UPPSALA
UNIVERSITET

# Nucleotide misincorporations, fragmentation and length patterns in ancient DNA on a temporal scale

Susanna Rankin

---

Degree project in biology, Master of science (2 years), 2010
Examensarbete i biologi 45 hp till masterexamen, 2010
Biology Education Centre, Uppsala University, and Max Planck Institute for Evolutionary Anthropology
Supervisors: Svante Pääbo and Anders Virtanen

## Summary

Ancient DNA has become a new tool for exploring evolutionary questions by looking directly at the genomes of extinct species. However ancient DNA has been plagued with problems such as the contamination with modern DNA. I examined 44 mitochondrial genomes of animals from an age range of 47 years to 2400 years before present and eight previously published 40,000-year-old Neandertal mitochondrial genomes to study three patterns of DNA damage that may be useful as tools to differentiate between modern and ancient DNA. These three patterns are: a small average length of around 50 base pairs of ancient DNA fragments; misincorporation of thymines instead of cytosine residues due to deamination of cytosines, and fragmentation due to depurination events. None of these patterns have been observed in modern DNA. Short fragment length and fragmentation due to depurination patterns were present in all samples to the same extent, showing no trend over time. In contrast, nucleotide misincorporations due to deamination showed a significant trend over time, with a frequency at or below 0.05 of deaminated cystosines in samples 100 years or younger and 0.1 to 0.58 in samples 500 years and older. This suggests that misincorporation due to deamination may be used to differentiate DNA that is younger than 100 years from older DNA in ancient DNA extracts that are contaminated with modern DNA.

## Introduction

Analyzing ancient DNA is useful to study organisms that have become extinct in the past 100,000 years and is particularly useful in understanding evolutionary

relationships, such as relationships between modern humans and Neandertals. Until recently the study of ancient DNA was limited partly due to contamination with modern DNA and damage to the ancient DNA.

DNA in living organisms is constantly damaged by biochemical processes. However, a living cell is equipped with enzymes and tools to correct or eliminate such damage (Lindahl 1993). When an organism dies the damages that occur to the DNA are no longer repaired. There are many forms of DNA damage. One such type of damage is strand breakage (Paabo *et al*. 2004). Ancient DNA fragments have repeatedly been found to be much shorter than modern DNA, averaging around 80 base pairs (bp) (Paabo *et al.* 1989). DNA can be broken up by many processes, such as depurination (Lindahl 1993). Depurination refers to the removal of a purine (a guanidine or adenine) from the phosphate-sugar backbone of DNA, which produces an abasic site and allows hydrolosis of the sugar-phosphate backbone (Lindahl 1993) (Figure 1a). Ancient DNA has also been found to develop cross linkage, which blocks DNA polymerases and therefore hinders PCR amplification (Paabo *et al.* 1989). A third form of DNA damage that occurs in ancient DNA is DNA lesions that can be due to factors such as background radiation that cause oxidative lesions through free radicals (Paabo *et al.* 2004).

Hydrolytic lesions cause the loss of amino groups from the bases adenine (A), guanine (G), cytosine (C) and 5-methylcytosine, in a process called deamination (Figure 1b). These bases change into hypoxanthine, xanthine, uracil and thymine (T), respectively, thus causing a misincorporation event when a DNA polymerase creates a new DNA strand (Paabo *et al.* 2004). The most common misincorporation event in

ancient DNA is the change from cytosine to thymine since deaminated cytosine becomes uracil, which is read as a thymine by DNA polymerases (3,5).

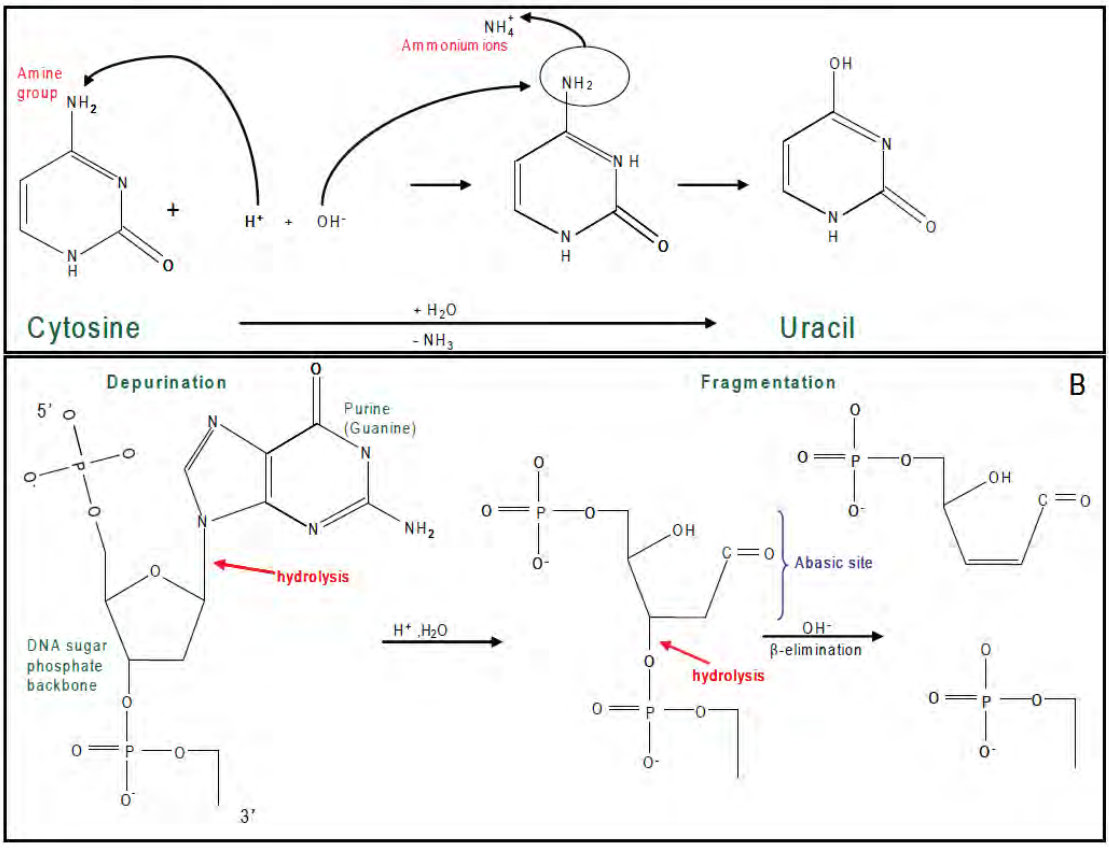


Figure 1. A: Molecular depiction of the deamination process. Hydrolysis causes the amine group to be replaced by a hydroxyl group. B: Molecular depiction of a two step process. First depurination and second the subsequent cleavage of the sugar-phosphate backbone through hydrolysis.

DNA damage can cause problems for amplifying and analyzing ancient DNA, either by blocking PCR amplification through crosslinkage and other lesions, or giving false sequences through nucleotide misincorporation. Another large problem of ancient DNA is the small amounts of DNA present in a fossil (Briggs *et al.* 2007). Since only small amounts of usable ancient DNA are present, the DNA has to be amplified through PCR to be able to sequence and analyze it. Additionally a dead organism is invaded by microbes that can contribute to the break down of macromolecules (Eglinton *et al.*1991).

The microbes not only contribute to DNA damage but often also contribute the majority of DNA extracted from the fossil (Briggs *et al.*2007). The amount of microbial DNA varies depending on the location of the fossil and how it was preserved. A further problem encountered in ancient DNA extraction is contamination with modern DNA. Due to the small amounts of DNA found in fossils, even small amounts of contamination can represent a large proportion of the DNA in the extract and can be difficult to differentiate from the endogenous DNA if they derive from the same species.

