gargammel: a sequence simulator for ancient DNA.

Supplementary Material

Gabriel Renaud, Kristian Hanghøj, Eske Willerslev and Ludovic Orlando

Contents

1 Supplementary Methods

1.1 Overall description

The algorithm works by taking a vector of three numbers representing the desired proportion of the final data set:

- microbial contamination
- endogenous DNA
- contamination from the same species.

For ancient hominin samples, the contamination from the same species can be viewed as contamination stemming from present-day humans involved in excavation, handling and/or extraction of DNA. Gargammel users are required to provide the genome files in fasta format of the genomic references representing all three sources of DNA fragments. Alternatively, when complex microbial communities are to be simulated, users can also provide a vector, '*all_taxa.tsv*', of microbial abundances from metaBIT [20] and closely-related genomes will be automatically fetched from NCBI using the *retrieveFromMetabit* script, which is provided with gargammel. Additionally, the software package provides a script named *ms2chromosomes.py*, which enable users to run Hudson's ms [13] to create an input in terms of contaminant and endogenous genomes to be used as input for gargammel. This script simulates a simple 2 population model and transforms the resulting coalescence tree into sequences using seq-ren [25]. The first population represents a diploid endogenous organism while the second represents the contaminating population source. The Supplemental sections 2.5 and 2.6 provide explicit examples on how the *retrieveFromMetabit* and *ms2chromosomes.py* scripts can be used.

The algorithm underpinning gargammel proceeds by calculating the number of fragments to extract from all three databases given their relative genome size to achieve the desired fragment composition. Fragments are then extracted from the genome files (if requested, with respect to desired size distributions - see section 1.3 below), ancient DNA (aDNA) damage is added, and sequencing adapters are appended at the ends. Finally, sequencing errors along with corresponding quality scores are added as to produce a set of simulated Illumina reads that could have been obtained for ancient material.

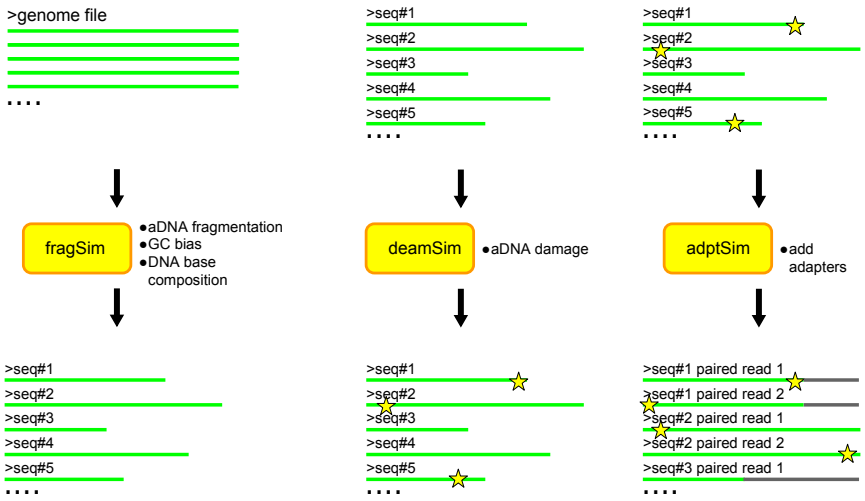
Gargammel is composed of various programs written in C++. The overall driver script is written in Perl. A flowchart representing the 3 main programs constituting gargammel is presented in Supplementary Figure 1. The various underlying programs can also be used independently by users to create their own custom workflow (see Supplementary Table 1). Our software is meant to be used via the command line and was tested on Linux and MacOS.

program	input	output
fragSim	fasta reference	aDNA fragments
deamSim	aDNA fragments	deaminated aDNA fragments
adptSim	deaminated aDNA fragments	raw Illumina sequences
ART	raw Illumina sequences	Illumina reads

Supplementary Table 1: Sub-programs embedded within the gargammel pipeline. The formats of input and output files are indicated. Illumina reads include the sequence with some potential sequencing errors and corresponding quality scores whereas raw Illumina sequences designates the raw sequencing templates (amplified aDNA fragment plus possible adapters)

1.2 Producing an example of a set of microbial genomes

To simulate an example of microbial sequences from an empirical dataset, we used DNA fragments prior to alignment from the CL32 and C28 libraries of the Clovis individual from [26] were used as input for metaBIT[20] to infer the metagenomic composition. A total of 32 microbial species were



Supplementary Figure 1: Flowchart for the 3 main sub-programs in the gargammel package. The initial fragments are generated using *fragSim*. Deamination can be added using *deamSim*. Finally, the Illumina sequencing adapters can be added using *adptSim* and are used as input for ART, an Illumina sequencer simulator.

identified using a 0.1% threshold of minimal abundance. The *retrieveFromMetabit* script provided with gargammel allows users to automatically download the reference genomes for the identified species to be used as a source of microbial contamination by gargammel.

The script *retrieveFromMetabit* takes the tabulated taxonomic abundances *'all_taxa.tsv'* file produced by metaBIT [20] as input. In addition to downloading the requested reference genomes of the identified microbial species (and required by gargammel), it also produces a file named *'list'* containing assembly names and relative abundances of the downloaded microbes. In case of unclassified taxa on the species level, relative abundances of identified species in the *'list'* file are scaled to sum to one. The *'list'* file is used by gargammel to reflect the simulated microbial content.

The microbial community from 2 empirical samples (the 12.8 kyr-old Native American Clovis individual from [26] and the 36.8 kyr-old Kostenki K14 individual from [34]) are available with gargammel to provide users with examples.

1.3 Simulating DNA fragmentation

As DNA degrades over time, the aDNA molecules that are extracted from subfossils are heavily fragmented. Gargammel allows users to select the size of fragments either by specifying a fixed fragment size or by providing an empirical distribution of fragment sizes. In the latter case, the distribution is provided as a text file where each line indicates the size of an individual fragment and its relative frequency. During simulations, fragments of a specific length will be generated with a probability corresponding to their relative frequency defined in the text file. Additionally, the user can also provide the location and the scale parameters of a log-normal distribution. These parameters can be obtained using a maximum-likelihood fit of a log-normal distribution on a set of empirical aDNA fragment lengths. For instance, this can be done using the "fitdistr" function in from the Fitdistrplus package[1] in R. Please note that this fit on the empirical distribution needs to be performed prior to length filtering and is not applicable if single-end reads were used (as the full size distribution of aDNA templates can then not be fully characterized). To simulate the computational processing that is routinely done in aDNA research, gargammel offers the possibility to discard fragments that are smaller than a certain length (e.g. 35bp as done in [16, 33]).

To illustrate the use of an empirical size distribution, we provide an example of a text file containing empirical fragments lengths and their frequency from the Ust'-Ishim study [5] as part of

[1]Marie Laure Delignette-Muller, Regis Pouillot, Jean-Baptiste Denis, and Christophe Dutang. Fitdistrplus: help to fit of a parametric distribution to non-censored or censored data. https://cran.r-project.org/web/packages/fitdistrplus/index.html Accessed online October 5th 2016

the package (file: src/sizedist.size.gz). The leftmost coordinate is taken at random using a uniform probability distribution. The length of the simulated fragment is then used to dictate the rightmost position. Finally, either the plus or minus strand is selected each with equal probabilities of 0.5.

The initial empirical fragment size distribution can depend on the software used for adapter removal and overlap merging as some procedures are more sensitive than others [32]. Some programs might skew the distribution towards shorter or longer fragment sizes and contain residual adapter sequences [37]. We therefore recommend that users apply the same trimming software on the final simulated data as the one originally used on the empirical data for obtaining the fragment size distribution to minimize any possible biases.

1.4 Simulating DNA base composition

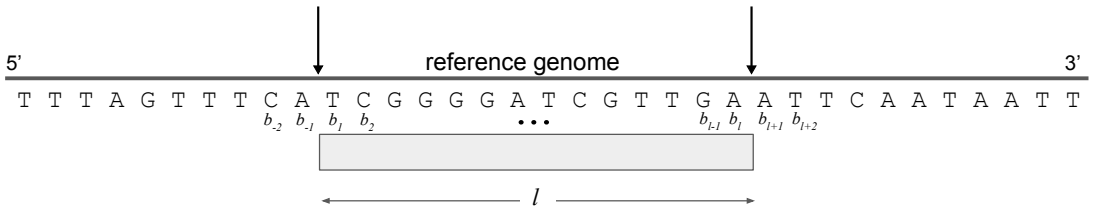
Several studies have reported that the base composition at 5' and 3' ends of aDNA fragments depart from the average genomic composition [1, 14]. Gargammel allows users to input a specific base composition profile for the 5' and 3' ends. Such base composition can be directly obtained from empirical data using the "dnacomp.txt" file produced by mapDamage2 [14]. These base compositions are simply the frequencies of the 4 nucleotides $\{A, C, G, T\}$ at a given position with respect to aDNA fragments. Empirically speaking, these frequencies represent the composition around DNA breaks.

The *fragSim* module attempts to model the desired base composition using the following procedure:

- randomly select a fragment of a certain length (see section 1.3 for more details about fragment length distribution).
- compute the probability p of observing this fragment under the empirical base composition.
- accept this fragment with probability p, reject with probability $1 - p$.

This heuristic allows *fragSim* to produce fragments with a base composition around the ends that matches the one observed empirically.

More specifically, let a potential fragment $\mathbb{F}$ have length l. Let bases $.., b_{-2}, b_{-1}$ be the ones preceding the 5' end and $b_1, b_2, ..$ be the bases after the 5' end (see Figure 2 for a schematic representation). Similarly, let bases $b_{l-1}, b_l, ..$ be the bases before the 3' end and $b_{l+1}, b_{l+2}, ..$ be the ones after the 3' end.



Supplementary Figure 2: A schematic representation of the notation used for the base composition around aDNA breakpoints. A certain fragment of length l, represented in the figure by a shaded rectangle, is randomly chosen on the reference genome. The probability of that specific DNA break, represented by the vertical arrows, occurring around the ends depends on the input base composition. In this example, only 2 bases adjacent to the breaks would be considered, these bases are represented in the figure by b_i.

Once fragment $\mathbb{F}$ has been selected, the probability of observing its base composition around the break points needs to be computed. To achieve this, we cannot directly use the frequencies for the different bases provided by the user as the base composition of the genome might be different. For instance, if a genome shows equal frequencies for bases A, C, G, T (i.e. $\{0.25, 0.25, 0.25, 0.25\}$ for each base, respectively) and the frequency at given certain position adjacent to a break is also $\{0.25, 0.25, 0.25, 0.25\}$, there is no preference for specific bases compared to the background. If however, the background genomic frequency is $\{0.30, 0.20, 0.20, 0.30\}$ and a base frequency of

$\{0.25, 0.25, 0.25, 0.25\}$ is found next to a break, this indicates an enrichment for Cs and Gs and a depletion of As and Ts compared to the background.

To account for the background genomic base frequency, the base frequency of base b at position i, denoted $F_i(b)$, has to be scaled. Let $f_i(b)$ be this new scaled frequency to be used for simulations. Let the background frequency of base b be $F(b)$, the new scaled frequency of base b at position i is given by:

$$f_i(b) = 0.25 + F_i(b) - F(b) \tag{1}$$

If a base is enriched for at position i, the expression $F_i(b) - F(b)$ will be positive and the scaled frequency $f_i(b)$ will be higher than 0.25. In the previous example where the background frequency $F(b)$ was $\{0.30, 0.20, 0.20, 0.30\}$ and the frequency a given position adjacent to the 5' end (such that $i = 1$) would be $F_1(b) = \{0.25, 0.25, 0.25, 0.25\}$, the scaled frequency for base A, denoted $f_1(A)$ would be $0.25 + 0.25 - 0.30 = 0.20$ thus showing a depletion for As at position $i = 1$. If the scaled frequencies are uniform such that $f_i(A) = f_i(C) = f_i(G) = f_i(T) = 0.25$, the base composition at the ends of the selected fragments will not have any preference for any particular base and every generated fragment will show an equal probability p of being accepted. These fragments will simply revert to the genomic background of the genome used as input.

