Genome Biology

SOFTWARE Open Access

AuthentiCT: a model of ancient DNA damage to estimate the proportion of present-day DNA contamination

Stéphane Peyrégne* and Benjamin M. Peter

* Correspondence: stephane.peyregne@gmail.com
Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany

Abstract

Contamination from present-day DNA is a fundamental issue when studying ancient DNA from historical or archaeological material, and quantifying the amount of contamination is essential for downstream analyses. We present AuthentiCT, a command-line tool to estimate the proportion of present-day DNA contamination in ancient DNA datasets generated from single-stranded DNA libraries. The prediction is based solely on the patterns of post-mortem damage observed on ancient DNA sequences. The method has the power to quantify contamination from as few as 10,000 mapped sequences, making it particularly useful for analysing specimens that are poorly preserved or for which little data is available.

Keywords: Contamination, Ancient DNA, Deamination, Damage patterns

Background

After the death of an organism, its DNA decays and is progressively lost through time [1, 2]. Under favourable conditions, DNA can preserve for hundreds of thousands of years and provide valuable information about the evolutionary history of organisms [3, 4]. Yet, only minute amounts of ancient DNA (aDNA) often remain in historical or archaeological material. In addition, most of the extracted DNA usually comes from microorganisms that spread in decaying tissues [5, 6]. Whereas microbial sequences rarely align to the reference genome used for identifying endogenous sequences if appropriate length cut-offs are used [7–9], contamination with DNA from closely related organisms represents a recurrent problem [10–12]. This is particularly true for the genomic analyses of ancient humans, as the individuals handling the specimens during excavation and at later times often leave their DNA behind [13, 14]. Because this contamination can substantially affect the results of population genetic or phylogenetic analyses, quantifying the level of contamination is crucial for downstream analyses. An estimate of the level of present-day DNA contamination is also desirable for making decisions when screening samples to identify those that can be further sequenced with reasonable effort and expenses.

Approaches to quantify the level of contamination can be divided into three categories. Some methods rely on prior knowledge of sequence differences between the contaminating and endogenous genomes [15–17]. Alternatively, if these differences are unknown a priori, other methods evaluate the excess of alleles compared to the expected ploidy [18–21]. The third set of methods uses patterns of chemical damage that are characteristic of aDNA [17, 22].

Amongst the approaches that rely on genetic differences, the most common strategy is to identify diagnostic positions expected to differ between the contaminating and endogenous sequences [16, 20]. The proportion of sequences that carry the contaminant allele at diagnostic positions represents an estimate of the level of contamination. This approach is particularly well-tailored for studying the mitochondrial genome, because it is an extensively studied, non-recombining locus that is often available at high coverage. In contrast, local genealogies along the nuclear genome may differ from the overall population relationship (incomplete lineage sorting), making the identification of diagnostic positions difficult. By leveraging differences in allele frequencies between populations, it is possible to estimate the proportion of present-day human DNA contamination amongst nuclear sequences from archaic hominins [21–23]. For the analysis of early modern humans, this approach remains challenging because of the lack of knowledge about rare sequence variants in the sample of interest that are unlikely to be shared with the present-day human contaminant. Thus, contamination estimates obtained from mitochondrial sequences are often used as a proxy for the level of nuclear DNA contamination [24–26]. However, the ratio of mitochondrial DNA to nuclear DNA may vary between the endogenous and contaminating DNA [27, 28], leading to potential differences in the level of contamination between the mitochondrial and nuclear genomes [29, 30].

Other approaches that compare the nuclear genetic differences between contaminating and endogenous genomes commonly exploit the ploidy of the sex chromosomes [18, 19, 31]. For instance, apparent heterozygous sites on the X chromosome of a male individual or sequences mapping to the Y chromosome for a female individual represent evidence of DNA contamination. Although these analyses do not rely on a prior knowledge about the ancestry of the ancient individual, they are either restricted to the X chromosome of male samples or cannot detect female contamination in female samples. Another concern is that the level of contamination may differ between the sex chromosomes and the autosomes if the sexes of the contaminant(s) and the ancient individual differ. Other approaches for the autosomes exist, e.g. methods using apparent alternative alleles at homozygous positions or an allelic imbalance at heterozygous positions [20, 21]. However, such approaches assume that high sequence coverage is available.

