Competitive mapping allows to identify and exclude human DNA contamination in ancient faunal genomic datasets

Tatiana R. Feuerborn[1,2,3,4,*], Elle Palkopoulou[3], Tom van der Valk[3,4], Johanna von Seth[3,4,5], Arielle Munters[6], Patrícia Pečnerová[7], Marianne Dehasque[3,4,5], Irene Ureña[8], Erik Ersmark[3,4], Vendela Kempe Lagerholm[2,4], Maja Krzewinska[2,4], Ricardo Rodríguez-Varela[2,4], Anders Götherström[2,4], Love Dalén[3,4], David Díez-del-Molino[3,4,5,*]

[1] Globe Institute, University of Copenhagen, Copenhagen, Denmark.

[2] Archaeological Research Laboratory, Department of Archaeology and Classical Studies, Stockholm University, Stockholm, Sweden.

[3] Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden.

[4] Centre for Palaeogenetics, Stockholm, Sweden.

[5] Department of Zoology, Stockholm University, Stockholm, Sweden.

[6] Department of Organismal Biology, Human Evolution, Uppsala University, Sweden.

[7] Department of Biology, University of Copenhagen, Denmark.

[8] CIBIO/InBio - Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Vairão, Portugal.

* Correspondence: Tatiana R. Feuerborn (tatianafeuerborn@palaeome.org), David Díez-del-Molino (diez.molino@gmail.com)

Abstract

After over a decade of developments in field collection, laboratory methods and advances in high-throughput sequencing, contamination remains a key issue in ancient DNA research. Currently, human and microbial contaminant DNA still impose challenges on cost-effective sequencing and accurate interpretation of ancient DNA data. Here we investigate whether human contaminating DNA can be found in ancient faunal sequencing datasets. We identify variable levels of human contamination, which persists even after the sequence reads have been mapped to the faunal reference genomes. This contamination has the potential to affect a range of downstream analyses. We therefore introduce a fast and simple method, based on competitive mapping, which

allows identifying and removing human contamination from ancient faunal DNA datasets with limited losses of true ancient data. This method could represent an important tool for the ancient DNA field.

Keywords

Ancient DNA; DNA contamination removal; palaeogenomics; competitive mapping

Introduction

Right after the death of an organism, microbial communities colonize the decomposing tissues and together with enzymes from the organism they start degrading the DNA molecules (Lindahl, 1993; Renaud, Schubert, *et al.*, 2019). DNA degradation is dependent on time and environmental variables such as temperature but also humidity and acidity (Kistler *et al.*, 2017). Even though the specific model for DNA decay is still debated and it is likely multifactorial (Kistler *et al.*, 2017), the consequence is that ancient remains typically contain very few quantities of endogenous DNA and these sequences are characterized by short fragment sizes (Dabney *et al.*, 2013).

A second major challenge of ancient DNA research is contamination from exogenous sources (Malmström *et al.*, 2005; Der Sarkissian *et al.*, 2015). Environmental DNA molecules in the soil matrix the ancient sample was taken from can easily overwhelm the small amounts of endogenous DNA of the ancient sample. This is also true for DNA from people who collected and handled the samples in the field and/or museum collections (Green *et al.*, 2006; Der Sarkissian *et al.*, 2014). While the use of Polymerase Chain Reaction (PCR) technology allowed ancient DNA research to overcome these low concentration problems, the sensitivity of the PCR has made it very difficult to avoid introducing modern contaminant sequences among the authentic ancient DNA (Willerslev and Cooper, 2005).

In the last decade, together with more refined DNA extraction and laboratory methods tailored to efficiently retrieve very short and scarce DNA sequences (Dabney *et al.*, 2013; Gamba *et al.*, 2016), it has become possible to obtain massive amounts of sequences from ancient material using high-throughput sequencing technologies. These technologies have allowed the recovery of hundreds of ancient human (reviewed in Slatkin and Racimo, 2016) and other high quality ancient faunal genomes such as those from horses (Orlando *et al.*, 2013), wooly mammoths (Palkopoulou *et al.*, 2015), and bears (Barlow *et al.*, 2020). However, the challenges from exogenous contamination remain and have sparked a search for computational methods to identify and monitor contaminant DNA sequences in ancient sequencing datasets.

Aside from the short fragment size, the other most notable characteristic of ancient DNA is post-mortem damage. After death, the repairing mechanisms of DNA damage such as hydrolysis and oxidation stop functioning, and this damage accumulates in predictable patterns (Renaud, Schubert, *et al.*, 2019). The most common ancient DNA damage is deamination of cytosines to uracils in the overhangs of fragmented DNA molecules (Gilbert *et al.*, 2003; Stiller *et al.*, 2006; Briggs *et al.*, 2007). This results in an excess of C to T substitutions in the 5' end (and G to A in the 3' end) of ancient DNA sequences. Since this feature is very common in sequences derived from ancient DNA sources and absent in younger samples, it has been widely used as a key criteria to authenticate ancient DNA experiments (Sawyer *et al.*, 2012; Dabney *et al.*, 2013).