Using the scaled frequencies, the probability of observing the base composition for fragment $\mathbb{F}$, denoted p, can be computed. If we only consider 1 base before and after the 5' and 3' ends, we only consider the following bases $b_{-1}, b_1, b_l, b_{l+1}$. The probability of observing the fragment given the base composition is obtained by multiplying the base frequencies:

$$p = f_{-1}(b_{-1}) f_1(b_1) f_l(b_l) f_{l+1}(b_{l+1}) \tag{2}$$

If more bases adjacent to the 5'/3' breaks are considered, further terms are added to equation 2. Finally, fragment $\mathbb{F}$ is retained with probability p and rejected with probability $1 - p$. This heuristic allows users to create simulated aDNA fragments with a base composition corresponding to those observed empirically. A drawback of this method is that the acceptance probability p will decrease with the total number of bases considered that are adjacent to the breaks. This entails that more fragments will be discarded thus resulting in greater runtimes but in more realistic aDNA breakpoints (see results in subsection 2.2 for greater details).

1.5 Simulating ancient DNA damage

After death, cytosine residues tend to lose their amine group and are converted into uracil residues [11]. As a result of nick fill-in steps during library construction and further template amplification, uracil residues will be observed as thymines [1]. The two main protocols for the sequencing of aDNA namely the double-strand [21] and single-strand protocols [7] result in different damage patterns [29].

Gargammel allows users to input the parameters of the mapDamage2[14] model to simulate damage consistent with a double strand library preparation. The software also allows users to provide a substitution matrix that is applied to the aDNA fragments to simulate empirically observed substitution rates due to post-mortem damage. Deamination is applied to each base independently of one another.

1.6 Simulating polymerase induced GC bias

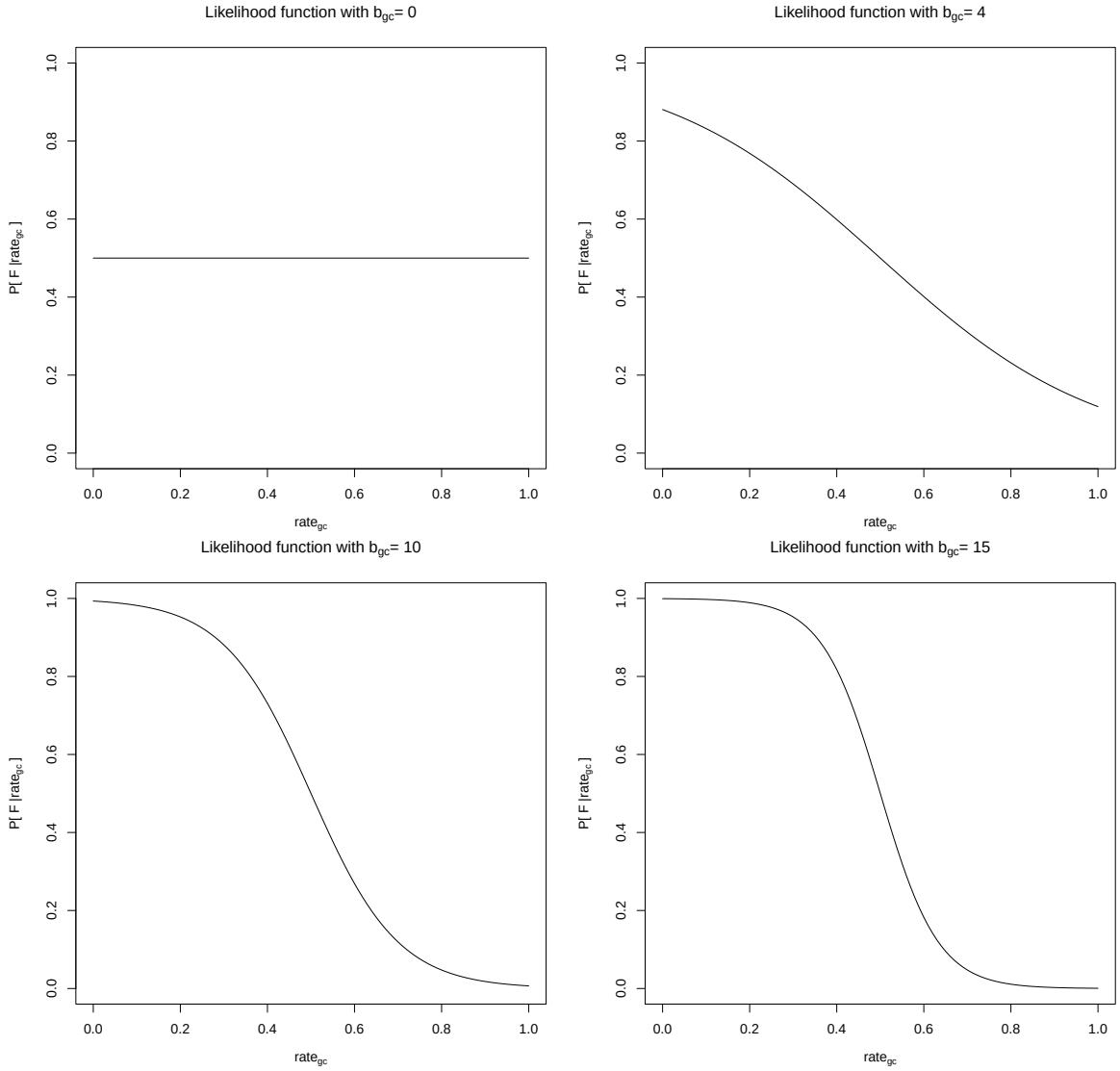
Multiple factors, including the type of DNA polymerase used during library amplification [3, 4] as well as the sequencing technology itself [24], can induce a GC bias in the recovered population of DNA sequences. We therefore added the possibility of modeling GC bias in the *fragSim* module. This bias is modeled at the initial stage where fragments are selected by computing the probability, denoted $P[\mathbb{F}]$, that a fragment $\mathbb{F}$ is finally amplified and sequenced in the final sequencing run. More specifically, this probability is quantified as follows:

$$P[\mathbb{F}] = P[\mathbb{F}|rate_{GC}]P[rate_{GC}] \tag{3}$$

where $rate_{GC}$ is the GC count for $\mathbb{F}$. The prior on the value of $rate_{GC}$ is obtained using a normal distribution of GC content for random fragments from the genome. The likelihood $P[\mathbb{F}|rate_{GC}]$ is computed using a logistic function:

$$P[\mathbb{F}|rate_{GC}] = \frac{1}{1 + e^{b_{GC}(rate_{GC} - \mu_{GC})}} \tag{4}$$

where the term b_{GC} is a model parameter to account for the GC bias. The resulting logistic function and various levels of GC bias are plotted in Supplementary Figure 3. Finally, the fragment is accepted with probability $P[\mathbb{F}]$ and produced as part of the final output.



Supplementary Figure 3: The resulting logistic function for various values for b_{GC}. The numeric values for the b_{GC} parameter are reported above each respective graph. At $b_{GC} = 0$, there is effectively no GC-bias but an example of a severe GC-bias is modeled at $b_{GC} = 15$.

1.7 Simulating the Illumina sequencing process

Currently, our package uses the Illumina platform as it is the most commonly used platform for aDNA projects [28]. Gargammel simulates the sequencing starting from the 5' end followed by paired-end turnaround and sequencing from the 3' end of the fragment. Our package also allows for single-end sequencing. If the fragment length is less than the desired read length, sequencing adapters are then appended at the end of the fragments. For longer fragments, it is also possible to have an overlapping portion of the forward and reverse sequencing reads (this only applies when simulated fragments are shorter than twice the read length).

To generate platform-specific sequencing errors, gargammel uses the ART simulator [12]. This software can generate error profiles consistent with many of the most commonly used Illumina platforms (e.g. Genome Analyzer II, HiSeq 2500). Each base is also assigned a quality score reflecting the probability of a sequencing error. The resulting fastq files are then produced as the final output of our simulation pipeline.

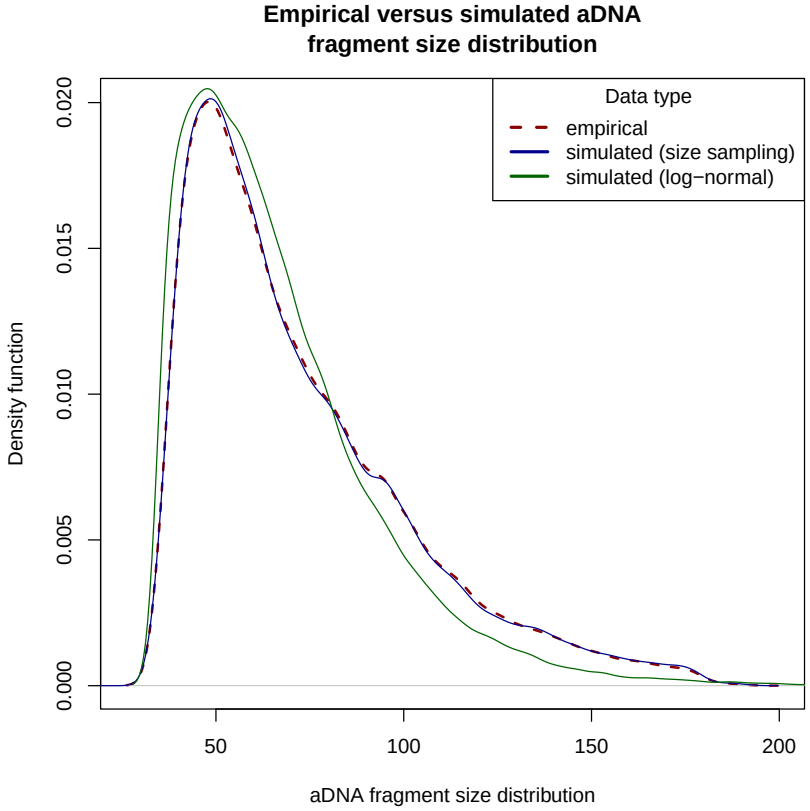
2 Supplementary Results

2.1 Simulated DNA fragmentation

To evaluate the ability of the *fragSim* subprogram to produce realistic fragment size distributions, the length of 50k empirical fragments from the Ust'-Ishim individual [5] was measured. The resulting length distribution was used to simulate aDNA fragments by either:

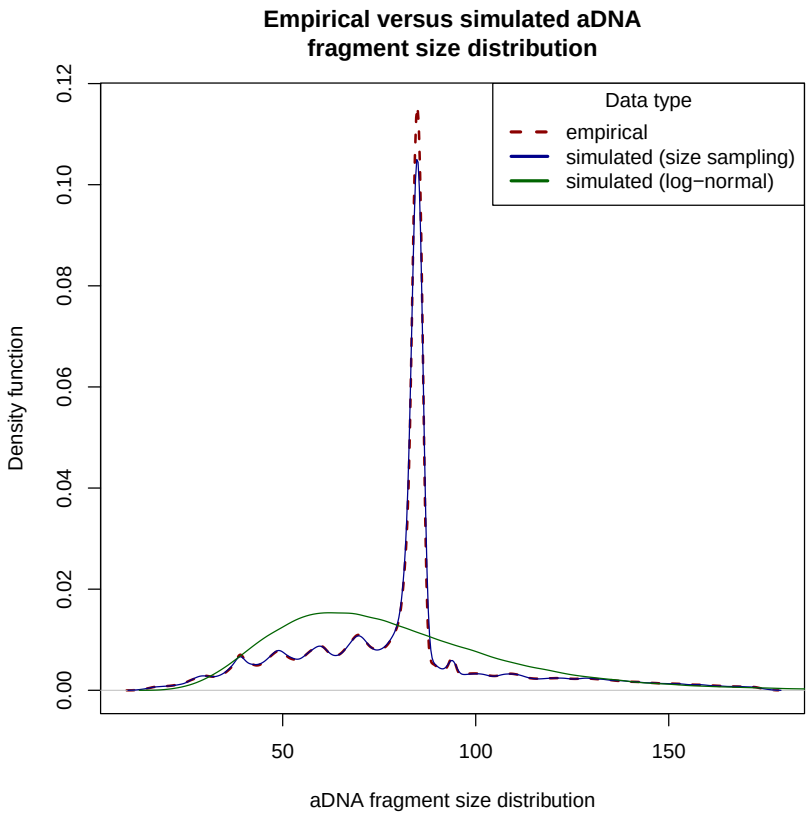
1. By sampling a fragment size from the empirical distribution (using option "-s")
2. By using a log-normal distribution that approximates the one observed empirically (using option "–loc 4.046626 –scale 0.42017143")

In both cases, 50k fragments were generated and the option "-m 35" was used to limit the fragments produced to those longer than 35bp, as performed in the original publication [5]. The fragment lengths distribution for both the empirical and simulated datasets are shown in Supplementary Figure 4. Our results show that *fragSim* can produce realistic fragment size distributions for simulated datasets.



