


METHODOLOGY ARTICLE

Open Access



# Competitive mapping allows for the identification and exclusion of human DNA contamination in ancient faunal genomic datasets

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

## Abstract

**Background:** 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.

**Results:** 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.

**Conclusions:** We propose 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

## Background

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 [1–3]. DNA degradation is dependent on time and environmental variables such as temperature but also humidity and acidity [4]. Even though the specific model for DNA decay is still debated and it is likely multifactorial

[4], the consequence is that ancient remains typically contain very few molecules of endogenous DNA and these sequences are characterized by short fragment sizes [5].

A second major challenge of ancient DNA research is contamination from exogenous sources [6, 7]. Environmental DNA molecules in the soil matrix the ancient sample was recovered from can easily overwhelm the small amounts of endogenous DNA. This is also true for DNA from people who collected and handled the samples in the field and/or museum collections [8, 9]. While the use of Polymerase Chain Reaction (PCR) technology allowed ancient DNA research to overcome low concentration problems, the sensitivity of the PCR has made it

\* Correspondence: [tatianafeuerborn@palaeome.org](mailto:tatianafeuerborn@palaeome.org); [diez.molino@gmail.com](mailto:diez.molino@gmail.com)

<sup>1</sup>Globe Institute, University of Copenhagen, Copenhagen, Denmark

<sup>3</sup>Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

very difficult to avoid introducing modern contaminant sequences among the authentic ancient DNA [10].

In the last decade, together with more refined DNA extraction and laboratory methods tailored to efficiently retrieve very short and scarce DNA sequences [5, 11], 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 [12]) and other high quality ancient faunal genomes such as those from horses [13], woolly mammoths [14], and bears [15]. 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 [16, 17]. The most common ancient DNA damage is deamination of cytosines to uracils in the overhangs of fragmented DNA molecules [16, 18, 19]. 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 [5, 20].

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 [21, 22]. 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 [23]. In order to avoid these, thresholds for minimum fragment length, that still allow for enough specificity of the alignments, are used [24–26].

Modern human contamination is especially problematic for human palaeogenomic studies since ancient, anatomically modern humans typically fall within the variation of modern humans [27, 28]. This has led to the development of a plethora of methods aimed at computationally quantifying and monitoring exogenous contamination in ancient human DNA datasets [29]. However, the number of methods that allow for the effective exclusion of this type of contamination remains

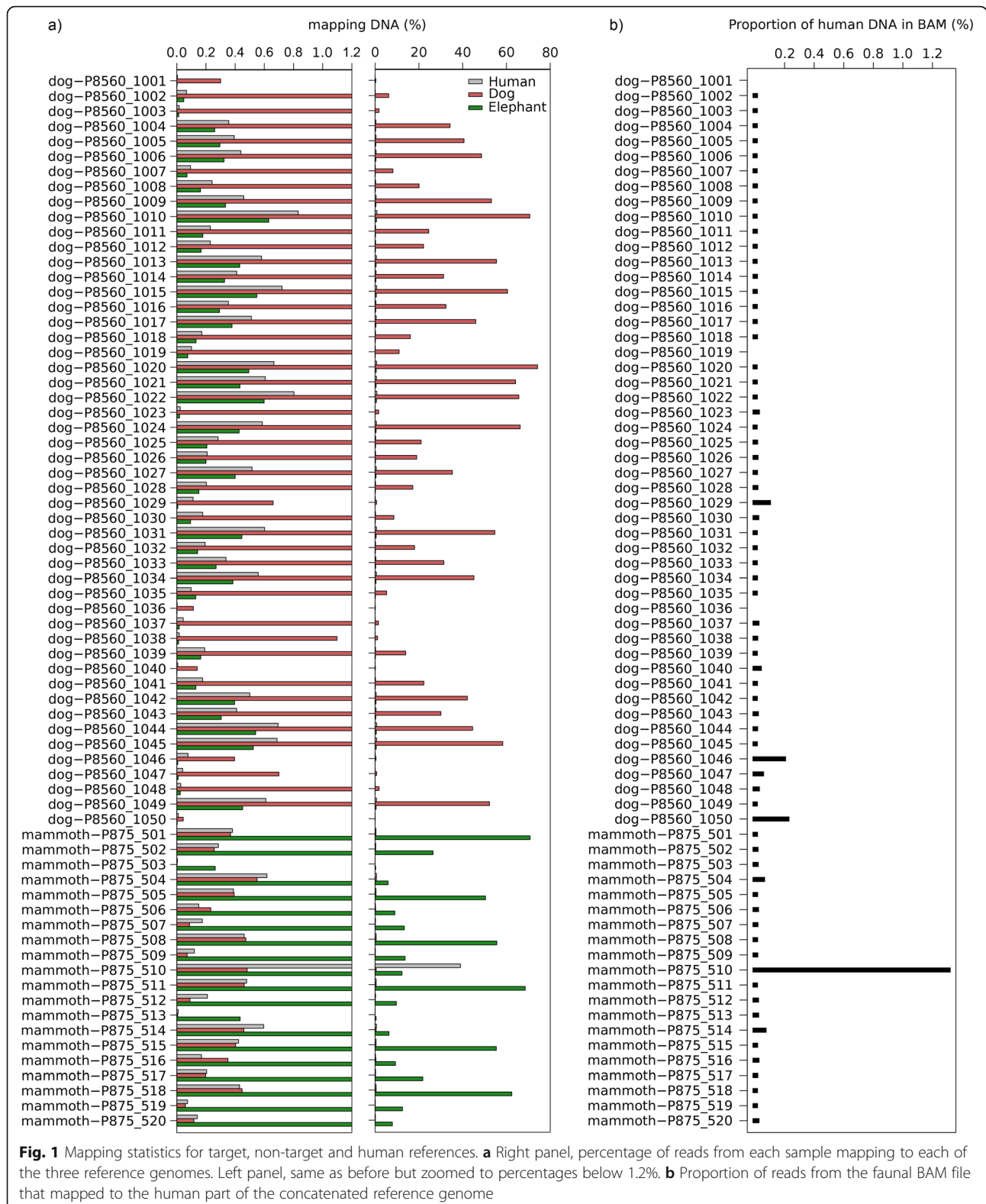
limited. For example, Skoglund et al. [30] used the differential empirical distributions of post-mortem damage (PMD) scores, based on both base quality scores and their level of polymorphism with respect to the reference genome, to differentiate DNA sequences from ancient and modern samples. The PMD 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 [30].

