentially affecting DNA preservation. This could be a possible cause for the improved preservation in bone powders compared to the increased degradation in extract samples.

Previous research developed a model (Kistler et al., 2017) analysing ancient DNA degradation based on sample age and environmental variables, establishing a classic thermal age model for the Cytosine to Thymine deamination. On the other hand, it was not possible to create a model for aDNA fragmentation. No correlation was found between DNA fragmentation and sample age across the analysed timespans, even by increasing the minimal age difference up to 1000 years, no statistical relationship can be established. This model is probably highly accurate, when considering the large sample size of 185 paleogenomic datasets.

In contrast, the model by Kistler et al. (2017) proposed that thermal fluctuations and precipitations were significant predictors of fragmentation. Given that all the samples in this experiment were stored at constant +4°to +6°C, is likely that thermal fluctuations were minimal and might have contributed to the samples in 2023 not further fragmenting. Therefore, implying that proper storage conditions, being constant temperature in this case, might have contributed to the good preservation of the Powder and Indexed Library samples. Although the same constant thermal storage conditions were applied to Extracts and for these samples the deamination increased.

The precipitation has not been assessed. The humidity levels reflect the storage of fridge and freezers in the laboratory facility. Possibly, the constant and cold temperature for the samples from the Indexed Library and Powder group might have contributed to the DNA being better preserved than 8 years ago. The importance of cold temperature on the stability of DNA is was already proposed in the nineties by (Lindahl, 1993) and the importance of cold and constant temperature is further supported since the oldest recovered aDNA stems from a permafrost sample (van der Valk et al., 2021).

The findings on fragmentation are supported by previous research (Sawyer et al., 2012), proposing that DNA fragmentation degrades rapidly soon after the onset of death and then subsequently slows down on the fragmentation process. The researchers Kistler et al. (2017) further propose by comparing the degradation processes of both Cytosine deamination and fragmentation, a time-dependent fragmentation process may be incompatible.

Despite the results showing a significantly higher degree of degradation for the Extract group, the mean fragment length and the fragment length distribution have increased for all samples from 2023 compared to 2015. As previous literature supports that fragmentation does not decrease consistently over time, this could explain why the fragmentation has not decreased consistently (Sawyer et al., 2012). However, this does not explain why the fragment length has increased, therefore there is probably an additional explanation for the improvement in read recovery. A possible cause for the increase in fragment length could also be traced back to short fragments degrading faster, contributing to an increase in fragment length for the 2023 samples. This was not tested in this experiment.

The rate of unique reads has also increased significantly, despite the higher degradation. It is proposed here that the increase in fragment length and the improvement in read recovery could be traced back due to the differences in Illumina sequencing technologies („Illumina sequencing history: a decade in

sequencing“, 2025). In 2015 the Illumina sequencing technology used was the Illumina MiSeq platform, and in 2023 NovaSeq and NextSeq Illumina platforms were used.

The observation, that the read count is significantly higher between MiSeq and NovaSeq has also been reported by previous research (Han et al., 2024) examining the difference in yielded reads between MiSeq and NovaSeq data from DNA stemming from DNA from the oral microbiome. This is the first time similar findings have been reported for ancient DNA data. The same findings have not been reported for NextSeq data yet, but for the samples of this experiment the same findings can be reported for the samples produced from NextSeq data. Moreover, most samples selected for the final analysis were sequenced with NextSeq. For the samples produced from NovaSeq the same significantly higher read count as in (Han et al., 2024) has been observed for all selected samples for final analysis. It has not been tested for the samples that have not been included in the final analysis. This is the first time similar findings have been reported for NextSeq data.

This observed difference in obtained reads from MiSeq and NovaSeq Illumina platforms can likely be attributed to the improvement in sequencing capacity (Han et al., 2024). The MiSeq platform generates up to 300 bp paired reads and NovaSeq Illumina platforms generates up to 150 bp paired reads. MiSeq being able to process up to 15 Gb in 56 hours and NovaSeq, in the paper referred to as Novate, being able to process up to 3000 Gb in 44 hours. The output from NovaSeq can reach up until 20 Billion reads (Zhong et al., 2021). In comparison, MiSeq has an output only reaching as far as 25 Million single reads or 50 Million paired-end reads (Ravi et al., 2018). These differences in Illumina platforms are probably responsible for the observed differences in recovered reads between the samples from 2023 to 2015.

To ensure reproducibility, the 2015 lab pipeline has been replicated for the 2023 samples, with minimal exceptions. Since the findings showed an improvement in deamination for the groups Indexed Library and Powder the laboratory pipeline was re-evaluated to assess if the pipeline was accurately replicated, and to confirm that the pipeline has not been inadvertently improved. No binding apparatus was used in 2023, for safety reasons to ensure the samples do not exit the machine when being rotated in the centrifuges. Moreover, the samples in 2023 were during the extraction process, after the incubation phase, centrifuged for 2 mins at 13,000rpm and the samples in 2015 for 2 minutes at 17.000 g. This difference in force used on the sample after centrifugation, is estimated not to influence the results.

Further confounding factors have been considered: the bioinformatics pipeline has been reevaluated to assess if reproducibility was maintained and assessed if any of the steps could have possibly influenced the samples increase in rate of unique reads or the fragment length. The only difference is that in 2023 the fragment length was standardized to a minimum of 30bp (-q30), but the reads in the 2015 samples that had a length less <=29 bp were excluded from the analysis. Therefore, this should not influence the results. Also, an increase in contamination between 2015 and 2023 samples produced from Extract, has been reported. There is a difference in sample size between the groups Extract (*n = 9*)., Indexed Library (*n = 11*)., Powder (*n = 11*) and 2015 (*n = 4*). This difference in sample size may have introduced a bias and could possibly explain difference in test results and differences in the significance of tests. This is highlighted by the observation, that the dataset is complete for the groups Powder and Indexed Library, but not for samples produced from Extract or samples from 2015.

Another possibility is that 8 years of storage are possibly not enough for a substantial reduction in fragment length that would be detectable with the improvement of obtained reads obtained from an improvement in sequencing technology. Possibly by repeating the test with an increased number of years in storage the fragmentation would have increased more and possibly impacted the quality of obtained reads further.

Interestingly, enough previous research had concluded that storage time does not have a major influence on a sample's preservation, but geological age being a far more important factor. This was concluded through the development of a decay kinetics model to primarily estimate the half-life of DNA. (Allentoft et al., 2012). More specifically, according to the model although storage time being a statistically significant factor contributing to a sample's preservation, it only accounts for 7.7% of the variation. Comparing this for example to the combining effect of storage time and geological age, which contribute together for 45.2% of the observed variation.

Further, through the DNA half-life model (Allentoft et al., 2012) two sets of samples were compared. Whereas one set was stored for about 7 years, like this experimental set-up, and another set of samples was stored for over 50 years, reporting surprising results: the samples stored for a shorter period exhibited more degradation than those stored for over 50 years. The study concluded that geological age is a much more reliable indicator of DNA preservation, than storage time.

As previously mentioned, a study examining the effect of modern DNA from blood samples, found no significant difference in DNA yield across different storage durations, with the samples being stored at stable -30°C with an average storage duration of 12 years. (Chen et al., 2018). Studies examining the successful retrieval of amplifiable DNA from frozen stored bones (Pruvost et al., 2007) have already been assessed, although the original study did not specify the storage period for the freshly excavated samples.

Further research could possibly examine the effect of long-term storage on frozen bone-powders (-20°C), this could further improve the DNA preservation of this type of specimen. Moreover, further studies could reassess the degradation of long-term storage on extracts by testing the effect of long-term storage on extracts by storing extracts below -20°C, for examples at -30°C and -80°C degrees, as this could possibly mitigate the degradation of extracts. Additionally, future studies could possibly re-examine the relationship between long-term storage and recovered aDNA fragment by incorporating un-indexed libraries into the analysis, as well as increasing the time of the long-term storage.

According to the conclusion and findings of this research project following recommendations for aDNA long-term storage protocols have been concluded:

- Samples should be kept at cold and constant temperatures.
- The best suited type of samples for storage are Powders and Indexed Libraries.
- Extracts are not as well-suited for storage and should therefore not be kept for an extended time period, since they show signs of degradation already after short periods of storage like 8 years.
- Alternatively, one could try to possibly store extracts for an extended period but by lowering the storage temperature below -20°C.
- Keeping bone powders in a fridge at constant 4-6°C for long-term storage is sufficient to preserve the samples for long-term storage.
- Similarly, it is sufficient to store indexed libraries at constant -20°C in a freezer to preserve the samples for long-term storage.

5 Conclusion

In summary, while the Extract samples showed a higher Cytosine to Thymine and a higher Guanine to Adenine deamination, indicating increased degradation, the overall fragmentation and rate of unique reads significantly increased in all sample types processed in 2023. Moreover, the deamination for samples produced from powders and indexed libraries improved over the long-term storage. These

findings indicate that, despite the observed degradation in extracts, advancements in sequencing technologies are so pronounced that the rate of recovered unique reads and increase in fragment length statistically significantly increased. The improved preservation of the Powder and Indexed Library samples could be attributed to the higher stability of powders and indexed libraries compared to extracts. Additionally, the constant and cold storage conditions, +4° to +6°C in the fridge and -20°C in the freezer, probably supported the improved preservation of the respective samples.

These findings suggest that although degradation processes in ancient DNA are inevitable, through the continuous improvement of sequencing technologies and the implementation of optimized storage conditions the negative consequences of aDNA degradation can be mitigated.

Additionally, through the continuous advancements of sequencing technologies and through maintaining constant and cold storage conditions, not only will the negative effect of degradation on aDNA be reduced, but the quality of recovered aDNA samples is expected to improve consistently.

References

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., Campos, P. F., Samaniego, J. A., Gilbert, M. T. P., Willerslev, E., Zhang, G., Scofield, R. P., Holdaway, R. N., & Bunce, M. (2012). The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1748), 4724–4733. https://doi.org/10.1098/rspb.2012.1745

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. F., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., & Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, *104*(37), 14616–14621. https://doi.org/10.1073/pnas.0704665104

Brundin, M., Figdor, D., Sundqvist, G., & Sjögren, U. (2013). DNA binding to hydroxyapatite: A potential mechanism for preservation of microbial DNA. *Journal of endodontics*, *39*(2), 211–216.

Callaway, E. (2023). „Truly gobsmacked“: Ancient-human genome count surpasses 10,000. *Nature*.

Campos, P. F., Craig, O. E., Turner-Walker, G., Peacock, E., Willerslev, E., & Gilbert, M. T. P. (2012). DNA in ancient bone–Where is it located and how should we extract it? *Annals of Anatomy-Anatomischer Anzeiger*, *194*(1), 7–16.

Chen, W. C., Kerr, R., May, A., Ndlovu, B., Sobalisa, A., Duze, S. T., Joseph, L., Mathew, C. G., & Babb de Villiers, C. (2018). The integrity and yield of genomic DNA isolated from whole blood following long-term storage at- 30° C. *Biopreservation and biobanking*, *16*(2), 106–113.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., & 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. https://doi.org/10.1073/pnas.1314445110

Dabney, J., & Meyer, M. (2019). Extraction of Highly Degraded DNA from Ancient Bones and Teeth. In B. Shapiro, A. Barlow, P. D. Heintzman, M. Hofreiter, J. L. A. Paijmans, & A. E. R. Soares (Hrsg.), *Ancient DNA* (Bd. 1963, S. 25–29). Springer New York. https://doi.org/10.1007/978-1-4939-9176-1_4

Dabney, J., Meyer, M., & Pääbo, S. (2013). Ancient DNA damage. *Cold Spring Harbor perspectives in biology*, *5*(7), a012567.

