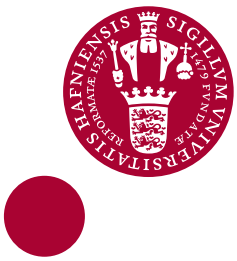

Authentication and assessment of contamination in ancient DNA

Renaud, Gabriel; Schubert, Mikkel; Sawyer, Susanna; Orlando, Ludovic

Published in:
Ancient DNA

DOI:
10.1007/978-1-4939-9176-1_17

Publication date:
2019

Authentication and assessment of contamination in ancient DNA

Gabriel Renaud[1], Mikkel Schubert[1], Susanna Sawyer[1], Ludovic Orlando[1,2*]

[1] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark

[2] Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, CNRS UMR 5288, Université de Toulouse, University Paul Sabatier, Toulouse, France 31000

[*] Corresponding author: Lorlando@snm.ku.dk

Running Head: Authentication of ancient DNA data

This is a nonstandard format chapter

Abstract

Contamination from both present-day humans and post-mortem microbial sources is a common challenge in ancient DNA studies. Here we present a set of tools to assist in the assessment of contamination of ancient DNA data. These tools perform several standard tests of authenticity of ancient DNA data, including detecting the presence of post-mortem damage signatures in sequence alignments and quantifying the amount of present-day human contamination.

Keywords

Contamination, Ancient DNA, Post mortem damage, Schmutzi, DICE, mapDamage2.0

1. Introduction

DNA extracted from preserved materials has enabled unprecedented insights into the history of humans (1-2), animals (3-6), plants (7), and microbes (8-11). This material, referred to as ancient DNA (aDNA), is obtained through DNA extraction from bones, teeth, and other materials such as hair, mummified tissues, and dental plaque derived from organisms that died up half a million years ago (12-15). Since DNA degrades over time after the death of an organism, aDNA is characterized by short fragment size (often less than 50 bp (13)) and post-mortem damage in the form of cytosine (C) > thymine (T) and the complementary guanine (G) > adenine (A) substitutions (16-17). After death, remaining tissues are colonized by microbial decomposers, which can introduce microbial DNA contamination that in some cases represents more than 99% of recovered DNA (18). Finally, as very little DNA tends to be preserved and much of this DNA can belong to exogenous sources (except for particular material such as petrosal bones, tooth cementum, and hair (19)), even small amounts of

present-day DNA can overwhelm the original and endogenous DNA, making contamination a significant challenge in DNA studies.

Contamination derived from living humans is particularly problematic in the study of ancient humans, due to the high genetic similarity been modern and ancient humans. Contamination can be introduced to ancient human samples in several ways. During excavation, for example, bones and teeth are typically handled in non-sterile environments by bare hands (with notable exceptions, such as excavations at El Sidrón Cave in Asturias, Spain (20)). Bones and teeth are also sometimes cleaned by washing them in water, which can contain skin flakes from humans. This is a problem because hydroxyapatite, which is the main mineral component of bone, tooth enamel, and dentin, absorbs DNA in a liquid environment (21). Ancient samples can also be contaminated during museum storage, both through variation in preservation conditions and through contact with other samples (22-23).

Once in the laboratory, the risk of contamination can be reduced by working in a sterile clean environment. Researchers often remove or chemically treat the outer surface of bones and teeth, which can reduce surface contamination. Many (although not all) reagents to be used in DNA extraction and genomic library preparation can be treated by UV radiation or with exonucleases (24).

Computational pipelines have been developed to detect the presence of contaminating DNA after DNA sequencing has been performed. In these approaches, the more closely related the source of the aDNA is to the present-day contaminants, the harder it is to distinguish between the authentic and contaminating DNA. In archaic hominins such as Neanderthals, all mitochondrial genome sequences published so far fall outside the variation of modern humans (19,25-26), making present-day human contamination estimates achievable if mitochondrial coverage is sufficiently high (18,27). However, such estimates do not necessarily reflect the amount of nuclear DNA contamination (18). To date, few high

coverage nuclear genomes from archaic hominins are published, and contamination estimates of nuclear DNA are challenging. For ancient, anatomically modern humans, differentiating between present-day human contaminants and endogenous DNA is even more difficult, since both the ancient mitochondrial and nuclear DNA can fall within the variation of present-day humans (28-32). This makes methods used for Neanderthals, such as identifying diagnostic positions in the mitochondrial genome (18), inapplicable to anatomically modern humans (29-30).

The molecular footprints of post-mortem DNA damage can be useful for differentiating between authentically ancient DNA and present-day contaminants (33). In a living organism, damage to the DNA strands, including via hydrolysis and oxidation, is continuously repaired (36). After death, however, DNA is left unrepaired, and damage accumulates in predictable patterns, the signal of which is a distinguishing characteristic of aDNA. The post mortem damage most commonly associated with aDNA is cytosine to uracil deamination (or to thymine, if the cytosine is methylated) (17). This deamination is more likely to occur in the single stranded overhangs of the fragmented aDNA (36), and results in a C>T (G>A) replacement signal at the 5' (3') ends of aDNA sequences (17). Different nucleotide misincorporation patterns are expected depending on the molecular tools used during library preparation (37-39), and library amplification (37,40), and software such as mapDamage (34-35) has been developed to automate detection of these signals. While observing these patterns in aDNA datasets is compatible with DNA that has originated from an ancient source, it does not rule out the possibility of contamination, as mixtures of aDNA templates and fresh modern contaminants can coexist and generate bona fide patterns.

As contamination can originate from many sources (e.g. cross-contamination, microbial colonization, modern human contamination), it is crucial to assess the authenticity of aDNA sequence data sets prior to using these data for downstream analyses. Here we

describe several tools that are useful in the assessment of data authenticity in ancient DNA data sets. We begin by discussing the basic prerequisites for performing these analyses. We then describe the automated PALEOMIX pipeline (section 2). Next, we describe how the signatures of post mortem damage may be examined using mapDamage2.0 (section 3). Finally, we discuss how the amount of contamination may be estimated for the mitochondrial genome using schmutzi (section 4) and for the nuclear genome using ANGSD and DICE (section 5).

2. Processing data using the PALEOMIX pipeline

In order to carry out the analyses described in the subsequent parts of this chapter, it is necessary to perform certain pre- and post-processing steps on the raw FASTQ data produced by the HTS instrument. For the purpose of this chapter, we will explain processing using the 'BAM pipeline', one of the pipelines included with the PALEOMIX suite of tools (44). This pipeline implements both the pre-processing and the post-processing steps outlined in detail in the sections below, and allows for easy mapping onto multiple reference sequences of interest. For detailed installation instructions, please refer to http://paleomix.readthedocs.io/

The steps can be summarized as follows:

1. Pre-processing of reads (section 2.1):
    a. Adapter sequences must be trimmed from short DNA inserts.
    b. Overlapping paired-end reads may be merged while taking into account individual per base quality scores.
2. During read mapping:
    a. Reads must be mapped to the mitochondrial (organellar) and nuclear genome separately.
3. Post-processing of alignments (section 2.3):

a. Mate information must be set for paired-end alignments.
b. The ‘MD’ tag must be calculated for every alignment.
c. Alignments must be sorted by coordinate.
d. PCR duplicates must be filtered.

4. Implementation with PALEOMIX (section 2.4)

2.1. Pre-processing of reads

Prior to mapping, raw reads must be trimmed of any residual adapter sequences used by the platform as sequencing primers, in order to eliminate the possibility that any such sequences interfere with downstream alignments and genotyping. This is particularly important due to the short size of aDNA fragments, which are typically shorter than the read length used by Illumina sequencing machines, and are consequently often terminated by platform specific adapter sequences (45). When performing paired-end sequencing, it is furthermore beneficial to merge overlapping portions of the read mates in order to reconstruct the full length of the original aDNA fragment. Merging read mates in a quality-aware manner also reduces the error rate in the resulting sequences (45). These two steps of read trimming and merging may be combined using tools specialized for aDNA, including AdapterRemoval v2 (46) and leeHom (47), as well as general tools for adapter trimming (48-50) and read merging (51-52). The steps aim at reconstructing the sequence of the original aDNA fragment onto which Illumina specific adapters were ligated and subsequently read by the sequencer.

2.2. Mapping to reference genomes

Once reads have been trimmed and (optionally) merged, mapping of the resulting sequences may be carried out using any of the numerous short-read alignment tools currently available (see e.g. (53) for an overview). These instructions make use of BWA (54), but any mapper may be substituted when carrying out this process (eg. BWA-PSSM (55) and Bowtie2 (56).