Here we use competitive mapping to investigate the presence of exogenous sequences in ancient sequencing files to evaluate the pervasiveness of human contamination in ancient faunal DNA studies. Previous ancient DNA studies have used similar strategies, i.e. mapping the sequenced ancient DNA data to several reference sequences at the same time, to identify target microbial species (e.g. [31, 32]). 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.

## Results

We first mapped the raw reads from all sequenced ancient samples (50 dogs, *Canis lupus familiaris*, and 20 woolly mammoths, *Mammuthus primigenius*) to three separate reference genomes: the African savannah elephant, dog and human. We found variable levels of sequences confidently mapped to foreign reference genomes (average 0.25% for non-target and 0.86% human) in these sequencing files (Fig. 1a). Most of the files (> 95%) contained less than 0.071% of sequences mapped to human and 0.054% the non-target species. We then estimated average read length (mRL) and post-mortem damage scores (PMD<sup>R</sup>) for all alignments. We detected some significant differences in these indices between sequences mapping to target and to non-target and human (Fig. S1). However, most comparisons between the sequences mapping to the non-target species and human references were not significant.

To investigate whether the target BAM files contain human contaminant sequences we remapped the aligned reads to a concatenated reference composed by the reference genome of the target species, dog or elephant,



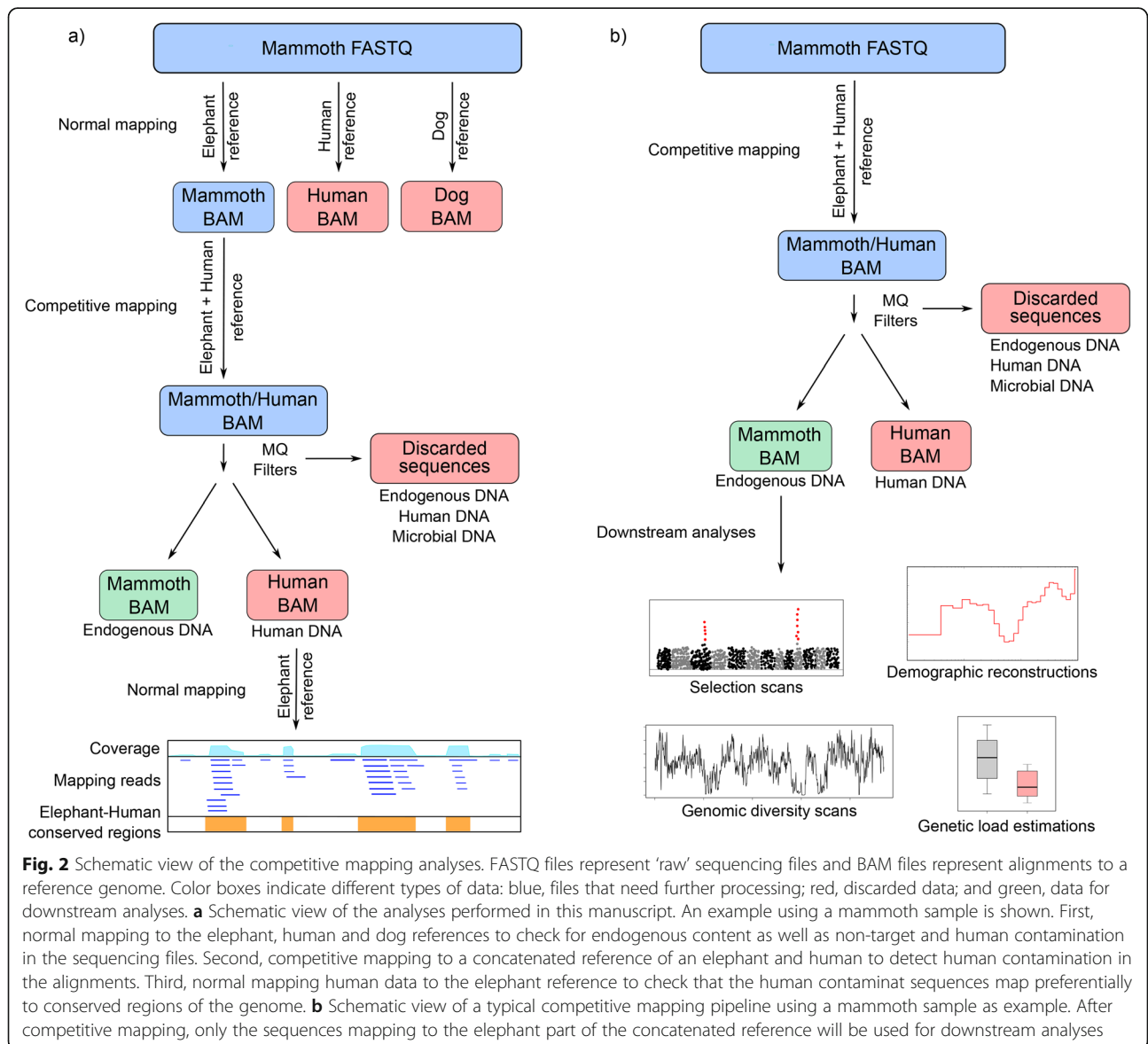
and the human reference genome (Fig. 2a). The concatenated reference was created by merging the two relevant reference genomes together to create one fasta

file containing all chromosomes for each species. This *competitive mapping* approach allowed us to differentiate between three kinds of reads contained in the target

species BAM files. First, reads which align 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 microbial contamination. Second, reads which align to the human reference genome and not to the target species reference genome. These sequences represent the fraction of human contamination in the faunal BAM files. And third, reads that align 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 to any mammalian genome by

random chance. In any case, because these sequences map to both target and human reference genomes at the same time they would thus be discarded when applying mapping quality filters (Fig. 2a).

For each sample, we extracted the reads aligned to the target species of the concatenated reference, representing the true ancient sequences, as well as the human, representing the amount of human contamination contained in the original target BAM file. We found that the alignment files from almost all samples contained sequencing reads that preferentially mapped to the human part of the reference genome than to the target part (average 0.03%; range 0–1.3%) (Fig. 1a, Supplementary Table 1). However, we caution that, 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. Finally, both mRL and PMD<sup>R</sup> were significantly lower in the sequences mapped to the human part than in the ones mapped to the target (Fig. 3).

When using competitive mapping, a fraction of sequences that align to both the target and the human parts of the concatenated reference, were lost (Fig. 2a). Our results indicated that this fraction was an average of 1.33% of the total number of reads per sample (range 0.6–4.3%, Fig. 4, Supplementary Table 1). However, when accounting only for conserved regions between the target species genome and the human genome, the amount of lost sequences was higher (average 3.65%; range 2–16.6%).