Dolenz, S., van der Valk, T., Jin, C., Oppenheimer, J., Sharif, M. B., Orlando, L., Shapiro, B., Dalén, L., & Heintzman, P. D. (2024). Unravelling reference bias in ancient DNA datasets. *Bioinformatics*, *40*(7), btae436.

Faller, A., & Schünke, M. (2016). *Der Körper des Menschen* (17. Aufl.). Georg Thieme Verlag.

Fernandes, D. M., Cheronet, O., Gelabert, P., & Pinhasi, R. (2021). TKGWV2: An ancient DNA relatedness pipeline for ultra-low coverage whole genome shotgun data. *Scientific Reports*, *11*(1), 21262.

Franklin, R. E., & Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature*, *171*(4356), 740–741.

Ginolhac, A., Rasmussen, M., Gilbert, M. T. P., Willerslev, E., & Orlando, L. (2011). mapDamage: Testing for damage patterns in ancient DNA sequences. *Bioinformatics*, *27*(15), 2153–2155.

Ginolhac, A., Schubert, M., & Jónsson, H. (2011). *mapDamage: Tracking and quantifying damage patterns in ancient DNA sequences*. https://ginolhac.github.io/mapDamage/

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., … Pääbo, S. (2010). A Draft Sequence of the Neandertal Genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Han, H., Lee, H.-J., Kim, K.-S., Chung, J., & Na, H. S. (2024). Comparison of the performance of MiSeq and NovaSeq in oral microbiome study. *Journal of Oral Microbiology*, *16*(1), 2344293. https://doi.org/10.1080/20002297.2024.2344293

Henneberger, K., Barlow, A., & Paijmans, J. L. A. (2019). Double-Stranded Library Preparation for Ancient and Other Degraded Samples. In B. Shapiro, A. Barlow, P. D. Heintzman, M. Hofreiter, J. L. A. Paijmans, & A. E. R. Soares (Hrsg.), *Ancient DNA* (Bd. 1963, S. 65–73). Springer New York. https://doi.org/10.1007/978-1-4939-9176-1_8

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.

Höss, M., Jaruga, P., Zastawny, T. H., Dizdaroglu, M., & Paabo, S. (1996). DNA damage and DNA sequence retrieval from ancient tissues. *Nucleic acids research*, *24*(7), 1304–1307.

Illumina sequencing history: A decade in sequencing. (2025). *Illumina*. https://www.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history/decade-in-sequencing.html

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.

Kistler, L., Ware, R., Smith, O., Collins, M., & Allaby, R. G. (2017). A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic acids research*, *45*(11), 6310–6320.

Krause, J., Adrian, W. B., Kircher, M., Maricic, T., Zwyns, N., Derevianko, A., & Pääbo, S. (2010). A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia. *Current Biology*. https://www.cell.com/current-biology/fulltext/S0960-9822(09)02139-3

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.

Lan, T., & Lindqvist, C. (2018). Technical Advances and Challenges in Genome-Scale Analysis of Ancient DNA. In C. Lindqvist & O. P. Rajora (Hrsg.), *Paleogenomics* (S. 3–29). Springer International Publishing. https://doi.org/10.1007/13836_2018_54

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *nature*, *362*(6422), 709–715.

Llamas, B., Valverde, G., Fehren-Schmitz, L., Weyrich, L. S., Cooper, A., & Haak, W. (2017). From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. *STAR: Science & Technology of Archaeological Research*, *3*(1), 1–14. https://doi.org/10.1080/20548923.2016.1258824

Love, C. C., Thompson, J. A., Lowry, V. K., & Varner, D. D. (2002). Effect of storage time and temperature on stallion sperm DNA and fertility. *Theriogenology*, *57*(3), 1135–1142.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., & Chen, Z. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380.

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb-prot5448.

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., & Stone, A. C. (2021). Ancient DNA analysis. *Nature reviews methods primers*, *1*(1), 14.

Pääbo, S. (1985). Molecular cloning of ancient Egyptian mummy DNA. *nature*, *314*, 644–645.

Pääbo, S. (1989). Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences*, *86*(6), 1939–1943. https://doi.org/10.1073/pnas.86.6.1939

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., & Hofreiter, M. (2004). Genetic Analyses from Ancient DNA. *Annual Review of Genetics*, *38*(1), 645–679. https://doi.org/10.1146/annurev.genet.37.110801.143214

Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., Spens, J., Thomsen, P. F., Bohmann, K., Cappellini, E., Schnell, I. B., Wales, N. A., Carøe, C., Campos, P. F., Schmidt, A. M. Z., Gilbert, M. T. P., Hansen, A. J., Orlando, L., & Willerslev, E. (2015). Ancient and

modern environmental DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1660), 20130383. https://doi.org/10.1098/rstb.2013.0383

Pinhasi, R. (2022). *Paleogenomics—Lecture (wintersemester 2022)*.

Pinhasi, R., Fernandes, D. M., Sirak, K., & Cheronet, O. (2019). Isolating the human cochlea to generate bone powder for ancient DNA analysis. *Nature protocols*, *14*(4), 1194–1205.

Pinhasi, R., Fernandes, D., Sirak, K., Novak, M., Connell, S., Alpaslan-Roodenberg, S., Gerritsen, F., Moiseyev, V., Gromov, A., & Raczky, P. (2015). Optimal ancient DNA yields from the inner ear part of the human petrous bone. *PloS one*, *10*(6), e0129102.

Pruvost, M., Schwarz, R., Correia, V. B., Champlot, S., Braguier, S., Morel, N., Fernandez-Jalvo, Y., Grange, T., & Geigl, E.-M. (2007). Freshly excavated fossil bones are best for amplification of ancient DNA. *Proceedings of the National Academy of Sciences*, *104*(3), 739–744. https://doi.org/10.1073/pnas.0610257104

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., & Gupta, R. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, *463*(7282), 757–762.

Ravi, R. K., Walton, K., & Khosroheidari, M. (2018). MiSeq: A Next Generation Sequencing Platform for Genomic Analysis. In J. K. DiStefano (Hrsg.), *Disease Gene Identification* (Bd. 1706, S. 223–232). Springer New York. https://doi.org/10.1007/978-1-4939-7471-9_12

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., & Johnson, P. L. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, *468*(7327), 1053–1060.

Rubio, L., Martinez, L. J., Martinez, E., & De Las Heras, S. M. (2009). Study of short-and long-term storage of teeth and its influence on DNA. *Journal of forensic sciences*, *54*(6), 1411–1413.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic Amplification of β-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia. *Science*, *230*(4732), 1350–1354. https://doi.org/10.1126/science.2999980

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, *94*(3), 441–448.

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

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., & Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PloS one*, *7*(3), e34131.

Shapiro, B., Barlow, A., Heintzman, P. D., Hofreiter, M., Paijmans, J. L. A., & Soares, A. E. R. (Hrsg.). (2019). *Ancient DNA: Methods and Protocols* (Bd. 1963). Springer New York. https://doi.org/10.1007/978-1-4939-9176-1

Skoglund, P. (2013, Juni 28). *ry_compute: Inferring the biological sex of ancient genomic shotgun samples* [Github.com]. https://github.com/pontussk/ry_compute

Skoglund, P., Storå, J., Götherström, A., & Jakobsson, M. (2013). Accurate sex identification of ancient human remains using DNA shotgun sequencing. *Journal of archaeological Science*, *40*(12), 4477–4482.

Smith, C. I., Chamberlain, A. T., Riley, M. S., Cooper, A., Stringer, C. B., & Collins, M. J. (2001). Not just old but old and cold? *Nature*, *410*(6830), 771–772.

Suzuki, S., Yasuda, T., Shiraishi, Y., Miyano, S., & Nagasaki, M. (2011). ClipCrop: A tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics*, *12*(S14), S7. https://doi.org/10.1186/1471-2105-12-S14-S7

van der Valk, T., Pečnerová, P., Díez-del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., & Sağlıcan, E. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, *591*(7849), 265–269.

Watson, J. D., & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, *171*(4356), Article 4356. https://doi.org/10.1038/171737a0

Zhong, Y., Xu, F., Wu, J., Schubert, J., & Li, M. M. (2021). Application of next generation sequencing in laboratory medicine. *Annals of laboratory medicine*, *41*(1), 25–43.

List of tables

List of figures

Appendix A

A.1 Qubit results

Table 27: Qubit results 2023

Sample ID	sequencing ID	Qubit (ng/ul)	TS passed
222482	S176_op1_oL1_oI1	1.38	yes
222483	S202A_op1_oL1_oI1	1.39	yes
222484	ARD13_op1_nL1_nI1	15.2	yes
222485	AUG175_op1_nL1_nI1	15.5	yes
222486	OI117_op1_nL1_nI1	14.1	yes
222487	OI157_op1_nL1_nI1	18.1	yes
222488	AUG87_op1_nL1_nI1	21	yes
222489	S176_op1_nL1_nI1	22.4	yes
222490	S202A_op1_nL1_nI1	11	yes
222491	AUG175_op1_oL1_nI1	10	yes
222492	OI157_op1_oL1_nI1	8.52	yes
222493	S176_op1_oL1_nI1	8.32	yes
222494	S202A_op1_oL1_nI1	12.5	yes
222495	LB1_op1_nL1_nI1	10.792	yes
222496	EB2_op1_nL1_nI1	0.296	yes
222497	LB2_op1_nL1_nI1	0.224	yes
222498	LB3_old_op1_oL1_nI1	0.156	yes
225353	ARD13_op1_oL1_oI1	1.39	yes
225354	AUG175_op1_oL1_oI1	3.4	yes
225355	OI117_op1_oL1_oI1	5.42	yes
225356	OI157_op1_oL1_oI1	3.74	yes
225357	AUG87_op1_oL1_oI1	3.48	yes
225358	OI89_op1_oL1_oI1	1.05	yes
225359	OI89_op1_nL1_nI1	1.38	yes
225360	OI89_op1_oL1_nI1	2.86	yes
225361	AUG182_op1_oL1_oI1	3.7	yes
225362	AUG182_op1_nL1_nI1	7.8	yes
225363	AUG182_op1_oL1_nI1	7.5	yes
225364	OI32_op1_oL1_oI1	4.4	yes
225365	OI32_op1_oL1_nI1	11.3	yes
225366	OI32_op1_nL1_nI1	5.08	yes
225367	OI95_op1_oL1_oI1	2.22	yes
225368	OI95_op1_oL1_nI1	11.6	yes
225369	OI95_op1_nL1_nI1	5.02	yes
225370	S111_op1_oL1_oI1	9.2	yes
225371	S111_op1_oL1_nI1	8.58	yes
225372	S111_op1_nL1_nI1	0.79	yes
225373	ARD13_op1_oL1_nI1	11	yes
225374	AUG87_op1_oL1_nI1	20.4	yes
225375	OI117_op1_oL1_nI1	21	yes