A recent publication by Briggs *et al.* 2007 has highlighted certain patterns of DNA damage that occur in 40,000-year-old Neandertal specimens. The first such pattern (Figure 2a) concerns how ancient DNA is fragmented when aligned to a reference genome. The short Neandertal fragments were aligned to the human reference sequence. Figure 2 shows the base frequency of the reference sequence and therefore enables one to look at the base composition outside the actual Neandertal fragment at the 5' end and the 3' end. There is a major spike in purine (A and G) bases one base pair outside the fragment at the 5' end while pyrimidine bases (C and T) are suppressed. One base pair outside the 3' end of the fragment, the opposite pattern is observed (a spike is seen in pyrimidines and dip in purines). Due to the preparation of the DNA for sequencing, the 3' end of the fragment shows the pattern of the 5' end of the complementary strand and therefore reflects the same actual damage pattern as seen in the 5' end of the fragment. This pattern suggests that a fragment is more likely to break after a purine, which could mean that the fragment broke due to depurination (Figure 1b).

Another pattern shown in Briggs *et al.* 2007 is a drastic increase in C to T misincorporations in the first few bases at the 5' end of ancient DNA fragments and an

increase of G to A misincorporations in the last few bases of the 3'end of the fragments (Figure 2b). This pattern can be explained by deamination of cytosine to uracil (Fig. 1a). The G to A misincorporations at the 3' ends of the fragments again represent the complementary strand due to the preparation of the DNA for sequencing.

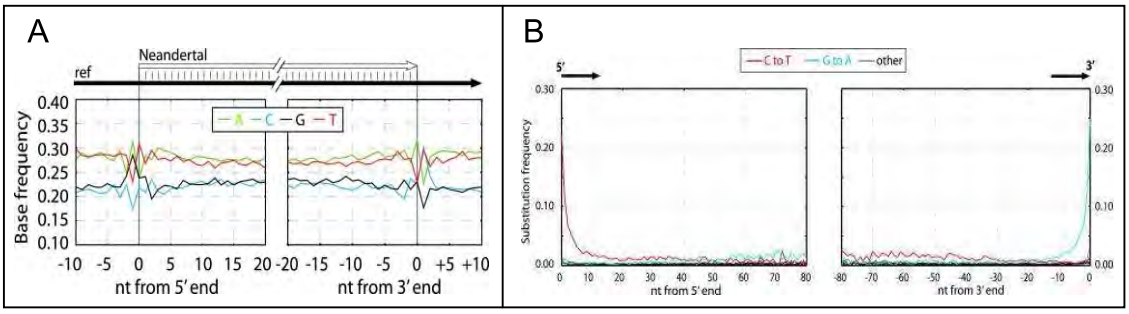


Figure 2. A: Fragmentation pattern shown in Briggs *et al.* 2007. B: Misincorporation pattern shown in Briggs *et al.* 2007.

These two patterns, along with short fragment length of ancient DNA as compared to modern DNA, have also been shown in ancient modern human bones from Kostenki, Russia (Krause *et al.* 2009), which is estimated to be about 40,000 years old. Surprisingly these patterns also exist in a 700-year-old French individual, suggesting that this pattern can arise within a few hundred years (unpublished data). Modern DNA present as human contamination in Neandertal bones, on the other hand, does not show fragmentation patterns, misincorporation patterns or short strand lengths (Briggs *et al.* 2007) making it potentially distinguishable from ancient DNA (Krause *et al.* 2009).

The aim of this project is to analyze fragmentation patterns, misincorporation patterns and fragment length in animal samples that vary in age from 47 years to 2400 years. These data can shed light on when in time these patterns arise, and if any or all of them are informative for determining whether DNA is ancient and if these patterns can be used to differentiate between ancient and contaminating modern DNA.

The ability to differentiate contamination from ancient DNA would be a major advancement for the study of ancient DNA. This is particularly relevant when working with human fossils, since contamination can come not only from the person extracting the DNA and sequencing it, but also from all of people who have handled the bone previously, such as archeologists. Therefore the ability to differentiate contaminating modern DNA from ancient DNA would have a huge impact on genetic studies of modern human remains, mainly by greatly advancing anthropological studies of thousands of human skeletons in museum collections. Forensic studies may also benefit from the results of this study, if forensic DNA is shown to have properties distinguishable from modern human contamination.

## *Results*

I gathered small samples from a total of 44 animal bones from museums in Germany and Holland from different time points across the last 2400 years (Appendix A). For my final analysis I also used data from Neandertals published in Briggs *et al.* 2007 and Briggs *et al.* 2009. I chose to use animal bones instead of human bones in order to curtail the problem of human contamination that may confuse the analysis.

I extracted ancient DNA in a clean room in order to avoid contamination since ancient DNA has such small amounts of DNA any contamination can easily overshadow the actual DNA. These tiny amounts of DNA also mean that ancient DNA cannot be visualized on a gel without amplification. Since I worked with many different animals, I included a positive extraction control, which could be amplified in a PCR after extraction and visualized on a gel. In addition I included a blank control in the extraction, which I

amplified in PCR along with the positive control and was treated as a sample for the entire process to be able to see after sequencing if any contamination occurred.

After extraction the DNA is converted into sequencing libraries. These libraries are created in order to be able to sequence each fragment and to have a reservoir from which more copies of the fragments can be created through PCR. This way the extract must only be used once to create the library and must not be used up for further sequencing. In order to create the Illumina libraries, I used the protocol described in Meyer *et al.* 2010, as this protocol was designed specifically with ancient DNA in mind.

Illumina library preparation consists of three steps. Many ancient DNA fragments have overhanging ends and these overhanging ends are blunt ended by filling in the 3' end of the shorter end and chewing off the 5' end of the longer end (Figure 3ii). This blunt ending process therefore causes the 3' end of fragments that had overhangs to be a copy of the sequence of the complimentary fragments' 5' end. For this reason if the 3' end shows a misincorporation of G to A, it is actually showing a C to T misincorporation of the complimentary strand's 5' end. Second an A and B adapter are ligated to the 5' and 3' ends of each fragment as they are needed for the Illumina sequencing process (Figure 3v). Last two oligonucleotides, called P5 and P7, are added to the 5' and 3' ends through a PCR in order for the fragments to stick to the flow cell during sequencing. The P7 tail has a unique sequence of six or seven bases called an index. Each sample receives its own index and the samples can therefore be pooled and then differentiated from each other after sequencing. This PCR is therefore called an indexing PCR. Indexing is also important since now the samples can be removed from the clean room for PCR amplifications and subsequent contamination will not be sequenced.