**Discussion**

**Contamination in raw sequencing files**

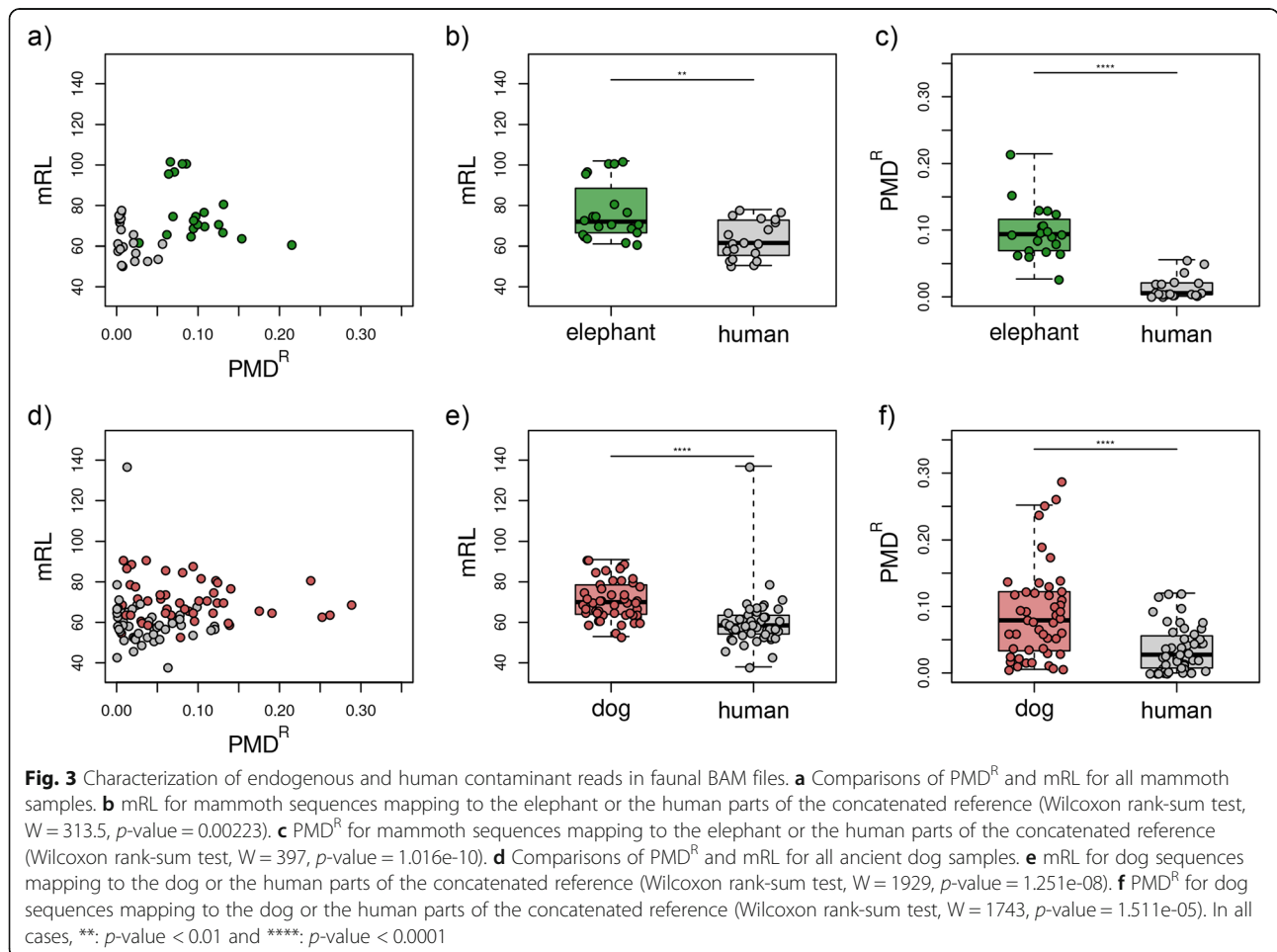
Overall, we found low levels of sequences mapped to foreign reference genomes in the raw sequencing files (Fig. 1a). The proportion of reads mapping to the non-target species and human for each sample were highly correlated (Fig. 5a), indicating that they mostly represent sequences