When mitochondrial genomes are analysed, it is important, regardless of the chosen mapper, that reads be mapped to the mitochondrial (by extension, organellar) genome in isolation. That is to say, that mapping must be carried out using only the mitochondrial genome as the target sequence, excluding any nuclear genome sequences. This is motivated by the presence of mitochondrial inserts in the nuclear genome (NUMTs), representing anything from small fragments to almost complete copies of the mitochondrial genome (57-58). The presence of these sequences may therefore result in both authentic and contaminant DNA sequences mapping to the nuclear genome, rather than to the mitochondrial reference sequence, or in a loss of sequence information as non-unique hits are generally discarded in downstream analyses (see section 2.3). This may hinder any attempts at calling a consensus sequence for the mitochondrial data due to the presence of gaps in the alignments (see Section 4).

2.3. Post-processing of alignments

Following sequence alignment, several additional steps have to be completed in order to prepare the results for the analyses described in subsequent sections. Incomplete mate-information in paired-end alignments such as insert size and mate related flags must be added. Secondly, the alignments must be sorted by genomic coordinates in order to allow indexing (e.g. using 'samtools index') and therefore efficient analyses of the data by the programs described below. Thirdly, the 'MD' tag, containing information about the presence of mismatches relative to the reference sequence, must be calculated using the SAMtools 'calmd' or 'fillmd' commands (59), as this information is used both by schmutzi and DICE to infer the sequence of the reference from the alignment alone (see section 4). In addition, alignments must be filtered for PCR duplicates, in particular samples with low DNA

contents, as such duplicates may greatly distort the results obtained from the subsequent analyses, through the overrepresentation of duplicated DNA inserts (60).

2.4. Implementation with PALEOMIX

Installation of PALEOMIX and its required Python modules may be accomplished running the 'pip' command as root:

% pip install paleomix

In addition to PALEOMIX itself, the following software also be installed: AdapterRemoval v2 (46), SAMTools (59), and BWA (54). In addition, the Picard Tools 'picard.jar' file must be downloaded and placed in ~/install/jar_root/. This software is used by the BAM pipeline, an automated pipeline for processing and mapping sequencing reads that is included as part of PALEOMIX. mapDamage2.0 (35) and the Genome Analysis ToolKit (GATK) (61) are optional components for the BAM pipeline, but for the purpose of these instructions we will not make use of these when running this pipeline.

To start a new mapping project, first create a default configuration file using output of the built-in 'makefile' command in the PALEOMIX 'BAM pipeline':

% bam_pipeline makefile > project.yaml

As indicated by the extension, the makefile is based on the YAML markup language (http://www.yaml.org/), a text format for organizing data that is both easy to read and easy to manipulate using any standard text editor.

The following instructions will focus on mapping to the mitochondrial genome, and assume that the mitochondrial genome is located in a read/writable directory at /genomes/mitochondria.fasta, but this may be substituted by any other path. To specify that the aDNA fragments should be mapped to this genome, open the 'project.yaml' file in a text editor (vim, emacs, nano, sublime, etc), and locate the section of the file starting with 'Prefixes:'. There specify the name and location of the genome that we wish to map to (named 'Mitochondria' in the following example):

```
Prefixes:
  Mitochondria:
    Path: /genomes/mitochondria.fasta
```

Any number of reference genomes (called 'Prefixes' in the pipeline), in the form of FASTA files containing one or more sequences, may be specified here (the .fasta extension is mandatory). Note that in YAML files, the indentation (none for the first line, one level for the second line, and two levels for the third line) is used to define the structure of the data. This indentation must consist purely of spaces, as the format prohibits the use of tabs for indentation.

Next, specify the samples to be mapped by specifying 5 pieces of information for each sample: 1) The filename to use for output files generated by the pipeline; 2) the name of the sample; 3) the name of the library from which reads have been sequenced; 4) the name of the sequencing run to be processed; and 5) the location of the files containing the sequencing data.

For example, assuming that we had downloaded the sample SRR123456 from the Sequence Read Archive, and saved the paired-end FASTQ reads in files

SRR123456_1.fastq.gz and SRR123456_2.fastq.gz, for mate 1 and mate 2 reads, respectively, these could be specified as follows, at the end of the configuration file:

```
SRR123456:
  SRR123456:
    Library1:
      ERR123456: /path/to/SRR123456_{Pair}.fastq.gz
```

The '{Pair}' value is used to signify that this is paired-end data, and the pipeline will expect the mate 1 and mate 2 reads to be located at the path generated by replacing this value by 1 and 2, respectively (here /path/to/SRR123456_1.fastq.gz and /path/to/SRR123456_2.fastq.gz). In the case of single-end reads, if the FASTQ reads are stored in file /path/to/SRR123456.fastq.gz for instance, this value would be omitted:

```
SRR123456:
  SRR123456:
    Library1:
      ERR123456: /path/to/SRR123456.fastq.gz
```

Multiple libraries and/or runs can be specified for each sample, as described in the documentation for the BAM pipeline.

Finally, locate the 'Features' section and disable the use of 'mapDamage' and the use of GATK, by replacing the 'yes'-value for 'mapDamage' and 'RealignedBAM' with 'no', as shown here:

Features:

RealignedBAM: no

mapDamage: no

...

Other options in the 'Features' section are not shown here for the sake of brevity, and these should be left as-is.

Once this has been done, the pipeline can be run using the 'bam_pipeline run' command:

% **bam_pipeline run project.yaml**

Running the BAM pipeline in this manner will

1. Trim remaining adapter sequences and merge overlapping reads.
2. Map the trimmed reads onto the listed reference sequences.
    a. If these references have not already been indexed, indexing is performed using the selected short read aligner (here BWA).
3. Remove any reads not mapping to the reference genome.
4. Process mapped reads using the SAMtools 'calmd' and 'fixmate' commands, to ensure that mate information is correct, and to ensure that the 'MD' is present.
5. Remove reads identified as PCR duplicates (here using the *Picard* 'MarkDuplicates' command).
6. Index the BAM for fast retrieval of arbitrary regions.

Based on the settings listed above, the resulting BAM file, containing alignments against the mitochondrial genome, will be located in the current directory with the filename 'SRR123456.mitochondria.bam'. This file is suitable for analysis following the instructions laid out in section 4.2 aiming at quantifying present-day human contamination from mitochondrial data and for obtaining a consensus call for the endogenous mitochondrial genome. A BAM file aligned to the nuclear genome using the same methodology can be used for the analyses described in section 4.3 to quantify present-day human contamination for the nuclear genome.

3. Estimating damage parameters and fragmentation patterns

As mentioned in the introduction, post mortem damage patterns are routinely used in assessing aDNA authenticity. Here we discuss mapDamage2.0 (35), which offers several tools for visualizing and modeling the patterns of post mortem damage observed in ancient samples. It also allows users to re-calculate base quality-scores to mitigate the impact of post mortem damage on downstream analyses. However, in this section, we focus solely on the basic plotting of error rates and fragmentation patterns in a BAM file. The modeling and plotting of post mortem DNA damage parameters, and the rescaling of base quality-scores, is not required for the purpose of authenticating aDNA. Basic plotting may be performed on any valid BAM file (as described in section 2), and makes no assumption about the presence or absence of post mortem DNA damage in the sample.

It is recommended to start any analyses of aDNA by performing a basic mapDamage plot; this not only serves to determine if aDNA is present, but may also help detect systematic sequencing errors and biases through visual inspection of the error and fragmentation plots.

Note that most of these plots are not as useful for extracts that have been treated with a 'USER' treatment: a combination of endonuclease VIII and uracil-DNA-glycosylase (UDG), as this treatment serves to erase the signature of post mortem damage (30,62).

3.1. Post mortem damage and fragmentation plots

Before proceeding, first install mapDamage2.0 and its dependencies as described at http://ginolhac.github.io/mapDamage/. To carry out basic plotting using mapDamage2.0, two files are required: Firstly, a reference sequence in FASTA format, and secondly, a BAM file containing alignments to that reference sequence. The BAM file must fulfill the requirements described in section 2, with the exception of the 'MD' tag information, which is not required. Once these requirements are met, basic plotting may be carried out using the following command (shown here for the example BAM produced in section 2):

% **mapDamage -r /genomes/mitochondria.fasta -i SRR123456.mitochondria.bam --no-stats**

The **--no-stats** option is required to disable the modeling of post mortem DNA damage parameters, which is otherwise carried out by default. Running this command will create a folder named **results_SRR123456.mitochondria/** in which the output files are placed (alternatively, this destination may be set using the -d command-line option). It is noteworthy that enabling the mapDamage feature by manually editing the Features section of the Paleomix makefile would result in running the same analysis:

```
Features:
  RealignedBAM: no
```

mapDamage: yes

...

% **bam_pipeline run project.yaml**

The primary output are the plots '**Fragmisincorporation_plot.pdf**' and '**Length_plot.pdf**'. The first of these, the '**Fragmisincorporation_plot.pdf**' file, plots the base composition around the 5' and 3' termini of aligned DNA sequences, as well as the rates of C>T and G>A mismatches observed relative to the reference genome across the alignments (the so-called nucleotide misincorporation patterns).