Supplementary Figure 4: Empirical versus simulated aDNA fragment size distributions for the Ust'Ishim anatomically modern human individual [5].

We repeated the same procedure using the size distribution of the aDNA fragments recovered from a ∼5.2 kyr-old horse [19], as the latter showed the strong 10-bp periodicity pattern that has been proposed to reflect nucleosomal DNA protection [23, 9]. As expected, our results show that for irregular distributions, modeling the fragment lengths with a log-normal distribution does not accurately match the original empirical one (see Supplementary Figure 5). However, selecting simulated fragments (options "-s") from the original empirical distribution results in an appropriate match to the empirical one. This shows the versatility of gargammel to emulate complex empirical size distributions.



Supplementary Figure 5: Empirical versus simulated aDNA fragment size distribution for aDNA fragment lengths with a high 10-bp periodicity. The spike around 85bp is unlikely to be caused by incorrectly merged reads since 98 cycles were used in the original Illumina sequencing run and since this enrichment is also observed for fragments of length 84bp and 86bp.

2.2 Simulated DNA composition at the end of fragments

The base compositions of aDNA fragments for two anatomically modern individuals, the 10.3 individual from [6] and the Ust'Ishim individual from [5], were evaluated using mapDamage2.0 [14]. For computational efficiency, only a random subset of one million fragments aligned against chromosome 21 was considered.

The resulting base composition was used by gargammel to model one million fragments from chromosome 21 using the *fragSim* subprogram. The program was rerun using the "-dist" option by considering between 0 and 10 flanking genomic bases, located prior to and following template termini. The generated fragments were then aligned using BWA v.0.5.10 [18] with the following options: "-n 0.01 -o 2 -l 16500" (see [31] for a review of the sensitivity of BWA for aDNA). Using mapDamage2.0, we then evaluated the base composition of the simulated fragments to test whether the simulated base composition would be similar to the empirical one. Furthermore, the time taken by the program was also evaluated when running *fragSim* on a computer with single AMD Opteron processor 2.3GHz CPU and 128G of RAM (see Supplementary Table 2). As the fasta references are memory mapped for speed, we recommend running gargammel on a computer with at least 10G of RAM.

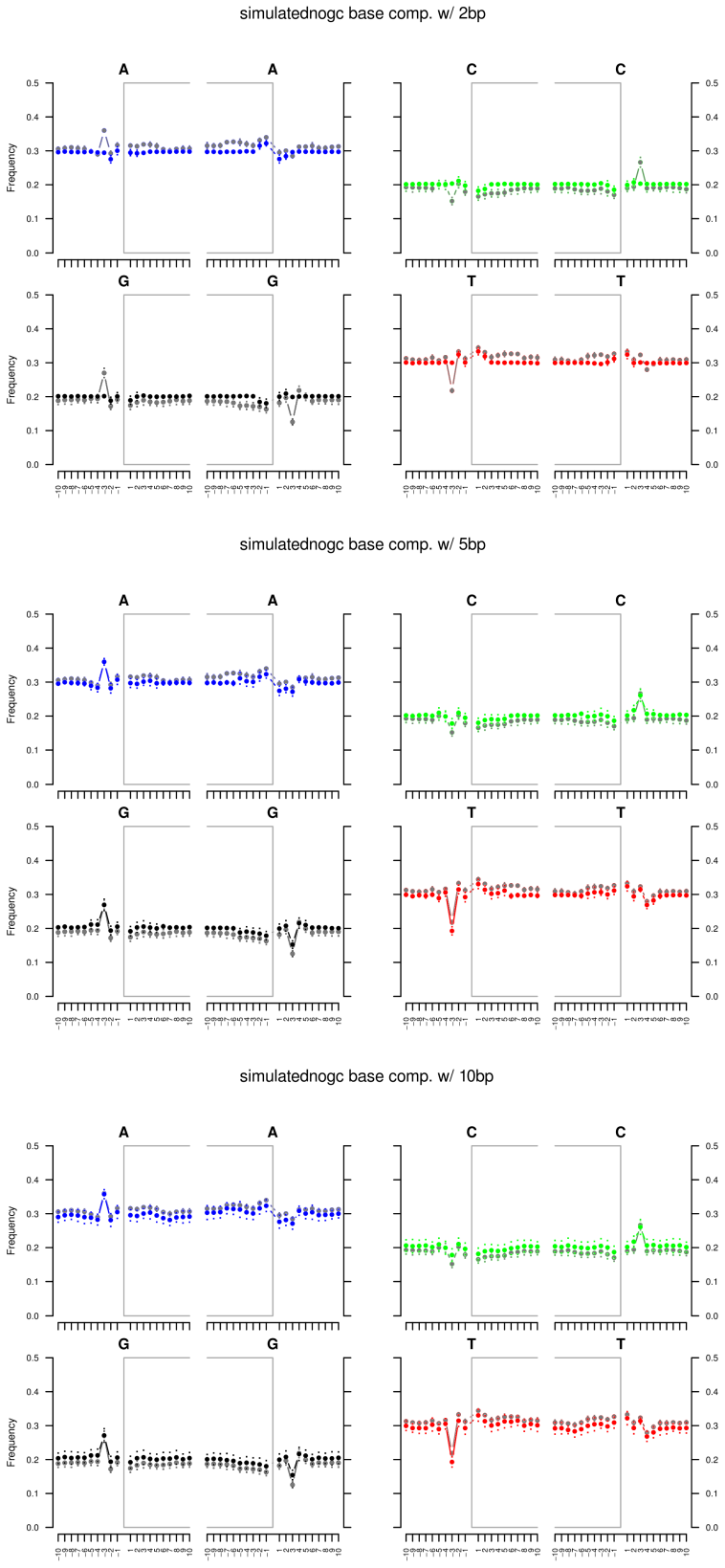
Visual inspection indicates that the base compositions of the simulated fragments look very similar to the empirical ones for the bases that were considered (see Supplementary Figure 6 for the 10.3 individual and Supplementary Figure 8 for Ust'-Ishim). As expected, both tests reveal that the greater amount of bases that are considered on genomic flanking regions, the closer the base composition of the simulated aDNA fragments matches the empirical one. Runtimes reveal that greater accuracy comes at a cost as the runtime increases with the number of bases considered (see Supplementary Table 2).

2.2.1 Accounting for GC biases

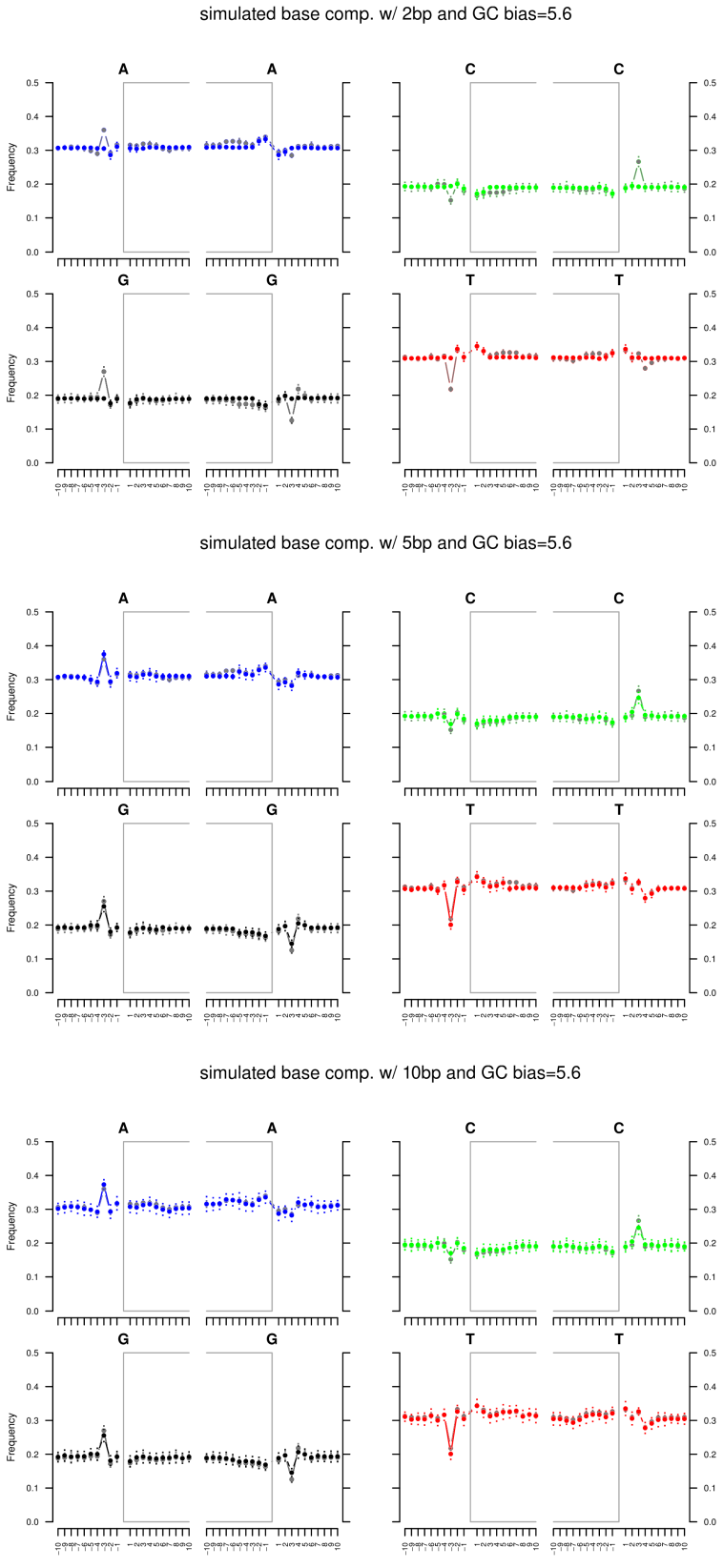
As a DNA polymerase biased against high GC-content was used in the Ust'Ishim study (the AccuPrime Pfx DNA polymerase; [5]), the background base composition is enriched for A/T and depleted for C/G thus resulting in a slight discrepancy [4].

A separate maximum-likelihood method was implemented to estimate b_{GC}, the GC bias parameter defined in Supplementary Section 1.6, for both empirical samples independently. Briefly, this method finds the most likely value for b_{GC} such that the GC content of a dataset simulated in absence of GC bias would be observed as the empirical GC distribution.

To model the bias induced by the polymerase, the maximum-likelihood estimate of the GC bias parameter for the Ust'Ishim data of 12.0 was used by *fragSim* and the simulation was repeated. Results show that the discrepancy between the simulated and empirical base composition is reduced when a simulated GC bias factor is used (see Supplementary Figure 9). The maximum-likelihood estimate for the data from the 10.3 individual (amplified with Accuprime Pfx DNA polymerase) gave a GC bias of 5.6. The simulation done for the base composition for the 10.3 individual was repeated using a GC bias of 5.6. This correction improved the fit of the simulated and observed data (see Supplementary Figure 7).