225376	EB1_op1_nL1_nI1	0.528	yes
233198	S176_op1_oL1_oI1	1.38	yes
233199	S202A_op1_oL1_oI1	1.39	yes
233200	ARD13_op1_nL1_nI1	15.2	yes
233201	AUG175_op1_nL1_nI1	15.5	yes
233202	OI117_op1_nL1_nI1	14.1	yes
233203	OI157_op1_nL1_nI1	18.1	yes
233204	AUG87_op1_nL1_nI1	21	yes
233205	S202A_op1_nL1_nI1	11	yes
233206	AUG175_op1_oL1_nI1	10	yes
233207	OI157_op1_oL1_nI1	8.52	yes
233208	S176_op1_oL1_nI1	8.32	yes
233209	S202A_op1_oL1_nI1	12.5	yes
233210	LB1_op1_nL1_nI1	10.792	yes
233211	EB2_op1_nL1_nI1	0.296	yes
233212	LB2_op1_nL1_nI1	0.224	yes
233213	LB3_old_op1_oL1_nI1	0.156	yes
233214	PCRB_1.1	0.216	yes
233215	PCRB_1.2	out of range too low	yes
233216	ARD13_op1_oL1_oI1	1.39	yes
233217	AUG175_op1_oL1_oI1	3.4	yes
233218	OI117_op1_oL1_oI1	5.42	yes
233219	OI157_op1_oL1_oI1	3.74	yes
233220	AUG87_op1_oL1_oI1	3.48	yes
233221	OI89_op1_oL1_oI1	1.05	yes
233222	OI89_op1_nL1_nI1	1.38	yes
233223	OI89_op1_oL1_nI1	2.86	yes
233224	AUG182_op1_oL1_oI1	3.7	yes
233225	AUG182_op1_nL1_nI1	7.8	yes
233226	AUG182_op1_oL1_nI1	7.5	yes
233227	OI32_op1_oL1_oI1	4.4	yes
233228	OI32_op1_oL1_nI1	11.3	yes
233229	OI32_op1_nL1_nI1	5.08	yes
233230	OI95_op1_oL1_oI1	2.22	yes
233231	OI95_op1_oL1_nI1	11.6	yes
233232	OI95_op1_nL1_nI1	5.02	yes
233233	S111_op1_oL1_oI1	9.2	yes
233234	S111_op1_oL1_nI1	8.58	yes
233235	S111_op1_nL1_nI1	0.79	yes
233236	ARD13_op1_oL1_nI1	11	yes
233237	AUG87_op1_oL1_nI1	20.4	yes
233238	OI117_op1_oL1_nI1	21	yes
233239	EB1_op1_nL1_nI1	0.528	yes

A.2 qPCR results

Table 28: qPCR results from 17.01.2023

Library ID	sequencing ID	Cq Raw	Copies raw	R^2 (ΔR)	Copies/uL	Copies total	Dye
L ARD13	ARD13_op1_nL1_nI1	14.2	1.01E+07	0,975	4.04E+08	1.62E+10	SYBR
L AUG175	AUG175_op1_nL1_nI1	12.22	3.16E+07	0,975	1.27E+09	5.06E+10	SYBR
L OI117	OI117_op1_nL1_nI1	14.72	7.48E+06	0,975	2.99E+08	1.20E+10	SYBR
L OI57 (L OI157)	OI57_op1_nL1_nI1	12.81	2.26E+07	0,975	9.03E+08	3.61E+10	SYBR
L AUG87	AUG87_op1_nL1_nI1	12.11	3.38E+07	0,975	1.35E+09	5.40E+10	SYBR
L OI89	OI89_op1_nL1_nI1	11.01	6.37E+07	0,975	2.55E+09	1.02E+11	SYBR
L S111	S111_op1_nL1_nI1	17.24	1.74E+06	0,975	6.98E+07	2.79E+09	SYBR
L S176	S176_op1_nL1_nI1	11.82	4.00E+07	0,975	1.60E+09	6.41E+10	SYBR
L S202A	S202A_op1_nL1_nI1	12.38	2.89E+07	0,975	1.16E+09	4.63E+10	SYBR
L AUG182	AUG182_op1_nL1_nI1	10.31	9.59E+07	0,975	3.84E+09	1.53E+11	SYBR
L OI32	OI32_op1_nL1_nI1	11.44	4.98E+07	0,975	1.99E+09	7.97E+10	SYBR
L OI95	OI95_op1_nL1_nI1	12.71	2.39E+07	0,975	9.57E+08	3.83E+10	SYBR
L EB1	Blank	27.05	6.00E+03	0,975	2.40E+05	9.60E+06	SYBR
LB1	Blank	26.91	6.52E+03	0,975	2.61E+05	1.04E+07	SYBR
L EB2	Blank	27.56	4.46E+03	0,975	1.79E+05	7.14E+06	SYBR
L Lib2	Blank	30.21	9.69E+02	0,975	3.88E+04	1.55E+06	SYBR

Table 29: qPCR values from 13.02.2023

Library ID	Well Type	Cq (ΔR)	Copies raw	R^2 (ΔR)	Copies/uL	Copies total	Dye
old_ARD13	ARD13_op1_oL1_nI1	11.46	1.22E+08	0.709	4.86E+09	1.95E+11	SYBR
old_AUG175	AUG175_op1_oL1_nI1	10.99	1.63E+08	0.709	6.54E+09	2.61E+11	SYBR
old_AUG87	AUG87_op1_oL1_nI1	12.59	6.05E+07	0.709	2.42E+09	9.68E+10	SYBR
old_OI117	OI117_op1_oL1_nI1	12.16	7.90E+07	0.709	3.16E+09	1.26E+11	SYBR
old_OI157	OI157_op1_oL1_nI1	13.62	3.18E+07	0.709	1.27E+09	5.09E+10	SYBR
old_S176	S176_op1_oL1_nI1	11.58	1.13E+08	0.709	4.52E+09	1.81E+11	SYBR
old_S202A	S202A_op1_oL1_nI1	11.72	1.04E+08	0.709	4.16E+09	1.66E+11	SYBR
LB3	Blank	25.22	2.34E+04	0.709	9.35E+05	3.74E+07	SYBR

Table 30: qPCR values from 02.05.2023

Library ID	Well Type	Cq (ΔR)	Quantity (copies)	R^2 (ΔR)	Copies/uL	Copies total	Dye
	AUG182_op1_oL1_nI1	7.47	2.36E+08	0.959	9.42E+09	3.77E+11	SYBR
*	ARD13_op1_oL1_oI1	8.17	1.35E+08	0.959	5.38E+09	2.15E+11	SYBR
	ARD13_op1_oL1_nI1	8.44	1.08E+08	0.959	4.34E+09	1.74E+11	SYBR
	OI32_op1_oL1_nI1	8.49	1.04E+08	0.959	4.16E+09	1.67E+11	SYBR
*	OI95_op1_oL1_oI1	8.76	8.36E+07	0.959	3.34E+09	1.34E+11	SYBR
	OI89_op1_oL1_nI1	8.83	7.92E+07	0.959	3.17E+09	1.27E+11	SYBR
	OI117_op1_oL1_nI1	9.04	6.69E+07	0.959	2.67E+09	1.07E+11	SYBR
	OI89_op1_nL1_nI1	9.26	5.61E+07	0.959	2.24E+09	8.98E+10	SYBR
	AUG182_op1_oL1_nI1	9.35	5.21E+07	0.959	2.08E+09	8.34E+10	SYBR
	OI32_op1_nL1_nI1	9.39	5.04E+07	0.959	2.02E+09	8.07E+10	SYBR

	OI95_op1_nL1_nI1	10.3	2.41E+07	0.959	9.66E+08	3.86E+10	SYBR
	AUG87_op1_oL1_nI1	10.38	2.27E+07	0.959	9.08E+08	3.63E+10	SYBR
*	OI89_op1_oL1_oI1	10.48	2.09E+07	0.959	8.37E+08	3.35E+10	SYBR
	S111_op1_oL1_nI1	12.4	4.46E+06	0.959	1.78E+08	7.14E+09	SYBR
	S111_op1_nL1_nI1	14.37	9.16E+05	0.959	3.66E+07	1.47E+09	SYBR
	EB1_op1_nL1_nI1	24.02	3.90E+02	0.959	1.56E+04	6.25E+05	SYBR
*	AUG182_op1_oL1_oI1	No Cq	No Cq	0.959			SYBR
*	AUG175_op1_oL1_oI1	No Cq	No Cq	0.959			SYBR
*	S111_op1_oL1_oI1	No Cq	No Cq	0.959			SYBR
*	OI117_op1_oL1_oI1	No Cq	No Cq	0.959			SYBR
	OI57_op1_oL1_oI1	No Cq	No Cq	0.959			SYBR
	OI32_op1_oL1_oI1	No Cq	No Cq	0.959			SYBR
*	AUG87_op1_oL1_oI1	No Cq	No Cq	0.959			SYBR
*	OI95_op1_oL1_oI1	No Cq	No Cq	0.959			SYBR

*=no library created - sample sequenced from already indexed library in 2015

Appendix B: Detailed DNA Extraction Protocol and Library preparation protocol.

B.1. DNA Extraction Protocol

In progress...

- PART 2: DNA EXTRACTION -

A. DIGESTION (1 DAY)

Extraction Buffer

	final []	1mL	9mL	12mL
EDTA (0.5M)	0.45M	900µL	8.1mL	10.8mL
Proteinase K (¬15mg/mL)	0.25mg/mL	16.7µL	150µL	200µL
dH$_2$0	up to XmL	83.3µL	750µL	1mL

Binding Buffer

	final []	for 50mL
Guanidine hydrochloride (MW 95.53)	5M	23.9gr
Isopropanol	40%	20mL
Tween-20 (10%)	0.05%	250µL
Sodium Acetate (3M)	90mM	1.5mL
Up to 50mL with H$_2$0	-	-

Should keep for 1 month. Making 3X 50mL Falcon tubes, is enough for 10 or so extracts. You might experience difficulty getting the Guanidine to dissolve completely. Make sure there are no large lumps of Guanidine present and as soon as possible add all the reagents up to 50mL.

TET Buffer

	final []	for 50mL
1M Tris-HCl (pH8.0)	10mM	500µL
EDTA (0.5M)	1mM	100µL
Tween-20 (10%)	0.05%	250µL
dH$_2$0	-	49mL 150µL

Can be stored indefinitely. Suggest volume is 50mL. 100% Tween-20 is basically impossible to pipette accurately in small amounts, so make a 10% stock solution by adding 1mL Tween-20 (check volume against graduation on pipette tip or Falcon tube) to 9mL dH$_2$0.