3.1.1. Post mortem damage plots

For aDNA libraries produced using blunt-end adapter ligation to double-stranded DNA templates and single strand fill-in following end-repair (63), this plot is expected to show an increase in C>T and G>A mismatches when approaching the 5' and 3' termini, respectively (Figure 1, b). Modern DNA, on the other hand, including contamination resulting from the handling of samples and/or introduced during library preparation, is expected to show C>T and G>A mismatches in line with other substitution rates (Figure 2, a). As such, the presence of post mortem damage provides powerful evidence of the presence of aDNA in the sample (64). This signal may, however, be influenced by the choice of library building protocol used during the sequencing of the ancient sample. In particular, the use of the A-tailing (AT) ligation protocol is known to reduce the rate of post mortem DNA damage observed at the first and the last position in aDNA fragments (Figure 1, c) (37), due to ligation bias against templates starting with thymine analogs, such as uracils.

It is furthermore notable G>A substitutions at the 3' termini of aDNA fragments are a product of the end-repair step, in which the missing DNA segments complementary to 5' overhangs are repaired, introducing adenines when facing uracils (instead of guanines at non-deaminated cytosines). However, for protocols targeting individual strands of aDNA and not involving 3'-end fill-in reactions (such as the single-stranded library preparation method (38)), only the 5' C>T pattern of post mortem DNA damage is observed. The 3' G>A pattern is then replaced by a 3' C>T pattern, almost symmetrical to that observed at 5' ends, as in absence of end-repair, 3' overhangs, which also accumulate uracils, are not removed.

While diagnostic for the presence of aDNA, mismatches resulting from the presence of post mortem DNA damage are also detrimental to efforts to genotype ancient specimens, and 'USER' treatment has been developed to eliminate the presence of such DNA damage in ancient samples, at the cost of some of the material, through the targeted excision of uracils (62,65). However, despite the elimination of uracils in the template molecules, the sequencing of ancient, USER-treated samples can still shows a slight signal of C>T at the 5' and 3' termini (Figure 1, d). This signal is mostly driven by the presence of methylated CpG epi-alleles in the sequence data (66-68).

It should be noted that contaminating sequences can also have appreciable amounts of post mortem DNA damage in samples that were excavated long ago and/or have been subject to chemical treatment (e.g. museum specimen and medical samples) (6,69). Similarly, hosts and their pathogens can show different levels of DNA damage, despite being exposed to identical preservation conditions (70). As such, while the presence of post mortem DNA damage is therefore only indicative of the presence of aDNA, it alone cannot establish data authenticity.

3.1.2. Base composition plot

As post mortem DNA fragmentation appears to be mainly driven through depurination (17), the base composition of the genomic positions immediately preceding aDNA fragment starts is non-random and enriched in purines (adenine and guanine residues). Users can check for the presence of this feature in their data, directly from the fragment misincorporation plot provided by mapDamage2.0 (35), where the top 4 plots provide base composition profiles within the 10 first and 10 last fragment positions, as well as in their respective flanking 10 bp regions in the reference genome. The exact mechanisms driving DNA fragmentation through depurination post-mortem are still unclear. However, tracking base compositional profiles within a sample set of 80 bones spanning the last 60,000 years, Sawyer and colleagues (71) have observed depurination mainly through adenine residues in their most recent samples (<500 years), instead of guanine residues in their older samples. These authors proposed that the action of various nucleases could drive such differences. Finally, it is important to remember that base composition profiles can be affected by the different molecular treatment used prior to, and during library construction. In particular, DNA ligases can show sequence-context dependent activities (37). More importantly, USER treatment of aDNA extracts results in the excision of most uracil residues present in aDNA templates (excepting those located at template termini) (62). First, the uracil DNA Glycosylase (UDG), present in the USER enzymatic mix, removes deaminated nucleotidic bases, leaving abasic sites instead of uracils. Abasic sites then provide suitable templates for the Endonuclease VIII (EndoVIII), also present in the mix and the DNA strand becomes further fragmented. Since this series of enzymatic is initiated at uracil residues (which are formed following post mortem cytosine deamination), the base composition profiles obtained from USER-treated DNA extracts are expected to show an excess of cytosine residues at the genomic position just preceding aDNA fragments starts.

4. Estimating contamination originating from present-day human mitochondrial DNA

4.1. Definitions and preliminaries

This section assumes that 1) the sequencing data has been prepared according to the methodology described in section 2 and 2) that the experimenters have verified the presence of post-mortem DNA damage consistent with the type of library preparation used, as described in section 3. Furthermore, the section is mostly aimed at the analysis of ancient hominin samples such as anatomically modern humans, Neanderthals and Denisovans.

A key step in ascertaining the amount of endogenous DNA is estimating the amount of present-day human contamination in the data, defined as the fraction of the total material that stems from the contemporaneous humans that were involved in either excavation, DNA extraction, or library preparation, or handling of the sample. More specifically, this fraction can be measured in two ways:

1) the fraction of the total DNA fragments that stem from present-day humans.
2) the fraction of the total bases that stem from present-day humans.

As previously documented (69), the average length of endogenous DNA fragments are expected to be shorter than the one observed for present-day human contaminants. An estimate of present-day human contamination might thus not be exactly the same depending on the definition used, and the definition used in the remainder of this chapter is indicated in each individual subsection

4.2. Contamination estimates for the mitochondrial genome

4.2.1 Described methods in the literature

Due to the relative abundance of mitochondrial to nuclear DNA and the haploid nature of the mitochondrial genome, mitochondrial DNA is routinely used in the field of aDNA. There are,

however, two main hurdles that impede the reconstruction of the mitochondrial genome from a set of aligned aDNA fragments. Firstly, the presence of post mortem damage can create false signals of mutation, especially at low coverage (see (72) for further discussion). Secondly, at high levels of present-day human contamination, a simple consensus call might predict the contaminant base due to its higher relative abundance (see Figure 2). Based on the hypothesis that a higher rate of post-mortem damage is expected in endogenous fragments, compared to present-day human contamination, algorithms have been designed to isolate endogenous DNA fragments showing signs of post mortem damage. These have been found to be able to reliably predict the endogenous mitochondrial genome for even highly contaminated data sets (73). The computational strategy to predict the endogenous mitochondrial genome depends on the level of present-day human contamination. At low levels of contamination, a simple consensus call is sufficient whereas, at higher levels of contamination, a more intricate algorithm is required. The importance of knowing whether a sample has high levels of present-day human contamination has therefore spurred the development of computational techniques for quantifying it.

One of the first methods to estimate the amount of present-day human contamination for Neanderthals samples to have been described in the literature relied on the principle of diagnostic positions (74). The same principle has since been extended to other cases, such as detecting probe contamination levels in bovine target-enrichment data (75). Originally, diagnostic positions were defined sites in the mitochondrial genome where most humans (≥99%) share one base and where every sequenced Neanderthals share a different base. The underlying assumption is that the reads matching the base observed for humans must represent contamination, whereas reads matching the Neanderthal base are endogenous in origin. Although this approach is suited for Neanderthal samples with high mitochondrial coverage, it suffers from a few shortcomings. First, it is impossible to include the probability

of seeing an endogenous fragment matching the contaminating base either due to sequencing errors or post mortem DNA damage, in the computation of the contamination estimate. Second, it is impossible to know, *a priori,* which bases on the mitochondrial genome are fixed in both Neanderthals and humans, which might bias any analyses of low-coverage data from an individual belonging to either an individual from a previously unsampled Neanderthal population or to a novel archaic group. Finally, this approach is not readily applicable to sequence data generated from anatomically modern humans.

A strategy for estimating present-day human contamination for ancient human samples is to look at private sites, which are defined to be unique to the individual of interest (18). The number of divergent bases found at private sites can provide an estimate of present-day human contamination. Again, this strategy is plagued by the inability to quantify error probabilities and include these in the computation. Furthermore, it is possible that a single individual might have very few unique variants, leaving large error bounds on the contamination, even if high coverage data are available.

In an attempt to quantify error probabilities and include them into the calculation of contamination rates, a maximum-likelihood method was proposed in (30) and is applicable both to modern humans and archaic hominins. This method used as input the predicted endogenous mitochondrial genome, the aligned aDNA fragments and a database of potential contaminants. This method has the advantage of allowing for multiple, different mitochondrial genomes as putative contaminants, each of which may contribute different amounts to the present-day human contamination. A caveat is that this method has a single error parameter, representing both deamination and sequencing errors. This error parameter is estimated on regions of the mitochondrial genomes that the program calls as monomorphic across all individuals including the endogenous one. This method also requires as input the

sequence of the endogenous mitochondrial genome, which cannot be derived from simple consensus calls for highly contaminated samples.