Alternatively, properties of aDNA molecules can be used to estimate contamination. Ancient DNA is typically fragmented into pieces shorter than 100 bp and exhibits miscoding base modifications that accumulate over time [32–35]. The most common miscoding lesions observed in aDNA are the results of cytosine deamination [36–39] that converts cytosine (C) into uracil (U), which is then misread as thymine (T), or 5-methylcytosine into thymine. These apparent C-to-T substitutions occur preferentially toward the ends of sequences [39], likely because single-stranded overhangs, which are common in aDNA, exhibit a rate of cytosine deamination about two orders of magnitude higher than double-stranded regions [1, 40]. To estimate present-day DNA contamination, these properties need to be formalised in a model of aDNA damage. The simplest approach ("conditional substitution analysis") is based on a model that assumes independence between C-to-T substitutions at both ends of sequences. Testing whether these substitutions are correlated between ends may reveal a set of undamaged sequences, which are likely contaminants [3]. This method works even for low sequence coverage but is primarily used to indicate the presence of contamination. Other methods extend this approach by considering the distribution of C-to-T substitutions along sequences, either assuming a parametric ([39], PMDtools [41] and mapDamage [42, 43]) or empirical distribution of these substitutions along sequences (contDeam [17] and aRchaic [44]). Notably, these methods assume that C-to-T substitutions in the aDNA sequences are independent of each other.

Here, we introduce a novel model for aDNA damage that does not assume independence between C-to-T substitutions. Our implementation, AuthentiCT, allows both estimation of the present-day DNA contamination rate and deconvolution of endogenous and contaminating sequences solely based on patterns of aDNA damage, which makes it applicable to any species, if a suitable reference genome is available for alignment. Applying this method to both simulated and existing aDNA datasets, we find that present-day DNA contamination can be estimated from as few as 10,000 sequences, making it a practical tool in the screening of samples for aDNA preservation.

Results

Method overview

In this section, we first motivate our approach by studying the features of aDNA damage. We then formalise a model of aDNA damage and develop a mixture model to describe and distinguish endogenous from putative contaminating sequences.

Ancient DNA deamination patterns used in this study

Deamination patterns in aDNA sequences depend on the DNA library preparation method used [45]. Some methods involve ligation of adapters to double-stranded DNA ("double-stranded libraries", [46]) while other methods convert the two DNA strands into separate library molecules ("single-stranded libraries", [47]). Here, we focus on the damage patterns in single-stranded libraries, as they fully preserve the strand orientation of the sequenced DNA fragments and are widely used in aDNA studies [48–56].

While C-to-T substitutions occur predominantly at the ends of DNA fragments, they are also common in the internal part [39, 57]. Our model is motivated by the finding that these internal C-to-Ts are not independent of each other (Fig. 1). Excluding the first and last five bases to mask potential overhangs, we found that C-to-Ts are particularly common in adjacent positions in many samples, with a significant deviation from the geometric distribution expected from independent events ($p < 10^{-15}$, chi-square goodness-of-fit test; see Additional file 1: Supplementary Note 1 for more details or for results excluding the first and last ten bases), and from a control using sheared present-day human DNA that was treated like aDNA [9]. Single-stranded regions inside aDNA fragments represent a possible cause.