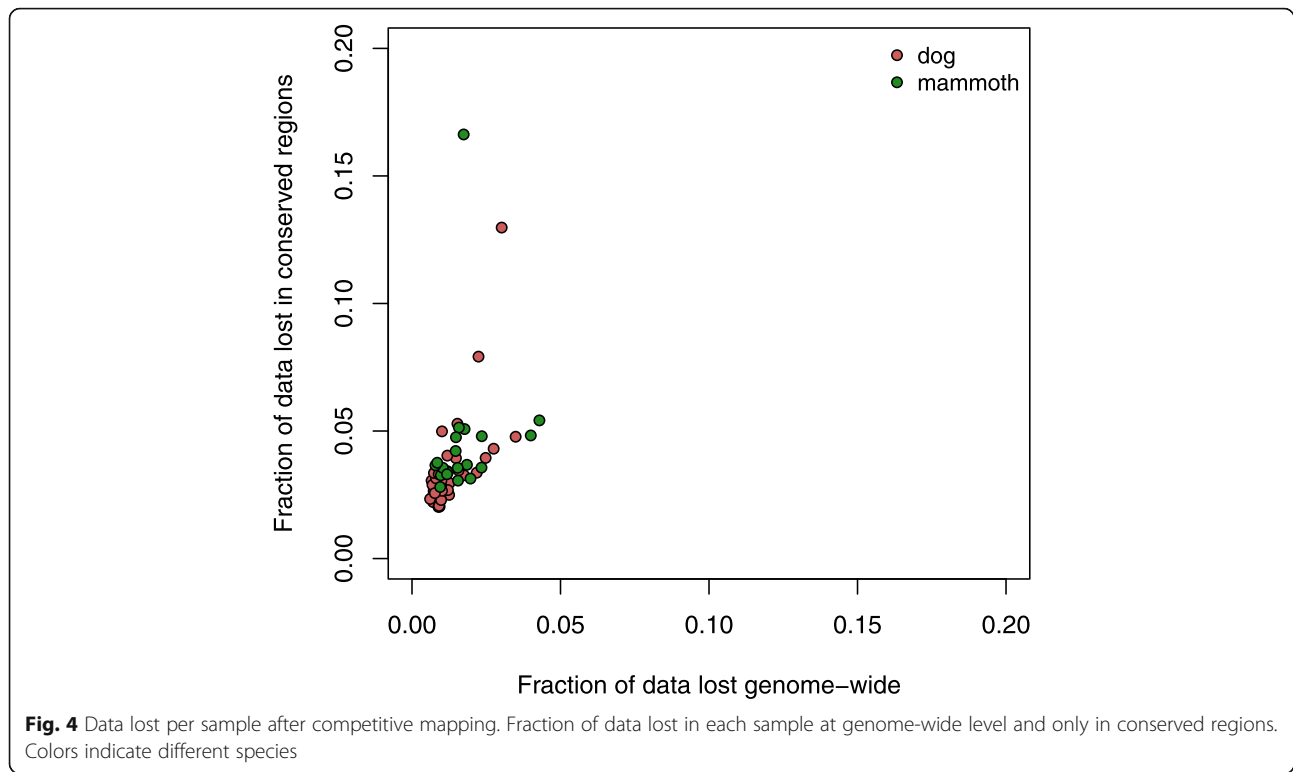
from the target species that map to conserved regions in the other two reference genomes. However, there were notable outliers in the amount of faunal sequences mapping to the human reference. For example, one sample contained a higher proportion of sequences mapped to the human (38.9%) than to the target species (12.3%). This suggested that there could be high levels of human DNA contamination in particular sequencing files.

When characterizing mRL and PMD<sup>R</sup> in the sequences mapping to the different reference genomes we found some differences between the sequences mapping to target compared to non-target and human (Fig. S1), in line with the latter being mostly composed by shorter sequences mapping to conserved regions and the former mostly true endogenous reads. In fact, our results suggest 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**

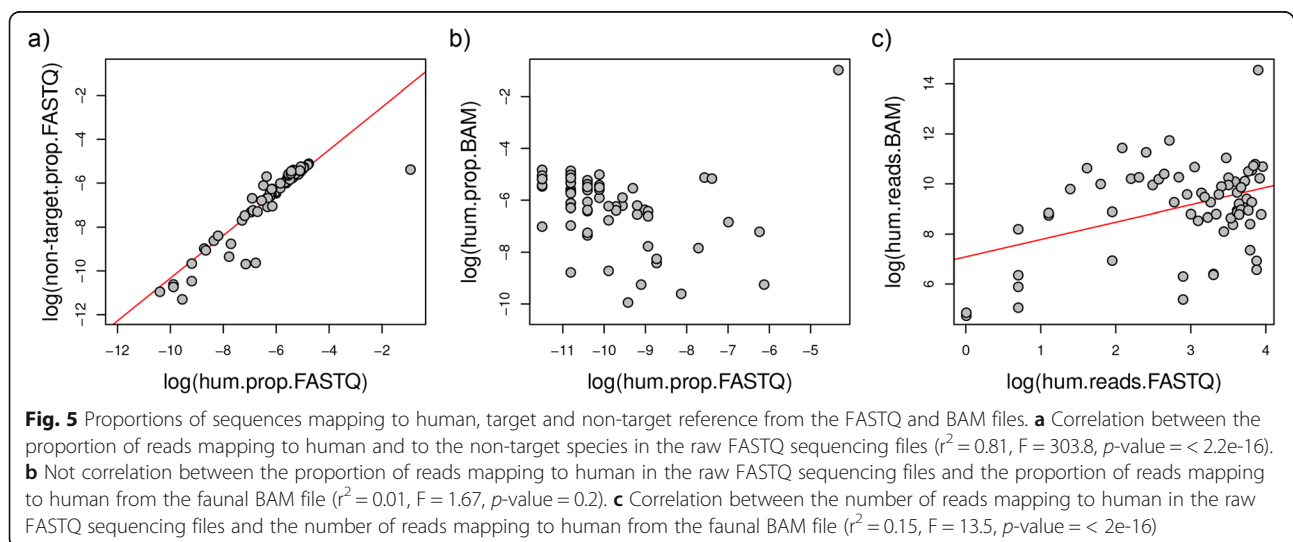
Given that we detected contaminant human sequences in all our ancient fauna sequencing files, we next used