4.2.2 Bayesian maximum *a posteriori* estimate

The task of estimating present-day human contamination is intertwined with the problem of inferring the endogenous mitochondrial genome. Conversely, the strategy used to infer the endogenous genome is highly dependent on the level of present-day human contamination. The software package schmutzi, freely available at http://grenaud.github.io/schmutzi/, contains two main modules to perform both tasks based on Bayesian maximum *a posteriori* algorithms (69). The first module, endoCaller, predicts the endogenous mitochondrial genome given a set of deamination parameters, a contamination prior and the fragment length distribution for both the endogenous and present-day human contamination. The second module, mtCont uses the endogenous mitochondrial genome as well as a representative database of known human mitochondrial genomes to predict the rate of present-day human contamination. schmutzi requires a Unix environment along with a C++ compiler, a Perl interpreter and a recent version of R with certain packages. This section also assumes that the user has a mapped, sorted and indexed BAM file prepared as described in section 2.

A wrapper-script allows the user to iteratively run endoCaller and mtCont. As endoCaller is the first program called as part of this iterative procedure, it requires a prior for the amount of present-day human contamination as well as rates of post mortem damage for the endogenous fragments. This prior information is supplied by another submodule called contDeam.pl, which is detailed in the next section.

Prior based on deamination patterns

Estimating the rate of present-day human contamination using the deamination patterns has first been suggested by (15). The schmutzi package contains an implementation under a probabilistic framework.

The principle undergirding an estimation of present-day contamination using the frequency of C to T misincorporations at the ends of the fragments is illustrated in Figure 3. Assuming that the misincorporation rates of the endogenous fragments are known, and that the fragments stemming from the present-day human contamination are free of post-mortem damage, it follows that the final observed frequency of C to T misincorporations could provide an estimate of the proportion of both sources. The misincorporation rate for the endogenous fragments can be obtained using a double conditioning procedure whereby fragments with deamination on their 5' end are retained, and the deamination on the other end is measured and vice versa. The estimate of present-day human contamination would be on a per-fragment instead of a per-base basis (see section 4.1).

To perform this initial estimate, use the following command:

% contDeam.pl --lengthDeam [length] --library [library type] --out [output prefix] [mt reference] [input bam file]

The script contDeam.pl has the following command line arguments that will modify how this estimate is computed:

- **--lengthDeam [length]**: The integer **[length]** is the number of nucleotides that will be considered by the algorithm when estimating present-day human contamination. This length can vary from 20-40 bases for double-stranded protocols to 2-5 bases for single-stranded protocols with UDG treatment.

- **--library [type]**: The type of library preparation protocol used to account for the type of damage patterns to expect. Currently, there are two possible values for **[type]**, either "single" for Meyer et al. (2012) procedures, or equivalent, and "double", for Meyer and Kircher (2010) procedures, or equivalent.

The present-day human contamination estimate will be written to a file called **[output prefix].cont.est** and the shape of the posterior probability distribution will be plotted to the file **[output prefix].cont.pdf**. The rates of post-mortem damage for the endogenous fragments will be produced for the 5' and 3' are written to the files called **[output prefix].endo.5p.prof** and **[output prefix].endo.3p.prof**, respectively. The assumptions and limitations of this procedure are discussed in (69).

Iterative consensus call and contamination estimates

The output of contDeam.pl, which was detailed in the previous subsection, is used as input for endoCaller. This sub-module aims at producing the sequence of the endogenous mitochondrial genomes given a prior on present-day human contamination, rates of post-mortem damage as well as fragment length distribution for the endogenous and contaminant independently. It is also worth noting that this module is also useful for calling non-hominin haploid genomes where the impact of post-mortem deamination on consensus called needs to be mitigated (e.g. animal mitochondrial genomes, chloroplasts, viruses).

If endoCaller is used to mitigate the impact of post-mortem damage on the consensus call, the following command can be used:

% endoCaller -seq [output prefix].fa -log [output prefix].log -deam5p [output prefix].endo.5p.prof -deam3p [output prefix].endo.3p.prof [mt reference] [input bam file]

where both **.prof** files were obtained from contDeam.pl, the fasta sequence of the consensus call can be found in **[output prefix].fa** and per base posterior probabilities can be found in this file **[output prefix].log**.

If both post mortem damage as well as the presence of present-day human contamination need to be mitigated, the following command can be used:

% endoCaller -cont [contamination prior] -single -seq [output prefix].fa -log [output prefix].log -deam5p [output prefix].endo.5p.prof -deam3p [output prefix].endo.3p.prof [mt reference] [input bam file]

The **-single** flag represents the prior assumption that contamination stems predominantly from a single mitochondrial genome (see description of the various options for endoCaller below).

The endoCaller program accepts the following options to modify the calculation of the endogenous base and (potentially) the contaminant base:

- **-deam5p [file]:** The **[file]** contains the frequency of misincorporations due to deamination starting from the 5' end for the endogenous fragments. Equivalent options are found for the 3' end of endogenous fragments. Again, equivalent options are found for contaminant fragments to account for the possibility of having deamination for contaminant fragments.
- **-cont [c]:** this option specifies the prior on the rate of present-day human contamination.

- **-single:** this flag tells the algorithm to assume that there is a single mitochondrial genome for the present-day human contaminant. This is especially useful at high levels of present-day human contamination. When this flag is not used, endoCaller can only mitigate the presence of low amounts of present-day human contamination (below ~30%).

Once the endogenous mitochondrial genome has been obtained, present-day human contamination can be estimated using *mtCont*, as follows:

% mtCont -deam5p [output prefix].endo.5p.prof -deam3p [output prefix].endo.3p.prof [output prefix].log [mt reference] [input bam file] [contaminant profile 1] [contaminant profile 2] ...

where **[output prefix].log** is produced by endoCaller, both **.prof** files were obtained from *contDeam.pl*. The files **[contaminant profile N]** are allele frequencies for different sub-haplogroups to be considered as potential contaminants. Allele frequencies for a non-redundant set of human haplogroups are provided with the software package. The estimate of present-day human contamination provided by *mtCont* is on a per nucleotide base (definition 2) in section 4.1.

The *mtCont* program can accept the following options to modify the calculation of the rate of present-day human contamination:

- **-deam5p [file]**: The **[file]** contains the misincorporation rates due to deamination starting from the 5' end for the endogenous fragments.
- **-deam3p [file]:** Same as above except for the 3' end of endogenous fragments.

To obtain accurate results, users should use endoCaller and mtCont iteratively until results are stable. To facilitate this process, a wrapper script calls both programs iteratively. This wrapper script can be invoked as such:

% schmutzi.pl --notusepredC --uselength --ref [mt reference] --out [out prefix]_npred [output prefix] [path to schmutzi]/eurasian/freqs/ [input bam file]

where options **--notusepredC** and **--uselength** instruct the program not to use the predicted contaminant as a contaminant profile (see comments about *mtCont* above) and used the length of the aDNA fragments to help the identification of the endogenous base. The term **[output prefix]** is the same used by *contDeam.pl*. The **[path to schmutzi]/eurasian/freqs/** is a path to a set of potential mitochondrial contaminants that is provided with the software package. As the name indicates, it focuses on Eurasian mitochondrial sequences. However, users can build their custom library of putative contaminants and the path to this new database can be modified.

The wrapper script should be run once again without the **--notusepredC** option. This will tell the program to use to predicted contaminant as a record in the set of potential contaminants:

% schmutzi.pl --uselength --ref [mt reference] --out [out prefix]_wpred [output prefix] [path to schmutzi]/eurasian/freqs/ [input bam file]

This option is preferable if there is a high level (above ~30%) of present-day human contamination. Since this is not known beforehand, running schmutzi.pl twice with and without the **--notusepredC** option is highly recommended.

Executing the commands above will create output files with the **[out prefix]_npred** and **[out prefix]_wpred** prefixes. If the iterative procedure is successful, it will create files with the suffix **_final.cont.est** for the final present-day human contamination where the posterior probability will be plotted as a file with the **_final.cont.pdf** suffix. The log for the final *mtCont* call will be stored in **_final_mtcont.out**, which allows users to see the most likely haplogroup for the contaminant. Finally, the file with the **_final_endo.fa** suffix will contain the unfiltered fasta prediction for the endogenous genome and **_final_endo.log** contains the per-base likelihood of being the endogenous base.

More generally, the wrapper script schmutzi.pl can accept the following options:

- **--contprior [c]**: This is the value in the first iteration for the rate of present-day human contamination to be used as prior for endoCaller. When unspecified, the value predicted by *contDeam.pl* is used as prior. This option is only recommended if the iterative procedure has not converged.
- **--uselength**: This flag instructs the algorithm to also use the length of the aDNA fragments in the computation to distinguish the endogenous base from the contaminant one. The use of this flag can lead to more accurate results for highly contaminated samples as present-day contaminants are expected to show longer size distributions than short aDNA fragments
- **--lengthDeam [bp]**: this option instructs the algorithm to consider the **[bp]** bases away from the 5'/3' ends to be the most deaminated. By default, this is given by the output from *contDeam.pl*.