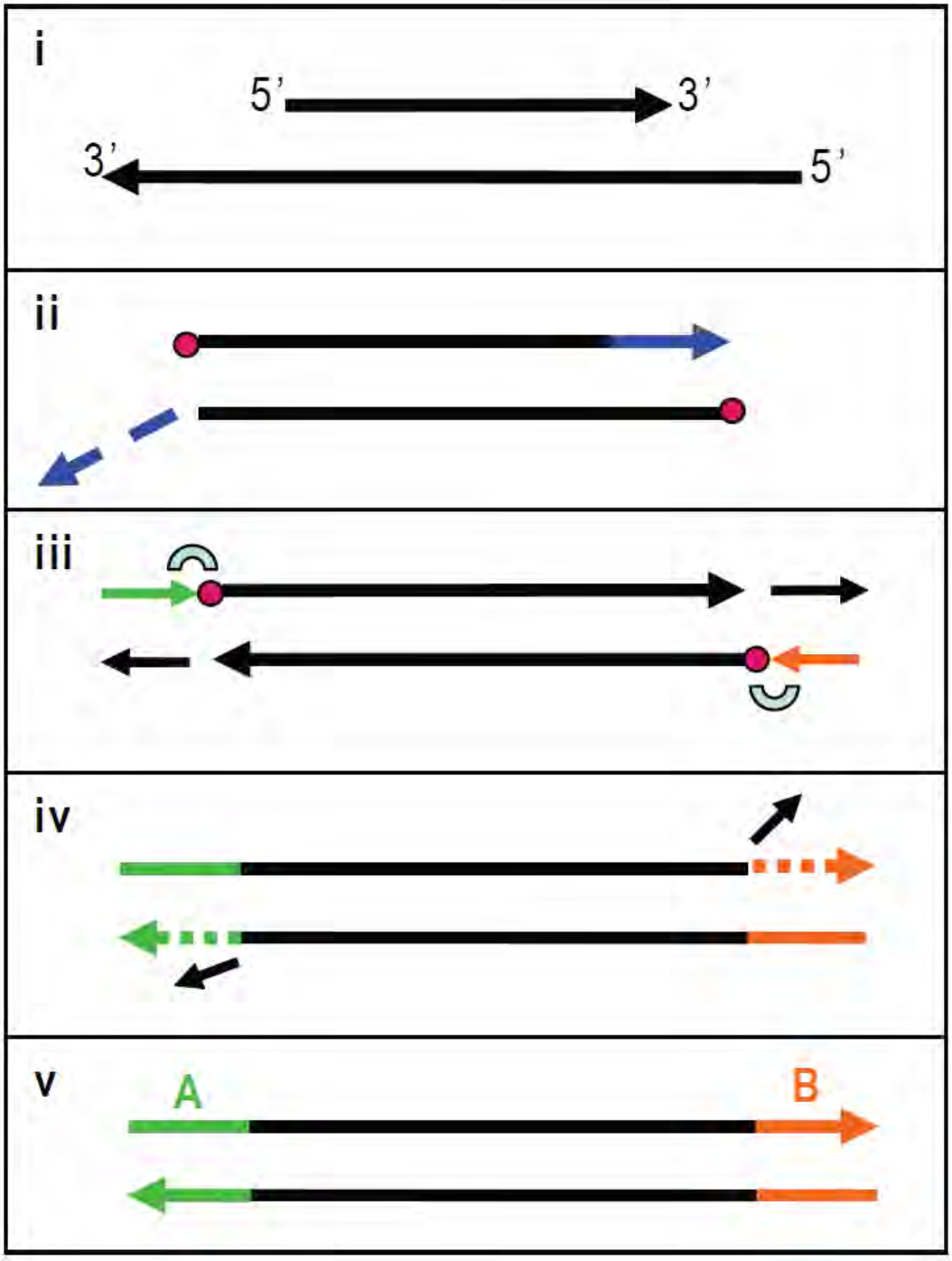


Figure 3. Illumina library preparation from Meyer *et al.*2010. Ancient fragments can have overhanging ends (i). The fragments are treated with *T4* DNA Polymerase, which has a 5' to 3' fill in activity and a 3' to 5' exonuclease activity and with *T4* polynucleotide kinase, which adds phosphates to the 5' ends (ii). A and B adapters are added and ligated to the 5' ends with *T4* ligase (iii). *Bst* polymerase is used to fill-in the adapters at the 3' ends (iv), giving a blunt ended fragment with A and B adapters to which the indexing adapters can be attached (v).

The indexing PCR that adds the P5 and P7 tails to the fragments is done with the Taq Gold ™ polymerase, since this polymerase can read over uracils. As previously discussed ancient DNA carries uracil residues as a result of deamination of cytosine residues. Not all enzymes can read uracils; one enzyme that is able to read over uracils is Taq polymerase. However Taq polymerase can only amplify samples to a maximum of $1*10^{12}$ molecules in total per 100ul reaction given 5 units polymerase. Since subsequent steps need higher copy numbers a second PCR amplification is therefore needed with a proof-reading enzyme called Phusion ™, which can amplify up to $1*10^{13}$ molecules total. Phusion ™ cannot read uracils and is therefore not suited for the initial indexing PCR. However, after the indexing PCR using Taq Gold ™ , the uracils have been converted to thymines, which can be read by the Phusion ™ enzyme. I checked each step with a quantitative PCR (qPCR): the library preparation after A and B adapter ligation but before indexing (to check the amount of molecules in the original sample before amplification), after the indexing PCR to check if indexing was successful and after the PCR with the Phusion ™ polymerase to check PCR efficiency using various primer pairs that anneal to the ligated adapters.

There are many methods to sequence DNA. One such method is shot gun sequencing, which is where the all of the extracted DNA is sequenced. This method is usually used to sequence entire genomes. However I have chosen to amplify only the mitochondrial genome (mt-genome), since there are more copies of the mt-genome than the nuclear genome, allowing me to gain more fragments for sequencing for each DNA extract. Since it would be a waste of sequencing capacity to sequence more than

necessary, I used a method, which involves capturing a part of the genome (in this case the mt-genome) and enriching it. Since I am working with more than one animal sample, I decided to use a bead-based method (Maricic 2010, submitted). This method relies on a biotinilated mt-DNA bait sequence of the modern equivalent of my sample, for example a modern cow, which is bound to magnetic streptavidine beads and when the sample is added, fishes out the mt-genome of the sample by hybridizing to complementary library molecules. The unbound library molecules are subsequently washed away. Thus the mtDNA fragments of my sample are enriched and sequencing is much more efficient and economical.

To make capturing of the samples more efficient, I pooled samples from the same species together before capturing mtDNA. Pooling is possible since each sample carries a unique index, allowing samples to be differentiated after sequencing. The pool was made so that each sample was in the same ratio to the other samples as it was before amplification, so that a sample that was more favorably amplified, and therefore had more copies of each unique molecule, was not sequenced more.

The Illumina platform is particularly well suited for working with ancient DNA samples that tend to be highly fragmented. On the Illumina platform the maximum length of molecules that can be sequenced is 190bp, which is considerably shorter than the traditional Sanger sequencing, but in the upper range of molecules that can be extracted from ancient organic material. The second big advantage of next generation sequencers compared to traditional Sanger sequencing is the up to$1*10^6$ times higher throughput per run simultaneously (Bentley *et al.* 2008).

In order to determine the length of the mtDNA fragments captured from my samples, I performed a paired-end sequencing run on the Illumina GAII platform, which means that each fragment was sequenced from both ends. To find the length of each fragment, I then used the alignment program BWA to align each fragment to its corresponding reference species mtDNA. As can be seen in Figure 4, the fragment length does not change over time from 100 years onwards and shows a sequence length observed before in up to 40,000-year-old fossil remains (Green *et al.* 2008, Briggs *et al.* 2009). However, the 47-year-old sample has a significantly greater fragment length distribution than the rest of the samples (Welch two sample t-test, $P<2.2e-16$). The p-value is unaffected if the possible outlier of the second 100-year-old sample (Sample 27) is removed. The data from the eight Neandertals included in subsequent analyses cannot be analyzed using the method I described above, since the Neandertals were sequenced in a single-end sequencing run, which means that they are only sequenced from one side so that the length of the molecules longer than the read length cannot be determined.