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 (Fig. 1b). Interestingly, the proportion of reads mapped to the human reference from the raw data and the fraction of reads mapping to the human part of the concatenated reference in the target BAM after competitive mapping

are not correlated (Fig. 5b). The reason for this is that the proportion of human reads in the BAM file also depends on the endogenous content of each sample. In fact, the total amount of human sequences that make it to the BAM files is proportional to the number of human sequences in the FASTQ (Fig. 5c). This indicates that the amount of human contamination that is retained in the target BAM files after alignment to the target reference genome can be roughly predicted from the amount of human contamination in the raw sequencing files.



We then estimated mRL and PMD<sup>R</sup> for the true ancient sequences and the contaminant sequences. For both mammoth and dog samples we found a clear distinction in PMD<sup>R</sup> of the sequences mapping to the target species and the ones mapped to human, with higher PMD<sup>R</sup> for the target species, representing true ancient sequences, and lower for the human sequences (Fig. 3c, f). However, we found that the contaminant human reads also displayed a lower mRL (Fig. 3b, e). 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 [26, 33].

#### Excluding contaminant reads from faunal BAM files

The presence of contaminant human sequences in ancient faunal BAM files 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. 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 [34] can be highly affected by the emergence of false variants caused by human contamination [35, 36]. This is especially true for analyses based on low to medium coverage samples, such as most ancient DNA studies. Additionally, since an unknown fraction of the reads discarded using competitive mapping can be of human origin, our detected levels of exogenous human sequences in ancient faunal alignments represent only the lower bound of contamination for these files.

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 (Fig. 2b). An additional advantage of this approach is that a portion of contamination from short microbial reads, common in ancient datasets [26], should also be excluded with this method as many of these short reads would align to both target and human parts of the concatenated reference and are filtered out using the mapping quality filters.

One relevant downside of using competitive mapping 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 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 very limited in a genome-wide context (average 1.3%), and slightly concentrated in conserved regions of the genome (average 3.65%). Unfortunately, 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.

#### 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 propose 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, 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 [36, 37]. There is a growing array of computational methods that allow to confidently identify contamination levels (reviewed in [29]), but few that allow to efficiently separate authentic ancient sequences from contaminating DNA [26, 30]. 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.

#### Methods

##### Materials

We analyzed genomic data from 70 ancient and historical mammalian specimens, 50 dogs and 20 woolly mammoths (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 and are radiocarbon dated.

### 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. [38]. 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 [39, 40]. 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/Seq-Prep](https://github.com/jstjohn/Seq-Prep)) with default settings (excluding sequences shorter than 30 bp after merging) and a slight modification of the source code to calculate the base qualities in the overlapping region [14]. We then mapped the merged reads to three separate reference genomes: the African savannah elephant genome (LoxAfr4, Broad Institute), the dog genome (CanFam3.1, [41]), and the human reference genome (Hg19). All mappings were performed using *BWA aln v0.7.8* [42] using settings adapted for ancient DNA as in Pečnerová et al. [43].

We removed PCR duplicates from the alignments using a script ([github.com/pontusssk/samremovedup](https://github.com/pontusssk/samremovedup)) that takes into account both starting and end coordinates of the reads to be identified as duplicates [44] and estimated the number of unique mapping reads using *samtools v1.8* [45]. In all cases, we refer to *mapped reads* to those sequences retained after filtering by mapping quality > 30. We consider true endogenous sequences are those mapping to the target species (i.e. dog reference for ancient dog samples and elephant reference for mammoth samples) and exogenous contaminant sequences are those mapping to the non-target reference (i.e. elephant and human references for ancient dog samples and dog and human references for mammoth samples). 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 [4, 46, 47] measured as median read length (mRL) and deamination patterns [48, 49] measured as post-mortem damage scores (PMD, [30]). For each sample, we define the PMD ratio (PMD<sup>R</sup>) as the fraction of sequences that display a PMD score > 5. Therefore, a higher PMD<sup>R</sup> value indicates that the sample contains more sequences with larger PMD scores, thus it contains more ancient DNA sequences.

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 overlapping 30 bp long sequences with a step size of 1 bp. We then mapped the obtained short sequences to the other two reference genomes, dog and elephant, using *BWA* [50]. 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.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-020-07229-y>.

**Additional file 1.** Supplementary Information which contains Extended results note 1, Figure S1 and Supplementary Table 1.