We also recommend running *log2fasta* on the file with the **_final_endo.log** suffix (endogenous mitochondrial genome for the final iteration) using different thresholds for the base prediction quality. The resulting sequences in fasta format can be used in a tool such as HaploGrep2.0 (76) to assign haplogroups. If the predicted haplogroup for the low-quality prediction as well as the higher-quality one is stable, this indicates that this prediction is likely free of the influence of contaminant bases.

4.2.3 Limitations

The schmutzi software package is able to quantify present-day human contamination as well as infer the endogenous mitochondrial genome even in highly contaminated datasets. However, this package faces certain limitations. First, there is no guarantee that the algorithm will converge. This is especially the case for highly contaminated samples with medium coverage where the algorithm can oscillate between two equally likely states. Furthermore, the algorithm can converge on an incorrect value especially for very low coverage samples (below ~5-fold). Second, this algorithm does not take demography into account when predicting bases. This would be highly useful for low coverage samples as the likelihood of a certain base changes depending on which haplogroup the endogenous mitochondrial belongs to. Third, unlike contamMix (30), this algorithm does not allow for the possibility of having multiple contaminants. Fourth, if the genetic distance between the endogenous and present-day human contaminant mitochondria is limited, on the order of a few divergent base pairs, the algorithms will have insufficient data to operate on which will result in the loss of statistical power. The consequences of this effect are worse at low coverage.

Finally, both contamMix and schmutzi are predicated on having the endogenous mitochondrial genome, where the former uses it as input and the latter jointly predicts it along with estimating contamination. However, for the prediction of the endogenous mitochondrial genome, a remaining challenge is to be able to distinguish between the endogenous and

contaminant bases if few chemical properties allow us to distinguish them. This could occur, for instance, if the contaminant DNA fragments have some levels of post mortem damage, e.g when a previously contaminated sample had been treated with harsh chemicals. Alternatively, his could be the case if the endogenous DNA fragments have little to no damage for highly contaminated samples, e.g. when one ancient sample is the contamination source for another one. Furthermore, it is worth mentioning that the potential contribution on contamination estimates of significant heteroplasmy rates in ancient samples has not been considered.

5. Estimating contamination in human nuclear genomics data sets

Estimation of present-day human contamination in nuclear genomic data faces several unique challenges compared to working in mitochondrial data sets. First, the relative per-site coverage of the nuclear genome is likely to be low compared to that of the mitochondrial genome. Second, heterozygous sites may exist in both the endogenous and contaminant individual(s), and this can complicate calculations.

An initial idea to quantify present-day human contamination for ancient nuclear genome data relied on sites where derived alleles were practically fixed in present-day humans (38). The method computed the likelihood of observing the data given a certain contamination rate and three different models of endogenous genotypes: homozygous derived, homozygous ancestral and heterozygous for both alleles. The model was later refined using a bivariate distribution of error probabilities (30).

Restricting the analysis to the X chromosome, which is found in only one copy in males, provides a convenient simplification. A probabilistic algorithm using monomorphic regions of the X chromosome and derived allele frequencies to predict the probability of seeing a certain base due to contamination was described (77) and implemented as part of

ANGSD (78). This procedure is the first of two described below. The second method described, called DICE (79), was originally developed to quantify present-day human contamination for nuclear data from ancient samples, and co-estimate some demographic parameters (such as drift times and admixture rates) linking the ancient genome to a present-day human population. While the method based on the X chromosome only works on male individuals, DICE can work on both males and females.

5.1. Contamination estimates using the X chromosome

The software ANGSD allows for the analysis of next-generation data by allowing users to compute various statistics for population genetics directly from BAM files. This software is freely available from https://github.com/ANGSD/angsd and is written predominantly in C++. Installation instructions are provided with the software. ANGSD requires a Unix environment along with a C++ compiler and a recent version of R with certain packages. The following instructions assume that the user has a mapped, sorted and indexed BAM file prepared as described in section 2.

Apart from statistical tests for population genetics, ANGSD can be used to obtain estimates of present-day human contamination using the X chromosome for males. We thus first need to ascertain if the ancient sample is likely a male, for instance using X-to-Autosome coverage ratios and the methodology described in (80). Then, we need to generate a binary count for the X chromosome before proceeding with the statistical estimation of contamination. These two steps are described below.

First, base counts can be generated using the following command on sorted and indexed BAM file:

% angsd -i [bam file] -r X:5000000-154900000 -doCounts 1 -iCounts 1 -minMapQ 30 -minQ 20

where **[bam file]** is the input BAM file, the **-r** specifies the region to consider. The -r flag limits the analysis to a certain range of coordinates on the chromosome to avoid telomeric and centromeric regions.The **-minMaqQ** and **-minQ** apply filters for the mapping quality and base quality respectively. A file called **angsdput.icnts.gz** should be produced.

The statistical estimate of present-day human contamination is performed by running the following command:

% misc/contamination -a angsdput.icnts.gz -h RES/HapMapChrX.gz

where **misc/contamination** is an executable found in ANGSD installation directory. The file **angsdput.icnts.gz** is produced by the previous command above whereas **RES/HapMapChrX.gz** is a mappability and HapMap file for the X chromosome found in the ANGSD installation directory. The output is produced on the console and includes the maximum-likelihood estimate of the present-day human contamination as well as the corresponding confidence interval.

5.2. Demographic inference and contamination estimates (DICE)

DICE calculates the probability of finding an allele as a heterozygous or homozygous state in the ancient genome, given the demographic history of the population to which the genome belongs. Since this demography is not known *a priori*, it is jointly inferred along with the contamination rate and another parameter for the error rate. This error rate is an all encompassing term representing mismatches due to either sequencing errors, mismappings or

nucleotide misincorporations due to deamination. This demography is inferred by defining an **anchor** population that shares a common ancestor with the ancient genome, while the contamination rate is estimated using a second putative **contaminant** population (which could also be the same as the **anchor**). Allele frequencies in suitable input format for DICE - obtained from data from the 1000 Genomes project (81) - and the source code for the program, together with installation instructions, are made available at: http://grenaud.github.io/dice/.

The sorted and indexed BAM file of reads aligned against the human reference genome has first to be transformed to the native *DICE* format. This is done through the **BAM2DICE** program and requires the user to select which present-day human population(s) should be considered as putative contaminants. The resulting files are ready to be used by the *DICE* program, which runs a Markov Chain Monte Carlo (MCMC) algorithm to produce posterior distributions of all parameters, with the output for each chain written to an output file. An R script called **logs2text.R** can be used to parse this file and produce a text file which contains the posterior mode and posterior confidence intervals for each putative present-day human contaminant population and also reports the population with the highest posterior probability of being the contaminant. The estimate of present-day human contamination produced by *DICE* is on a per nucleotide basis (definition 2) in section 4.1.

The process of format conversion, executing *DICE,* and parsing the resulting output can be streamlined using the following script:

% dice2Makefile.pl --anch [population code of anchor] --out [output prefix] --reg [regions to include] --alfr [path to dice folder]/alleleFreqNuc/ [human genome reference] [bam file 1] [bam file 2].... > Makefiledice

where **[path to dice folder]/alleleFreqNuc/** contains the allele frequency files and **[population code of anchor]** is the file prefix for the population used as anchor in the allele frequency folder**.** For instance, to use the West African Yoruba as the anchor population, one can use **--anch YRI**. The regions to include are specified by **--reg** and we recommend to use for example **[path to dice folder]/mapability/all.1kregions.gz** as these are regions with a high mappability score (see the methodology described at http://lh3lh3.users.sourceforge.net/snpable.shtml). The **--alfr** option specifies the allele frequency folder. A readily available set of allele frequencies for various populations can be downloaded using the Makefile in *DICE*'s main directory (**[path to dice folder]/alleleFreqNuc/**)**.** The **[human genome reference]** should be the same as that used for mapping.

Currently, the *dice2Makefile.pl* script accepts the following options that can modify the computations:

- **--cont:** This option allows the user to specify the populations to be used as putative sources of present-day human contaminant. By default, the program takes all the populations available in a directory. If computational resources are limited, restricting the number of putative contaminants might be an adequate option.
- **--tau**: This refers to the two drift parameters - tauA and tauC - separating the ancestral population from the population to which the ancient genome belongs and the present-day population. The drift parameter is equal to the time in generations divided by the effective population size in each population. By default, DICE explores a parameter space between a minimum of 0 and a maximum 1.It is possible that the MCMC chains have reached the upper bound in the parameter space, when one of the daughter populations has a very high drift time separating it from the ancestral population. This can be seen if the estimates for either one of the tau parameters reach 0.99 and remain

blocked in this state. If that is the case, the value of the tau parameters can be increased above 1 using this option.