In modern-day ancient DNA studies, exogenous sequences are differentiated from real ancient sequences from the source organism by mapping all sequences to a reference genome and keeping only those that result in alignments with less than a defined number of differences (Prüfer *et al.*, 2010; Kircher, 2012). This approach to circumvent environmental contamination has gained general acceptance, and currently exogenous contaminants are at most considered problematic due to their consumption of sequencing capacity. However, the probability of spurious alignments from exogenous sequences occurring by chance increases with decreasing sequence length (Smith, Waterman and Burks, 1985). In order to avoid these, thresholds for minimum fragment length, that still allow for enough specificity of the alignments, are used (Green *et al.*, 2010; Meyer *et al.*, 2016; de Filippo, Meyer and Prüfer, 2018).

Modern human contamination is especially problematic for human palaeogenomic studies since ancient, anatomically modern humans typically fall within the variation of modern humans (Lazaridis *et al.*, 2014; Allentoft *et al.*, 2015). This has led to the development of a plethora of methods aimed at computationally quantifying and monitoring exogenous contamination in ancient human DNA datasets (Meyer *et al.*, 2012; Fu *et al.*, 2014; Rasmussen *et al.*, 2015; Racimo, Renaud and Slatkin, 2016). However, the number of methods that allow for the effective exclusion of this type of contamination remains limited. For example, Skoglund *et al.*, (2014) used the differential empirical distributions of post-mortem damage (PMD) scores, based on both base quality scores and their level of polymorphism with respect to the reference genome, to differentiate DNA sequences from ancient and modern samples. The PMD scores in a contaminated ancient sample could then be used to successfully identify and separate the sequences that are most likely to have originated from an ancient template molecule from the contaminant ones. Even though this method can allow for the enrichment of the proportion of ancient sequences several-fold in respect to the contaminant sequences, the amount of data lost in the process is very large (45%-90% depending on the age of the ancient sample, Skoglund *et al.*, 2014).

Here we investigate the presence of exogenous sequences in ancient sequencing files to evaluate the pervasiveness of human contamination in ancient faunal DNA studies. We use competitive

mapping to identify the levels of contamination in ancient faunal sequencing files and characterize the exogenous sequences by using summary statistics to compare them to those of authentic ancient DNA. We then present this strategy as a simple and fast method that enables the conservative removal of human contamination from ancient faunal datasets with a limited loss of true ancient DNA sequences.

Materials and Methods

Samples

We analyzed genomic data from 70 ancient and historical mammalian specimens, 50 dogs (*Canis lupus familiaris*) and 20 woolly mammoths (*Mammuthus primigenius*, Supplementary Table 1). The materials derived from dogs originate from a variety of contexts (ethnographic collections and archaeological excavations) and materials (teeth and bones) which have been stored in museum collections for up to 125 years after collection/excavation. The twenty mammoth samples were all collected in Wrangel Island in several expeditions along the last 30 years.

Laboratory procedures

For all samples, the outer layers of bones, teeth and tusk were removed using an electric powered drill (Dremel, USA) in order to minimize external contamination. Approximately 50 mg of bone powder was recovered from inside the bone, tooth or tusk using an electric drill operated at low speed. We then extracted DNA from all samples using the silica-based protocol described in (Ersmark *et al.*, 2015). Thirty-four of the dog samples were additionally subjected to a pre-digestion step, incubated with EDTA, urea, and proteinase K for one hour at 55°C, to further reduce the amount of contamination within the extract by removing the superficial DNA. We did not treat any of the extracts with USER enzyme in order to enable assessment of post-mortem damage rates following DNA sequencing.

We constructed Illumina genomic libraries for sequencing from the DNA extracts using established ancient DNA protocols (Meyer and Kircher, 2010; Carøe *et al.*, 2017). All libraries were amplified using indexes unique for each sample and were subsequently pooled and sequenced on a total of 4 lanes on the Illumina HiSeq2500 platform at the National Genomics Infrastructure (Science for Life Laboratory, Stockholm), using paired-end 2x150bp settings.

Data analyses

We trimmed sequencing adapters and merged paired-end reads using *SeqPrep v.1.1* (github.com/jstjohn/SeqPrep) with default settings (excluding reads shorter than 30bp) and a slight modification of the source code to calculate the base qualities in the overlapping region (Palkopoulou *et al.*, 2015). We then mapped the merged reads to three separate reference

genomes: the African savannah elephant genome (LoxAfr4, Broad Institute), the dog genome (CanFam3.1, Lindblad-Toh *et al.*, 2005), and the human reference genome (Hg19). The mapping was performed using *BWA aln v0.7.8* (Li and Durbin, 2009) using settings adapted for ancient DNA as in Pečnerová *et al.*, (2017).

After mapping, we removed PCR duplicates using a custom script which takes into account both starting and end coordinates of the reads to be identified as duplicates (Palkopoulou *et al.*, 2018), then excluded reads with mapping quality < 30 and estimated the number of correctly unique mapping reads using *samtools v1.8* (Li *et al.*, 2009). We consider true endogenous sequences those correctly mapping to the target species (i.e dog reference for ancient dog samples and elephant reference for mammoth samples) and exogenous contaminant sequences those correctly mapping to the non-target reference (i.e elephant and human references for ancient dog samples and dog and human references for mammoth samples). The read length (RL) and PMD scores for each sequence in each file were extracted from the filtered BAM files metadata and estimated with PMDtools (Skoglund *et al.*, 2014), respectively, and were used to characterize true ancient and contaminant sequences.