Binding Apparatus

- Remove column and submerge the reservoir for 20 minutes in bleach OR clean well with DNAoff, rinse well with dH$_2$0 and UV for 30 minutes before use
- Assemble the binding apparatus by forcibly fitting the MinElute column on to the extension reservoir
- Remove the MinElute collection tube and cut off the lid, but remember to keep these for the later stages. The lid can be pushed into the top of the collection tube to maintain sterility. Labelling the side of the spin column as well as the lid of the Falcon tube is useful to ensure samples are not mixed up
- Put the reservoir plus spin column inside the 50mL Falcon tube. It fits nicely but the lid of the falcon tube no longer fits properly. Tape the lid on prior to centrifugation to prevent leakage

A.1 PROCEDURE

- Clean the benches and the hood with DNAoff, and ethanol on metal surfaces
- Prepare fresh Extraction Buffer according to the table, without the Proteinase-K, and UV it for 30 minutes
- Transfer the Proteinase-K to the falcon tube and pipette-mix
- Prepare an Extraction Blank in a 1,5mL tube
- Transfer 1mL into each tube with the 50mg of bone powder, extraction blanks and air control
- Vortex the tubes to ensure every bit of powder is dissolved
- Incubate at 37ºC for ~18 hours in the Thermomixer (i.e. with rotation)
- Passed the 18h at 37ºC, centrifuge the tubes (2 mins @ 13,000rpm), transfer the supernatant to 13 mL in the binding apparatus reservoir. Remember to have the apparatus labeled
- Centrifuge the binding apparatus for 4mins at 2.500rpm. Rotate 90º and centrifuge again for 2mins at 3.000rpm
- Unassemble the binding apparatus and place the MinElute column in the collection tube and replace the original cap (you should have retained these from earlier), then dry spin for 1min at 6.000rpm

- Add **650µL** of PE washing buffer to the Qiagen column and centrifuge again for 1 minute at 6.000rpm. discarding the flowthrough. Repeat washing step
- Dry spin the MinElute column at maximum speed for 1m, place in clean low retention 1.5mL tube
- Elute by adding **12.5µL** TET buffer to the silica membrane, incubate for 10m, and centrifuge at maximum speed for 30s. Repeat to give a total of 25µL DNA extract

B.2 Library preparation protocol for Illumina Sequencing

In progress…

- PART 3: LIBRARY PREPARATION -

FIRST, CHECK IF:

- All the reagents above exist in the correct concentrations (especially the primers). If not, you need to prepare them (relate to separate 'Recipes & Dilutions' document);
- Thermostats are at 12ºC and 25ºC, for the first two steps.

D. BLUNT-END REPAIR

- Get the tubes directly from the extraction step (should have 50µL). This step is done directly in them. Include a library blank
- Prepare the Blunt-End Repair MasterMix according to the following table. Volume of ddH_2O depends on the amount of DNA extract. Total reaction volume should be 70µl. Avoid vortexing after addition of enzymes

Blunt-End Repair MasterMix

	From Yang (50µl)	From Dabney (25µl)	**13** samples
	1x	**1x**	**?x (add 2 extra)**
End Repair Enzime Mix	3.5 µl	3.5 µl	45.5
End Repair Buffer	7 µl	7 µl	91
ddH_2O	9.5 µl	34.5 µl	448,5
Total	20 µl	45 µl	585

- Mix the mastermix by pipetting up and down
- Transfer 1 volume of mastermix (20 or 45µL) to 1 volume of DNA extract (50 or 25µL) and pipette-mix
- Incubate for 15 minutes at 25ºC followed by 5 minutes at 12ºC

A.1. SAMPLE CLEAN-UP

- Get the Qiagen Min-Elute purification kit: columns (fridge), PB, EB and EBT buffers. Label each column
- Add **350µL** of PB binding buffer to each Qiagen column and then add the 70µL of each end-repaired sample
- Centrifuge 1 minute at maximum speed (13.000 rpm) and discard the flowthrough
- Add **750µL** of PE washing buffer to the Qiagen column and centrifuge again for 1 minute at 13.000 rpm
- Discard the flowthrough and centrifuge again to remove all ethanol (included in the PE buffer), 1 minute, 13.000 rpm
- Place the spin column in a new 1.5mL tube
- Elute in **20µL** of EBT elution buffer and after 1 minute of wait, centrifuge for 1 minute at maximum speed. To centrifuge the tubes with the lids open you must leave 2 holes in between each tube and turn the tubes so the lids face clockwise direction, which is opposite of the centrifugation direction
- After centrifuging keep the 1.5ml tubes with the elute and throw away the 68olumn
- Put the samples on ice or proceed to the next step

B. ADAPTER LIGATION

- Prepare the mastermix according to the following table:
- **Note:** If white precipitate is present in the 10X DNA ligase buffer after thawing, warm the buffer to 37°C and vortex until the precipitate has dissolved
- **Note:** Since PEG is highly viscous, vortex the mastermix before adding the T4 DNA ligase and mix gently thereafter.

Adapter Ligation MasterMix

	1x	**13x (add 3 extra)**
ddH_2O	10 µl	130

T4 DNA ligase buffer (10x)	4 µl	52	
PEG-4000 (50%)	4 µl		52
Adapter mix (100uM each)	1 µl		13
T4 DNA ligase (5U/ul)	1 µl		13
Total	20 µl		260

- Mix the mastermix by pipetting up and down
- Add 20µl of mastermix to each eluate from step A.1 (Total reaction volume = 40µl) and pipette-mix
- Incubate for 30 minutes at 22˚C

B.1. SAMPLE CLEAN-UP

- Get the Qiagen Min-Elute purification kit: columns (fridge), PB, EB and EBT buffers. Label each column
- Add **200µL** of PB binding buffer to each Qiagen column and then add the 40µL of each sample
- Centrifuge 1 minute at maximum speed (13.000 rpm) and discard the flowthrough
- Add **750µL** of PE washing buffer to the Qiagen column and centrifuge again for 1 minute at 13.000 rpm
- Discard the flowthrough and centrifuge again to remove all ethanol (included in the PE buffer), 1 minute, 13.000 rpm
- Place the spin column in a new 1.5mL tube
- Elute in **20µL** of EBT elution buffer and after 1 minute of wait, centrifuge for 1 minute at maximum speed. To centrifuge the tubes with the lids open you must leave 2 holes in between each tube and turn the tubes so the lids face clockwise direction, which is opposite of the centrifugation direction
- After centrifuging keep the 1.5ml tubes with the elute and throw away the 69olumn
- Put the samples on ice or proceed to the next step

C. ADAPTER FILL-IN

- Prepare the mastermix according to the following table:

Adapter Fill-In MasterMix

	1x	13x (add 3 extra)
ddH_2O	13.5 µl	175,5
Thermopol reaction buffer (10x)	4 µl	52
dNTPs (10mM each)	1 µl	13
Bst polymerase, large fragment (8 U/ul)	1.5 µl	19,5
Total	20 µl	260

- Mix the mastermix by pipetting up and down
- Add 20µl of mastermix to each eluate from step B.1 (Total reaction volume = 40µl) and pipette-mix
- Incubate for 30 minutes at 37˚C followed by 20 minutes at 80˚C
- Put the samples on ice or proceed to the next step

D. INDEXING PCR (VALUES FOR SECOND INDEXING)

- NOTE: If using Accuprime's Pfx Supermix, use following steps. If not using the Supermix, refer to next page.
- Prepare the mastermix according to the following table (include a PCR blank):
-

Amplification MasterMix (with Pfx Supermix)

	1x	16x (add 1 extra)
Accuprime Pfx Supermix	20.5 µl	328
Primer IS4 (10mM)	0.5 µl	8
Total	21 µl	

- To clean PCR tubes, add 21µl of mastermix, 1µl of appropriate indexing adapter (5µM) & 3µl of DNA sample (total reaction volume = 25µl). Pipette-mix

- Take the samples to the 70olumn70ycler and perform the following cycle

95ºC	5 mins	
95ºC	15 sec	|
60ºC	30 sec	| 12x
68ºC	30 sec	|
68ºC	5 min	
4ºC	Forever	

- Take the samples to the **modern lab** and perform the clean-up

D.1. SAMPLE CLEAN-UP

- Get the Qiagen Min-Elute purification kit: columns (fridge), PB, EB and EBT buffers. Label each column
- Add **125µL** of PB binding buffer to each Qiagen column and then add the 25µL of each sample
- Centrifuge 1 minute at maximum speed (13.000 rpm) and discard the flowthrough
- Add **750µL** of PE washing buffer to the Qiagen column and centrifuge again for 1 minute at 13.000 rpm
- Discard the flowthrough and centrifuge again to remove all ethanol (included in the PE buffer), 1 minute, 13.000 rpm
- Place the spin column in a new 1.5mL tube
- Elute in **25µL** of EBT elution buffer and after 1 minute of wait, centrifuge for 1 minute at maximum speed. To centrifuge the tubes with the lids open you must leave 2 holes in between each tube and turn the tubes so the lids face clockwise direction, which is opposite of the centrifugation direction
- After centrifuging keep the 1.5ml tubes with the elute and throw away the 70olumn
- Put the samples on ice or proceed to the next step

Appendix C

C1. Bioinformatic customized scripts to create the final version of the .bam files and extract some information.

Step1

```
cd /lisc/scratch/anthropology/Pinhasi_group/raimo
for filename in ./Step0d/*.fastq.gz;
  do cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -m 30 "${filename}" > "${filename}.fastq" 2>
"${filename}.log";
  echo ${filename:9:6}
done
cd Step0d
rename .fastq.gz.fastq .fastq *
rename .fastq.gz.log .log *
mv *.log  $HOME/Cutadaptlogs
mv *.fastq ../Step1d
cd  $HOME
```

Step2

```
ref="/lisc/scratch/anthropology/Pinhasi_group/raimo/hg37/human_g1k_v37.fasta"
module load bwa
cd /lisc/scratch/anthropology/Pinhasi_group/raimo
for filename in ./Step1d/*.fastq;
"${filename}" > "${filename}.sam" 2> "${filename}.sam.log";
  do bwa mem -t 8 -R '@RG\tID:RaimoMasterarbeit\tSM:'${filename:9:6}'\tPL:ILLUMINA' $ref "${filename}" >
"${filename}.sam" 2> "${filename}.sam.log";
  echo ${filename:9:6}
done
cd Step1d
rename .fastq.sam .sam *
rename .fastq.sam.log .sam.log *
mv *.sam.log ../Step2d
mv *.sam ../Step2d
```

Step31

```
module load samtools
cd /lisc/scratch/anthropology/Pinhasi_group/raimo
for filename in ./Step2d/*.sam;
  do samtools view -b -q30  "${filename}" > "${filename}_q30.bam";
  echo ${filename:9:6}
done
cd Step2d
rename .sam_q30.bam _q30.bam *
mv *.bam ../Step3d
```

Step32

```
module load samtools
cd /lisc/scratch/anthropology/Pinhasi_group/raimo

for filename in ./Step3d/*q30.bam;
  do
```

```
    samtools sort "${filename}"  -m 4G -@ 8 -o "${filename}"_sorted.bam;
    echo ${filename:9:6}
done
cd Step3d
rename _q30.bam_sorted.bam _q30_sorted.bam *
```

Step33

```
module load samtools
cd /lisc/scratch/anthropology/Pinhasi_group/raimo

for filename in ./Step3d/*q30_sorted.bam;
do
   samtools rmdup -s "${filename}" "${filename}"_rmdup.bam;
   echo ${filename:9:6}
done
cd Step3d
rename q30_sorted.bam_rmdup.bam q30_sorted_rmdup.bam  *
samtools index -M *q30_sorted_rmdup.bam
```

Step34

```
module load samtools
cd /lisc/scratch/anthropology/Pinhasi_group/raimo

for filename in ./Step3d/*sorted_rmdup.bam;
do
   samtools flagstat -O tsv "${filename}" > /home/user/raimo/Summaries/"${filename:9:6}"_tsvsummary.txt;
   echo ${filename:9:6}
done
```