The commands can be now launched using the following:

% make -f Makefiledice -j [number of cores]

where the **[number of cores]** is the number of threads to run simultaneously on a multi-core machine. This process will create, for each BAM file, an output file with the following suffix **_Cont_[CNT]_Anch_[ACH].dice.out.gz** containing the posterior samples from the MCMC chain. The **[CNT]** and **[ACH]** are the population codes for the present-day human contaminant and for the anchor. Another file is produced with the suffix **.dice.txt,** which contains posterior modes for each contaminant population and reports the contaminant population showing the highest posterior probability.

Another program can be used to summarize this information as barplots with whiskers to represent the 95% posterior confidence intervals:

% logs2plot.R [prefix]_Cont_*_Anch_[ACH].dice.out.gz

Where **[ACH]** is the population code for the anchor population and **[prefix]** is the file prefix for all output files for a given BAM file. This commands creates files **[prefix]_c.pdf** for the estimates of the contaminant sorted with respect to the posterior probability. Comparisons between the different contaminant populations should be done when the same anchor population is used hence the use of a fixed **[ACH]** parameter. This command is part of the Makefile and is run automatically.

5.3. Limitations

The probabilistic model implemented as part of *DICE* will produce posterior samples of the drift, error and contamination parameter under a given set of anchor and contaminant populations. The user can try different contaminant panels and find which panel yields the highest posterior probability. Though the panel with the highest posterior probability is not statistically guaranteed to be the most likely contaminant, simulations show that, for a range of demographic scenarios tests, this is the case (79). *DICE* can return reliable posterior samples with an ancient genome of at least 3-fold coverage, so long as the drift times separating the endogenous and present-day human contaminant are sufficiently large, for example as in archaic hominins such as Neanderthals. However, it is unlikely to produce reliable posterior samples if the ancient genome and the contaminant are closely related.

6. Conclusions

The goal of this subsection is to give end-users tools to ascertain the authenticity of their data from hominin samples where downstream analyses can be affected by present-day human contamination. We should point out that the tools described in this section are by no means a panacea and have several limitations. As the field aDNA expands the need for cutting edge data authentication methods also increases. It is important to keep in mind that such contamination estimates are not a replacement for negative controls during data production.

Additionally, metaBIT (82) can be used to validate the quality of libraries from the same extract by comparing their microbial diversities. Ideally, different libraries from the same extract should have a similar microbial diversity. However, contamination events during

laboratory work could shift the microbial diversity towards the contaminating source, which can be detected by metaBIT.

The aforementioned techniques for estimating present-day human contamination can yield false positives under certain circumstances. Furthermore, the use of certain molecular tools during sample preparation as well as sample age and/or molecular preservation levels lead to little or no nucleotide misincorporations due to deamination (71). If no fragment length filters are used, highly fragmented samples may align equally well to bacterial genomes and the one of the endogenous species. Such factors can lead to false positives whereby the material is endogenous but flagged as contaminant.

This chapter has mostly focused on the estimation of present-day human contamination in ancient hominin samples. Although this extends beyond our scope, it is important to mention the potential contribution of contaminant DNA regardless of the source on downstream analyses for various types of endogenous organisms. Viruses, for example, tend to have few genome reference sequences or custom assembly algorithsm for ancient DNA, and have a rapid rate of molecular evolution that is likely to complicate assembly. Bacteria are present in nearly all aDNA extracts, but it is challenging to determine whether the becteria were present when the organism was alive or at what point they infiltrated the samples. In addition, the representation of the diversity of microbial life in existing genomic reference databases is poor compared to the actual diversity of microbes across the planet, which complicates identification and assembly. Plants can have large polyploid genomes that make the presence of contaminants, particularly from closely related species, challenging to identify. These and other challenges need to be taken into consideration when interpreting results from aDNA experiments, even when care is taken to authenticate data using these or other pipelines.

7. Notes

Mapping and mapDamage

- Due to the circularity of the mitochondrial genome, fragments that span the junction in the reference in fasta format might not be aligned properly. This can lead to a spurious drop in coverage in that region and a lack of resolution of those regions. A possible way to mitigate this is to extend the mitochondrial reference by copying the initial 1000bp bases at the end of the fasta reference and mapping the aDNA fragments against this new reference. A tool like bam-rewrap (https://bitbucket.org/ustenzel/biohazard) can allow users to specify the initial length of the mitochondria and copy the alignments exceeding the original length back to their original location at the beginning of the reference.
- It is important that the correct adapter sequences be specified when running the PALEOMIX pipeline or when trimming sequences using another program. PALEOMIX will, by default, trim standard Illumina paired-end adapter sequences. Please refer to the PALEOMIX and AdapterRemoval documentation for how correctly specify the adapters used during sequencing.
- It is possible that there is a low number of aDNA fragments that align to the endogenous reference. There three possibilities: 1) if these alignments are spread across the genomic reference, it is possible that there is low endogenous content or very low complexity. 2) if coverage of the endogenous reference is very unequal, it is possible that these alignments are simply microbial fragments aligning randomly to the reference. Especially for short microbial fragments, the probability of aligning to a reference by sheer serendipity is a lot higher than for longer fragments (83). 3) if the coverage is higher in exonic regions for instance, it is possible the reference might be too distant in terms of evolution time.

- If the fragment length distribution drops precipitously around the read length and single-end reads were used, this is expected as it is impossible to measure longer fragment lengths without the use of paired-end reads.
- If a large peak is observed at a particular length in the fragment length distribution, it is recommended to look at the fragments of that very specific length. It is possible that such sequences are merely chimeric adapters, excepting the case where this corresponds to the raw read length.
- If no deamination on either end is seen and no library preparation protocol can explain this lack of deamination, it is possible that the aligned fragments are merely present-day human contamination. When working with hominins, present-day human contamination can align at the same rate as the endogenous material. In that case, it is advised to look at the mitochondrial haplogroup for instance, and determine whether the assumed demography of the endogenous sample jibes with the determine haplogroup. However, assumptions about the demography of the endogenous sample can often be misleading.
- As of mapDamage2.0, mapDamage will generate a table of per-position mismatches for each sequence in the reference genome. For poorly assembled genomes, this may result in exceedingly large tables and even program failure, and using the --merge-reference-sequences option is recommended. This will record all per-position mismatches in a single table.

Estimating contamination

- As mentioned above, the lack of post mortem damage can indicate the absence of endogenous aDNA. If the USER treatment was performed, it is possible to have very little residual deamination. If that is the case, it will be difficult for endoCalller to

have sufficient statistical power to predict the endogenous base versus the contaminant one. If the endogenous mitochondrial genome is not predicted accurately, estimates of present-day human contamination will be unreliable. It is therefore recommended to try to *(i)* predict a consensus using endoCaller, *(ii)* ascertain the phylogenetic placement of that consensus using a tool such as HaploGrep2.0 (76), and; *(iii)* look forprivate mutations (i.e. unique to that sample or defining mutations for the alleged subhaplogroup).

- It is also possible that the iterative function of schmutzi does not converge. In such a case, it is possible that the algorithm oscillates between two almost equally likely models. It is recommended to use endoCaller with a low estimate of present-day human contamination and rerun endoCaller again with a higher one. Both predictions should be compared for divergent sites and phylogenetic information should be used to determine which consensus appears as the most likely.
- When using endoCaller, the program can predict the contaminant (using option **-single**) and the iterative procedure can use this prediction as a record in the database of putative contaminants (usually specified via **[path to schmutzi]/eurasian/freqs/**). This is normally done by default by schmutzi.pl. The final present-day human contamination obtained using default parameters might differ from the one obtained without the prediction of the contaminant (see option **--notusepredC)**. If the estimate obtained using the option **--notusepredC** is higher when enabling contaminant prediction, the first estimate is more likely correct as using the **--notusepredC** option tends to underestimate present-day human contamination. If the prediction of contaminant is enabled and produces a high estimate (above ~20%), there are two options. If the estimate obtained using the option **--notusepredC** is very low (below ~5%), the first estimate, without the use of this option, is likely an overestimate and

probably wrong. If the estimate obtained using the option **--notusepredC** is high (above ~10%), then it is likely an underestimate and the estimate obtained without using the option **--notusepredC** is likely correct.

- The methods to quantify present-day human contamination using the X chromosome implemented in ANGSD can work at a depth of coverage as low as 0.5X on the chromosome X if contamination is lower than 10%. Please note that this method is aimed at male individuals only. This method also underestimates contamination rates above 15% unless the depth for chromosome X is greater than 1X.
- The contamination estimate implemented in ANGSD is underestimated when the assumption about the 'contaminant population' is wrong. For instance, if the empirical contaminant is of European ancestry and the method is run using African frequencies. Therefore, it is recommended to use different panels which can provide some insight as to the about the ancestry of the contaminant.
- When using DICE, it is possible that there are major discrepancies between the estimates of contamination when using different populations as the putative contaminant source. Ideally, when ordering the contamination rates with respect the posterior mode, the populations that are closest to the true contaminant should have higher posterior modes. If, for instance, the contaminant individual had Han ancestry, then CHB should be the contaminant with the highest posterior mode, followed by CHS, JPT, followed by other Asians, Europeans and finally Africans. It is likely that, if there are major discrepancies for the present-day human contamination estimate (in the order of 15-20%) between various populations, the algorithm has not converged to the true posterior distribution and the estimates may be unreliable. This can be the case with very low coverage samples.