Competitive mapping

In order to investigate whether the target BAM files contain any human contaminant sequences we remapped the aligned reads to a concatenated reference composed by the reference genome of the target species, dog or elephant, and the human reference genome. All mappings and filtering were performed as described above. This competitive mapping allowed us to differentiate between three kinds of reads contained in the target species BAM files. First, reads which align correctly to the target reference genome and not to the human reference genome. These sequences represent the endogenous alignments that originate from the sample and not from human or bacteria contamination. Second, reads which align better to the human reference genome than to the target species reference genome. These sequences represent the fraction of human contamination in the faunal BAM files. And third, reads that align equally well to both the target reference and the human reference genomes. These sequences could have three origins, 1) true endogenous sequences from regions of the genome highly conserved or identical to the human genome, 2) human contaminant sequences from regions of the genome highly conserved or identical to the target genome, or 3) microbial contaminant sequences that would align equally well to any mammalian genome by random chance. In any case, because these sequences would map to both target and human reference genomes at the same time their mapping quality will be lower, and would thus be discarded when applying the mapping quality filters.

For each sample, we extracted the reads aligned to the target part of the concatenated reference, representing the true ancient sequences, as well as the human part, representing the amount of human contamination in the original target BAM file. Because an unknown fraction of the reads

discarded due to the mapping quality filters should also be human contaminant, the fraction of reads in the human part of the concatenated reference represents only a lower bound for the amount of contamination in the original faunal BAM file. We characterized all true endogenous sequences, as well as the human contaminant sequences using RL and PMD scores, as described above.

Finally, in order to estimate the amount of data lost using competitive mapping we identified conserved regions between the elephant and human genomes as well as the dog and human genomes. We first used a custom script to split the human reference genome into non-overlapping 35bp long sequences. We then mapped the obtained short sequences to the other two reference genomes, dog and elephant, using *BWA aln* with settings optimized for mapping short reads (Schubert *et al.*, 2012; Li, 2013). For each mapping, we filtered out reads with mapping quality below 30 and identified all genomic regions with at least one read mapped. The resulting BED files were used together with *samtools flagstat* to estimate the number of reads mapping to conserved regions before and after competitive mapping.

Results and Discussion

<u>Contamination in raw sequencing files</u>

Overall, we found low levels of sequences mapped to foreign reference genomes (average 0.25% for non-target and 0.86% human) in the raw sequencing files (Fig. 1A). Even though this was variable among samples, most of the sequencing files (>95%) contained less than 0.071% of sequences mapped to human and 0.054% the non-target species. In fact, the proportion of reads mapping to the non-target species and human were highly correlated (Fig. 2A), suggesting that they had a common origin. Given that human DNA is the most common mammal contamination source in ancient DNA studies (Cooper and Poinar, 2000; Hofreiter, Serre, *et al.*, 2001; Malmström *et al.*, 2005; Korlević *et al.*, 2015), it is then likely that a variable amount of contaminant human reads map to the two reference genomes used here, elephant and dog. In fact, there were notable exceptions to the amount of faunal sequences mapping to human, for example one sample displayed a higher proportion of sequences mapped to the human (38.9%) than to the target species (12.3%). This suggested that there could be high levels of human DNA contamination in some particular sequencing files.

Next, we sought to characterize the sequences mapping to the target reference genome as well as the ones mapping to the non-target and human references using two characteristics of ancient DNA: short fragment size (Rogaev *et al.*, 2006; Allentoft *et al.*, 2012; Kistler *et al.*, 2017) measured as median read length (mRL) and deamination patterns (Hofreiter, Jaenicke, *et al.*, 2001; Briggs *et al.*, 2010) measured as post-mortem damage scores (PMD, Skoglund *et al.*, 2014). For each

sample, we define the PMD ratio (PMD^R) as the fraction of sequences that display a PMD score > 5, and it should represent the 'ancientness' of each sample. Therefore, a higher PMD^R indicates the sample contains more sequences with larger PMD scores, thus it is more 'ancient'. We find some differences in mRL and PMD^R between sequences mapping to target and to non-target and human (Fig. S1), in line with the latter being mostly composed by contaminant sequences and the former mostly true ancient reads. Interestingly, we find almost no differences between the sequences mapping to the non-target species and human references, reinforcing the idea that these two files are composed of sequences with a common origin.

Human contamination in faunal BAM files

Given that we detected contaminant human sequences in all our ancient fauna sequencing files, we next used competitive mapping to explore whether these contaminant reads can be also found in the BAM file of the target species that would be used for downstream genomic analyses. We found that the BAM files from almost all samples contained sequencing reads that preferentially mapped to the human part of the concatenated reference genome, but the proportion was generally low (average 0.03%), the highest value being 1.3% (Fig. 1B, Supplementary Table 1). Interestingly, the proportion of reads mapped to the human reference from the raw data and in the target BAM files are not correlated (Fig. 2B). This indicates that the amount of human contamination in raw sequencing files is not a good predictor for the amount of human contamination that is retained in the target BAM files.

We then characterized both the true ancient sequences and the contaminant ones using mRL and PMD^R, as above. For both mammoth and ancient dog samples we found a clear distinction in PMD^R of the sequences mapping to the target species and the ones mapped to human, with higher PMD^R for the target species, representing true ancient sequences, and lower for the human sequences (Fig. 3C, 3F). However, we found that the contaminant human reads also displayed a lower mRL than the sequences aligned to the target species (Fig. 3B, 3E). This was contrary to the expectation of modern human contaminant sequences being longer than true ancient sequences, but can be explained by the fact that shorter contaminant sequences align easier to evolutionary conserved regions of the target species reference genome than longer sequences (Lee and Schatz, 2012; de Filippo, Meyer and Prüfer, 2018).