CreateReport1

```
echo 'ID ; Total Reads' > DATA_Reads.txt;
echo 'ID ; Trimmed Reads ; Percentual ' > DATA_TrimmedReads.txt;
for filename in ./Cutadaptlogs/*;
do
        (echo -n ${filename:15:6}; echo -n ' ; '; grep 'Total reads processed' "${filename}";)>> DATA_Reads.txt;
        (echo -n ${filename:15:6}; echo -n ' ; '; grep 'Reads with adapters' "${filename}"; )>>
DATA_TrimmedReads.txt;
        echo ${filename:15:6}
done
(sed 's/Total reads processed:/ /' DATA_Reads.txt) > DATA_Reads2;
(sed 's/Reads with adapters:/ /' DATA_TrimmedReads.txt) > DATA_TrimmedReads21;
(sed 's/,//g'  DATA_Reads2) > DATA_Reads.txt;
(sed 's/(/; /'  DATA_TrimmedReads21) > DATA_TrimmedReads22;
(sed 's/)//'  DATA_TrimmedReads22) > DATA_TrimmedReads23;
(sed 's/%//'  DATA_TrimmedReads23) > DATA_TrimmedReads24;
(sed 's/,//g'  DATA_TrimmedReads24) > DATA_TrimmedReads25;
(sed 's/\./,/'  DATA_TrimmedReads25) > DATA_TrimmedReads.txt;
rm DATA_Reads2;
rm DATA_TrimmedReads2*;
mv DATA_Reads.txt ./DATAnew;
mv DATA_TrimmedReads.txt ./DATAnew;
```

CreateReport2

```
echo 'ID ; Endogenous' > DATA_Endogenous.txt;
echo 'ID ; Unique' > DATA_Unique.txt;
for filename in ./Summaries/*tsvsummary.txt;
do
        (echo -n ${filename:12:6};echo -n ';';sed '7!d' "${filename}";)>> DATA_Endogenous.txt;
        (echo -n ${filename:12:6};echo -n ';';sed '9!d' "${filename}";)>> DATA_Unique.txt;
        echo ${filename:12:6}
done
(sed 's/mapped/ /' DATA_Endogenous.txt) > DATA_End21;
(sed 's/primary mapped/ /' DATA_Unique.txt) > DATA_Uni21;
(sed 's/\t/;/' DATA_End21) > DATA_End22;
(sed 's/\t/;/' DATA_Uni21) > DATA_Uni22;
(sed 's/;0//' DATA_End22) > DATA_Endogenous.txt;
(sed 's/;0//' DATA_Uni22) > DATA_Unique.txt;

rm DATA_End2*;
rm DATA_Uni2*;
mv DATA_Endogenous.txt ./DATAnew;
mv DATA_Unique.txt ./DATAnew;
```

MergeData

```
cd DATAmerged;
cp ../DATAold/DATA_Reads.txt DATA_Reads1.txt;
cp ../DATAold/DATA_TrimmedReads.txt DATA_TrimmedReads1.txt;
cp ../DATAold/DATA_Endogenous.txt DATA_Endogenous1.txt;
cp ../DATAold/DATA_Unique.txt DATA_Unique1.txt;
cp ../DATAnew/DATA_Reads.txt DATA_Reads2.txt;
cp ../DATAnew/DATA_TrimmedReads.txt DATA_TrimmedReads2.txt;
cp ../DATAnew/DATA_Endogenous.txt DATA_Endogenous2.txt;
cp ../DATAnew/DATA_Unique.txt DATA_Unique2.txt;
sed -i '1d' DATA_Reads2.txt;
sed -i '1d' DATA_TrimmedReads2.txt;
sed -i '1d' DATA_Endogenous2.txt;
sed -i '1d' DATA_Unique2.txt;
cat DATA_Reads1.txt DATA_Reads2.txt > DATA_Reads.txt;
cat DATA_TrimmedReads1.txt DATA_TrimmedReads2.txt > DATA_TrimmedReads.txt;
cat DATA_Endogenous1.txt DATA_Endogenous2.txt > DATA_Endogenous.txt;
cat DATA_Unique1.txt DATA_Unique2.txt > DATA_Unique.txt;
rm *1.txt;
rm *2.txt;
cd .. ;
```

C2. Bioinformatic customized scripts for Kinship analysis

To create text Plink files from the .bam files

```
#!/bin/bash
date
# Display usage if the script is not run with required options
usage() {
  echo "Usage: $0 -B <bam_file1> [-B <bam_file2> ...]"
  exit 1
}

# Check if no arguments were provided
```

```
if [ $# -eq 0 ]; then
   usage
fi

# Initialize an array to hold BAM files
declare -a bam_files

# Parse command-line options
while getopts "B:" opt; do
   case $opt in
      B) bam_files+=("$OPTARG")  # Add each BAM file to the array
      ;;
      *) usage
      ;;
   esac
done

# Check if at least one BAM file is specified
if [ ${#bam_files[@]} -eq 0 ]; then
   echo "Error: At least one BAM file must be specified."
   usage
fi

# Load required modules
module load samtools
module load sequencetools
echo "modules loaded"

# Define paths to tools and files
src="/lisc/scratch/anthropology/Pinhasi_group/raimo/hg19/v54.1.p1_1240K_public.snp"
ref="/lisc/scratch/anthropology/Pinhasi_group/raimo/hg19/hg19.fa"
InputBam="/lisc/scratch/anthropology/Pinhasi_group/raimo/junBam"
OutputDir="/lisc/user/raimo/Software/PlinkPedMap"

cd $OutputDir

# Process each BAM file
for bam_file in "${bam_files[@]}"; do
   sample=${bam_file:0:6}  # Extract the first 6 characters from each bam_file
   pileup_file="$OutputDir/${sample}_output.pileup"
   cleaned_pileup_file="$OutputDir/${sample}_cleaned_output.pileup"

   # Step 1: Create pileup file
   echo "samtools mpileup started"
   samtools mpileup -R -B -q30 -Q30 -f $ref $InputBam/$bam_file > $pileup_file
   echo "samtools mpileup done for $sample"

   # Step 2: Make substitutions: 's/chrXY/chr25/g', 's/chrX/chr23/g', 's/chrY/chr24/g', 's/chrM/chr26/g'
   sed 's/chrXY/chr25/g' $pileup_file | sed 's/chrX/chr23/g' | sed 's/chrY/chr24/g' | sed 's/chrM/chr26/g' >
$cleaned_pileup_file
   echo "substitution done for $sample"

   # Step 3: Use PileupCaller to convert pileup to SNP data and create .bed .bim .fam
   pileupCaller --sampleNames $sample --samplePopName $sample -f $src -p $sample --randomHaploid <
$cleaned_pileup_file
```

```
    echo "pileupCaller completed for $sample"

    # Load required modules
    module load conda
    # conda activate plink
    conda activate plink2-2.00a5.12
    echo "plink2 module loaded - conversion started"


    # Step 4: Use plink2 to export PED and MAP files
    plink2 --bfile $OutputDir/$sample --export ped --out $OutputDir/$sample --threads 4
    echo "Conversion completed for $sample. Output files are located in $OutputDir"
done
date
```

PKin.sh: Kinship analysis (a modified version of tkgwv2.py by Daniel Fernandes rewritten as bash shell script)

```
#!/bin/bash
# Author: Alexandra Raimo (this script runs plink2tkrelated created by Daniel Fernandes)
#The name PKin means PlinkKinship, it uses the Plink text format files
# plink2tkrelated.R has to be in the same directory of this script
# launch as : /lisc/user/raimo/Software/tkg-v1.sh -f /lisc/user/raimo/Software/FreqFile/1000GP3_EUR_1240K.frq -
P samples.txt
# launch this script from the directory hosting the .ped , the .map and the sample.txt files
# to do: check how /lisc/user/raimo/Software/tkgwv2-master/helpers/distSimulations.R works
date
# load the needed modules. R needs data.table. Check if data.table is installed or install.packages("data.table")
module load R
module load conda
conda activate plink
#needs plink 1.9 and doesn´t work for now with plink2

## a lot to be be improved still.
## freqFile hardcoded in /lisc/user/raimo/Software/tkgwv2-master/scripts/plink2tkrelated.R
echo -e "\n### Starting PKin.sh ###"
echo "Current working directory: $(pwd)"

cwd=$(pwd)
scriptwd=$(dirname "$(realpath "$0")")
echo "Script directory: $scriptwd"

# Initialize variables
argList=("$@")
plink2tkrelatedArgs=()
prefixes=()
prefix_file=""
verbose=false

# Function to display help message
show_help() {
    echo "Usage: PKin.sh [options]"
    echo "Options:"
    echo "  -h, --help              Show help"
    echo "  -f, --freqFile FILE     Set freqFile"
```

```
    echo "  -i, --ignoreThresh VALUE     Set ignoreThresh"
    echo "  -v, --verbose                Enable verbose output"
    echo "  -P, --prefixFile FILE        Specify file containing prefixes"
}

# Process arguments
while [[ $# -gt 0 ]]; do
    case "$1" in
        -h|--help)
            show_help
            exit 0
            ;;
        -f|--freqFile)
            plink2tkrelatedArgs+=("--freqFile=$2")
            shift 2
            ;;
        -i|--ignoreThresh)
            plink2tkrelatedArgs+=("--ignoreThresh=$2")
            shift 2
            ;;
        -v|--verbose)
            plink2tkrelatedArgs+=("--verbose")
            verbose=true
            shift
            ;;
        -P|--prefixFile)
            prefix_file="$2"
            shift 2
            ;;
        *)
            echo "Unknown option: $1"
            show_help
            exit 1
            ;;
    esac
done

# Read prefixes from file if provided
if [[ -n "$prefix_file" ]]; then
    echo "Reading prefixes from file: $prefix_file"
    if [[ -f "$prefix_file" ]]; then
        while IFS= read -r line || [[ -n "$line" ]]; do
            prefixes+=("$line")
        done < "$prefix_file"
    else
        echo "Error: The file $prefix_file was not found."
        exit 1
    fi
fi

# Error if no prefixes were found
if [[ ${#prefixes[@]} -eq 0 ]]; then
    echo "Error: No prefixes found in the provided file."
    exit 1
fi
```

```
# Add PED/MAP file arguments for each prefix
for prefix in "${prefixes[@]}"; do
   plink2tkrelatedArgs+=("--pedFile=${prefix}.ped" "--mapFile=${prefix}.map")
done
echo "plink2tkrelatedArgs: ${plink2tkrelatedArgs[@]}"

# Run the plink2tkrelated command with Rscript
echo -e "\nRunning plink2tkrelated with the following command:"
command="Rscript ${scriptwd}/plink2tkrelated.R ${plink2tkrelatedArgs[@]} $scriptwd"
echo "$command"
eval "$command"
date
echo -e "\n### PKin script completed ###"
```