- When running *DICE*, the user has to pick a population as anchor. Please note that different putative contaminant populations should be compared when using the same anchor population. The anchor population is used to compute the drift parameters linking it to the ancient sample. Ideally, a population that has not received gene flow from the ancient sample should be used. For instance, the Yoruba (YRI) anchor can be used if the ancient sample is a Neanderthal with potential European contamination. It is also possible that the default boundaries for the drift parameters might not be suitable for a particular ancient sample when using *DICE* (please see comment about the tau parameter in the *dice2make.pl* section). This will be seen when the MCMC chain reaches the upper bound for one of the tau parameters, which is set by default at 1.0. This problem can be circumvented by increasing the range of the parameter space for tau via the **--tau** option in *dice2make.pl*.
- There is also the possibility of admixture from the archaic population, to which the ancient sample belongs, into the anchor population, which might result in incorrect estimates of present-day human contamination. To account for this, *DICE* also offers the possibility of estimating admixture from the ancient samples via parameters **-3p, -aR** and **-aT** of the main executable program. Please refer to the software manual for further information regarding these parameters.

Acknowledgment

We would like to thank Fernando Racimo for comments and suggestions and José Victor Moreno Mayar and Thorfinn Sand Korneliussen for their insights into the contamination

method using the X chromosome. This work was supported by the Danish Council for Independent Research, Natural Sciences (Grant 4002-00152B); the Danish National Research Foundation (Grant DNRF94); Initiative d'Excellence Chaires d'attractivité, Université de Toulouse (OURASI); the Villum Fonden miGENEPI research project;, and; the European Research Council (ERC-CoG-2015-681605).

Figure 1. mapDamage2.0 misincorporation plots for modern DNA (a); aDNA sequenced using end-repair, blunt-end adapter ligation, and nick fill-in (b); aDNA sequenced using end-repair, adapter ligation at AT-overhangs, and nick fill-in, characterized by a decrease in the observed post mortem DNA damage rate at the first and last position (c); and aDNA that has been USER-treated (d). Red represents the rate of C>T substitutions relative to the reference, blue represents the rate of G>A substitutions. Samples (a) to (c) were sourced from (37), sample (d) was sourced from (38).

Figure 2. Schematic representation of the problem of determining the haplotype of the endogenous mitochondrial genome in the presence of present-day human contamination for aDNA data sets. The left column represents the true haplotype of the endogenous and contaminant mitochondrial genomes. The rightmost column shows the observed aDNA fragments aligned to the mitochondrial reference.

Figure 3. Representation of the effect of having 50% present-day human contamination on the final, observed set of aDNA fragments. In this example, the frequency of C to T misincorporations at the 5' end for the endogenous DNA is about 40% whereas the present-day human contamination does not have any deamination. Giving an equal mixture from both sources, the final observed dataset contains a rate of deamination equal to half the original endogenous one.

References:

1. Ermini L, Der Sarkissian C, Willerslev E, Orlando L (2015) Major transitions in human evolution revisited: a tribute to ancient DNA. J Hum Evol 79:4–20. doi: 10.1016/j.jhevol.2014.06.015
2. Llamas B, Fehren-Schmitz L, Valverde G, et al (2016) Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. Sci Adv. doi: 10.1126/sciadv.1501385
3. Librado P, Der Sarkissian C, Ermini L, et al (2015) Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. Proc Natl Acad Sci United States Am 112:E6889–E6897. doi: 10.1073/pnas.1513696112
4. Frantz LAF, Mullin VE, Pionnier-Capitan M, et al (2016) Genomic and archaeological evidence suggest a dual origin of domestic dogs. Sci 352:1228–1231. doi: 10.1126/science.aaf3161
5. MacHugh DE, Larson G, Orlando L (2016) Taming the Past: Ancient DNA and the Study of Animal Domestication. Annu Rev Anim Biosci. doi: 10.1146/annurev-animal-022516-022747
6. Der Sarkissian C, Ermini L, Schubert M, et al (2015) Evolutionary Genomics and Conservation of the Endangered Przewalski's Horse. Curr Biol : CB 25:2577–2583. doi: 10.1016/j.cub.2015.08.032
7. Da Fonseca RR, Smith BD, Wales N, et al (2015) The origin and evolution of maize in the Southwestern United States. Nat plants. doi: 10.1038/nplants.2014.3
8. Bos KI, Schuenemann VJ, Golding GB, et al (2011) A draft genome of Yersinia pestis from victims of the Black Death. Nature. doi: 10.1038/nature10675
9. Wagner MR, Lundberg DS, Coleman-Derr D, et al (2015) Corrigendum to Wagneret al.:

Natural soil microbes alter flowering phenology and the intensity of selection on flowering time in a wild Arabidopsis relative. Ecol Lett. doi: 10.1111/ele.12400

10. Ramos-Madrigal J, Smith BD, Moreno-Mayar JV, et al (2016) Genome Sequence of a 5,310-Year-Old Maize Cob Provides Insights into the Early Stages of Maize Domestication. Curr Biol : CB 26:3195–3201. doi: 10.1016/j.cub.2016.09.036
11. Rasmussen S, Allentoft ME, Nielsen K, et al (2015) Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. Cell 163:571–582. doi: 10.1016/j.cell.2015.10.009
12. Orlando L, Ginolhac A, Zhang G, et al (2013) Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499:74–78. doi: 10.1038/nature12323
13. Dabney J, Knapp M, Glocke I, et al (2013) Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci United States Am 110:15758–15763. doi: 10.1073/pnas.1314445110
14. Meyer M, Arsuaga J-L, de Filippo C, et al (2016) Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature 531:504–507. doi: 10.1038/nature17405
15. Meyer M, Fu Q, Aximu-Petri A, et al (2014) A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature 505:403–406. doi: 10.1038/nature12788
16. Hofreiter M, Serre D, Poinar HN, et al (2001) Ancient DNA. Nat Rev Genet 2:353–9. doi: 10.1038/35072071
17. Briggs AW, Stenzel U, Johnson PL, et al (2007) Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci U S 104:14616–21. doi: 10.1073/pnas.0704665104
18. Green RE, Malaspinas A-S, Krause J, et al (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell 134:416–426. doi:

10.1016/j.cell.2008.06.021

19. Gilbert MTP, . Wilson, A. S. ,. Bunce, M. ,. Hansen, A. J. ,. Willerslev, E. ,. Shapiro, B. ,. Higham, T. F. G. ,. Richards, M. P. ,. O'Connell, T. C. ,. Tobin, D. J. ,. Janaway, R. C. ,. Cooper, A. (2004) Ancient mitochondrial DNA from hair [1]. Curr. Biol. 14:
20. Pilli E, Modi A, Serpico C, et al (2013) Monitoring DNA contamination in handled vs. directly excavated ancient human skeletal remains. PloS one. doi: 10.1371/journal.pone.0052524
21. Korlević P, Gerber T, Gansauge M-T, et al (2015) Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. BioTechniques 59:87–93. doi: 10.2144/000114320
22. Guschanski K, Krause J, Sawyer S, et al (2013) Next-generation museomics disentangles one of the largest primate radiations. Syst Biol 62:539–554. doi: 10.1093/sysbio/syt018
23. Pruvost M, Schwarz R, Correia VB, et al (2007) Freshly excavated fossil bones are best for amplification of ancient DNA. Proc Natl Acad Sci United States Am 739–44. doi: 10.1073/pnas.0610257104
24. Champlot S, Berthelot C, Pruvost M, et al (2010) An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. PloS one. doi: 10.1371/journal.pone.0013042
25. Serre D, Langaney A, Chech M, et al (2006) No Evidence of Neandertal mtDNA Contribution to Early Modern Humans. In: Early Mod. Humans at Morav. Gate. Springer Vienna, pp 491–503
26. Brown S, Higham T, Slon V, et al (2016) Identification of a new hominin bone from Denisova Cave, Siberia using collagen fingerprinting and mitochondrial DNA analysis. Sci reports. doi: 10.1038/srep23559
27. Briggs AW, Good JM, Green RE, et al (2009) Targeted retrieval and analysis of five