Considering species by species, the mammoth samples displayed a clearer distinction in PMD^R than the dog samples when comparing the reads mapped to the target and to the human parts of the concatenated reference (Fig. 3A, 3D). This may be related to both the age of the samples post-mortem and the age since collection. PMD scores are roughly proportional to the sample's age (Skoglund *et al.*, 2014), and while the mammoth samples are thousands of years old, they have been housed in collections for less than 30 years. The dog samples on the other hand were only a maximum of 1,000 years old (Supplementary Table 1), but were housed in museum collections

since their excavation or collection for up to 125 years. Since the conditions in museum collections are usually far from ideal for DNA preservation (Burrell, Disotell and Bergey, 2015; Díez-Del-Molino *et al.*, 2018), this extended period of storage therefore could have had an impact on the preservation of both endogenous dog and contaminant human DNA sequences in the ancient dog samples comparison to the mammoth samples.

Excluding contaminant reads from faunal BAM files

Despite being limited, the detected levels of exogenous human sequences in ancient faunal BAM files represent only the lower bound of contamination. Importantly, the presence of contaminant human sequences in ancient faunal BAM files, as those detected here, can be challenging for any downstream analyses that are based on evolutionary conserved parts of the genome, such as coding regions, since the contaminant sequences are concentrated in these regions. Additionally, other downstream analyses based on genome-wide scans such as estimations of heterozygosity, estimation of inbreeding levels using runs-of-homozygosity, or analyses focused on the presence of rare variants (Schiffels *et al.*, 2016) can be highly affected by the emergence of false variants caused by human contamination (Llamas *et al.*, 2017; Renaud, Hanghøj, *et al.*, 2019).

We therefore propose that the method applied here, using competitive mapping of the raw data to a concatenated reference genome composed by the reference genome of the target species and the human genome, represents a fast and simple approach to effectively exclude contaminating human DNA from ancient faunal BAM files. An additional advantage of this approach is that a large part of contamination from short microbial reads, common in ancient datasets (de Filippo, Meyer and Prüfer, 2018), should also be excluded with this method as many of these short reads would align equally well to both target and human parts of the concatenated reference and thus can be filtered out using the mapping quality filters.

One relevant downside of our proposed method could be the loss of data. True ancient sequences from the target species that belong to conserved regions of the genome and are identical between the target species and human, would align equally well to both parts of the concatenated reference, and thus be lost when using the mapping quality filters. However, our results indicate that the amount of data lost this way is limited in a genome-wide context (range 0.6 - 4.3%, with an average of 1.33%), and concentrated in conserved regions of the genome (range 2.7 - 17.8%, average 4.53%, Figure 4, Supplementary Table 1). Unfortunately, since these lost reads are discarded because they map equally well to both references, we do not have a practical way to estimate what fraction of those sequences are true target sequences and how many are of human or microbial origin. This slight loss of data can be potentially detrimental for samples that have medium to low coverage, and especially for studies focusing on conserved regions of the genome.

Conclusions

We show that variable levels of contaminant human sequences exist in ancient faunal datasets. To some extent, this human contamination persists even after sequence reads have been mapped to faunal reference genomes, and is then characterized by short fragment lengths that are concentrated in evolutionary conserved regions of the genome. This results in human contaminant sequences being included in ancient faunal alignment files and thus have the potential to affect a range of downstream analyses. To address this, we here present a fast and simple strategy: competitive mapping of raw sequencing data to a concatenated reference composed of the target species genome and a human genome, where only the sequences aligned to the target part of the concatenated reference genome are kept for downstream analyses. This approach leads to a small loss of data concentrated in conserved regions of the genome, but allows for the effective removal of the putative human contaminant sequences.

Contamination is a key issue in ancient DNA studies. Preventive measures both during field collection and in the laboratory therefore remain a critical aspect of ancient DNA research (Korlević *et al.*, 2015; Llamas *et al.*, 2017). There is a growing array of computational methods that allow to confidently identify contamination levels (reviewed in Renaud, Schubert, *et al.*, 2019), but few that allow to efficiently separate authentic ancient sequences from contaminating DNA (Skoglund *et al.*, 2014; de Filippo, Meyer and Prüfer, 2018). Thus, the method we propose here represents an important addition to the selection of tools aimed at computationally reducing the effects of human contamination in ancient faunal DNA research.

Acknowledgements

The authors would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure, and UPPMAX (project numbers: b2014312 and SNIC2020/5-3) for providing assistance in massive parallel sequencing and computational infrastructure. Genetic analyses were funded through a grant from the Swedish Research Council (VR grant 2012-3869) awarded to L.D.. J.v.S. and LD. acknowledge support from FORMAS (project 2015-676), T.R.F. acknowledges support from the EU-funded ITN project ArchSci2020 (grant no. 676154) and for the Qimmeq Project funding from the Velux Foundations, the Aage og Johanne Louis-Hansens Fond and the Wellcome Trust (grant no. 210119/Z/18/Z). D.D.dM. was supported through a Carl Tryggers scholarship (grant CTS17:109).

References

1. Allentoft, M. E. *et al.* (2012) ‘The half-life of DNA in bone: measuring decay kinetics in 158

dated fossils', *Proceedings of the Royal Society B: Biological Sciences*, pp. 4724–4733. doi: 10.1098/rspb.2012.1745.