C3. Bioinformatic customized scripts for Calico

```
#!/bin/bash
# Load all needed modules
module load samtools
module load conda
conda activate python2
# Print start date
date

# Define paths to tools and files
ref="/lisc/scratch/anthropology/Pinhasi_group/raimo/hg19/hg19.fa"
InputBam="/lisc/scratch/anthropology/Pinhasi_group/raimo/junBam"
OutputDir="/lisc/user/raimo/Software/calico/CalicoOutput"

# Create or clear the report file
report_file="$OutputDir/reportJun.txt"
echo -e "Sample\tSITES\tMAJ\tMIN\tCONT\tCI_low\tCI_up" > "$report_file"

# Loop through all BAM files in InputBam
for bam_file in "$InputBam"/*.bam; do
  # Extract the sample name (first six characters of the BAM file name)
  sample=$(basename "$bam_file" | cut -c1-6)
  # Define the output file for each sample
  output_file="$OutputDir/${sample}_calico.txt"

  # Run samtools mpileup and the calico script for each BAM file
  samtools mpileup -E -f "$ref" "$bam_file" | python /lisc/user/raimo/Software/calico/calico.0.2.py -o
"$output_file" --outputsites

  # Extract the second row of the output file and append it to the report file
  # Adding sample name at the beginning of the line
  contamination_data=$(sed -n '3p' "$output_file" | awk '{$NF=""; print $0}')
  echo $sample
  echo $contamination_data
  echo -e "$sample\t$contamination_data" >> "$report_file"
done

# Print end date
date
```

C4. Bioinformatic customized scripts for Amber

```
#!/bin/bash
# Load all needed modules
#I need python3, which is already installed in the software
#python -m pip install -U pysam
#python -m pip install -U matplotlib
#python -m pip install -U numpy

date

# Define paths to directories and files
#src="/lisc/scratch/anthropology/Pinhasi_group/raimo/hg19/v54.1.p1_1240K_public.snp"
#ref="/lisc/scratch/anthropology/Pinhasi_group/raimo/hg19/hg19.fa"
InputBam="/lisc/scratch/anthropology/Pinhasi_group/raimo/junBam"
OutputDir="/lisc/user/raimo/Software/AmberDir/AmberOutput/Results"
BamList="/lisc/user/raimo/Software/AmberDir/BamList.tsv"

# Create output directory if it doesn't exist
mkdir -p "$OutputDir"

# Go to the directory with BAM files
cd "$InputBam" || exit

# Loop through each BAM file in the directory
for bam_file in *.bam; do
   # Extract the sample name (first 6 characters of the file name)
   sample=${bam_file:0:6}

   # Write the sample name and BAM file path to BamList.tsv
   echo -e "${sample}\t${InputBam}/${bam_file}" > "$BamList"

   # Define the output file name
   sample_output="$OutputDir/${sample}-results"

   # Run AMBER.py for the current BAM file
   python3 /lisc/user/raimo/Software/AmberDir/AMBER.py --bamfiles "$BamList" --output "$sample_output"

   # Check if the output file was successfully created
   if [[ -f "$sample_output" ]]; then
      echo "Success: Results for sample $sample written to $sample_output"
   else
      echo "Error: Results for sample $sample not found in $sample_output"
   fi
done
date
```

C5. Bioinformatic customized scripts for mapdamage GtoA (same script available for CtoT)

```
#!/bin/bash

# Define the base directory for input files
BASE_DIR="/lisc/user/raimo/mapDamage"
```

```
# Define the output directory and file
OUTPUT_DIR="/lisc/user/raimo/Software/Degradation/Results"
OUTPUT_FILE="$OUTPUT_DIR/mapdamage-table-3GtoA.txt"

# Ensure the output directory exists
mkdir -p "$OUTPUT_DIR"

# Initialize the output file (clear if it exists)
> "$OUTPUT_FILE"

# Find all files named 3pGtoA_freq.txt within directories starting with 'results_'
find "$BASE_DIR" -mindepth 2 -maxdepth 2 -type f -name "3pGtoA_freq.txt" | while read -r file; do
  # Extract exactly 6 characters after 'results_'
  sample_name=$(basename "$(dirname "$file")" | sed -E 's/^results_([0-9]{6}).*/\1/')

  # If the output file is empty, add the headers and the first sample
  if [ ! -s "$OUTPUT_FILE" ]; then
    # Add the first column (pos) and the sample name, tab-separated
    awk -v sample="$sample_name" 'NR==1 {print "pos\t" sample} NR>1 {print $1 "\t" $2}' "$file" >
"$OUTPUT_FILE"
  else
    # Add just the sample column to the file, tab-separated
    awk -v sample="$sample_name" 'NR==1 {print sample} NR>1 {print $2}' "$file" > temp_col.txt

    # Append the new column with the sample name as the header
    paste "$OUTPUT_FILE" temp_col.txt > temp_output.txt && mv temp_output.txt "$OUTPUT_FILE"
    rm temp_col.txt
  fi
done

# Print success message
echo "Table created at $OUTPUT_FILE"
```

C6. Bioinformatic customized scripts for Sex analysis

```
#!/bin/bash

##the original script was written using python2; I adjusted the last two lines of the original program so that it works
with python3, which is used on the cluster.
date

# Define input and output directories
InputDir="/lisc/scratch/anthropology/Pinhasi_group/raimo/marBam"
OutputDir="/lisc/user/raimo/Software/Sex/Results"

# Create the output directory if it doesn't exist
mkdir -p "$OutputDir"

# Load necessary modules (adjust versions as needed)
module load samtools

# Loop through all BAM files in the InputDir
for bamfile in "$InputDir"/*.bam; do
  # Extract sample name from the BAM file
  sample_name=$(basename "$bamfile" .bam)
```

```
    # Run the Python script on the current BAM file and save output
    echo "Processing $sample_name..."
    samtools view -q 30 "$bamfile" | python ry_compute.py --chrXname X --chrYname Y >
"$OutputDir/${sample_name}_sex_determination.txt"
done

date
echo "Sex determination for all BAM files from March completed!"
```

C7. A selection of produced R programs to create grouped tables

Deamination create transpose

```
Sys.setenv(LANG = "en_US.UTF-8")

# Load necessary libraries
if (!requireNamespace("readxl")) install.packages("readxl")
if (!requireNamespace("writexl")) install.packages("writexl")
if (!requireNamespace("dplyr")) install.packages("dplyr")
library(readxl)
library(writexl)
library(dplyr)

# Step 1: Load the Excel file
input_file <- "Deamination.xlsx"  # Replace with your actual file name
sheet_name <- "3GtoA"             # Replace with your sheet name
data <- read_excel(input_file, sheet = sheet_name)

# Step 2: Filter rows to include only relevant SampleTypes
filtered_data <- data %>%
  filter(SampleType %in% c("2015", "Extract", "IndexedLibrary", "Powder"))

# Step 3: Create new columns for each SampleType
filtered_data <- filtered_data %>%
  mutate(
    pos1_2015 = ifelse(SampleType == "2015", pos1, NA),
    pos1_Extract = ifelse(SampleType == "Extract", pos1, NA),
    pos1_IndexedLibrary = ifelse(SampleType == "IndexedLibrary", pos1, NA),
    pos1_Powder = ifelse(SampleType == "Powder", pos1, NA)
  )

# Step 4: Aggregate data by LabID (Maximum values)
max_aggregated_data <- filtered_data %>%
  group_by(LabID) %>%
  summarise(
    pos1_2015 = max(pos1_2015, na.rm = TRUE),
    pos1_Extract = max(pos1_Extract, na.rm = TRUE),
    pos1_IndexedLibrary = max(pos1_IndexedLibrary, na.rm = TRUE),
    pos1_Powder = max(pos1_Powder, na.rm = TRUE)
  ) %>%
  ungroup() %>%
  mutate(across(where(is.numeric), ~ ifelse(is.infinite(.), -999, .)))  # Replace -Inf with -999

# Step 5: Aggregate data by LabID (Minimum values)
min_aggregated_data <- filtered_data %>%
```

```
 group_by(LabID) %>%
 summarise(
  pos1_2015 = min(pos1_2015, na.rm = TRUE),
  pos1_Extract = min(pos1_Extract, na.rm = TRUE),
  pos1_IndexedLibrary = min(pos1_IndexedLibrary, na.rm = TRUE),
  pos1_Powder = min(pos1_Powder, na.rm = TRUE)
 ) %>%
 ungroup() %>%
 mutate(across(where(is.numeric), ~ ifelse(is.infinite(.), -999, .)))  # Replace Inf with -999

# Step 6: Aggregate data by LabID (Mean values)
mean_aggregated_data <- filtered_data %>%
 group_by(LabID) %>%
 summarise(
  pos1_2015 = mean(pos1_2015, na.rm = TRUE),
  pos1_Extract = mean(pos1_Extract, na.rm = TRUE),
  pos1_IndexedLibrary = mean(pos1_IndexedLibrary, na.rm = TRUE),
  pos1_Powder = mean(pos1_Powder, na.rm = TRUE)
 ) %>%
 ungroup()

# Step 7: Export each aggregated dataset as a separate Excel file
write_xlsx(max_aggregated_data, "max_aggregated_data.xlsx")
write_xlsx(min_aggregated_data, "min_aggregated_data.xlsx")
write_xlsx(mean_aggregated_data, "mean_aggregated_data.xlsx")

cat("Script completed successfully! Check the output files:\n",
   "max_aggregated_data.xlsx\n",
   "min_aggregated_data.xlsx\n",
   "mean_aggregated_data.xlsx\n")
```

Fragmentation create transpose

```
Sys.setenv(LANG = "en_US.UTF-8")

# Load necessary libraries
if (!requireNamespace("readxl")) install.packages("readxl")
if (!requireNamespace("writexl")) install.packages("writexl")
if (!requireNamespace("dplyr")) install.packages("dplyr")
library(readxl)
library(writexl)
library(dplyr)

# Step 1: Load the main Excel file
input_file <- "Deamination.xlsx"  # Replace with your actual file name
sheet_name <- "5CtoT"            # Replace with your sheet name 3GtoA or 5CtoT
data <- read_excel(input_file, sheet = sheet_name)

# Step 2: Load the mean_fragment_length.txt file
fragment_length <- read.delim("mean_fragment_length.txt", sep = "\t", header = TRUE)

# Step 3: Ensure SampleID is consistent across tables
data <- data %>%
  mutate(SampleID = as.character(SampleID))

fragment_length <- fragment_length %>%
  mutate(SampleID = as.character(SampleID))
```

```
# Step 4: Add SampleType and LabID to fragment_length
fragment_length <- fragment_length %>%
  left_join(data %>% select(SampleID, SampleType, LabID), by = "SampleID")

# Step 5: Remove rows where SampleType is NA
fragment_length <- fragment_length %>%
  filter(!is.na(SampleType))

# Step 6: Create columns for each SampleType and group by LabID
fragment_length_grouped <- fragment_length %>%
  mutate(
    fragment_length_2015 = ifelse(SampleType == "2015", Mean.Fragment.Length..bp., NA),
    fragment_length_Powder = ifelse(SampleType == "Powder", Mean.Fragment.Length..bp., NA),
    fragment_length_Extract = ifelse(SampleType == "Extract", Mean.Fragment.Length..bp., NA),
    fragment_length_IndexedLibrary = ifelse(SampleType == "IndexedLibrary", Mean.Fragment.Length..bp., NA)
  ) %>%
  group_by(LabID) %>%
  summarise(
    fragment_length_2015 = mean(fragment_length_2015, na.rm = TRUE),
    fragment_length_Powder = mean(fragment_length_Powder, na.rm = TRUE),
    fragment_length_Extract = mean(fragment_length_Extract, na.rm = TRUE),
    fragment_length_IndexedLibrary = mean(fragment_length_IndexedLibrary, na.rm = TRUE)
  ) %>%
  ungroup()

# Step 7: Save the fragment_length_grouped table to an Excel file
write_xlsx(fragment_length_grouped, "fragment_length_grouped.xlsx")

cat("Script completed successfully! Check the output file: fragment_length_grouped.xlsx\n")
```