Neandertal mtDNA genomes. Sci 325:318–321. doi: 10.1126/science.1174462

28. Sawyer S, Renaud G, Viola B, et al (2015) Nuclear and mitochondrial DNA sequences from two Denisovan individuals. Proc Natl Acad Sci United States Am 112:15696–15700. doi: 10.1073/pnas.1519905112
29. Lazaridis I, Patterson N, Mittnik A, et al (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409–413. doi: 10.1038/nature13673
30. Fu Q, Li H, Moorjani P, et al (2014) Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514:445–449. doi: 10.1038/nature13810
31. Allentoft ME, Sikora M, Sjögren K-G, et al (2015) Population genomics of Bronze Age Eurasia. Nature 522:167–172. doi: 10.1038/nature14507
32. Haak W, Lazaridis I, Patterson N, et al (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522:207–211. doi: 10.1038/nature14317
33. Krause J, Briggs AW, Kircher M, et al (2010) A complete mtDNA genome of an early modern human from Kostenki, Russia. Curr Biol : CB 20:231–236. doi: 10.1016/j.cub.2009.11.068
34. Ginolhac A, Rasmussen M, Gilbert MTP, et al (2011) mapDamage: testing for damage patterns in ancient DNA sequences. Bioinforma 27:2153–2155. doi: 10.1093/bioinformatics/btr347
35. Jónsson H, Ginolhac A, Schubert M, et al (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinforma 29:1682–1684. doi: 10.1093/bioinformatics/btt193
36. Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362:709–715. doi: 10.1038/362709a0

37. Seguin-Orlando A, Schubert M, Clary J, et al (2013) Ligation bias in illumina next-generation DNA libraries: implications for sequencing ancient genomes. PloS one. doi: 10.1371/journal.pone.0078575
38. Meyer M, Kircher M, Gansauge M-T, et al (2012) A high-coverage genome sequence from an archaic Denisovan individual. Sci 338:222–226. doi: 10.1126/science.1224344
39. Wales N, Ramos Madrigal J, Cappellini E, et al (2016) The limits and potential of paleogenomic techniques for reconstructing grapevine domestication. J Archaeol Sci. doi: 10.1016/j.jas.2016.05.014
40. Seguin-Orlando A, Hoover CA, Vasiliev SK, et al (2015) Amplification of TruSeq ancient DNA libraries with AccuPrime Pfx: consequences on nucleotide misincorporation and methylation patterns. STAR: Sci & Technol Archaeol Res. doi: 10.1179/2054892315Y.0000000005
41. Wall JD, Kim SK (2007) Inconsistencies in Neanderthal genomic DNA sequences. PLoS Genet 3:1862–1866. doi: 10.1371/journal.pgen.0030175
42. Prüfer K, Meyer M (2015) Anthropology. Comment on “Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern Native Americans”. Sci. doi: 10.1126/science.1260617
43. Weiß CL, Dannemann M, Prüfer K, Burbano HA (2015) Contesting the presence of wheat in the British Isles 8,000 years ago by assessing ancient DNA authenticity from low-coverage data. eLife. doi: 10.7554/eLife.10005
44. Schubert M, Ermini L, Der Sarkissian C, et al (2014) Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat Protoc 9:1056–1082. doi: 10.1038/nprot.2014.063
45. Kircher M (2012) Analysis of high-throughput ancient DNA sequencing data. Methods Mol Biol 840:197–228. doi: 10.1007/978-1-61779-516-9_23

46. Schubert M, Lindgreen S, Orlando L (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res notes. doi: 10.1186/s13104-016-1900-2
47. Renaud G, Stenzel U, Kelso J (2014) leeHom: adaptor trimming and merging for Illumina sequencing reads. Nucleic acids Res. doi: 10.1093/nar/gku699
48. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma 30:2114–2120. doi: 10.1093/bioinformatics/btu170
49. O'Connell J, Schulz-Trieglaff O, Carlson E, et al (2015) NxTrim: optimized trimming of Illumina mate pair reads. Bioinforma 31:2035–2037. doi: 10.1093/bioinformatics/btv057
50. Sturm M, Schroeder C, Bauer P (2016) SeqPurge: highly-sensitive adapter trimming for paired-end NGS data. BMC Bioinforma. doi: 10.1186/s12859-016-1069-7
51. Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinforma 30:614–620. doi: 10.1093/bioinformatics/btt593
52. Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinforma 27:2957–2963. doi: 10.1093/bioinformatics/btr507
53. Mielczarek M, Szyda J (2016) Review of alignment and SNP calling algorithms for next-generation sequencing data. J Appl Genet 57:71–79. doi: 10.1007/s13353-015-0292-7
54. Li HH, Durbin RR (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. doi: 10.1093/bioinformatics/btp324
55. Kerpedjiev P, Frellsen J, Lindgreen S, Krogh A (2014) Adaptable probabilistic mapping of short reads using position specific scoring matrices. BMC Bioinforma. doi: 10.1186/1471-2105-15-100
56. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat methods 9:357–359. doi: 10.1038/nmeth.1923

57. Nomiyama H, Fukuda M, Wakasugi S, et al (1985) Molecular structures of mitochondrial-DNA-like sequences in human nuclear DNA. Nucleic acids Res 13:1649–1658. doi: 10.1093/nar/13.5.1649

58. Lopez JV, Yuhki N, Masuda R, et al (1994) Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J Mol Evol 39:174–190.

59. Li H, Handsaker B, Wysoker A, et al (2008) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.

60. Dozmorov MG, Adrianto I, Giles CB, et al (2015) Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data. BMC Bioinforma. doi: 10.1186/1471-2105-16-S13-S10

61. McKenna A, Hanna M, Banks E, et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. doi: 10.1101/gr.107524.110

62. Briggs AW, Stenzel U, Meyer M, et al (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic acids Res. doi: 10.1093/nar/gkp1163

63. Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. doi: 10.1101/pdb.prot5448

64. Krause J, Unger T, Noçon A, et al (2008) Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. BMC Evol Biol. doi: 10.1186/1471-2148-8-220

65. Rohland N, Harney E, Mallick S, et al (2015) Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos Trans R Soc London Ser B, Biol Sci. doi:

10.1098/rstb.2013.0624

66. Pedersen JS, Valen E, Velazquez AMV, et al (2014) Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. Genome Res 24:454–466. doi: 10.1101/gr.163592.113
67. Gokhman D, Lavi E, Prüfer K, et al (2014) Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. Sci 344:523–527. doi: 10.1126/science.1250368
68. Hanghøj K, Seguin-Orlando A, Schubert M, et al (2016) Fast, Accurate and Automatic Ancient Nucleosome and Methylation Maps with epiPALEOMIX. Mol Biol Evol 33:3284–3298. doi: 10.1093/molbev/msw184
69. Renaud G, Slon V, Duggan AT, Kelso J (2015) Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biol. doi: 10.1186/s13059-015-0776-0
70. Schuenemann VJ, Singh P, Mendum TA, et al (2013) Genome-wide comparison of medieval and modern Mycobacterium leprae. Sci 341:179–183. doi: 10.1126/science.1238286
71. Sawyer S, Krause J, Guschanski K, et al (2012) Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. PloS one. doi: 10.1371/journal.pone.0034131
72. Parks M, Lambert D (2015) Impacts of low coverage depths and post-mortem DNA damage on variant calling: a simulation study. BMC Genomic-. doi: 10.1186/s12864-015-1219-8
73. Skoglund P, Northoff BH, Shunkov MV, et al (2014) Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc Natl Acad Sci United States Am 111:2229–2234. doi: 10.1073/pnas.1318934111
74. Green RE, Briggs AW, Krause J, et al (2009) The Neandertal genome and ancient DNA

authenticity. EMBO J 28:2494–2502. doi: 10.1038/emboj.2009.222

75. Zhang H, Paijmans JLA, Chang F, et al (2013) Morphological and genetic evidence for early Holocene cattle management in northeastern China. Nat Commun. doi: 10.1038/ncomms3755

76. Weissensteiner H, Pacher D, Kloss-Brandstätter A, et al (2016) HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic acids Res 44:W58–W63. doi: 10.1093/nar/gkw233

77. Rasmussen M, Sikora M, Albrechtsen A, et al (2015) The ancestry and affiliations of Kennewick Man. Nature 523:455–458. doi: 10.1038/nature14625 http://www.nature.com/nature/journal/vnfv/ncurrent/abs/nature14625.html#supplementary-information

78. Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinforma. doi: 10.1186/s12859-014-0356-4

79. Racimo F, Renaud G, Slatkin M (2016) Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans. PLoS Genet. doi: 10.1371/journal.pgen.1005972

80. Skoglund P, Storå J, Götherström A, Jakobsson M (2013) Accurate sex identification of ancient human remains using DNA shotgun sequencing. J Archaeol Sci. doi: 10.1016/j.jas.2013.07.004

81. Abecasis GR, Auton A, Brooks LD, et al with 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65. doi: 10.1038/nature11632

82. Louvel G, Der Sarkissian C, Hanghøj K, Orlando L (2016) metaBIT, an integrative and automated metagenomic pipeline for analysing microbial profiles from high-throughput sequencing shotgun data. Mol Ecol Resour 16:1415–1427. doi: 10.1111/1755-

0998.12546

83. Renaud G, Hanghøj K, Willeslev E, Orlando L (2016) gargammel: a sequence simulator for ancient DNA. Bioinformatics btw670. doi: 10.1093/bioinformatics/btw670