2. Allentoft, M. E. *et al.* (2015) 'Population genomics of Bronze Age Eurasia', *Nature*, 522(7555), pp. 167–172.

3. Barlow, A. *et al.* (2020) 'Middle Pleistocene Cave Bear Genome Calibrates the Evolutionary History of Palaearctic Bears'. doi: 10.2139/ssrn.3523359.

4. Briggs, A. W. *et al.* (2007) 'Patterns of damage in genomic DNA sequences from a Neandertal', *Proceedings of the National Academy of Sciences of the United States of America*, 104(37), pp. 14616–14621.

5. Briggs, A. W. *et al.* (2010) 'Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA', *Nucleic acids research*, 38(6), p. e87.

6. Burrell, A. S., Disotell, T. R. and Bergey, C. M. (2015) 'The use of museum specimens with high-throughput DNA sequencers', *Journal of human evolution*, 79, pp. 35–44.

7. Carøe, C. *et al.* (2017) 'Single-tube library preparation for degraded DNA', *Methods in Ecology and Evolution*.

8. Cooper, A. and Poinar, H. N. (2000) 'Ancient DNA: do it right or not at all', *Science*. American Association for the Advancement of Science, p. 1139.

9. Dabney, J. *et al.* (2013) 'Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments', *Proceedings of the National Academy of Sciences of the United States of America*, 110(39), pp. 15758–15763.

10. Der Sarkissian, C. *et al.* (2014) 'Shotgun microbial profiling of fossil remains', *Molecular ecology*, 23(7), pp. 1780–1798.

11. Der Sarkissian, C. *et al.* (2015) 'Ancient genomics', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1660), p. 20130387.

12. Díez-Del-Molino, D. *et al.* (2018) 'Quantifying Temporal Genomic Erosion in Endangered Species', *Trends in ecology & evolution*, 33(3), pp. 176–185.

13. Ersmark, E. *et al.* (2015) 'Population demography and genetic diversity in the Pleistocene cave lion', *Open Quaternary*, 1(4), pp. 1–14.

14. de Filippo, C., Meyer, M. and Prüfer, K. (2018) 'Quantifying and reducing spurious alignments for the analysis of ultra-short ancient DNA sequences', *BMC biology*, 16(1), p. 121.

15. Fu, Q. *et al.* (2014) 'Genome sequence of a 45,000-year-old modern human from western Siberia', *Nature*, 514(7523), pp. 445–449.

16. Gamba, C. *et al.* (2016) 'Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing', *Molecular ecology resources*, 16(2), pp. 459–469.

17. Gilbert, M. T. P. *et al.* (2003) 'Distribution patterns of postmortem damage in human mitochondrial DNA', *American journal of human genetics*, 72(1), pp. 32–47.

18. Green, R. E. *et al.* (2006) 'Analysis of one million base pairs of Neanderthal DNA', *Nature*, 444(7117), pp. 330–336.

19. Green, R. E. *et al.* (2010) 'A Draft Sequence of the Neandertal Genome', *Science*, 328(5979), pp. 710–722.

20. Hofreiter, M., Serre, D., *et al.* (2001) 'Ancient DNA', *Nature reviews. Genetics*. Nature Publishing Group, 2(5), pp. 353–359.

21. Hofreiter, M., Jaenicke, V., *et al.* (2001) ‘DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA’, *Nucleic acids research*, 29(23), pp. 4793–4799.

22. Kircher, M. (2012) ‘Analysis of high-throughput ancient DNA sequencing data’, *Methods in molecular biology* , 840, pp. 197–228.

23. Kistler, L. *et al.* (2017) ‘A new model for ancient DNA decay based on paleogenomic meta-analysis’, *Nucleic acids research*, 45(11), pp. 6310–6320.

24. Korlević, P. *et al.* (2015) ‘Reducing microbial and human contamination in DNA extractions from ancient bones and teeth’, *BioTechniques*, 59(2), pp. 87–93.

25. Lazaridis, I. *et al.* (2014) ‘Ancient human genomes suggest three ancestral populations for present-day Europeans’, *Nature*, 513(7518), pp. 409–413.

26. Lee, H. and Schatz, M. C. (2012) ‘Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score’, *Bioinformatics* , 28(16), pp. 2097–2105.

27. Li, H. *et al.* (2009) ‘The Sequence Alignment/Map format and SAMtools’, *Bioinformatics* , 25(16), pp. 2078–2079.

28. Li, H. (2013) ‘Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM’, *arXiv [q-bio.GN]* 1303.3997.

29. Li, H. and Durbin, R. (2009) ‘Fast and accurate short read alignment with Burrows-Wheeler transform’, *Bioinformatics* , 25(14), pp. 1754–1760.

30. Lindahl, T. (1993) ‘Instability and decay of the primary structure of DNA’, *Nature*, 362(6422), pp. 709–715.

31. Lindblad-Toh, K. *et al.* (2005) ‘Genome sequence, comparative analysis and haplotype structure of the domestic dog’, *Nature*, 438(7069), pp. 803–819.

32. Llamas, B. *et al.* (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*. Routledge, 3(1), pp. 1–14.

33. Malmström, H. *et al.* (2005) ‘Extensive human DNA contamination in extracts from ancient dog bones and teeth’, *Molecular biology and evolution*, 22(10), pp. 2040–2047.