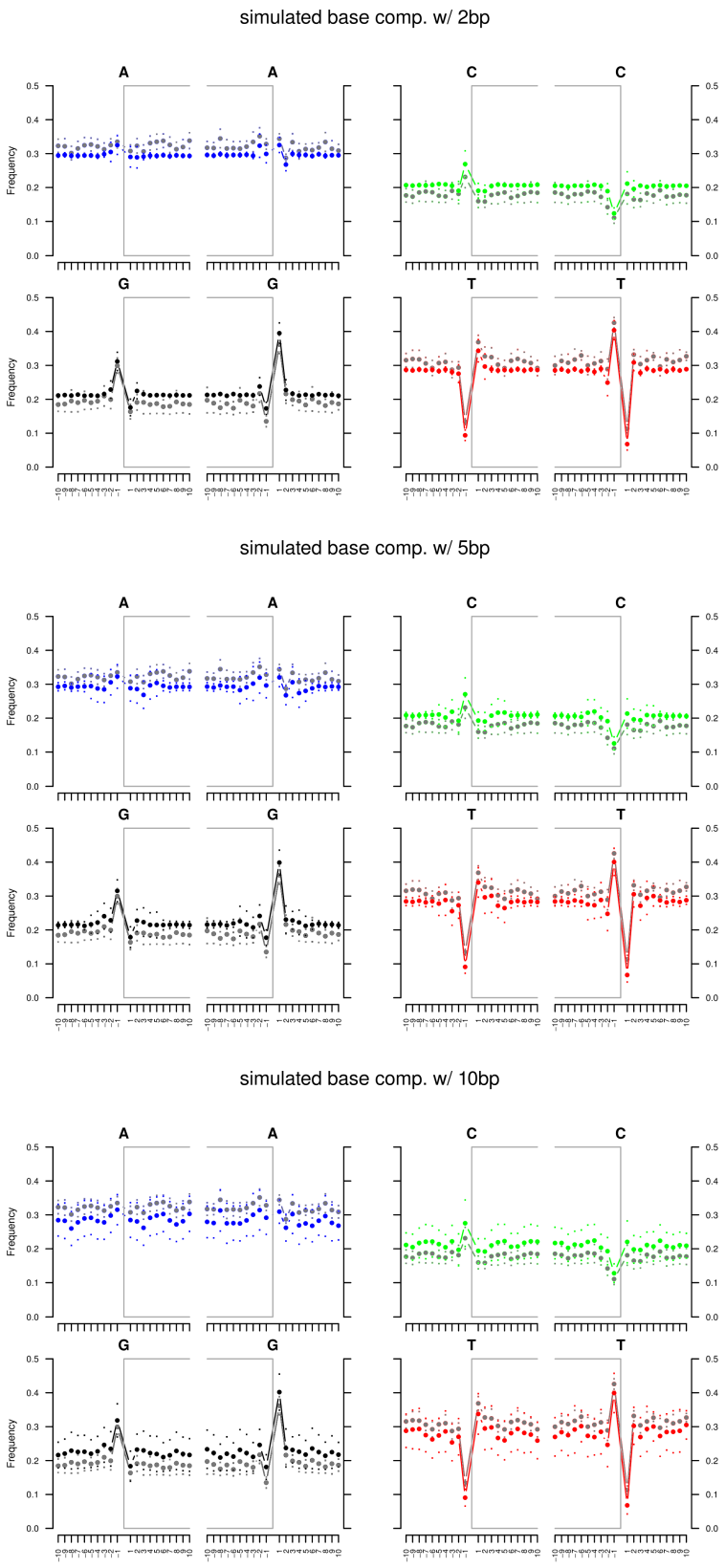
Supplementary Figure 6: The simulated (colored datapoints) versus the empirical (gray) base composition for the 10.3 individual from [6].



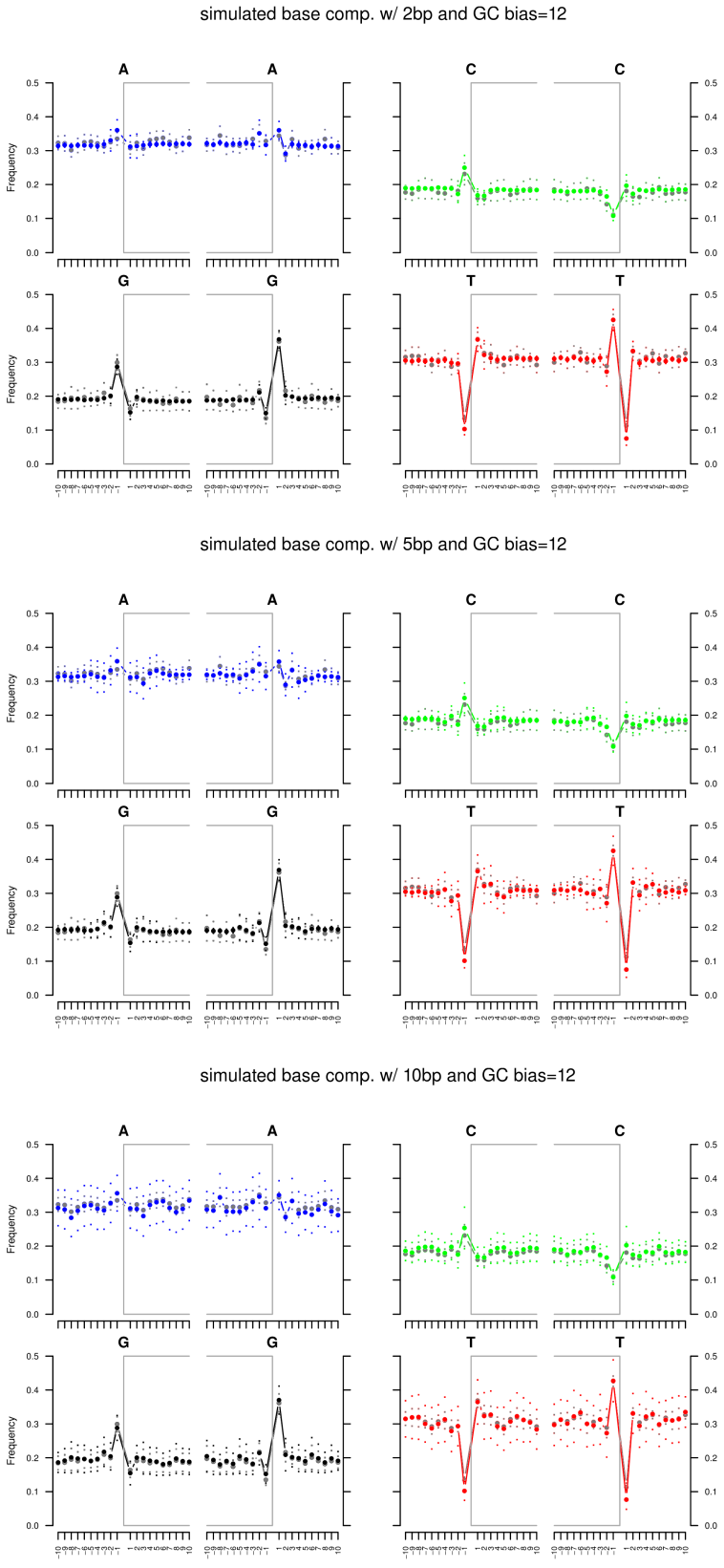
Supplementary Figure 7: The simulated (colored datapoints) versus the empirical (gray) base composition for the 10.3 individual from [6] using a simulated GC bias factor of 5.6.

bases considered	runtime (s)
0	22.0
1	88.0
2	147.8
3	215.1
4	362.7
5	551.6
6	929.5
7	1411.7
8	2668.9
9	3513.5
10	5568.0

Supplementary Table 2: Runtime as a function of the number of bases that were considered. The probabilistic algorithm used discards a higher fraction of fragments as the number of considered bases increases thus resulting in higher runtimes.



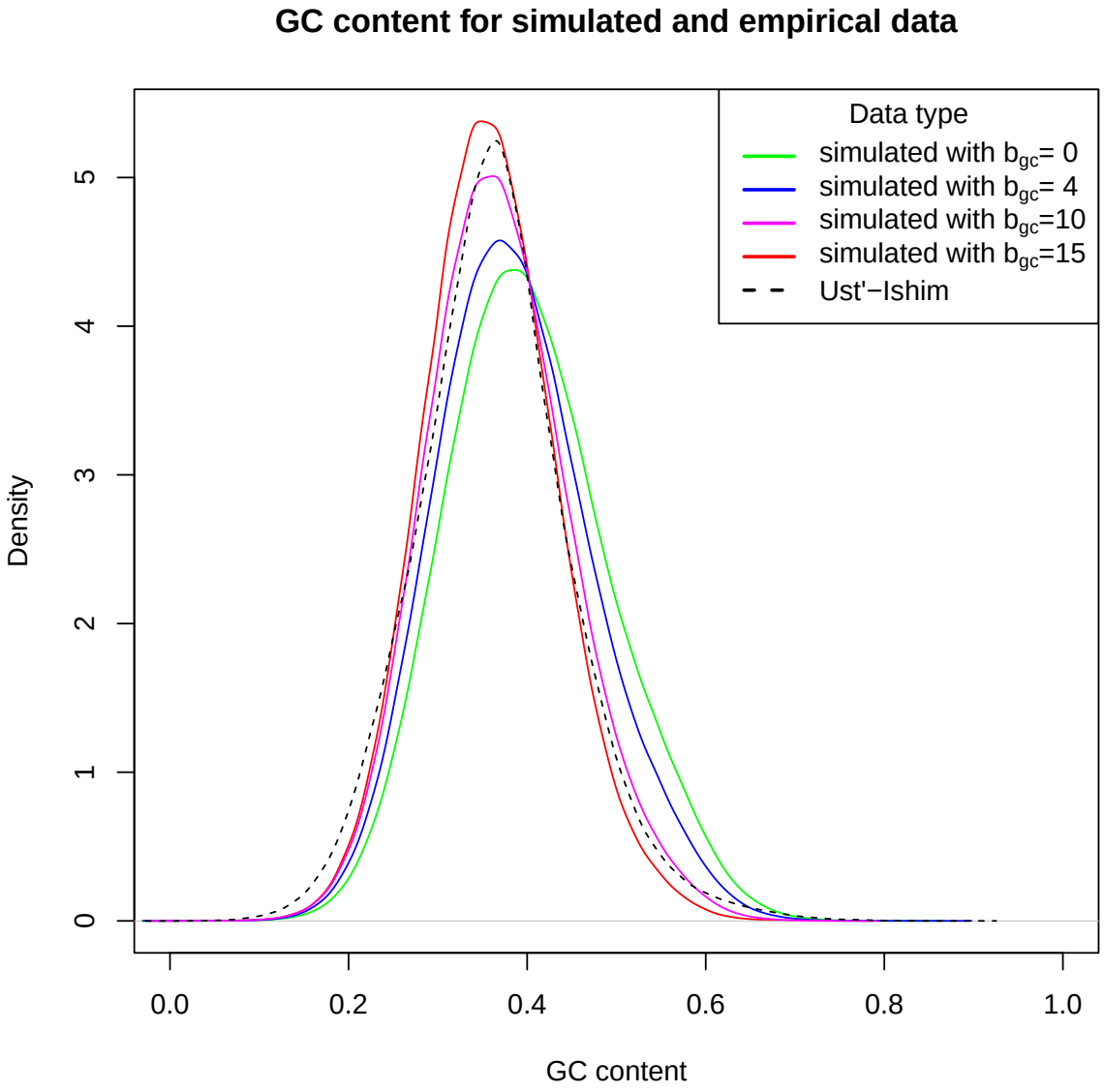
Supplementary Figure 8: The simulated (colored datapoints) versus the empirical (gray) base composition for the Ust'-Ishim individual from [5]. The polymerase used for this study showed a stronger GC bias than the data reported in Supplementary Figure 6.