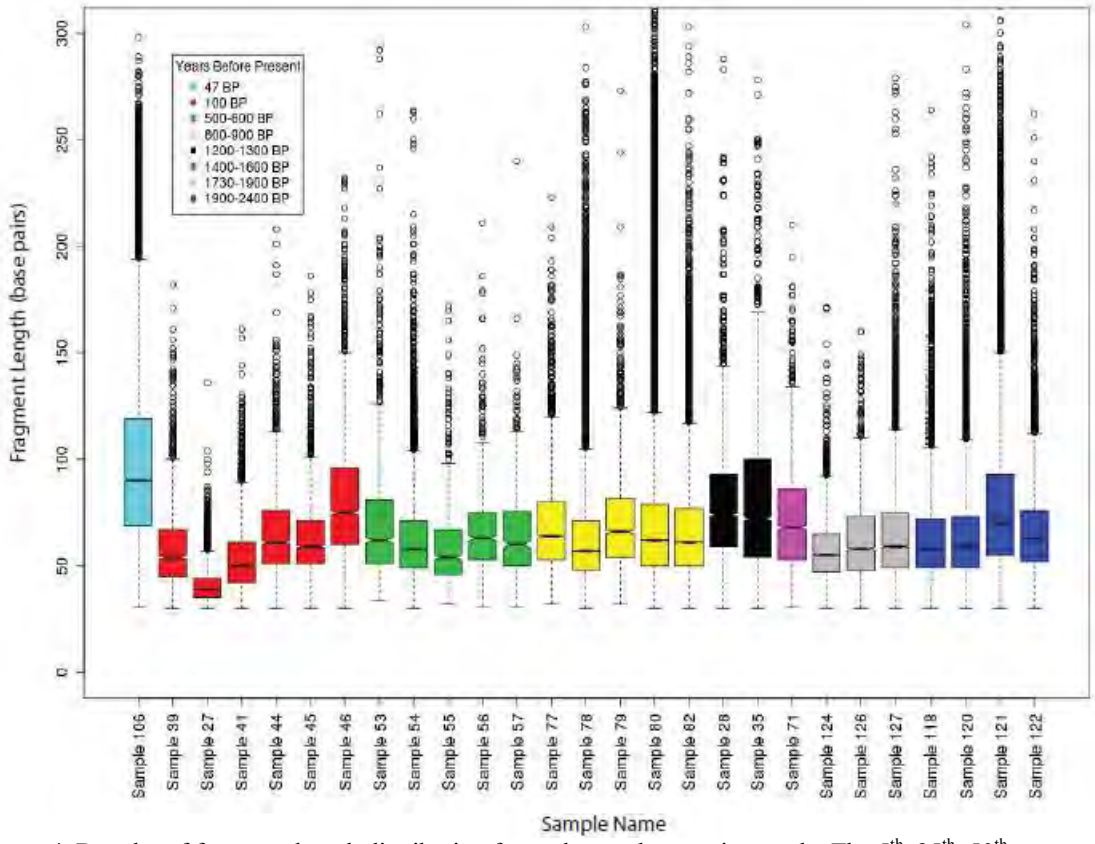


Figure 4. Boxplot of fragment length distribution for each sample on a time scale. The 5th, 25th, 50th (Median), 75th, and 95th percentiles are indicated.

Fragmentation patterns, caused by depurination events, were determined by comparing each fragment to its corresponding reference sequence. Two positions were then analyzed. First, I analyzed the reference position one position upstream of the first base of the 5' end of the fragments. This position showed an excess of purines in Briggs *et al.* 2007. I calculated the frequency of purines at this base for each sample (Figure 5). The second position is the reference position one base downstream from the last 3' base in the sample fragment (Figure 5). For this base I calculated the frequency of pyrimidines for each sample. Since the pyrimidine frequency of the 3' end actually represents the frequency of purines of the 5' end of the complimentary strands, I took the average of the two frequencies to get one purine frequency at the 5' of both DNA strands from all

fragments for each sample. I plotted the frequencies of purines from the samples that range in age from 47 to 2400 years as well as eight Neandertal samples that are around 40,000 years old. All samples are elevated compared to the normal purine frequency of 0.5 along the rest of the fragment (Figure 6b). A plot of each base frequency for ten bases upstream of the 5' end of the fragment and ten bases into the 5' end of the fragment of sample 53 (a 500 – 600 year old horse) shows a very similar pattern to the pattern observed previously (Briggs *et al.* 2007) (Figure 6a, Figure 2a). No trend is apparent in this plot as the purine frequency has an $R^2$ correlation close to zero when compared to time ($R^2$ =0.04995, P=0.1045) (Figure 6b).

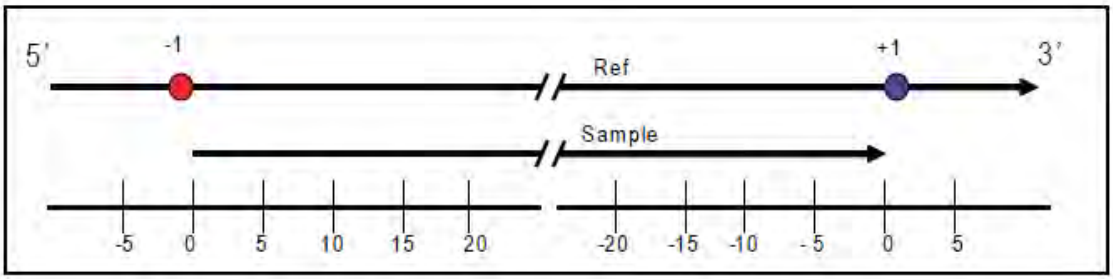


Figure 5. Description of the calculation of purine and pyrimidine frequency. Purine frequency (red) is calculated at the -1 position of the 5' fragment end; pyrimidine frequency (blue) is calculated at the +1 position of the 3' fragment end.

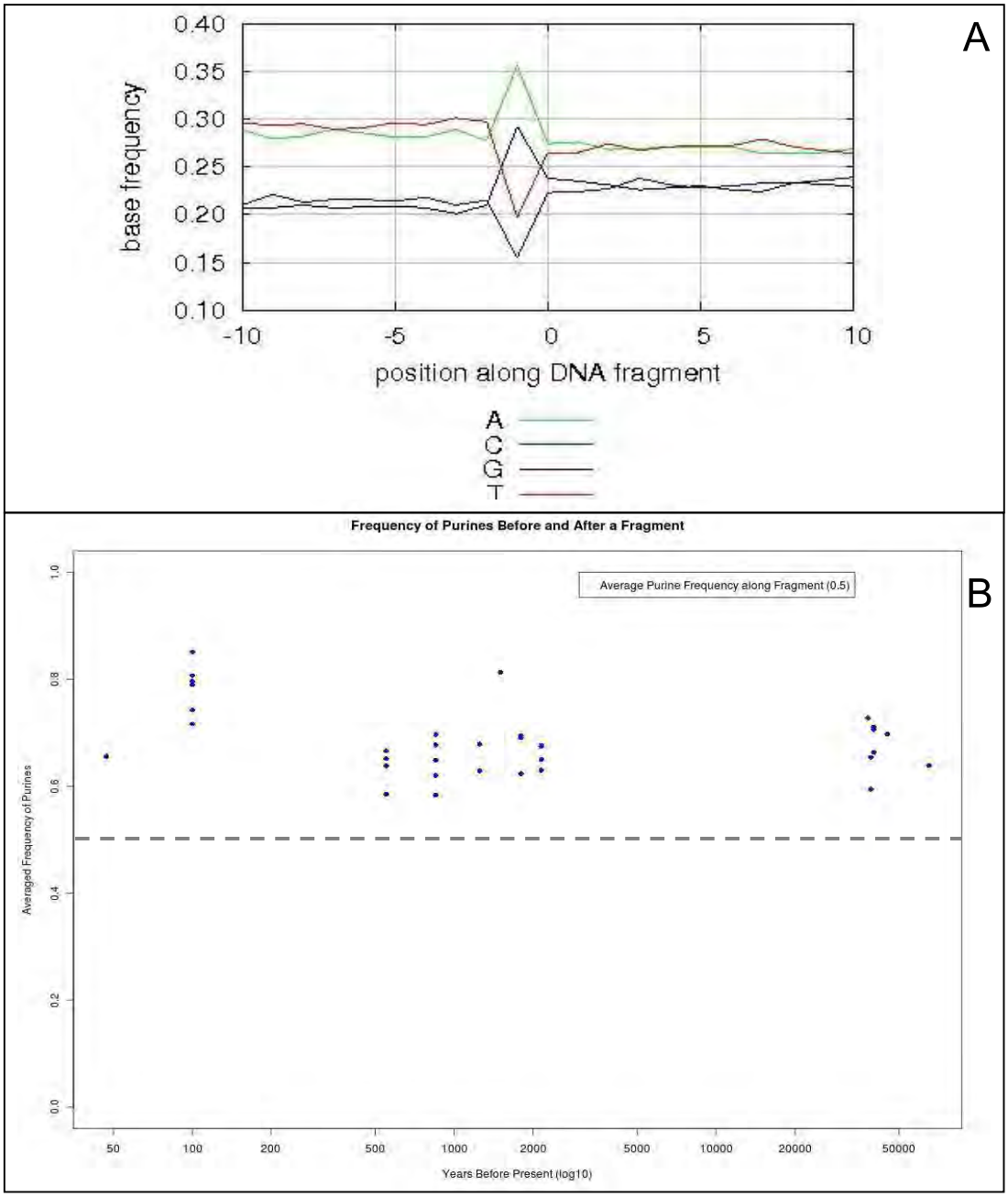


Figure 6. A: Base frequencies for ten bases of the reference DAN sequence 5' of the fragment and the first ten bases of the 5' end of the fragment for sample number 53. B: Plot of the average frequency of purines at the -1 position upstream of the 5' end and pyrimidines at the +1 position downstream of the 3' fragment end versus time for 43 samples from 47 years to 60,000 years before present. Pyrimidine frequency shows the frequency of purines in the 5' end of the complimentary strand. The normal purine and pyrimidine frequency for each position is 0.5.