### 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.

### Authors' contributions

T.R.F. and D.D.d.M. conceived the study with input from E.P., J.v.S., P.P., I.U., E.E., V.K.L., M.K., R.R.V., A.G. and L.D. T.R.F., E.P., and J.v.S. performed lab procedures. T.R.F., A.M., T.v.d.V., M.D. and D.D.d.M. analyzed the data. T.R.F. and D.D.d.M. wrote the manuscript with contributions from all other coauthors. All authors contributed to and approved the final version of the manuscript.

### Funding

Genetic analyses were funded through a grant from the Swedish Research Council (VR grants 2012–3869 and 2017-04647) awarded to L.D. J.v.S. and L.D. 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.d.M. was supported through a Carl Tryggers scholarship (grant CTS 17:109). Open Access funding provided by Stockholm University.

### Availability of data and materials

All the data generated in the current study are available in the European Nucleotide Archive (ENA), accession number PRJEB41038.



**Ethics approval and consent to participate**

All samples analyzed in this study originate from ethnographic collections, archaeological excavations and arctic expeditions. No ethics approval was required.

**Consent for publication**

'Not applicable'.

**Competing interests**

The authors declare that they have no competing interests.

**Author details**

<sup>1</sup>Globe Institute, University of Copenhagen, Copenhagen, Denmark. <sup>2</sup>Archaeological Research Laboratory, Department of Archaeology and Classical Studies, Stockholm University, Stockholm, Sweden. <sup>3</sup>Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden. <sup>4</sup>Centre for Palaeogenetics, Stockholm, Sweden. <sup>5</sup>Department of Zoology, Stockholm University, Stockholm, Sweden. <sup>6</sup>Department of Organismal Biology, Human Evolution, Uppsala University, Uppsala, Sweden. <sup>7</sup>Department of Biology, University of Copenhagen, Copenhagen, Denmark. <sup>8</sup>Department of Animal Breeding, INIA, Madrid, Spain.