Supplementary Figure 9: The simulated (colored datapoints) versus the empirical (gray) base composition with GC correction for the Ust'-Ishim individual from [5] where the DNA polymerase used for this study showed a GC bias. The GC bias correction factor used was 12 as determined via a maximum-likelihood estimate performed on the empirical sample.

2.3 Simulated GC-bias

To measure the effect of using a GC-bias term as described in section 1.6, *fragSim* produced 1M fragments from chromosome 21 with various values of the bias parameter b_{GC}. The result of this parameter on the GC distribution can be seen in Supplementary Figure 10. As the GC-bias parameter increases, the more the distribution shifts towards a lower GC content. To provide a comparison with an empirical sample, the distribution of the GC content for the Ust'-Ishim sample was also plotted. For this sample, the maximum-likelihood estimate of the GC bias parameter was b_{GC} =12.



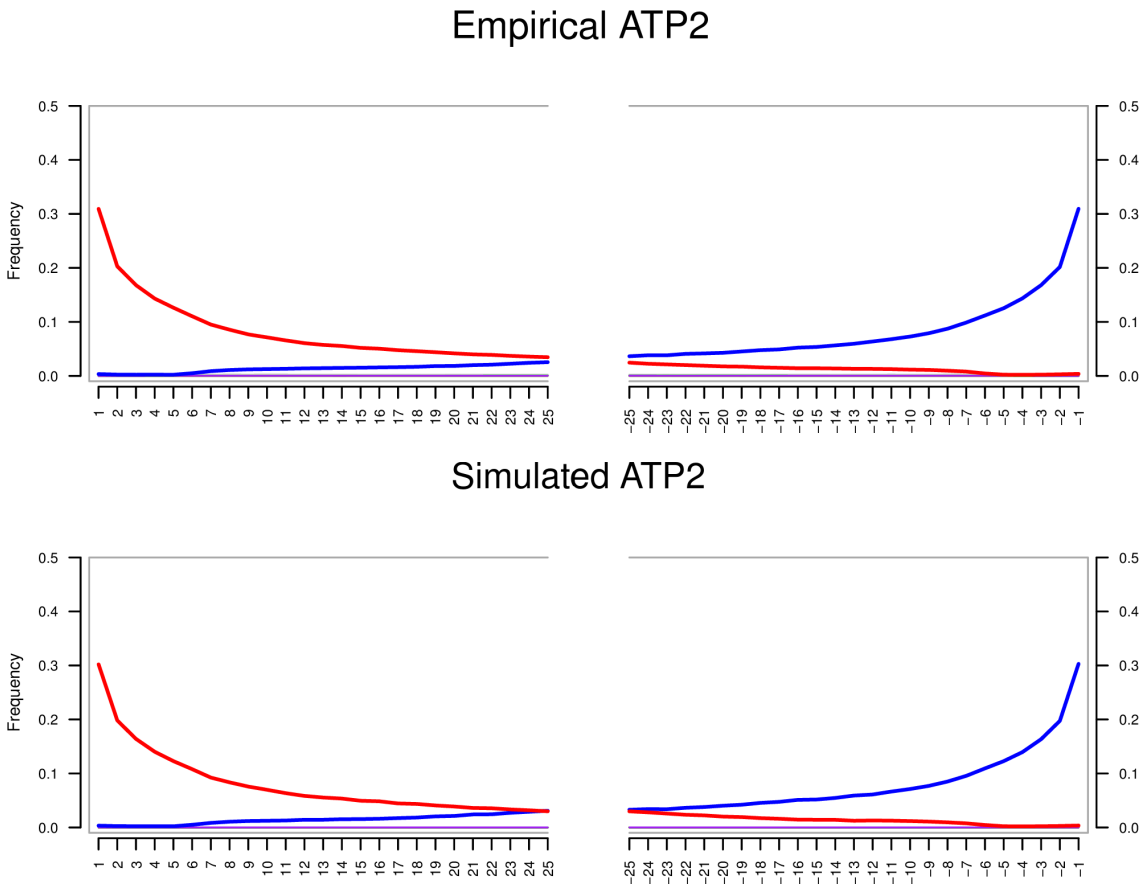
Supplementary Figure 10: Distribution of the GC content for 4 datasets produced with various values of the b_{GC} parameter. For each condition, one million DNA fragments were produced. At $b_{GC} = 0$ (green) the GC content is the one of the reference genome. The GC bias and the preference for AT-rich fragments increase with increasing values for the b_{GC} parameter. To provide a comparison for those GC content distributions for simulated sets to an empirical one, the GC content distribution (one million fragments) of the Ust'Ishim individual is also plotted.

2.4 Simulated DNA damage

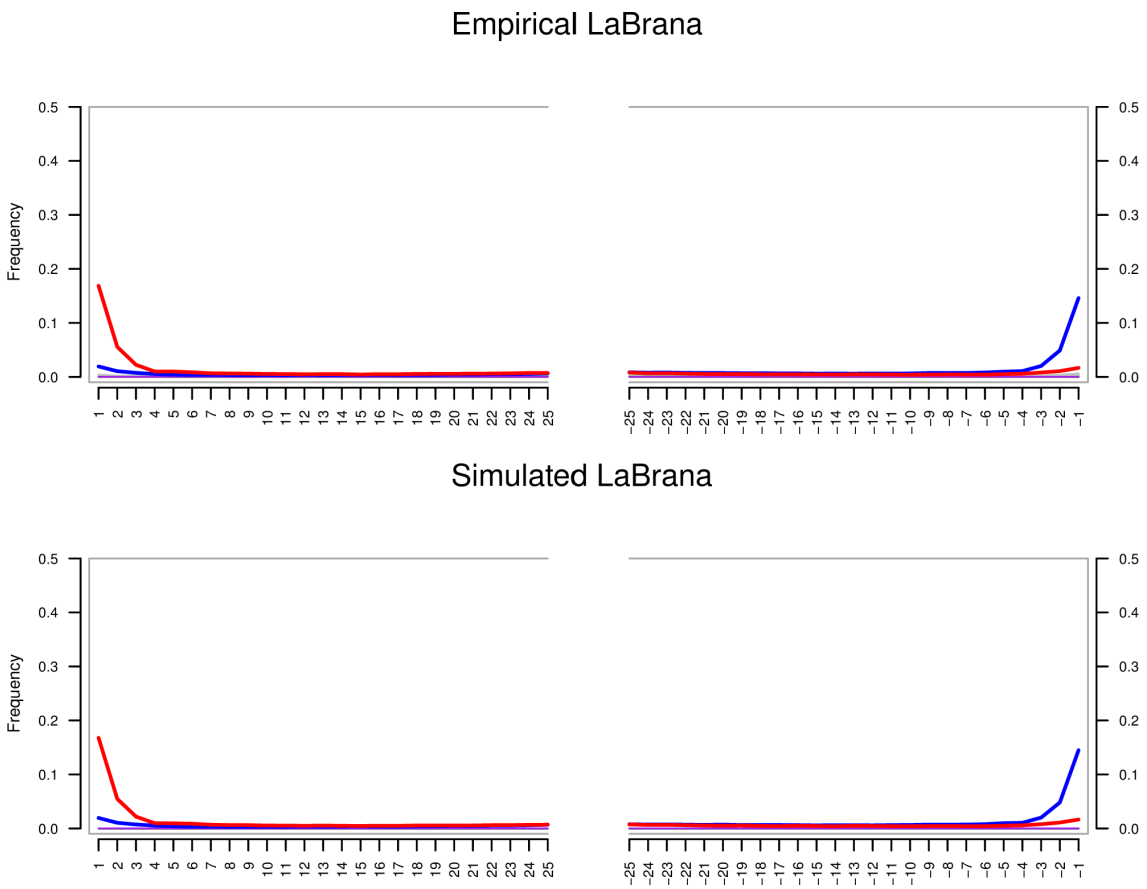
To evaluate the accuracy of the deamination produced by the *deamSim* subprogram, one million fragments were randomly selected from chromosome 21 and deamination was added *in silico*. Deamination was added according to the empirical rates observed in three different studies to show the versatility of the program. First, we used the sequence data underlying the genome sequence of a heavily deaminated sample, ATP2 from [8], that had been sequenced using a double-strand protocol. Secondly, The La Braña individual [30], also sequenced using a double-stranded DNA library construction protocol but showing less DNA damage. Finally, we also considered the sequence data underlying the Ust'Ishim individual [5], a sample with low levels of damage that had been treated using the USER enzyme mix [2] (a mixture of Uracil DNA Glycosylase and endonuclease VIII) and sequenced with the single-strand protocol described in [7].

For all three samples, the rate of base substitutions due to aDNA damage was evaluated for each individual separately, using mapDamage2 [14]. These rates were then used by *fracSim* to add damage to the one million fragments. Subsequently, the 3 different simulated sequence sets were mapped back to chromosome 21 using BWA v.0.5.10 [18] with the following options: "-n 0.01 -o 2 -l 16500". Empirical and simulated nucleotide misincorporation profiles are shown in Supplementary Figures 11-13.

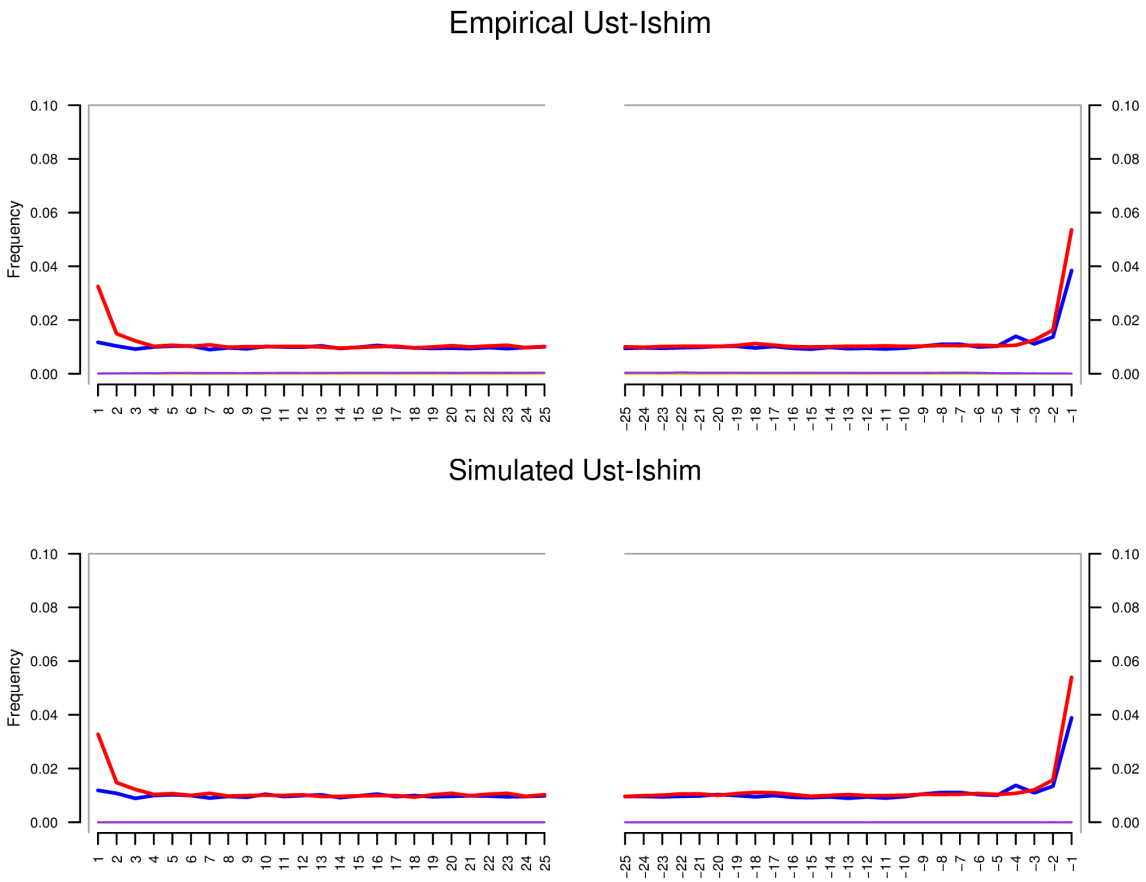
Our results show a high level of correlation between the empirical rates of damage and the simulated ones on a per sample basis . The plots were obtained using mapDamage2.0 [14].