In order to determine misincorporation patterns due to deamination, I calculated the frequency of deamination of cytosine at the first position of the 5'end and last position of the 3' end of each sample. I used the 5' deamination frequency for my subsequent analysis as the pattern observed at the 3' end showed the same degree of deamination (data not shown). As with the fragmentation data, I added the information from eight  Neandertals. When plotted against time, the misincorporation patterns at the first position on the 5' fragments ends show a significant positive correlation ($R^2$ = 0.5918, P = 4.05e-08) (Figure 7b). A graph of the misincorporation of C to T for the first 15 bases of the 5' end and G to A of the last 15 bases of the 3' end of the fragment for sample 53 shows a similar pattern to the pattern shown previously (Briggs *et al.* 2007) (Figure 7a, Figure 2b).

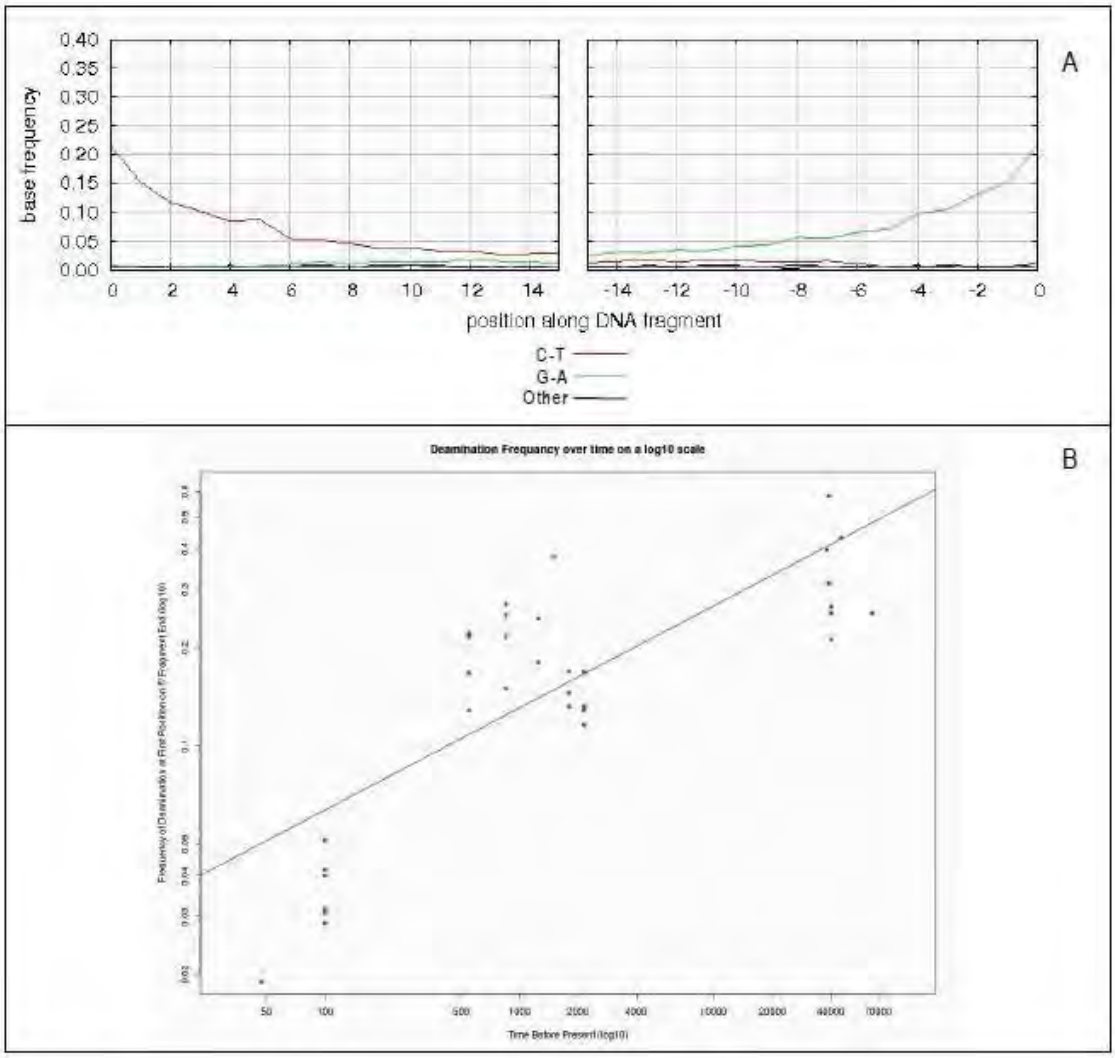


Figure 7. A: Misincorporation patterns of C to T for the first 15 bases of the 5' end and of G to A for the last 15 bases of the 3' end of the fragment of sample 53. B: Deamination rate at the first position on the 5' end of the fragments for 35 samples ranging in age from 47 to 60,000 years. A regression line ($R^2$ = 0.5918) is shown.

## *Discussion*

Patterns that differentiate ancient DNA from modern DNA can shed light on how DNA degrades as well as give scientists the possibility to separate contaminating modern DNA from endogenous ancient DNA. In 2009, Krause *et al.* highlighted three patterns that could be used in such a way: short fragment length, particular sequence contexts at strand breaks, hypothesized to result from depurination (Briggs *et al.* 2007) and high

levels of C to T misincorporations due to deamination, particularly at the beginning and end of a strand (Briggs *et al.* 2007), which were found to be present in 40,000-year-old Neandertals (Briggs *et al.* 2007) and early modern humans (Krause *et al.* 2009). My aim in this study was to study how predictable these patterns are on a temporal scale, gathering samples from 47 to 2400 years BP and including data on eight Neandertals ranging in age from 40,000 to 60,000 years old.

Fragmentation of ancient DNA can occur through many processes such as exposure of DNA to nucleases shortly after death, degradation by microorganisms and chemical processes such as depurination through hydrolysis (Paabo *et al.* 2004). My data suggests that DNA is fragmented already after 50 years and does not fragment any further over time after that (Figure 4). I did not include any fragments that were smaller than 30 base pairs since these fragments are too small to be of a high enough quality to be mapped to the reference sequence. Therefore it could be that older DNA extracts have smaller fragments and that I am only looking at the fragments that are longer than 30 base pairs. However, in such a case the mean length of all fragments would be right at the 30 base pair mark. Since the means of the older samples are higher than 30 base pairs and comparable to the younger samples, the fragment size does not decrease over time with in the first 2400 years. The fragment sizes seem to be larger in the sample dating only 47 years BP, suggesting that fragmentation is occurring 50 years after death and then stops being prevalent when DNA reaches a mean of 80 base pairs. Nevertheless the 47 year old sample is only one sample and more data for this age group is needed to make a statistical claim about this observation. Therefore short fragment lengths could only be used to differentiate ancient from modern contaminating DNA if the contamination is extremely

fresh since contaminating DNA can come from older DNA, such as from archaeologists or museum staff that handled bones up to or even beyond 100 years ago.

It seems that DNA is fragmented very quickly within the first few decades after death and then does not fragment significantly more as years pass. There are a few possible explanations for this. One explanation is that DNA with a mean length around 80 base pairs may have properties that make it more stable and less susceptible to outside processes. Another explanation may be that most DNA is fragmented to a point where it is no longer measurable but a small amount of DNA may adhere to the bones in such a way making it less open to enzymatic and chemical processes. The reason for this pattern however is still unclear.