34. Meyer, M. *et al.* (2012) ‘A high-coverage genome sequence from an archaic Denisovan individual’, *Science*, 338(6104), pp. 222–226.

35. Meyer, M. *et al.* (2016) ‘Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins’, *Nature*. Nature Publishing Group, pp. 1–15.

36. Meyer, M. and Kircher, M. (2010) ‘Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing’, *Cold Spring Harbor Protocols*, p. db.prot5448–pdb.prot5448.

37. Orlando, L. *et al.* (2013) ‘Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse’, *Nature*, 499, pp. 74–78.

38. Palkopoulou, E. *et al.* (2015) ‘Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth’, *Current biology: CB*, 25(10), pp. 1395–1400.

39. Palkopoulou, E. *et al.* (2018) ‘A comprehensive genomic history of extinct and living elephants’, *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), pp. E2566–E2574.

40. Pečnerová, P. *et al.* (2017) ‘Genome-Based Sexing Provides Clues about Behavior and Social Structure in the Woolly Mammoth’, *Current biology: CB*, 27(22), pp. 3505–3510.e3.

41. Prüfer, K. *et al.* (2010) ‘Computational challenges in the analysis of ancient DNA’, *Genome biology*, 11(5), p. R47.

42. Racimo, F., Renaud, G. and Slatkin, M. (2016) ‘Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans’, *PLOS Genetics*, p. e1005972.

43. Rasmussen, M. *et al.* (2015) ‘The ancestry and affiliations of Kennewick Man’, *Nature*, 523(7561), pp. 455–458.

44. Renaud, G., Schubert, M., *et al.* (2019) ‘Authentication and Assessment of Contamination in Ancient DNA’, *Methods in molecular biology* , 1963, pp. 163–194.

45. Renaud, G., Hanghøj, K., *et al.* (2019) ‘Joint Estimates of Heterozygosity and Runs of Homozygosity for Modern and Ancient Samples’, *Genetics*, 212(3), pp. 587–614.

46. Rogaev, E. I. *et al.* (2006) ‘Complete mitochondrial genome and phylogeny of Pleistocene mammothMammuthus primigenius’, *PLoS biology 4(3): e73*.

47. Sawyer, S. *et al.* (2012) ‘Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA’, *PloS one*, 7(3), p. e34131.

48. Schiffels, S. *et al.* (2016) ‘Iron Age and Anglo-Saxon genomes from East England reveal British migration history’, *Nature communications*, 7, p. 10408.

49. Schubert, M. *et al.* (2012) ‘Improving ancient DNA read mapping against modern reference genomes’, *BMC genomics*, 13, p. 178.

50. Skoglund, P. *et al.* (2014) ‘Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal’, *Proceedings of the National Academy of Sciences of the United States of America*, 111(6), pp. 2229–2234.

51. Slatkin, M. and Racimo, F. (2016) ‘Ancient DNA and human history’, *Proceedings of the National Academy of Sciences*, 2016(23), pp. 1–8.

52. Smith, T. F., Waterman, M. S. and Burks, C. (1985) ‘The statistical distribution of nucleic acid similarities’, *Nucleic acids research*, 13(2), pp. 645–656.

53. Stiller, M. *et al.* (2006) ‘Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA’, *Proceedings of the National Academy of Sciences of the United States of America*, 103(37), pp. 13578–13584.

54. Willerslev, E. and Cooper, A. (2005) ‘Ancient DNA’, *Proceedings. Biological sciences / The Royal Society*, 272(1558), pp. 3–16.