Supplementary Figure 11: Simulated rates of aDNA damage (bottom) versus the original empirical rate of damage (top). Empirical rates of damage were taken from a 4.5k-year-old human ("ATP2") sequenced with a double-strand library protocol. The empirical sample had high levels of aDNA damage. The C to T substitutions are represented in red, the G to A substitutions are shown in blue and the remaining substitutions in grey.



Supplementary Figure 12: Simulated rates of aDNA damage (bottom) versus the original empirical rate of damage (top). Empirical rates of damage were taken from a 7k-year-old human ("LaBrana") sequenced with a double-strand library protocol. The empirical sample had medium levels of aDNA damage. The C to T substitutions are represented in red, the G to A substitutions are shown in blue and the remaining substitutions in grey.



Supplementary Figure 13: Simulated rates of aDNA damage (bottom) versus the original empirical rate of damage (top). Empirical rate of damage were taken from a 45k year old human ("Ust-Ishim") sequenced with a single-strand library protocol with USER treatment. The empirical sample had low levels of aDNA damage. Please note the Y-axis scale is different from Supplementary Figures 11 and 12 as DNA damage rates are less pronounced. The C to T substitutions are represented in red, the G to A substitutions are shown in blue.

2.5 Test case 1: impact of contamination on D-statistics

As a test case, we evaluated the impact of present-day human contamination on the outcome of admixture tests based on the D-statistics [22]. It is important to highlight that the goal of the test is not to make any general claims about the amount of present-day human contamination needed to create a spurious signal of admixture. To achieve this, a greater number of variables will still need to be considered such as the depth of coalescence, rates of contamination and the total number of alignment fragments. Our goal here is simply to illustrate one of the many possible utilizations of gargammel, namely its ability to allow the user to measure the impact of aDNA idiosyncrasies on downstream analyses.
Briefly, we used ms to simulate genomic data for 5 lineages, including one outgroup, two ancient individuals and two modern-day contaminants. More specifically, we have:

- "endogenous": Representing the actual endogenous or archaic individual
- "n_endogenous": A neighboring or closely related individual to the endogenous one
- "contaminant": The individual that contaminates the ancient sample
- "n_contaminant": A neighboring or closely related individual to the contaminant one
- "outgroup": An outgroup for all 4 individuals

For all lineages except the "outgroup", 2 individuals were generated as to simulate a diploid genome per population. Within a given population, lineages were joined after 0.15 units of coalescence. To put this framework in the perspective of human history, previous studies have used 0.1125 for the joining of Africans and non-Africans lineages and 0.3 for present-day humans and Neanderthal populations [36].
First, we generated set **A** by joining the contaminant and endogenous lineages at 0.32 units of coalescence and setting the outgroup at 6 units of coalescence. We also set θ to 20 as in [36] without any growth parameters using the following command:

```
ms 9 1 -T -t 20 -I 5 1 2 2 2 2  -ej 0.15 2 3 -ej 0.15 4 5
-ej 0.32 3 5 -ej 6 5 1
```

We created set **B** using the same conditions as for set A, except that we joined the outgroup at 3 units of coalescence:

```
ms  9 1 -T -t 20 -I 5 1 2 2 2 2  -ej 0.15 2 3 -ej 0.15 4 5
-ej 0.32 3 5 -ej 3 5 1
```

We created set **C** by joining the outgroup at 6 units of coalescence as in **A** but joined the contaminant and endogenous lineages much later, at 0.55 units of coalescence:

```
ms 9 1 -T -t 20 -I 5 1 2 2 2 2  -ej 0.15 2 3 -ej 0.15 4 5
-ej 0.55 3 5 -ej 6 5 1
```

For each set, we then generated a total of 3,000 sequences using seq-gen [25] with an HKY model[10], a length of 10kb per sequence and a branch length scaling factor of 0.00045.

The resulting trees computed using a maximum-likelihood criterion under the HKY model for the sequences generated for set **A**, **B** and **C** can be seen in Supplementary Figure 14.
We then used gargammel on the resulting data to simulate various levels of contamination:

```
gargammel.pl -c 3 --misince dnacomp.txt --comp
0,[cont. rate],[1-cont. rate]   --minsize 35 --loc
4.106487474 --scale  0.358874723 -damage 0.03,0.4,0.01,0.2
-o /path to output directory/ /path to input directory/
```

The command above was executed for various values of "[cont. rate]", representing the rate of present-day human contamination, ranging from 0 to 0.5 (ie. 50% contamination). The "minsize" parameter eliminates fragments with length less than 35bp. The location and scale parameters generate aDNA fragments with an average size of 64.7bp. Deamination was added using the Briggs-Johnson model implemented in mapDamage2.0 with nick frequency = 0.03, geometric parameter for the length of overhanging ends = 0.4, probability of deamination in double-stranded parts=0.01 and 0.2 in single-stranded parts.

A single simulation to generate an average coverage of 3X for an approximate total of 1.4M DNA fragments took 131 minutes on a single AMD Opteron processor 2.3GHz CPU on a machine with 128G of RAM.

D-statistics in the form of (Outgroup,n_contaminant; endogenous, n_endogenous) were computed for each simulated dataset, aiming at measuring the possible impact of contamination levels. Negative statistics indicate an excess of shared derived polymorphisms between n_contaminant and endogenous whereas positive statistics indicate such an excess between n_contaminant and n_endogenous. The D-statistics was computed using a base sampled at random without any base quality cutoffs. To mitigate the impact of simulated damage, the experiment is repeated using transversions only. Boostrap replicates were used to compute a Z-score.

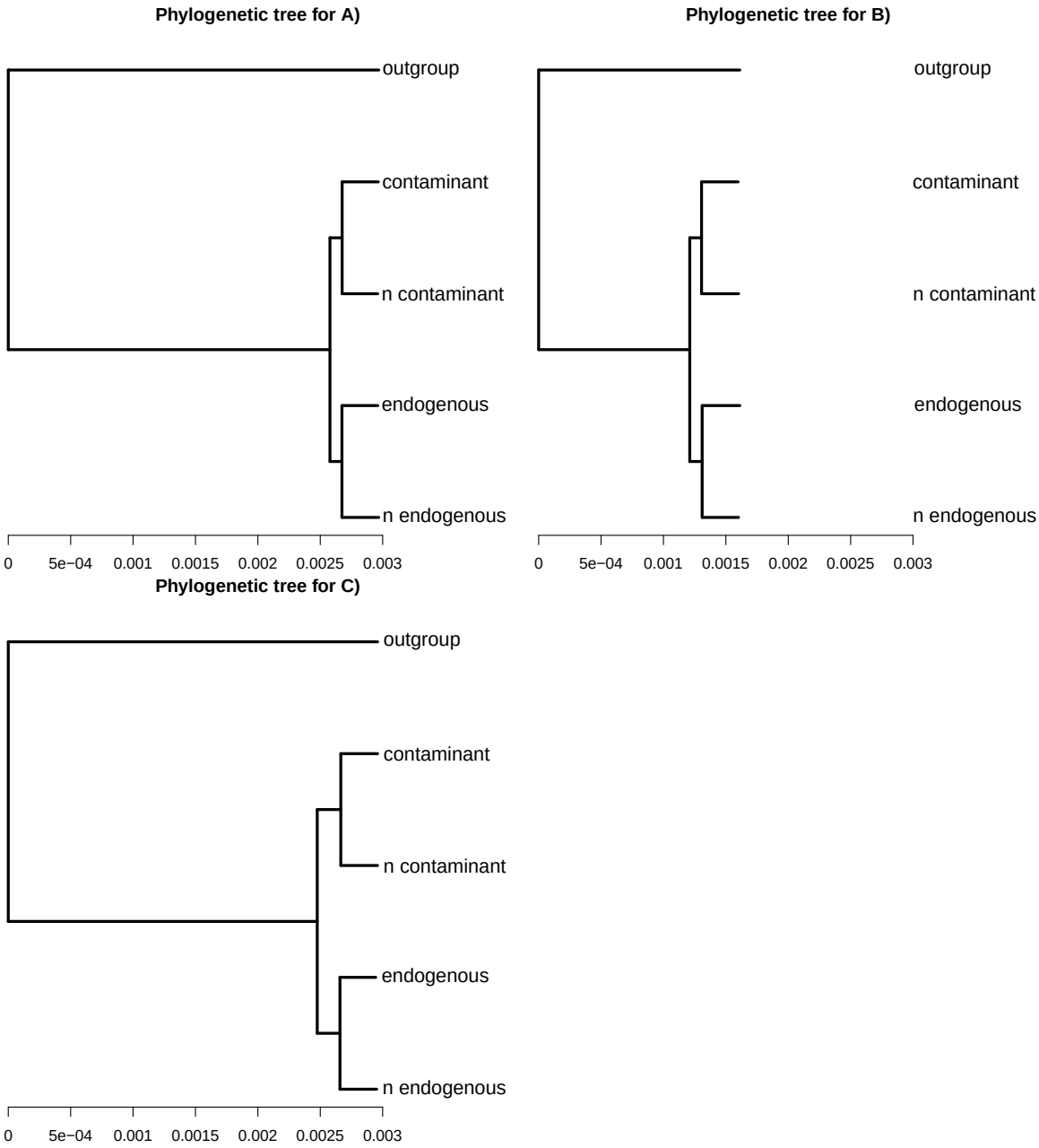
Using set **A** (see Table 3), with no contamination, if all sites are considered, there is a spurious signal of admixture from the individual serving as proxy for the contaminant ("n_contaminant") to the neighboring population of the endogenous ("n_endogenous"). This is likely due to the fact that the overwhelming majority of segregating sites are ancestral in the outgroup (by definition) and all 3 remaining chromosomes (2 endogenous and 1 contaminant) are derived (ie. sites are in the form of AB;BB, where A corresponds to the ancestral allele and B the derived one). Ancient DNA damage can convert the endogenous chromosome, which carried at such sites the derived allele, into the ancestral allele, which introduces a spurious excess of one of the patterns considered in the calculation of the D-statistics (namely, AB;AB). This effect is only present when all mutation types are considered, as expected, and leads to significantly positive D-statistics at $\sim$15% of contamination.

For the given number of sites and given phylogenetic topology in the simulation, it is only from approximately 35% present-day contamination, that the D-statistics ($Z \leq -3$) started showing evidence of gene flow between the contaminant and the endogenous sample. This effect was present when considering transversions only from approximately 30% contamination levels.

We then repeated the D-statistics calculation but restricting the analysis to fragments filtered for a post-mortem damage (deamination) score of 3 according to the methodology described in [35]. As no deamination was added to the contaminant fragments, such a procedure should enrich for endogenous fragments. If all types of segregating sites are included, the effect of having spurious signals of admixture from the n_contaminant proxy to the neighboring population of the endogenous described above (n_endogenous) is even more exacerbated due to an enrichment of deaminated fragments ($Z > 10.7$). However, when restricting the analysis to transversions only, the previous signal of admixture observed when all fragments are considered is lost, leading to genuine outcomes for the admixture test. Thus, having even moderate amounts of present-day contamination can result in signals of admixture if all fragments are considered but restricting the analysis to likely-damaged fragments and transversions only has the potential to minimize this risk.

We then measured the impact of demographic parameters in set **B**, joining the outgroup twice as early with the remaining 4 lineages (see Supplementary Table 4). No significantly positive D-statistics values were found considering all fragments and all mutation types, due to the strong reduction of ABBB patterns simulated. Contamination rates from 20% and above affected the outcome of the D-statistics test, supporting some gene flow between the n_contaminant and the n_endogenous population. Additionally, the predictable loss of segregating sites under the simulated model resulted in D-statistics that are less consistent as the D- and Z-scores are more uneven when considering transversions only.