Depurination contributes to fragmentation of DNA (1,6) (Figure 1b), and could be used as another signal of ancient DNA. The normal frequency of purines and pyrimidines are equal in a fragment of DNA, each being 0.5 (Figure 2b, Figure 6b) (Briggs *et al.* 2007). However, in 2007, Briggs *et al.* discovered that directly before a strand break, the frequency of purines is elevated and the frequency of pyrimidines is suppressed (Briggs *et al.* 2007). The authors hypothesized that the explanation is depurination (Briggs *et al.* 2007), since depurination causes the formation of an abasic site, which is susceptible to strand breakage 3' to the abasic site (Figure 1b).

Depurination-driven strand breaks are due to a chemical process (hydrolysis). I hypothesized that a trend in depurination-driven strand breaks would be visible over time, with the amount of depurination-driven strand breaks increasing. My data show that this elevation in purine frequency prior to strand breaks is already present in a 47 year old bone and that the frequency does not change significantly over the next 60,000 years,

which indicates that the same processes are responsible for strand breakages of all ages (Figure 6). The depurination-driven strand breaks detected in the Neandertal samples could be new depurination-driven strand breaks but given the observation that fragment length is not reduced over time over the first 2400 years, it may be more reasonable that these depurination-driven strand breaks occurred early after death and the signal has since stayed. This would explain why the signal has not increased over time. As with the fragmentation of DNA, a possible explanation may be that short DNA fragments are not as susceptible to hydrolysis or that the DNA that is found in ancient DNA was protected from hydrolysis by its position in the bone. The constant purine frequency over time correlates with the fragment length pattern since depurination-driven strand breaks can lead to fragmentation, although it is unclear how much depurination-driven strand breaks contributes to fragmentation.

However this method of detecting depurination-driven strand breaks is indirect and it does not detect depurination that occurs within a fragment where the abasic site has not been cleaved. In addition, since there is only one sample that represents an age of ~50 years BP, more sampling would be needed from this time period and earlier to confirm that purine frequency is indeed already elevated within 50 years and at what point this elevation occurs. It is relevant to note that all of the samples that are up to 100 years old are samples that were shot in Africa and then stored in a museum and never spent any time buried in the ground. Museum storage may have had an effect on the depurination-driven strand breakage rate, possibly due to humidity or due to chemicals used to treat the skeletons. Samples of this age that spent time buried in soil would therefore contribute to this study as well as samples that are younger than 50 years.

When cytosine is exposed to water, hydrolysis can occur and cause the cytosine to turn into uracil (Figure 1a). When a DNA polymerase reads a fragment that has been deaminated, it adds a thymine to the complimentary strand being formed at the position opposite of the uracil, since thymine base pairs with uracil. This causes a misincorporation at that base. In 2007, Briggs *et al.* discovered that C to T misincorporations are highly concentrated at the ends of ancient fragments from 40,000 year old Neandertals (Figure 2a). I analyzed misincorporation patterns at the first base of DNA fragments for samples ranging in age from 47 to 60,000 years old. Unlike depurination-driven strand breaks and fragment length patterns in ancient DNA, misincorporation patterns due to deamination do show a temporal trend (Figure 7). It seems that in the case of deamination, the hydrolytic process that causes cytosine to become uracil (Figure 1a) is a continuous process over time unlike the hydrolytic processes that causes strand breakage due to depurination. Briggs *et al.* 2007 hypothesized that the deamination at the fragment ends in ancient DNA must occur after the strand has broken, leaving single stranded ends open to hydrolytic attack (Briggs *et al.* 2007). This hypothesis is based on the knowledge that double stranded DNA, as found in the center of an ancient fragment, is more protected from hydrolysis causing deamination than single stranded DNA, such as the DNA found in the single stranded overhangs of ancient DNA fragments (Briggs *et al.* 2007). Therefore hydrolysis seems to act continuously over time on the single stranded overhangs of fragmented DNA. Another explanation could have been that the deamination in younger samples was limited by the lack of fragmented DNA, since the more DNA is fragmented the more deamination could occur. However since DNA fragment size is already as small after 100

years as after 60,000 years (Figure 4), this limiting factor cannot be an explanation for the lack of deamination in younger samples.

The trend in misincorporation patterns due to deamination could allow scientists to distinguish ancient DNA from contaminating DNA, possibly even contaminating DNA that is a few decades old and. However caution is needed since due to the variance seen in samples of a similar age but different locations, it may be possible that the location and burial conditions of an organism can have an influence on the deamination rate.

There are many open questions that have been uncovered by this study. One such question is how fast strand breaks and reduction of fragment size occur in samples that are younger than 50 years old, and if more sampling of samples in the 50 to 100 years BP age range is necessary. It would be of interest to know when depurination-driven strand breaks and short fragment lengths occur, whether this is within a few months of an organism's death or after a few years. Another question would be to assess misincoporation rates from samples from the 100 to 500 year period. In addition it may be good to gather samples from 500 to 40,000 year old animals to affirm when misincorporation rates reach a plateau. Due to the variance in misincorporation patterns in samples from the same age but from different locations, it would be good to test how much location has an effect on misincorporation rates. The samples for this study are not sufficient, since there are too many confounding variables. Samples of the same age and different locations would be needed to answer this question of how much location has an effect on misincorporation.

From my data, I conclude that of the three patterns I analyzed, misincorporation due to deamination would be the most usable to differentiate between newer, possibly

contaminating DNA, and ancient DNA. It seems that a cutoff at a deamination frequency of cytosine of 0.1 at the first position on the 5' or 3' fragment end would be a good measure of whether DNA is older than 500 years or whether it might potentially represent modern contamination. However this would only be an option if the sample comes from one biological source (Krause *et al.* 2009), since a sample that is 50% from an ancient source and 50% from a non-ancient source could show a misincorporation due to deamination frequency of around 0.15, but this could be an average of the contaminating DNA with a low frequency and the ancient DNA with high frequency. This cutoff of a frequency of 0.1 could change though if further sampling between 100 and 400 years shows different results.

Neither depurination-driven strand breaks nor fragment length patterns seem to be useful for separating up to 100 year old contaminating DNA from ancient samples. However, the discovery that fragmentation occurs so quickly after death does open up many interesting questions about how DNA degrades. It is useful to the ancient DNA community to be able to use a measure such as misincorporation patterns to separate contaminating DNA even if the sample being studied is younger than 40,000 years old.

## *Materials and Methods*

### Bone procurement

Before sampling ancient bones in the museums, I set up a clean working area on a table to emulate a clean-room environment. The table was disinfected with bleach and

covered in aluminum foil. I wore a lab coat, face mask, hair net and two pairs of gloves. After each sampling, I changed the outer gloves. All materials had been UV radiated and packed into sterile plastic bags in a clean room at the Max Planck Institute. All samples and bones for later sampling were individually stored in sterile falcon tubes or plastic bags.

**DNA extraction methods**

All ancient DNA samples were handled in a clean room facility specifically built to avoid contamination. The facility is UV radiated when not in use and under negative air pressure. I changed shoes and into a lab coat, face mask, hair net and plastic shield for my face before entering the clean room. No PCR products are allowed into the facility and movement from a PCR room to the clean room is prohibited. I extracted bones under a hood in a special extraction room within the facility. I changed gloves and cleaned the bench with bleach between each sample.

I drilled each bone with a dentist drill to produce up to 400 mg of bone powder. The powder was then extracted using an EDTA and proteinase K based extraction buffer, a high slat binding buffer and a silica suspension as described in Rohland 2007. I used a column extraction and washing method as described in Rohland 2009. DNA extracts were kept in siliconized Eppendorf tubes to avoid DNA sticking to the walls of the tubes and were kept in a freezer at -20 C.

Modern cow blood was obtained from the Leipzig Veterinary School from a healthy cow and DNA was extracted using the Qiagen DNeasy Blood and Tissue Kit

(Qiagen, Germany). Modern horse DNA was already available for use. All extracted modern DNA concentrations were determined on a nanodrop and loaded into an agarose gel to check quality.