Received: 1 July 2020 Accepted: 16 November 2020

Published online: 30 November 2020

**References**

- Lindahl T. Instability and decay of the primary structure of DNA. *Nature*. 1993;362:709–15.
- Dabney J, Meyer M, Pääbo S. Ancient DNA damage. *Cold Spring Harb Perspect Biol*. 2013;5:a012567.
- Pääbo S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc Natl Acad Sci U S A*. 1989;86:1939–43.
- Kistler L, Ware R, Smith O, Collins M, Allaby RG. A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic Acids Res*. 2017;45:6310–20.
- Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, et al. Complete mitochondrial genome sequence of a middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc Natl Acad Sci U S A*. 2013;110:15758–63.
- Malmström H, Storå J, Dalén L, Holmlund G, Götherström A. Extensive human DNA contamination in extracts from ancient dog bones and teeth. *Mol Biol Evol*. 2005;22:2040–7.
- Der Sarkissian C, Allentoft ME, Ávila-Arcos MC, Barnett R, Campos PF, Cappellini E, et al. Ancient genomics. *Philos Trans R Soc Lond Ser B Biol Sci*. 2015;370:20130387.
- Der Sarkissian C, Ermini L, Jónsson H, Alekseev AN, Crubezy E, Shapiro B, et al. Shotgun microbial profiling of fossil remains. *Mol Ecol*. 2014;23:1780–98.
- Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, et al. Analysis of one million base pairs of Neanderthal DNA. *Nature*. 2006;444:330–6.
- Willerslev E, Cooper A. Ancient DNA. *Proc Biol Sci*. 2005;272:3–16.
- Gamba C, Hanghøj K, Gaunitz C, Alfarhan AH, Alquraishi SA, Al-Rasheid KAS, et al. Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Mol Ecol Resour*. 2016;16:459–69.
- Slatkin M, Racimo F. Ancient DNA and human history. *Proc Natl Acad Sci*. 2016;2016:1–8.
- Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, et al. Recalibrating Equus evolution using the genome sequence of an early middle Pleistocene horse. *Nature*. 2013;499:74–8.
- Palkopoulou E, Mallick S, Skoglund P, Enk J, Rohland N, Li H, et al. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr Biol*. 2015;25:1395–400.
- Barlow A, Pajjmans JLA, Alberti F, Gasparyan B, Bar-Oz G, Pinhasi R, et al. Middle Pleistocene cave bear genome calibrates the evolutionary history of Palaeartic bears; 2020.
- Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, et al. Patterns of damage in genomic DNA sequences from a Neanderthal. *Proc Natl Acad Sci U S A*. 2007;104:14616–21.
- Renaud G, Schubert M, Sawyer S, Orlando L. Authentication and assessment of contamination in ancient DNA. *Methods Mol Biol*. 1963;2019:163–94.
- Gilbert MTP, Willerslev E, Hansen AJ, Barnes I, Rudbeck L, Lynnerup N, et al. Distribution patterns of postmortem damage in human mitochondrial DNA. *Am J Hum Genet*. 2003;72:32–47.
- Stiller M, Green RE, Ronan M, Simons JF, Du L, He W, et al. Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. *Proc Natl Acad Sci U S A*. 2006;103:13578–84.
- Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One*. 2012;7:e34131.
- Prüfer K, Stenzel U, Hofreiter M, Pääbo S, Kelso J, Green RE. Computational challenges in the analysis of ancient DNA. *Genome Biol*. 2010;11:R47.
- Kircher M. Analysis of high-throughput ancient DNA sequencing data. *Methods Mol Biol*. 2012;840:197–228.
- Smith TF, Waterman MS, Burks C. The statistical distribution of nucleic acid similarities. *Nucleic Acids Res*. 1985;13:645–56.
- Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neanderthal genome. *Science*. 2010;328:710–22.
- Meyer M, Arsuaga J-L, de Filippo C, Nagel S, Aximu-Petri A, Nickel B, et al. Nuclear DNA sequences from the middle Pleistocene Sima de los Huesos hominins. *Nature*. 2016;1:1–15.