Conversely, set **C** sought to evaluate the impact on the D-statistics of a deeper joining between the contaminant and endogenous lineages. The expected greater number of segregating sites yields a much clearer picture than set **A** (and especially set **B**). When restricting the analysis to transversions only, gene flow is inferred from $\sim$7% present-day human contamination. As the greater number of segregating sites gives a more consistent evaluation of the effect of present-day human contamination on the D-statistics, we thus recommend that, when shallower coalescence



Supplementary Figure 14: Phylogenetic trees of the sequences generated using simulations for sets **A**, **B** and **C**. The ancient sample (labeled "endogenous") has a neighboring population (labeled "n_endogenous"). The present-day human contaminant (labeled "contaminant") has also a neighboring population (labeled "n_contaminant") which was used as reference genome. A total of 2 individuals were simulated for every population as to simulated a diploid organism. Finally, a single individual (labeled "outgroup") was simulated to provide an outgroup.

times are to be explored, a greater number of simulations should be performed to measure the effect of contamination on the D-statistics.

contamination rate	all fragments		PMD filtered fragments	
	TS+TV	TV only	TS+TV	TV only
0	0.108 (6.395)	0.024 (1.107)	0.301 (11.875)	0.023 (0.608)
1	0.095 (5.655)	0.017 (0.785)	0.288 (10.895)	0.013 (0.340)
2	0.101 (5.870)	0.034 (1.563)	0.347 (13.064)	0.133 (3.312)
3	0.081 (4.672)	0.016 (0.766)	0.268 (10.739)	-0.006 (-0.154)
4	0.095 (5.564)	-0.011 (-0.509)	0.335 (13.633)	0.040 (1.042)
5	0.079 (4.559)	0.035 (1.587)	0.328 (13.357)	0.052 (1.389)
7	0.063 (3.795)	0.004 (0.179)	0.326 (13.468)	0.067 (1.817)
10	0.045 (2.556)	-0.022 (-0.967)	0.306 (11.932)	0.023 (0.622)
15	0.057 (3.330)	-0.014 (-0.645)	0.286 (11.825)	0.006 (0.155)
20	0.018 (1.092)	-0.059 (-2.812)	0.289 (11.444)	0.017 (0.452)
25	0.005 (0.279)	-0.051 (-2.502)	0.299 (11.959)	0.057 (1.500)
30	-0.043 (-2.611)	-0.110 (-5.330)	0.289 (11.364)	-0.027 (-0.672)
35	-0.050 (-3.080)	-0.121 (-6.035)	0.347 (13.415)	0.095 (2.471)
40	-0.076 (-4.636)	-0.136 (-6.667)	0.308 (11.912)	0.054 (1.378)
45	-0.094 (-6.004)	-0.164 (-8.507)	0.308 (12.115)	0.042 (1.115)
50	-0.123 (-8.111)	-0.158 (-8.451)	0.305 (12.088)	0.020 (0.505)

Supplementary Table 3: Value of the D-statistics for D(outgroup,n_contaminant,endogenous,n_endogenous) for set **A** at various levels of contamination. A negative D indicates gene flow from "n_contaminant" to "endogenous" (see Figure 14). The D-statistics is either computed on both transitions and transversions (TS+TV) or transversions only (TV only). The latter is performed to mitigate the impact of deamination on the D-statistics. The Z-score is reported in parentheses. A Z-score between -3 and 3 is generally is considered non-significant.

contamination rate	all fragments		PMD filtered fragments	
	TS+TV	TV only	TS+TV	TV only
0	-0.024 (-1.356)	-0.050 (-2.302)	0.162 (5.770)	0.023 (0.602)
1	0.026 (1.502)	-0.001 (-0.070)	0.214 (7.598)	0.049 (1.269)
2	0.005 (0.266)	-0.021 (-0.983)	0.200 (7.479)	0.082 (2.295)
3	0.026 (1.527)	-0.012 (-0.536)	0.185 (6.754)	0.033 (0.873)
4	0.000 (0.000)	-0.020 (-0.912)	0.127 (4.606)	-0.030 (-0.819)
5	-0.017 (-0.971)	-0.043 (-1.952)	0.165 (6.182)	-0.019 (-0.507)
7	0.011 (0.618)	-0.017 (-0.789)	0.161 (5.568)	0.026 (0.664)
10	-0.037 (-2.162)	-0.060 (-2.838)	0.169 (6.196)	0.017 (0.448)
15	-0.016 (-0.908)	-0.041 (-1.951)	0.182 (6.711)	0.027 (0.738)
20	-0.055 (-3.255)	-0.074 (-3.582)	0.132 (4.683)	-0.040 (-1.025)
25	-0.056 (-3.387)	-0.060 (-2.842)	0.114 (4.044)	-0.032 (-0.845)
30	-0.090 (-5.574)	-0.121 (-6.025)	0.145 (5.059)	-0.037 (-0.916)
35	-0.086 (-5.163)	-0.100 (-4.993)	0.192 (6.865)	0.020 (0.531)
40	-0.127 (-7.737)	-0.136 (-6.625)	0.160 (5.949)	-0.013 (-0.357)
45	-0.155 (-9.850)	-0.199 (-10.324)	0.158 (5.716)	-0.032 (-0.843)
50	-0.160 (-10.306)	-0.167 (-8.563)	0.153 (5.637)	-0.077 (-2.020)

Supplementary Table 4: Value of the D-statistics for D(outgroup,n_contaminant,endogenous,n_endogenous) for set **B** at various levels of contamination. Joining the outgroup before results in a loss of statistical power.

contamination rate	all fragments		PMD filtered fragments	
	TS+TV	TV only	TS+TV	TV only
0	0.087 (4.124)	0.019 (0.720)	0.324 (11.185)	-0.002 (-0.047)
1	0.085 (3.975)	-0.031 (-1.084)	0.378 (13.027)	0.042 (0.822)
2	0.054 (2.582)	-0.034 (-1.229)	0.276 (9.279)	-0.116 (-2.442)
3	0.040 (1.977)	-0.047 (-1.665)	0.374 (13.134)	0.031 (0.623)
4	-0.003 (-0.126)	-0.073 (-2.627)	0.340 (12.137)	0.026 (0.544)
5	0.027 (1.386)	-0.067 (-2.579)	0.335 (11.566)	0.026 (0.512)
7	0.013 (0.676)	-0.108 (-4.135)	0.335 (11.502)	-0.053 (-1.117)
10	-0.031 (-1.559)	-0.127 (-4.997)	0.338 (11.999)	-0.031 (-0.631)
15	-0.116 (-5.773)	-0.195 (-7.769)	0.364 (12.211)	0.015 (0.319)
20	-0.166 (-8.720)	-0.237 (-9.996)	0.298 (9.987)	-0.029 (-0.596)
25	-0.190 (-10.495)	-0.271 (-11.745)	0.339 (11.838)	0.027 (0.566)
30	-0.259 (-15.781)	-0.311 (-14.457)	0.292 (9.697)	-0.021 (-0.458)
35	-0.297 (-17.962)	-0.366 (-17.750)	0.359 (12.473)	0.007 (0.151)
40	-0.330 (-20.743)	-0.407 (-21.220)	0.351 (11.982)	0.059 (1.225)
45	-0.389 (-24.921)	-0.442 (-23.507)	0.363 (12.822)	0.051 (1.077)
50	-0.417 (-28.920)	-0.466 (-25.682)	0.340 (11.948)	0.047 (0.961)

Supplementary Table 5: Value of the D-statistics for D(outgroup,n_contaminant,endogenous,n_endogenous) for set **C** at various levels of contamination. Joining the contaminant and the endogenous sub-trees later results in a greater number of segregating sites and an increase of statistical power.

2.6 Test case 2: impact of microbial contamination on DNA fragment alignments

As a second test case, the impact of having large amounts of microbial contamination on DNA fragment alignments was evaluated. As endogenous genome, we used the genome of the Clovis individual from [26]. Genotypes were called from the aligned BAM files using samtools/bcftools[17]. Fasta files were inferred using the "consensus" command from bcftools and used as the endogenous source of DNA.

The taxonomic profile of the microbial communities present in the DNA extract was obtained as indicated in Supplementary Section 1.2, using metaBIT [20]. The resulting microbial vector indicating detected microbial species and their relative abundance was used as the source of microbial contamination. Various levels of microbial contamination (denoted b) were simulated across a range of fragment lengths (denoted l).

A total of 2M DNA fragments were simulated using gargammel to generate a dataset representing an Illumina HiSeq 2500 sequencing run for those 2M fragments. The resulting simulated reads were treated similarly to an empirical aDNA dataset. More precisely, the simulated pairs of reads showing significant overlap were merged into a single DNA fragment using leeHom [27] using the "–ancientdna" option. The resulting data were mapped to the human reference using BWA v.0.5.10 [18] with the following options: "-n 0.01 -o 2 -l 16500".

As expected, the alignment statistics show that for very short fragments (e.g. 20bp), the rate of misalignment to the human reference is quite high even when filtering for mapping quality greater than 30 (see Table 6). However, having larger fragments (e.g. greater than 35bp) reduces the chance of misalignment to the human reference due to microbial contamination. The mapping quality reported by BWA does not take account aDNA damage when computing mapping qualities (see [15] for a discussion how to include this damage in a probabilistic framework while mapping). However, the objective of this section is to evaluate the impact of bacterial DNA fragments on mapping using the most common methodologies in aDNA research. The procedure described here can be easily implemented to select the minimal size threshold that best matches the molecular complexity found in a given sample for a specific mapping methodology.