**Positive and negative controls**

When extracting ancient DNA, I performed a positive control each time I had made new reagents. This positive control consisted of a cave bear bone sample with previous high DNA extraction success rates. After extraction, I PCR amplified the extract using the H221 and L164 primers (Appendix A). I also included a blank or negative control for each extraction event, which was then treated as an individual sample throughout the indexing and PCR amplification process and was sequenced in the end. Each qPCR also had another blank control in addition to the extraction negative.

**Library Preparation and Purification**

The library preparation was performed after extraction of ancient DNA. I used the protocol detailed in Meyer 2010 (submitted). Library preparation consists of two steps. First two adapters are attached to each end (an A (5') and B (3') adapter) of the fragment. Then a so-called indexing step takes place by amplifying the DNA to attach two tails, a P5 tail on the 5' end and a P7 tail on the 3' end. The P7 tail has a 7-basepair barcode at the 5' end called the index.

**PCRs**

All PCRs of the prepared libraries were performed in a non-clean room lab. Reagents and amounts are as seen in Table 1. Indexing PCRs were performed using the AmpliTaq Gold system by Applied Biosystems and the Phusion PCRs were performed using the Phusion High-Fidelity DNA Polymerase 2x Master Mix system by Finnzymes. Quantitative PCR (qPCR) was performed on a MX3005P qPCR machine using the 2X DyNAmo HS SYBR Green Mastermix by Fermentas. For the qPCR of non-indexed samples, I used a primer pair that binds to the Illumina A and B adaptors, Sol_PCR_p5 and Sol_PCR_p7. For qPCRs performed after indexing, I used a primer pair that binds to the index add-ons, Sol_bridge_p5 and Sol_bridge_p7. A qPCR Illumina standard was used for each qPCR that measured $1*10^8$ down to $1*10^3$ copies per uL. All long range PCRs (LR-PCR) were performed using the 720U Expand Long Template PCR System by Roche using primer pairs as listed in Appendix B.

Table 1. Explanation of PCRs. Reagents, amounts and times are given. Minutes are denoted with ‘ and seconds with “.

| PCR Name | Reagent | Amount per reaction (uL) | Cycles and times |
|---|---|---|---|
| **Indexing PCR** | Water | 45 | 94 °C – 12’ |
| | Thermapol Buffer (10x) | 6 | 94°C -20” |
| | dNTPs (25mM) | 1 | 60°C – 30” } 10 cycles |
| | Forward Primer (10uM) | 4 | 72°C – 40” |
| | Index Primer (10 uM) | 4 | 72°C – 5’ |
| | Taq Gold (5U/uL) | 1 | |
| | Template | 38 | |
| **Phusion PCR** | 2X Phusion Master Mix | 50 | 98 °C – 30” |
| | Sol_bridge_P5 (10 uM) | 4 | 98°C -20” |
| | Sol_bridge_P7 (10uM) | 4 | 60°C – 30” } 10 cycles |
| | Water | 37 | 72°C – 40” |
| | Template | 5 | 72°C – 5’ |
| **Long Range - PCR** | Water | | 92 °C – 2’ |
| | Buffer 2 (10X) * = (Expand Long Range Pack) | 5 | 92°C - 10” |
| | dNTP Mix (10uM)* | 2.5 | 65°C – 15” } 10 cycles |
| | BSA (10mg/mL) | 2 | 68°C – 9’ |
| | Enzyme* | 0.7 | 92°C - 10” |
| | DMSO* | 1.5 | 65°C – 15” } 30 cycles |
| | Primer 1 (10uM) | 1.5 | 68°C – 9’ |
| | Primer 2 (10uM) | 1.5 | +20sec/cycle |
| | Water | 34.3 | 68°C – 7’ |
| | Template | 1 | |
| **Positive Control** | Water | 5.8 | 94 °C – 12’ |
| | Buffer (10x) * = AmpliTaq Gold Kit | 2 | 94°C -20” |
| | $MgCl_2$ (25mM) * | 3.2 | 55°C – 30” 50 cycles |
| | AmpliTaq Gold * | 0.2 | 72°C – 40” |
| | dNTPs (25mM) | 0.2 | 72°C – 4’ |
| | L164 (10uM) | 1 | |

| | | | |
|---|---|---|---|
| | H221 (10uM) | 1 | |
| | Template | 5 | |
| **qPCR** | DyNAmo HS SYBR Green Master Mix (2x) | 12.5 | 95 °C – 12’<br>95°C -20”<br>60°C – 30”<br>72°C – 40”<br>} 40 cycles |
| | Water | 10.5 | |
| | Primer 1 (10uM) | 1 | |
| | Primer 2(10uM) | 1 | |
| | Template | 1 | |

**Gel Analysis**

I used a 2% agarose gel for all gels measuring DNA fragment lengths smaller than 1 kilobase. For these gels I used a O’GeneRuler 50bp ladder by Fermentas. Gels were performed to check all qPCR results for primer dimmer formation, the positive control for ancient DNA extraction and if called for in a protocol used. I also used a 0.8% agarose gel with a O’GeneRuler 1 kb plus ladder by Fermentas to check my long range PCR results. All 2% gels were run at 180 Volts for 30 minutes in either 100 mL or 300 mL gel boxes. The 0.8% gels were run at 120 volts for 30 minutes in 100 mL gel boxes.

**Mitochondrial Enrichment**

Mitochondrial enrichment of my ancient DNA was performed using the targeted region capture on the dynal streptavidine beads as detailed in Maricic *et al.* 2010 (submitted) on pooled DNA. Sequences for the blocking oligos and biotinylated adapters are shown in Appendix B. Elution from the beads occurred through the NaOH melting option and MinElute purification and elution with Buffer EB with 0.05% tween.

**Sequencing and Analysis**

Sequencing was performed on the Illumina $GA_{II}$ platform as described by Bentley *et al.* 2008. Each pool of DNA samples of the same animal species was again pooled with other samples with differing indexes and a total of three lanes were run to sequence the DNA. A paired end run was done for 76 cycles per lane to allow the length of each fragment to be determined. After sequencing the sequences were base-called and merged using the Ibis base caller (Kircher *et al.* 2009).

I used merged reads to align the sequences using MIA (Briggs *et al.* 2009) and the closest possible reference. If the reference sequence showed a higher pair-wise difference from the sample sequence than 200 positions, then a new sequence was produced from the sample sequence where only bases that had at least a five-fold coverage were included. This new sequence was then used as a new reference and MIA was run again.

I looked at the following positions in order to analyze data: whether a cytosine existed at the first position of the 5' end or the last position of the 3' of each fragment, whether these Cs were deaminated, the frequency of purines in the position in the reference sequence before the first position of the 5' end and frequency of pyrimidines at the position after the last base of the 3' end of each fragment. The frequency of deamination was calculated by dividing the number of deaminated Cs at a position by the number of total Cs at this position.

The alignment program BWA (Li *et al.* 2009) was used to align merged and non-merged reads from sequencing. The length of each fragment was determined by either

determining the length of the merged read, or using the reference as a basis for the length of the fragment from two unmerged reads.

Data was analyzed using the statistical platform R (R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.). Graphs were produced by using the boxplot function for fragment length and the plot function was used for the fragmentation due to depurination and misincorporation due to deamination by time graphs. A regression line with $R^2$ values and p-values was determined using the linear model options.

## *Acknowledgments*

I would like to thank my mentor for my master's project, Johannes Krause, for helping me every step of the way through my project. Also thank you to Martin Kircher, Udo Stenzel and Michael Dannemann for help with bioinformatics, Matthias Meyer, Knut Finstermeier and Qiaomei Fu for help with ideas and lab work, and Tomislav Maricic for help on my analysis. A big thank you to Svante Pääbo for supervising my project and for accepting me to work on this project at the Max Planck Institute.

## *References*

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS,

Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456**: 53-59.

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prufer K, Meyer M, Krause K, Ronan M, Lachmann M, Pääbo S. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. PNAS **104**: 14616-14621.