- de Filippo C, Meyer M, Prüfer K. Quantifying and reducing spurious alignments for the analysis of ultra-short ancient DNA sequences. *BMC Biol*. 2018;16:121.
- Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of bronze age Eurasia. *Nature*. 2015;522:167–72.
- Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. 2014;513:409–13.
- Peyrégne S, Prüfer K. Present-day DNA contamination in ancient DNA datasets. *Bioessays*. 2020;42:2000081.
- Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Pääbo S, Krause J, et al. Separating endogenous ancient DNA from modern day contamination in a Siberian Neanderthal. *Proc Natl Acad Sci U S A*. 2014;111:2229–34.
- Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjögren K-G, et al. Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell*. 2015;163:571–82.
- Valtueña AA, Mittnik A, Key FM, Haak W, Allmäs R, Belinskij A, et al. The stone age plague and its persistence in Eurasia. *Curr Biol*. 2017;27:3683–91.e8.
- Lee H, Schatz MC. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics*. 2012;28:2097–105.
- Schiffels S, Haak W, Paajanen P, Llamas B, Popescu E, Loe L, et al. Iron age and Anglo-Saxon genomes from East England reveal British migration history. *Nat Commun*. 2016;7:10408.
- Renaud G, Hanghøj K, Korneliusen TS, Willerslev E, Orlando L. Joint estimates of heterozygosity and runs of homozygosity for modern and ancient samples. *Genetics*. 2019;212:587–614.
- Llamas B, Valverde G, Fehren-Schmitz L, Weyrich LS, Cooper A, Haak W. From the field to the laboratory: controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. *STAR*. 2017;3:1–14.
- Korlević P, Gerber T, Gansauge M-T, Hajdinjak M, Nagel S, Aximu-Petri A, et al. Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *Biotechniques*. 2015;59:87–93.
- Ersmark E, Orlando L, Sandoval-Castellanos E, Barnes I, Barnett R, Stuart A, et al. Population demography and genetic diversity in the Pleistocene cave lion. *Open Quaternary*. 2015;1:1–14.
- Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*. 2010;2010.pdb.prot5448.
- Carøe C, Gopalakrishnan S, Vinner L, Mak SST, Sinding M-HS, Samaniego JA, et al. Single-tube library preparation for degraded DNA. *Methods Ecol Evol*. 2017;9:1.
- Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*. 2005;438:803–19.
- Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*. 2009;25:1754–60.

43. Pečnerová P, Díez-Del-Molino D, Dussex N, Feuerborn T, von Seth J, van der Plicht J, et al. Genome-based sexing provides clues about behavior and social structure in the woolly mammoth. *Curr Biol*. 2017;27:3505–10.e3.
44. Palkopoulou E, Lipson M, Mallick S, Nielsen S, Rohland N, Baleka S, et al. A comprehensive genomic history of extinct and living elephants. *Proc Natl Acad Sci U S A*. 2018;115:E2566–74.
45. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
46. Rogaev EI, Moliaka YK, Malyarchuk BA, Kondrashov FA, Derenko MV, Chumakov I, et al. Complete mitochondrial genome and phylogeny of Pleistocene mammoth *Mammuthus primigenius*. *PLoS Biol*. 2006;4:1.
47. Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc R Soc B Biol Sci*. 2012;279:4724–33.
48. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res*. 2010;38:e87.
49. Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Pääbo S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res*. 2001;29:4793–9.
50. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013; <http://arxiv.org/abs/1303.3997>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