Figures

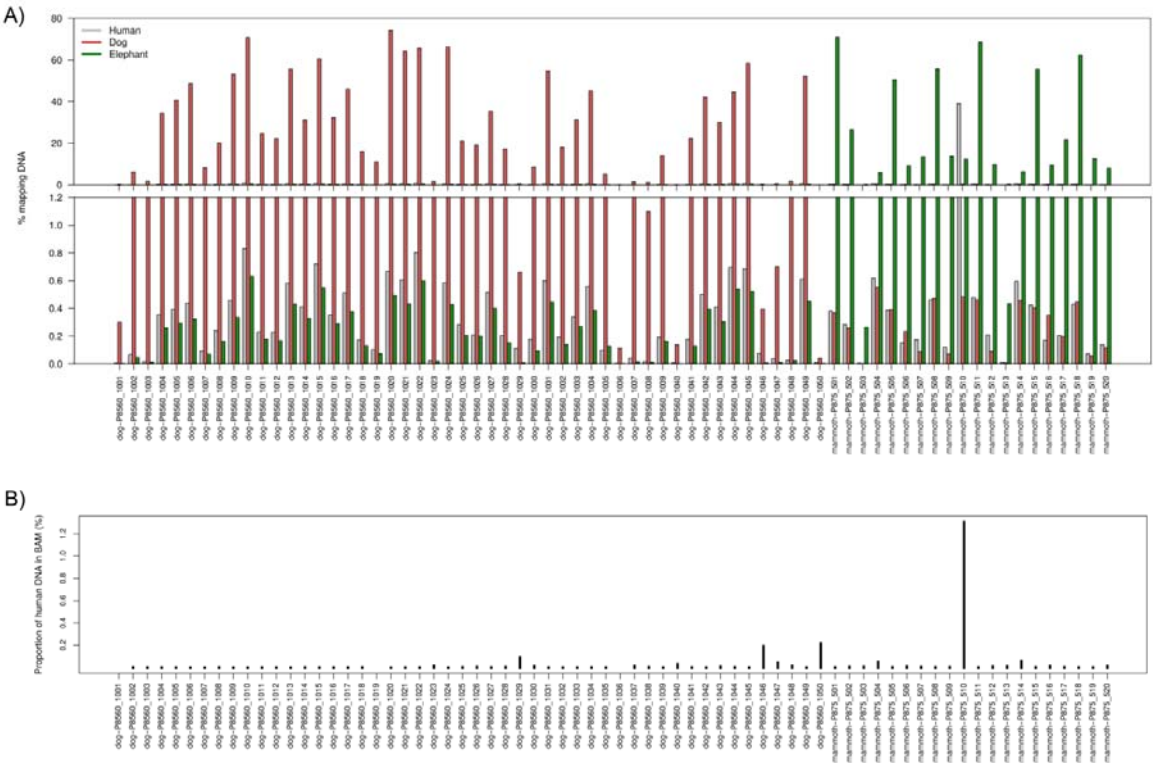

Figure 1. Mapping statistics for target, non-target and human references. A) Upper panel, percentage of reads from each sample mapping to each of the three reference genomes. Lower panel, same as before but zoomed to percentages below 1.2%. B) Proportion of reads from the faunal BAM file that mapped correctly to the human part of the concatenated reference genome.

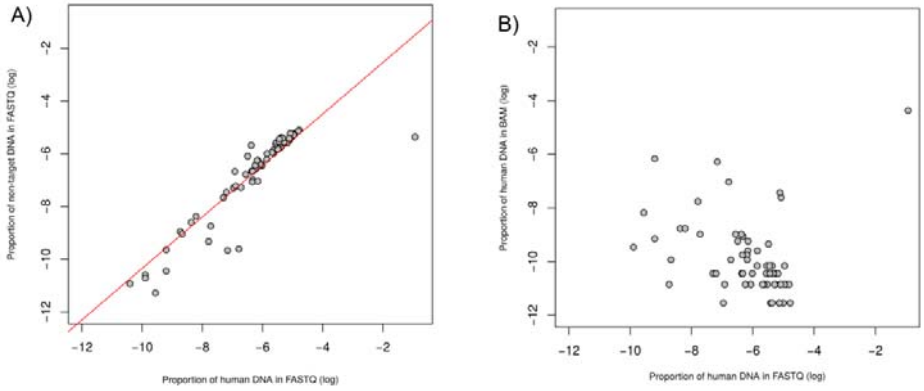

Figure 2. Proportions of sequences mapping to human, target and non-target reference from the FASTQ and BAM files. A) Correlation between the proportion of reads mapping to human and to the non-target species in the raw FASTQ sequencing files (r^2 = 0.81, p-value: < 2.2e-16). B) Correlation between the proportion of reads mapping to human in the raw FASTQ sequencing files and the proportion of reads mapping to human from the faunal BAM file (r^2 = 0.01, p-value: 0.2).