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Schmitz R, Doronichev V, Golovanova L, Rasilla M, Fortea J, Rosas A, Pääbo S. 2009. Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. Science **325**: 318-320.

Eglinton G, Logan GA.1991. Molecular preservation. Philos Trans R Soc Lond B Biol Sci **333**: 315-327.

Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, Prüfer K, Siebauer M, Burbano HA, Ronan M, Rothberg JM, Egholm M, Rudan P, Brajković D, Kućan Z, Gusić I, Wikström M, Laakkonen L, Kelso J, Slatkin M, Pääbo S. 2008. A Complete Neandertal Mitochondrial Genome Sequence Determined by Hihg-Throuput Sequencing. Cell **134**: 416-426.

Kircher M, Stenzel U, Kelso J. 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biology **10**: R83.

Krause J, Briggs AW, Kircher M, Maricic T, Zwyns N, Derevianko A, Pääbo S. A complete mtDNA genome of an early modern human from Kostenki, Russia. 2009. Current Biology **20**: 231-236.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, **25**: 1754-1760.

Lindahl T. 1993. Instability and decay of the primary structure of DNA. Nature **362**: 709-715.

Pääbo S. 1989. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. PNAS **86**: 1939–1943.

Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M. 2004. Genetic Analysis from Ancient DNA. Annu Rev Genet **38**: 645-679.

## *Appendix*

Appendix A. List of all bones used in this experiment.

| **Sample** | **Original Museum** | **Location of Excavation** | **Approximate date** | **Grams used for extraction** |
|---|---|---|---|---|
| Horse 118 | Earth Integrated Archeology, Amersfoort, Holland (EIA) | Sittard, Holland | 400 BC – 100 AD | 297 |
| Horse 119 | EIA | Sittard, Holland | 400 BC – 100 AD | 326 |
| Horse 120 | EIA | Sittard, Holland | 400 BC – 100 AD | 314 |
| Horse 121 | EIA | Sittard, Holland | 400 BC – 100 AD | 279 |
| Horse 122 | EIA | Sittard, Holland | 400 BC – 100 AD | 301 |
| Horse 123 | EIA | Sittard, Holland | 100AD – 270AD | 218 |
| Horse 124 | EIA | Sittard, Holland | 100AD – 270AD | 269 |
| Horse 125 | EIA | Sittard, Holland | 100AD – 270AD | 322 |
| Horse 126 | EIA | Sittard, Holland | 100AD – 270AD | 277 |
| Horse 127 | EIA | Sittard, Holland | 100AD – 270AD | 287 |
| Horse 128 | EIA | Sittard, Holland | 1300-1400 AD | 277 |
| Horse 129 | EIA | Sittard, Holland | 1300-1400 AD | 350 |
| Horse 130 | EIA | Sittard, Holland | 1300-1400 AD | 295 |

| | | | | |
|---|---|---|---|---|
| Horse 131 | EIA | Sittard, Holland | 1300-1400 AD | 316 |
| Horse 52 | EIA | Sittard, Holland | 1300-1400 AD | 309 |
| Horse 53 | EIA | Sittard, Holland | 1400-1500 AD | 192 |
| Horse 54 | EIA | Sittard, Holland | 1400-1500 AD | 278 |
| Horse 55 | EIA | Sittard, Holland | 1400-1500 AD | 288 |
| Horse 56 | EIA | Sittard, Holland | 1400-1500 AD | 295 |
| Horse 57 | EIA | Sittard, Holland | 1400-1500 AD | 317 |
| Cow 77 | Museum fuer Haustierkunde Halle, Germany (MHK) | Halle Markt, Germany | 12th Century | 236 |
| Cow 78 | MHK | Halle Markt, Germany | 12th Century | 362 |
| Cow 79 | MHK | Halle Markt, Germany | 12th Century | 246 |
| Cow 80 | MHK | Halle Markt, Germany | 12th Century | 324 |
| Cow 81 | MHK | Halle Markt, Germany | 12th Century | 413 |
| Cow 82 | MHK | Halle Markt, Germany | 12th Century | 253 |
| Horse 1 | MHK | Deersheim, Germany | 8th Century | 93 |
| Horse 4 | MHK | Deersheim, Germany | 8th Century | 55 |
| Horse 28 | MHK | Deersheim, Germany | 8th Century | 103 |
| Horse 35 | MHK | Deersheim, Germany | 8th Century | 74 |
| Horse 41 | MHK | Deersheim, Germany | 8th Century | 79 |
| Horse 44 | MHK | Loebnitz, Germany | 5/6th Century | 63 |
| Horse 45 | MHK | Loebnitz, Germany | 5/6th Century | 138 |
| Horse 62 | MHK | Loebnitz, Germany | 5/6th Century | 67 |
| Horse 64 | MHK | Loebnitz, | 5/6th Century | 88 |

| | | | | |
|---|---|---|---|---|
| | | Germany | | |
| Horse 71 | MHK | Loebnitz, Germany | 5/6th Century | 80 |
| Horse 86 | MHK | Loebnitz, Germany | 5/6th Century | 110 |
| Cercopithecus ascanius 106 | Smithsonian, Washington DC | Kenya | 1963 | 4.8 |
| Cercopithecus ascanius 27 | Berlin Naturkundemuseum (NKM) | Congo | 1908 | > 4 |
| Cercopithecus petaurista 39 | NKM | Togo | 1903 | > 4 |
| Cercopithecus mitis 41 | NKM | Tanzania | 1894 | > 4 |
| Cercopithecus cephus 44 | NKM | Cameroon | 1913 | > 4 |
| Cercopithecus neglectus 45 | NKM | Ethiopia | 1927 | > 4 |
| Madrillus leucophaeus 46 | NKM | Cameroon | 1896 | > 4 |

Appendix B.

| **Primer/Oligonucleotide Name** | **Sequence 5' to 3'** | **Use** |
|---|---|---|
| Sol_PCR_Short_P5 | ACACTCTTTCCCTACACGAC | Binds to Solexa A adaptor |
| Sol_PCR_Short_P7 | GTGACTGGAGTTCAGACGTGTG | Binds to Solexa B adaptor |
| Sol_iPCR_Forward | AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT | Index PCR |
| Sol_iPCR_Reverse | CAAGCAGAAGACGGCATACGAGATnnnnnnGTGACTGGAGTTCAGACGTGT | Index PCR |
| Sol_Bridge_P5 | AATGATACGGCGACCACCGA | Binds to index tail P5 |
| Sol_Bridge_P7 | CAAGCAGAAGACGGCA | Binds to index tail P7 |

| | | |
|---|---|---|
| | TACGA | |
| Horse_for_1 | CCTAATCAACTGGCTTC AATCTACTTCTCC | LR-PCR primer of mt -- side 1 |
| Horse_rev_1 | AGGATTGAGGCCAGTT GGCCGAT | LR-PCR primer of mt – side 1 |
| Horse_for_2 | TACCTCCATGCCAACG GAGCATC | LR-PCR primer of mt – side 2 |
| Horse_rev_2 | TCCAGCTCAAGCGCCG AATAGG | LR-PCR primer of mt – side 2 |
| Cow_for_1 | CCTCGCCGTGCGATTA ACAGCCAACATC | LR-PCR primer of mt -- side 1 |
| Cow_rev_1 | GGTGGCTGGCACGAGA TTTACCAACCCT | LR-PCR primer of mt -- side 1 |
| Cow_for_2 | GGATGCTTGGACTCAG CTATGGCCGTCA | LR-PCR primer of mt – side 2 |
| Cow_rev_2 | TGGATCCAATGATGAC GTGGAGGCCGTG | LR-PCR primer of mt – side 2 |
| Biotin adaptor (ds) | Bio-TCAAGGACATCCG GTT CCT GTAGGC | Biotinilate bait to attach to streptavidine beads |
| H221 | CGGACTAAGTGAAATA CATGCT | PCR for positive extraction control |
| L164 | GCATATAAGCATGTAC ATATTATGC | PCR for positive extraction control |