Unique Read Percent create transpose

```
Sys.setenv(LANG = "en_US.UTF-8")

# Load necessary libraries
if (!requireNamespace("readxl")) install.packages("readxl")
if (!requireNamespace("writexl")) install.packages("writexl")
if (!requireNamespace("dplyr")) install.packages("dplyr")
library(readxl)
library(writexl)
library(dplyr)

# Step 1: Load the main Excel file
input_file <- "Deamination.xlsx"  # Replace with your actual file name
sheet_name <- "5CtoT"            # Replace with your sheet name 3GtoA or 5CtoT
data <- read_excel(input_file, sheet = sheet_name)

# Step 2: Load the mean_fragment_length.txt file
UniqueReads <- read.delim("UniqueReadsPercent.txt", sep = "\t", header = TRUE)

# Step 3: Ensure SampleID is consistent across tables
data <- data %>%
  mutate(SampleID = as.character(SampleID))
```

```
UniqueReads <- UniqueReads %>%
  mutate(SampleID = as.character(SampleID))

# Step 4: Add SampleType and LabID to fragment_length
UniqueReads <- UniqueReads %>%
  left_join(data %>% select(SampleID, SampleType, LabID), by = "SampleID")

# Step 5: Remove rows where SampleType is NA
UniqueReads <- UniqueReads %>%
  filter(!is.na(SampleType))

# Step 6: Create columns for each SampleType and group by LabID
UniqueReads_grouped <- UniqueReads %>%
  mutate(
    UniqueReads_2015 = ifelse(SampleType == "2015", UniquePercent, NA),
    UniqueReads_Powder = ifelse(SampleType == "Powder", UniquePercent, NA),
    UniqueReads_Extract = ifelse(SampleType == "Extract", UniquePercent, NA),
    UniqueReads_IndexedLibrary = ifelse(SampleType == "IndexedLibrary", UniquePercent, NA)
  ) %>%
  group_by(LabID) %>%
  summarise(
    UniqueReads_2015 = mean(UniqueReads_2015, na.rm = TRUE),
    UniqueReads_Powder = mean(UniqueReads_Powder, na.rm = TRUE),
    UniqueReads_Extract = mean(UniqueReads_Extract, na.rm = TRUE),
    UniqueReads_IndexedLibrary = mean(UniqueReads_IndexedLibrary, na.rm = TRUE)
  ) %>%
  ungroup()

# Step 7: Save the fragment_length_grouped table to an Excel file
write_xlsx(UniqueReads, "UniqueReadsPercent_43.xlsx")
write_xlsx(UniqueReads_grouped, "UniqueReadsPercent_grouped.xlsx")

cat("Script completed successfully! Check the output file: UniqueReadsPercent_grouped.xlsx\n")
```

C8. A selection of produced SPSS Syntax scripts

Mann Whitney Independent Samples 3 Tests

```
* Encoding: UTF-8.
*Kruskall Wallis for unrelated samples.
 *Nonparametric Tests: Independent Samples.
* Filter the data to include only the four groups.
*USE ALL.
*COMPUTE filter_$=(SampleType = "2015" OR SampleType = "Extract" OR SampleType = "IndexedLibrary" OR
SampleType = "Powder").
*FILTER BY filter_$.
*EXECUTE.

* Filter the data to include only 2015 and Extract.
USE ALL.
COMPUTE filter_$=(SampleType = "2015" OR SampleType = "Extract").
FILTER BY filter_$.
EXECUTE.

NPTESTS
```

```
  /INDEPENDENT TEST (pos1) GROUP (SampleType)
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.

EXAMINE
  VARIABLES=pos1
  BY SampleType
  /PLOT BOXPLOT
  /STATISTICS NONE.
EXECUTE.

* Filter the data to include only 2015 and IndexedLibrary.
USE ALL.
COMPUTE filter_$=(SampleType = "2015" OR SampleType = "IndexedLibrary").
FILTER BY filter_$.
EXECUTE.

NPTESTS
  /INDEPENDENT TEST (pos1) GROUP (SampleType)
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.

EXAMINE
  VARIABLES=pos1
  BY SampleType
  /PLOT BOXPLOT
  /STATISTICS NONE.
EXECUTE.


* Filter the data to include only 2015 and Powder.
USE ALL.
COMPUTE filter_$=(SampleType = "2015" OR SampleType = "Powder").
FILTER BY filter_$.
EXECUTE.

NPTESTS
  /INDEPENDENT TEST (pos1) GROUP (SampleType)
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.

EXAMINE
  VARIABLES=pos1
  BY SampleType
  /PLOT BOXPLOT
  /STATISTICS NONE.
EXECUTE.

Create Transpose For Related Test CtoT
* Encoding: UTF-8.
* Step 1: Filter to include only rows with desired SampleType values.
*if used on the final table with just included samples, is not needed.
*USE ALL.
*SELECT IF SampleType = "2015" OR SampleType = "Extract" OR SampleType = "IndexedLibrary" OR SampleType =
"Powder".
```

```
*EXECUTE.
* till here.


* Step 2: Create separate variables for each SampleType.
COMPUTE pos1_2015 = pos1 * (SampleType = "2015").
COMPUTE pos1_Extract = pos1 * (SampleType = "Extract").
COMPUTE pos1_IndexedLibrary = pos1 * (SampleType = "IndexedLibrary").
COMPUTE pos1_Powder = pos1 * (SampleType = "Powder").
EXECUTE.

* Step 3: Aggregate data by LabID using MAX.
AGGREGATE
  /OUTFILE=* MODE=ADDVARIABLES OVERWRITE=YES
  /BREAK=LabID
  /pos1_2015 = MAX(pos1_2015)
  /pos1_Extract = MAX(pos1_Extract)
  /pos1_IndexedLibrary = MAX(pos1_IndexedLibrary)
  /pos1_Powder = MAX(pos1_Powder).

* Step 4: Delete all variables named pos1 through pos25.
DELETE VARIABLES pos1 TO pos25.
DELETE VARIABLES SampleID.
DELETE VARIABLES SequencingID.
DELETE VARIABLES SampleCond.
DELETE VARIABLES SeqDate.
EXECUTE.

* Step 5: Sort the dataset by LabID and SampleType.
SORT CASES BY LabID SampleType.

* Step 6: Identify the first case for each LabID.
MATCH FILES /FILE=*
  /BY LabID
  /FIRST=FirstCase.

* Step 7: Keep only the first case for each LabID.
SELECT IF FirstCase = 1.
EXECUTE.

* Step 8: Remove the helper variable used for identification and SampleType.
DELETE VARIABLES SampleType.
DELETE VARIABLES FirstCase.
EXECUTE.

* Step 9: Save the modified dataset back to the same file.
SAVE OUTFILE='D:\01 Uni Wien\01 Master Evo Anthropologie\00 Master thesis\SPSS tables\trasposedTable.sav' .


Wilcoxon Related Rank Transposed
* Encoding: UTF-8.
*Alalyse transposed table

*Nonparametric Tests: Related Samples.
NPTESTS
```

```
  /RELATED TEST(pos1_2015 pos1_Extract) WILCOXON
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.

NPTESTS
  /RELATED TEST(pos1_2015 pos1_IndexedLibrary) WILCOXON
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.

NPTESTS
  /RELATED TEST(pos1_2015 pos1_Powder) WILCOXON
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.
```

Appendix D: Supplementary Data Tables

Table 31: Sex determination results

SampleID	SequencingID	LabID	Nseqs	NchrY+NchrX	NchrY	R_y	SE	95% CI	Assignment
225353	ARD13_op1_oL1_oI1	ARD13-RAD13	4.50E+05	22535	150	0.0067	5.00E-04	0.0056-0.0077	XX
222484	ARD13_op1_nL1_nI1	ARD13-RAD13	1.50E+06	76541	457	0.006	3.00E-04	0.0054-0.0065	XX
222491	ARD13_op1_oL1_nI1	ARD13-RAD13	1.25E+06	63007	367	0.0058	3.00E-04	0.0052-0.0064	XX
233224	ARD13_op1_nL1_nI1	ARD13-RAD13	4.05E+06	208376	1102	0.0053	2.00E-04	0.005-0.0056	XX
233230	ARD13_op1_oL1_nI1	ARD13-RAD13	4.21E+06	216093	1217	0.0056	2.00E-04	0.0053-0.0059	XX
RAD13-2015	RAD13-2015	ARD13-RAD13	3.60E+04	1793	9	0.005	1.70E-03	0.0017-0.0083	XX
222485	AUG175_op1_nL1_nI1	AUG175	2.18E+06	61338	5489	0.0895	1.20E-03	0.0872-0.0917	XY
225354	AUG175_op1_oL1_oI1	AUG175	3.10E+06	86557	7640	0.0883	1.00E-03	0.0864-0.0902	XY
225373	AUG175_op1_oL1_nI1	AUG175	8.25E+05	23091	2041	0.0884	1.90E-03	0.0847-0.0921	XY
233199	AUG175_op1_oL1_oI1	AUG175	1.64E+06	46257	4048	0.0875	1.30E-03	0.0849-0.0901	XY
233218	AUG175_op1_oL1_nI1	AUG175	2.60E+06	74175	6332	0.0854	1.00E-03	0.0834-0.0874	XY
233225	AUG175_op1_nL1_nI1	AUG175	5.93E+06	171428	14445	0.0843	7.00E-04	0.0829-0.0856	XY
AUG175-2015	AUG175-2015	AUG175	3.49E+05	9890	868	0.0878	2.80E-03	0.0822-0.0933	XY
225361	AUG182_op1_oL1_oI1	AUG182	6.29E+06	313273	1942	0.0062	1.00E-04	0.0059-0.0065	XX
225362	AUG182_op1_nL1_nI1	AUG182	1.95E+06	97570	648	0.0066	3.00E-04	0.0061-0.0072	XX
225363	AUG182_op1_oL1_nI1	AUG182	1.27E+06	63795	379	0.0059	3.00E-04	0.0053-0.0065	XX
233206	AUG182_op1_oL1_oI1	AUG182	2.56E+07	1289323	6519	0.0051	1.00E-04	0.0049-0.0052	XX
233207	AUG182_op1_nL1_nI1	AUG182	6.66E+06	338712	1873	0.0055	1.00E-04	0.0053-0.0058	XX
233208	AUG182_op1_oL1_nI1	AUG182	4.28E+06	221506	1121	0.0051	2.00E-04	0.0048-0.0054	XX
AUG182-2015	AUG182-2015	AUG182	5.16E+05	26122	116	0.0044	4.00E-04	0.0036-0.0052	XX
222488	AUG87_op1_nL1_nI1	AUG87	2.21E+06	62167	5634	0.0906	1.20E-03	0.0884-0.0929	XY
225357	AUG87_op1_oL1_oI1	AUG87	3.43E+06	95468	8557	0.0896	9.00E-04	0.0878-0.0914	XY
225374	AUG87_op1_oL1_nI1	AUG87	4.16E+06	176745	4409	0.0249	4.00E-04	0.0242-0.0257	Not Assigned
233202	AUG87_op1_oL1_oI1	AUG87	3.50E+06	98255	8632	0.0879	9.00E-04	0.0861-0.0896	XY
233228	AUG87_op1_nL1_nI1	AUG87	1.13E+07	325820	27693	0.085	5.00E-04	0.084-0.086	XY
233239	AUG87_op1_oL1_nI1	AUG87	1.57E+05	4770	367	0.0769	3.90E-03	0.0694-0.0845	consistent with XY, not XX
AUG87-2015	AUG87-2015	AUG87	3.09E+05	8766	753	0.0859	3.00E-03	0.08-0.0918	XY
222486	OI117_op1_nL1_nI1	OI117	1.20E+06	62206	450	0.0072	3.00E-04	0.0066-0.0079	XX
225355	OI117_op1_oL1_oI1	OI117	2.74E+06	137659	1388	0.0101	3.00E-04	0.0096-0.0106	XX
233200	OI117_op1_oL1_oI1	OI117	8.82E+06	449765	3933	0.0087	1.00E-04	0.0085-0.009	XX
233226	OI117_op1_nL1_nI1	OI117	3.71E+06	193539	1343	0.0069	2.00E-04	0.0066-0.0073	XX
OI117-2015	OI117-2015	OI117	1.53E+05	7834	56	0.0071	1.00E-03	0.0053-0.009	XX
225364	OI32_op1_oL1_oI1	OI32	6.72E+06	185600	16218	0.0874	7.00E-04	0.0861-0.0887	XY
225365	OI32_op1_oL1_nI1	OI32	1.27E+06	36125	3091	0.0856	1.50E-03	0.0827-0.0884	XY
225366	OI32_op1_nL1_nI1	OI32	2.17E+06	60957	5464	0.0896	1.20E-03	0.0874-0.0919	XY
233209	OI32_op1_oL1_oI1	OI32	2.16E+07	604232	51623	0.0854	4.00E-04	0.0847-0.0861	XY
233210	OI32_op1_oL1_nI1	OI32	3.87E+06	111126	9501	0.0855	8.00E-04	0.0839-0.0871	XY
233211	OI32_op1_nL1_nI1	OI32	6.76E+06	191989	16389	0.0854	6.00E-04	0.0841-0.0866	XY
OI32-2015	OI32-2015	OI32	6.04E+05	16992	1464	0.0862	2.20E-03	0.0819-0.0904	XY
233219	OI57_op1_oL1_oI1	OI57	1.15E+07	535583	7903	0.0148	2.00E-04	0.0144-0.0151	XX
233227	OI57_op1_nL1_nI1	OI57	1.13E+07	583472	2855	0.0049	1.00E-04	0.0047-0.0051	XX
233231	OI57_op1_oL1_nI1	OI57	4.03E+06	209408	1130	0.0054	2.00E-04	0.0051-0.0057	XX
OI57-2015	OI57-2015	OI57	1.24E+05	6297	26	0.0041	8.00E-04	0.0025-0.0057	XX