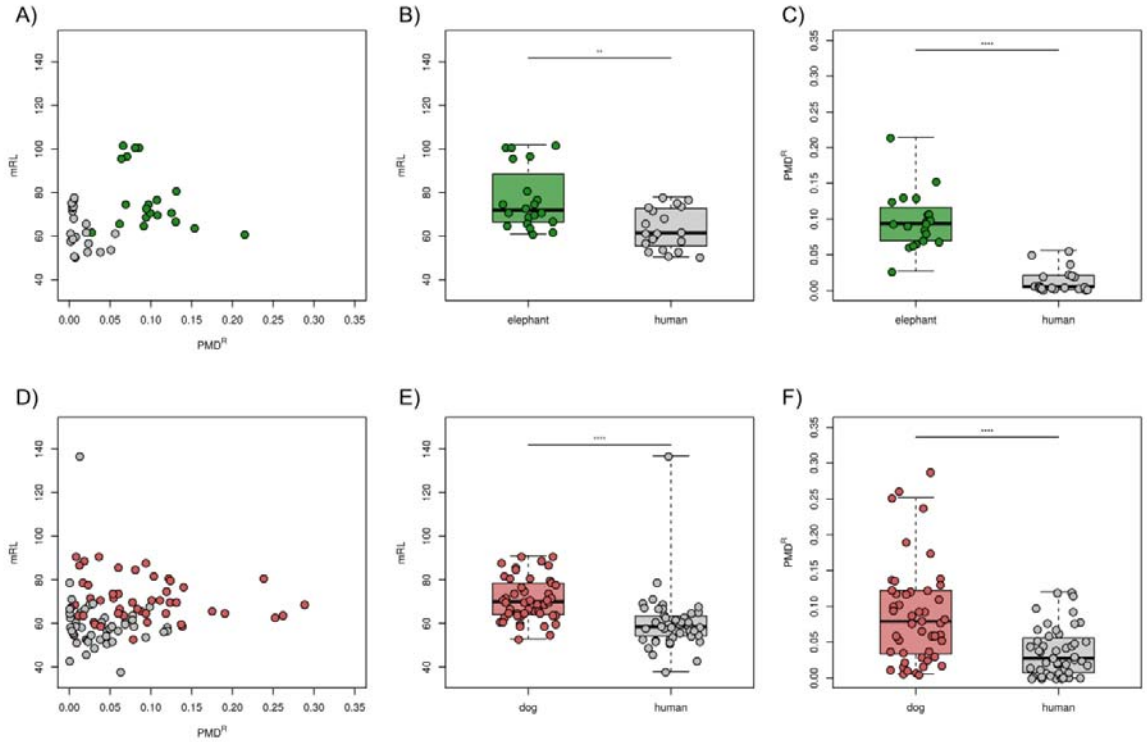

Figure 3. Characterization of endogenous and human contaminant reads in faunal BAM files. A) Comparisons of PMD^R and mRL for all mammoth samples. B) mRL for mammoth sequences mapping to the elephant or the human parts of the concatenated reference. C) PMD^R for mammoth sequences mapping to the elephant or the human parts of the concatenated reference. D) Comparisons of PMD^R and mRL for all ancient dog samples. E) mRL for dog sequences mapping to the dog or the human parts of the concatenated reference. F) PMD^R for dog sequences mapping to the dog or the human parts of the concatenated reference. In all cases, **: p-value < 0.01 and ****: p-value < 0.0001.

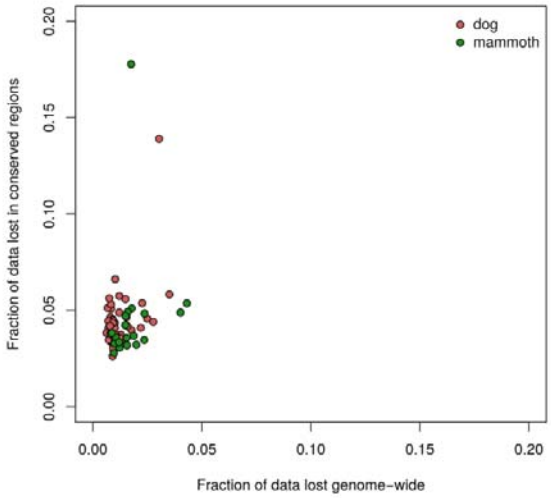

Figure 4. Data lost after competitive mapping. Fraction of data lost in each sample at genome-wide level and only in conserved regions. Colors indicate different species.

Supplementary Figures

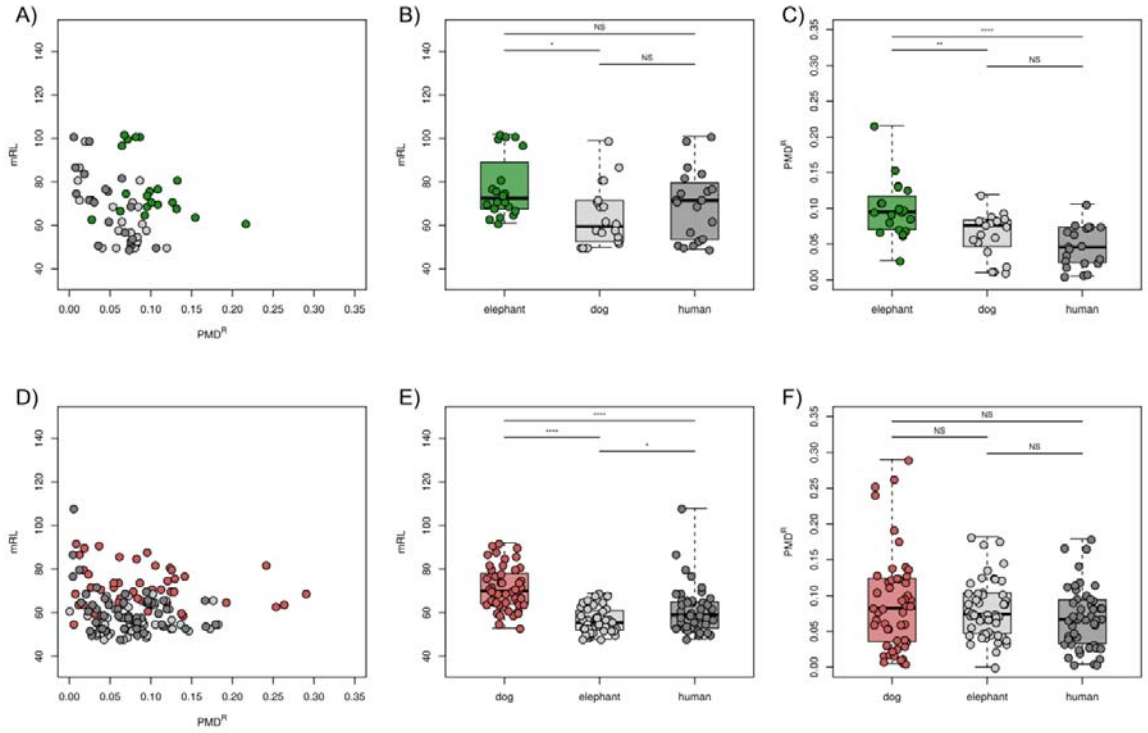

Figure S1: Characterization of sequences mapping to the target, non-target and human references. A) Comparisons of PMD^R and mRL for all mammoth samples. B) mRL for mammoth sequences mapping to the elephant, dog and human references. B) PMD^R for mammoth sequences mapping to the elephant, dog and human references. D) Comparisons of PMD^R and mRL for all ancient dog samples. B) mRL for dog sequences mapping to the elephant, dog and human references. B) PMD^R for dog sequences mapping to the elephant, dog and human references. In all cases, NS: p-value >0.05, *: p-value <0.05, **: p-value < 0.01 and ****: p-value < 0.0001.