l	m (%)	merged reads into fragment	aligned fragments		aligned microbial fragments		aligned endogenous fragments	
			all	MQ>30	all	MQ>30	all	MQ>30
20	0	1,999,747 (99.99%)	1,999,747 (99.99%)	569,473 (28.47%)	0 (0.00%)	0 (0.00%)	1,999,747 (99.99%)	569,473 (28.47%)
20	25	1,999,756 (99.99%)	1,957,972 (97.90%)	454,600 (22.73%)	458,155 (22.91%)	28,021 (1.40%)	1,499,817 (74.99%)	426,579 (21.33%)
20	50	1,999,771 (99.99%)	1,916,870 (95.84%)	340,769 (17.04%)	916,991 (45.85%)	56,332 (2.82%)	999,879 (49.99%)	284,437 (14.22%)
20	75	1,999,767 (99.99%)	1,875,272 (93.76%)	226,788 (11.34%)	1,375,318 (68.77%)	84,620 (4.23%)	499,954 (25.00%)	142,168 (7.11%)
20	100	1,999,761 (99.99%)	1,833,934 (91.70%)	113,026 (5.65%)	1,833,934 (91.70%)	113,026 (5.65%)	0 (0.00%)	0 (0.00%)
25	0	1,999,716 (99.99%)	1,999,708 (99.99%)	1,438,477 (71.92%)	0 (0.00%)	0 (0.00%)	1,999,708 (99.99%)	1,438,477 (71.92%)
25	25	1,999,714 (99.99%)	1,777,842 (88.89%)	1,153,636 (57.68%)	278,067 (13.90%)	75,159 (3.76%)	1,499,775 (74.99%)	1,078,477 (53.92%)
25	50	1,999,698 (99.98%)	1,555,735 (77.79%)	870,870 (43.54%)	555,886 (27.79%)	151,082 (7.55%)	999,849 (49.99%)	719,788 (35.99%)
25	75	1,999,738 (99.99%)	1,334,497 (66.72%)	585,662 (29.28%)	834,574 (41.73%)	225,487 (11.27%)	499,923 (25.00%)	360,175 (18.01%)
25	100	1,999,723 (99.99%)	1,114,216 (55.71%)	301,105 (15.06%)	1,114,216 (55.71%)	301,105 (15.06%)	0 (0.00%)	0 (0.00%)
30	0	1,999,667 (99.98%)	1,999,559 (99.98%)	1,546,319 (77.32%)	0 (0.00%)	0 (0.00%)	1,999,559 (99.98%)	1,546,319 (77.32%)
30	25	1,999,674 (99.98%)	1,549,805 (77.49%)	1,172,250 (58.61%)	50,124 (2.51%)	11,742 (0.59%)	1,499,681 (74.98%)	1,160,508 (58.03%)
30	50	1,999,685 (99.98%)	1,099,665 (54.98%)	797,221 (39.86%)	99,853 (4.99%)	23,593 (1.18%)	999,812 (49.99%)	773,628 (38.68%)
30	75	1,999,694 (99.98%)	649,441 (32.47%)	421,292 (21.06%)	149,545 (7.48%)	35,310 (1.77%)	499,896 (24.99%)	385,982 (19.30%)
30	100	1,999,699 (99.98%)	198,399 (9.92%)	46,648 (2.33%)	198,399 (9.92%)	46,648 (2.33%)	0 (0.00%)	0 (0.00%)
35	0	1,999,681 (99.98%)	1,999,529 (99.98%)	1,615,425 (80.77%)	0 (0.00%)	0 (0.00%)	1,999,529 (99.98%)	1,615,425 (80.77%)
35	25	1,999,671 (99.98%)	1,512,450 (75.62%)	1,214,278 (60.71%)	12,797 (0.64%)	3,290 (0.16%)	1,499,653 (74.98%)	1,210,988 (60.55%)
35	50	1,999,641 (99.98%)	1,025,113 (51.26%)	813,792 (40.69%)	25,348 (1.27%)	6,677 (0.33%)	999,765 (49.99%)	807,115 (40.36%)
35	75	1,999,655 (99.98%)	538,021 (26.90%)	412,593 (20.63%)	38,142 (1.91%)	9,652 (0.48%)	499,879 (24.99%)	402,941 (20.15%)
35	100	1,999,661 (99.98%)	51,186 (2.56%)	13,335 (0.67%)	51,186 (2.56%)	13,335 (0.67%)	0 (0.00%)	0 (0.00%)
40	0	1,999,628 (99.98%)	1,999,423 (99.97%)	1,668,895 (83.44%)	0 (0.00%)	0 (0.00%)	1,999,423 (99.97%)	1,668,895 (83.44%)
40	25	1,999,642 (99.98%)	1,504,291 (75.21%)	1,253,374 (62.67%)	4,713 (0.24%)	926 (0.05%)	1,499,578 (74.98%)	1,252,448 (62.62%)
40	50	1,999,616 (99.98%)	1,009,004 (50.45%)	836,166 (41.81%)	9,304 (0.47%)	1,836 (0.09%)	999,700 (49.98%)	834,330 (41.72%)
40	75	1,999,614 (99.98%)	513,899 (25.69%)	420,226 (21.01%)	14,067 (0.70%)	2,786 (0.14%)	499,832 (24.99%)	417,440 (20.87%)
40	100	1,999,643 (99.98%)	18,972 (0.95%)	3,671 (0.18%)	18,972 (0.95%)	3,671 (0.18%)	0 (0.00%)	0 (0.00%)
45	0	1,999,606 (99.98%)	1,999,567 (99.98%)	1,714,028 (85.70%)	0 (0.00%)	0 (0.00%)	1,999,567 (99.98%)	1,714,028 (85.70%)
45	25	1,999,611 (99.98%)	1,503,705 (75.19%)	1,285,569 (64.28%)	4,030 (0.20%)	856 (0.04%)	1,499,675 (74.98%)	1,284,713 (64.24%)
45	50	1,999,592 (99.98%)	1,007,930 (50.40%)	858,275 (42.91%)	8,151 (0.41%)	1,725 (0.09%)	999,779 (49.99%)	856,550 (42.83%)
45	75	1,999,602 (99.98%)	512,204 (25.61%)	431,272 (21.56%)	12,311 (0.62%)	2,735 (0.14%)	499,893 (24.99%)	428,537 (21.43%)
45	100	1,999,598 (99.98%)	16,543 (0.83%)	3,539 (0.18%)	16,543 (0.83%)	3,539 (0.18%)	0 (0.00%)	0 (0.00%)

Supplementary Table 6: Tally of simulated aDNA fragments from either an endogenous or a microbial source aligning to the human genome reference. The number of simulated reads that were merged into a single reconstructed fragment is in the third column. Only reads that were merged into a fragment were considered. The number in parentheses represents the fraction of the initial number of simulated fragments. 'l' and 'm' represent the fragment length and microbial contamination, respectively.

References

[1] Adrian W. Briggs, Udo Stenzel, Philip L.F. Johnson, Richard E. Green, Janet Kelso, Kay Prüfer, Matthias Meyer, Johannes Krause, Michael T. Ronan, Michael Lachmann, et al. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, 104(37):14616–14621, 2007.

[2] Adrian W Briggs, Udo Stenzel, Matthias Meyer, Johannes Krause, Martin Kircher, and Svante Pääbo. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Research*, 38(6):e87–e87, 2010.

[3] Yen-Chun Chen, Tsunglin Liu, Chun-Hui Yu, Tzen-Yuh Chiang, and Chi-Chuan Hwang. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLOS ONE*, 8(4):e62856, 2013.

[4] Jesse Dabney and Matthias Meyer. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques*, 52(2):87–94, 2012.

[5] Qiaomei Fu, Heng Li, Priya Moorjani, Flora Jay, Sergey M. Slepchenko, Aleksei A. Bondarev, Philip L.F. Johnson, Ayinuer Aximu-Petri, Kay Prüfer, Cesare de Filippo, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 514(7523):445–449, 2014.

[6] Cristina Gamba, Eppie R Jones, Matthew D Teasdale, Russell L McLaughlin, Gloria Gonzalez-Fortes, Valeria Mattiangeli, László Domboróczki, Ivett Kővári, Ildikó Pap, Alexandra Anders, et al. Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, 5, 2014.

[7] Marie-Theres Gansauge and Matthias Meyer. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols*, 8(4):737–748, 2013.

[8] Torsten Günther, Cristina Valdiosera, Helena Malmström, Irene Ureña, Ricardo Rodriguez-Varela, Óddny Osk Sverrisdóttir, Evangelia A Daskalaki, Pontus Skoglund, Thijessen Naidoo, Emma M Svensson, et al. Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proceedings of the National Academy of Sciences*, 112(38):11917–11922, 2015.

[9] Kristian Hanghøj, Andaine Seguin, Mikkel Schubert, Tobias Madsen, Jakob Skou Pedersen, Eske Willerslev, and Ludovic Orlando. Fast, accurate and automatic ancient nucleosome and methylation maps with epiPALEOMIX. *Molecular Biology and Evolution*, page msw184, 2016.

[10] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[11] Michael Hofreiter, Viviane Jaenicke, David Serre, Arndt von Haeseler, and Svante Pääbo. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research*, 29(23):4793–4799, 2001.

[12] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 2012.

[13] Richard R Hudson. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[14] Hákon Jónsson, Aurélien Ginolhac, Mikkel Schubert, Philip L.F. Johnson, and Ludovic Orlando. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, 29(13):1682–1684, 2013.

[15] Peter Kerpedjiev, Jes Frellsen, Stinus Lindgreen, and Anders Krogh. Adaptable probabilistic mapping of short reads using position specific scoring matrices. *BMC Bioinformatics*, 15(1):1, 2014.

[16] Petra Korlević, Tobias Gerber, Marie-Theres Gansauge, Mateja Hajdinjak, Sarah Nagel, Aximu Ayinuer-Petri, and Matthias Meyer. Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *BioTechniques*, 59(2):87–93, 2015.

[17] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[18] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[19] Pablo Librado, Clio Der Sarkissian, Luca Ermini, Mikkel Schubert, Hákon Jónsson, Anders Albrechtsen, Matteo Fumagalli, Melinda A Yang, Cristina Gamba, Andaine Seguin-Orlando, et al. Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. *Proceedings of the National Academy of Sciences*, 112(50):E6889–E6897, 2015.

[20] Guillaume Louvel, Clio Der Sarkissian, Kristian Hanghøj, and Ludovic Orlando. metaBIT, an integrative and automated metagenomic pipeline for analyzing microbial profiles from high-throughput sequencing shotgun data. *Molecular Ecology Resources*, 2016.

[21] Matthias Meyer and Martin Kircher. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 2010(6):pdb.prot5448, 2010.

[22] Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

[23] Jakob Skou Pedersen, Eivind Valen, Amhed M Vargas Velazquez, Brian J Parker, Morten Rasmussen, Stinus Lindgreen, Berit Lilje, Desmond J Tobin, Theresa K Kelly, Søren Vang, et al. Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Research*, 24(3):454–466, 2014.

[24] Michael A Quail, Iwanka Kozarewa, Frances Smith, Aylwyn Scally, Philip J Stephens, Richard Durbin, Harold Swerdlow, and Daniel J Turner. A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, 5(12):1005–1010, 2008.

[25] Andrew Rambaut and Nicholas C Grass. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences: CABIOS*, 13(3):235–238, 1997.

[26] Morten Rasmussen, Sarah L Anzick, Michael R Waters, Pontus Skoglund, Michael DeGiorgio, Thomas W Stafford Jr, Simon Rasmussen, Ida Moltke, Anders Albrechtsen, Shane M Doyle, et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*, 506(7487):225–229, 2014.

[27] Gabriel Renaud, Udo Stenzel, and Janet Kelso. leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Research*, 42(18):e141, 2014.

[28] Ermanno Rizzi, Martina Lari, Elena Gigli, Gianluca De Bellis, and David Caramelli. Ancient DNA studies: new perspectives on old samples. *Genetics Selection Evolution*, 44(1):1, 2012.

[29] Nadin Rohland, Eadaoin Harney, Swapan Mallick, Susanne Nordenfelt, and David Reich. Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 370(1660):20130624, 2015.

[30] Federico Sánchez-Quinto, Hannes Schroeder, Oscar Ramirez, María C Ávila-Arcos, Marc Pybus, Iñigo Olalde, Amhed MV Velazquez, María Encina Prada Marcos, Julio Manuel Vidal Encinas, Jaume Bertranpetit, et al. Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. *Current Biology*, 22(16):1494–1499, 2012.

[31] Mikkel Schubert, Aurelien Ginolhac, Stinus Lindgreen, John F. Thompson, Khaled A.S. Al-Rasheid, Eske Willerslev, Anders Krogh, and Ludovic Orlando. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, 13(1):178, 2012.

[32] Mikkel Schubert, Stinus Lindgreen, and Ludovic Orlando. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes*, 9(1):1, 2016.

[33] Andaine Seguin-Orlando, Cristina Gamba, Clio Der Sarkissian, Luca Ermini, Guillaume Louvel, Eugenia Boulygina, Alexey Sokolov, Artem Nedoluzhko, Eline D Lorenzen, Patricio Lopez, et al. Pros and cons of methylation-based enrichment methods for ancient DNA. *Scientific Reports*, 5, 2015.

[34] Andaine Seguin-Orlando, Thorfinn S Korneliussen, Martin Sikora, Anna-Sapfo Malaspinas, Andrea Manica, Ida Moltke, Anders Albrechtsen, Amy Ko, Ashot Margaryan, Vyacheslav Moiseyev, et al. Genomic structure in Europeans dating back at least 36,200 years. *Science*, 346(6213):1113–1118, 2014.

[35] Pontus Skoglund, Bernd H. Northoff, Michael V. Shunkov, Anatoli P. Derevianko, Svante Pääbo, Johannes Krause, and Mattias Jakobsson. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, 111(6):2229–2234, 2014.

[36] Melinda A Yang, Anna-Sapfo Malaspinas, Eric Y Durand, and Montgomery Slatkin. Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. *Molecular Biology and Evolution*, 29(10):2987–2995, 2012.

[37] Xiaofan Zhou and Antonis Rokas. Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. *Molecular Ecology*, 23(7):1679–1700, 2014.