225358	OI89_op1_oL1_oI1	OI89	1.60E+04	701	16	0.0228	5.60E-03	0.0118-0.0339	consistent with XX, not XY
225359	OI89_op1_nL1_nI1	OI89	5.37E+04	2094	35	0.0167	2.80E-03	0.0112-0.0222	consistent with XX, not XY
225360	OI89_op1_oL1_nI1	OI89	8.93E+05	39125	358	0.0092	5.00E-04	0.0082-0.0101	XX
233204	OI89_op1_nL1_nI1	OI89	4.23E+04	2096	7	0.0033	1.30E-03	0.0009-0.0058	XX
233205	OI89_op1_oL1_nI1	OI89	1.77E+05	9176	60	0.0065	8.00E-04	0.0049-0.0082	XX
OI89-2015	OI89-2015	OI89	6.51E+02	39	0	0	0.00E+00	0.0-0.0	consistent with XX
225367	OI95_op1_oL1_oI1	OI95	5.74E+06	292660	1916	0.0065	1.00E-04	0.0063-0.0068	XX
225368	OI95_op1_oL1_nI1	OI95	1.01E+06	52659	357	0.0068	4.00E-04	0.0061-0.0075	XX
225369	OI95_op1_nL1_nI1	OI95	1.72E+06	89630	485	0.0054	2.00E-04	0.0049-0.0059	XX
233212	OI95_op1_oL1_oI1	OI95	6.25E+06	322958	1890	0.0059	1.00E-04	0.0056-0.0061	XX
233213	OI95_op1_oL1_nI1	OI95	3.18E+06	169926	883	0.0052	2.00E-04	0.0049-0.0055	XX
233214	OI95_op1_nL1_nI1	OI95	5.30E+06	279796	1435	0.0051	1.00E-04	0.0049-0.0054	XX
OI95-2015	OI95-2015	OI95	2.79E+05	14445	70	0.0048	6.00E-04	0.0037-0.006	XX
225370	S111_op1_oL1_oI1	S111	2.32E+06	116302	819	0.007	2.00E-04	0.0066-0.0075	XX
225371	S111_op1_oL1_nI1	S111	5.35E+05	27461	180	0.0066	5.00E-04	0.0056-0.0075	XX
225372	S111_op1_nL1_nI1	S111	1.75E+05	8933	78	0.0087	1.00E-03	0.0068-0.0107	XX
233215	S111_op1_oL1_oI1	S111	5.99E+06	306915	1825	0.0059	1.00E-04	0.0057-0.0062	XX
233216	S111_op1_oL1_nI1	S111	2.20E+06	114142	671	0.0059	2.00E-04	0.0054-0.0063	XX
233217	S111_op1_nL1_nI1	S111	7.60E+04	3943	34	0.0086	1.50E-03	0.0057-0.0115	XX
S111-2015	S111-2015	S111	1.80E+05	9335	53	0.0057	8.00E-04	0.0042-0.0072	XX
222482	S176_op1_oL1_oI1	S176	1.30E+05	6373	54	0.0085	1.10E-03	0.0062-0.0107	XX
222489	S176_op1_nL1_nI1	S176	1.99E+06	98339	707	0.0072	3.00E-04	0.0067-0.0077	XX
222493	S176_op1_oL1_nI1	S176	9.74E+06	495159	2856	0.0058	1.00E-04	0.0056-0.006	XX
233222	S176_op1_oL1_oI1	S176	6.98E+05	34710	256	0.0074	5.00E-04	0.0065-0.0083	XX
233232	S176_op1_oL1_nI1	S176	6.83E+06	351903	1895	0.0054	1.00E-04	0.0051-0.0056	XX
S176-2015	S176-2015	S176	3.83E+04	1821	17	0.0093	2.30E-03	0.0049-0.0138	XX
222483	S202A_op1_oL1_oI1	S202A	7.41E+04	3659	36	0.0098	1.60E-03	0.0066-0.013	XX
222490	S202A_op1_nL1_nI1	S202A	2.38E+06	119088	721	0.0061	2.00E-04	0.0056-0.0065	XX
222494	S202A_op1_oL1_nI1	S202A	6.22E+06	316130	1880	0.0059	1.00E-04	0.0057-0.0062	XX
233223	S202A_op1_oL1_oI1	S202A	6.03E+05	30012	231	0.0077	5.00E-04	0.0067-0.0087	XX
233229	S202A_op1_nL1_nI1	S202A	6.87E+06	349506	1866	0.0053	1.00E-04	0.0051-0.0056	XX
233233	S202A_op1_oL1_nI1	S202A	3.17E+06	163057	988	0.0061	2.00E-04	0.0057-0.0064	XX
S202A-2015	S202A-2015	S202A	3.11E+04	1537	16	0.0104	2.60E-03	0.0053-0.0155	XX
225376	EB1_op1_nL1_nI1	Blank-EB1	2.32E+03	85	6	0.0706	2.78E-02	0.0161-0.125	consistent with XY, not XX
233221	EB1_op1_nL1_nI1	Blank-EB1	3.88E+02	15	0	0	0.00E+00	0.0-0.0	consistent with XX
222496	EB2_op1_nL1_nI1	Blank-EB2	1.75E+03	57	1	0.0175	1.74E-02	-0.0165-0.0516	consistent with XX , not XY
233235	EB2_op1_nL1_nI1	Blank-EB2	2.38E+04	1157	10	0.0086	2.70E-03	0.0033-0.014	XX
222495	LB1_op1_nL1_nI1	Blank-LB1	1.70E+02	9	0	0	0.00E+00	0.0-0.0	consistent with XX
233234	LB1_op1_nL1_nI1	Blank-LB1	1.48E+03	69	2	0.029	2.02E-02	-0.0106-0.0686	consistent with XX, not XY
222497	LB2_op1_nL1_nI1	Blank-LB2	2.02E+03	77	4	0.0519	2.53E-02	0.0024-0.1015	Not Assigned
233236	LB2_op1_nL1_nI1	Blank-LB2	8.68E+03	374	7	0.0187	7.00E-03	0.005-0.0325	consistent with XX, not XY
222498	LB3_old_op1_oL1_nI1	Blank-LB3	2.20E+03	71	1	0.0141	1.40E-02	-0.0133-0.0415	consistent with XX, not XY
233237	LB3_old_op1_oL1_nI1	Blank-LB3	1.02E+03	52	0	0	0.00E+00	0.0-0.0	consistent with XX
222526	PCRB_1.1	Blank-PCRB_1.1	8.20E+02	30	0	0	0.00E+00	0.0-0.0	consistent with XX
233238	PCRB_1.1	Blank-PCRB_1.1	1.89E+02	10	0	0	0.00E+00	0.0-0.0	consistent with XX
222527	PCRB_1.2	Blank-PCRB_1.2	1.34E+03	52	1	0.0192	1.90E-02	-0.0181-0.0566	consistent with XX, not XY

222487	OI157_op1_nL1_nI1	EXCL	1.93E+06	99126	528	0.0053	2.00E-04	0.0049-0.0058	XX
222492	OI157_op1_oL1_nI1	EXCL	1.35E+06	68512	442	0.0065	3.00E-04	0.0059-0.0071	XX
225356	OI157_op1_oL1_oI1	EXCL	1.44E+04	690	2	0.0029	2.00E-03	-0.0011-0.0069	XX
225375	OI117_op1_oL1_nI1	EXCL	1.55E+06	44045	3880	0.0881	1.40E-03	0.0854-0.0907	XY
233198	ARD13_op1_oL1_oI1	EXCL	1.31E+04	560	13	0.0232	6.40E-03	0.0107-0.0357	consistent with XX, not XY
233201	OI157_op1_oL1_oI1	EXCL	5.28E+04	2751	21	0.0076	1.70E-03	0.0044-0.0109	XX
233203	OI89_op1_oL1_oI1	EXCL	2.35E+03	92	3	0.0326	1.85E-02	-0.0037-0.0689	consistent with XX, not XY
233220	OI117_op1_oL1_nI1	EXCL	4.99E+06	143382	12453	0.0869	7.00E-04	0.0854-0.0883	XY