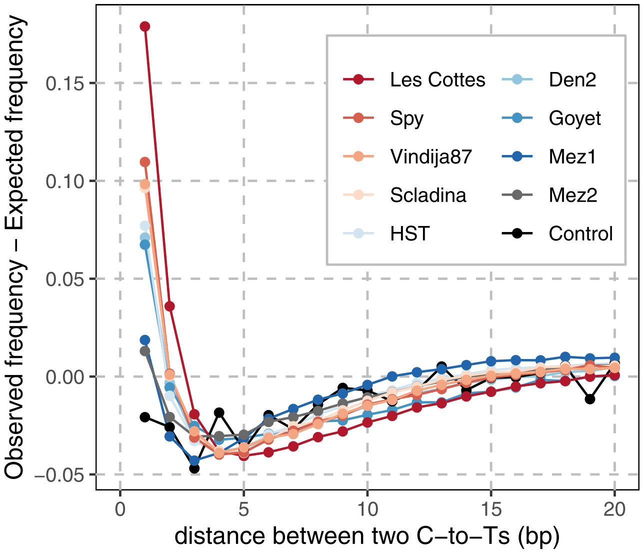

Fig. 1 Excess of directly adjacent C-to-T substitutions inside aDNA sequences. The observed and expected frequencies of the distance between pairs of C-to-T substitutions are compared amongst sequences that exhibit two internal C-to-T substitutions (excluding the first and last 5 bases). The colours correspond to sequence data from different archaic humans (> 500,000 sequences each; Les Cottés Z4-1514, Spy 94a, Vindija 87, Mezmaiskaya 2, Goyet Q56-1 from [58]; Scladina I-4A and Hohlenstein-Stadel from [23]; Denisova 2 from [59]; Mezmaiskaya 1 from [60]). The present-day human DNA control dataset is represented in black [9]. All sequences are derived from single-stranded DNA libraries

Model of ancient DNA damage

Motivated by this finding, we developed a model of aDNA damage that jointly models all C-to-T substitutions, accounting for the observed clustering of C-to-T substitutions within a sequence. We used a hidden Markov model (HMM), where each potentially deaminated site in the reference is an informative site, i.e. Cs or Gs for sequences aligning to the forward or reverse strand, respectively. Other positions will give the same likelihood for the endogenous and contaminating DNA models and are therefore ignored in both models. At the C or G positions of the reference genome, we classify

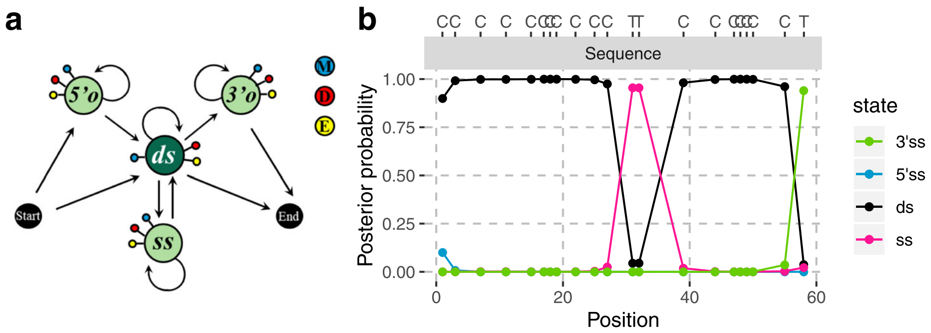

Fig. 2 Graphical representation of the model (a) and illustration of the posterior decoding (**b**). **a** States are depicted by nodes and transitions by edges. Each state emits a match to the reference M (blue) or a mismatch, which can either be compatible with cytosine deamination, D (red), or an error (or polymorphism), E (yellow). Single-stranded states (*5'o*, *3'o* and *ss*) and the double-stranded state (*ds*) are in light and dark green, respectively. **b** The posterior probability for each state is shown with different colours. (see Additional file 1: Supplementary Note 2 for an evaluation of the posterior decoding). We show bases on top for positions in the sequence (grey bar) that align to a C in the reference

observations either as emitting “M” (matches the reference allele), “D” (differs from the reference allele; compatible with a deamination) or “E” (other mismatches corresponding to sequencing errors or polymorphisms) (Fig. 2).

The model attempts to infer which parts of the underlying DNA molecule were single-stranded and double-stranded. It uses four hidden states corresponding to double-stranded (*ds*) or single-stranded stretches. We further separate internal single-stranded regions (*ss*) from 5′ (*5’o*) and 3′ (*3’o*) single-stranded overhangs. At the first position of the sequence alignment, the chain starts in a 5′ single-stranded overhang with probability *o* or in a double-stranded state with probability 1-*o*. Then, the lengths of the different regions follow geometric distributions, with parameters l_o, l_{ss} or l_{ds} for the overhangs, single-stranded and double-stranded regions, respectively. We therefore assumed the following matrix of transition probabilities:

$$\begin{array}{c} \\ 5'o \\ ds \\ ss \\ 3'o \end{array} \begin{array}{c} \begin{array}{cccc} 5'o & ds & ss & 3'o \end{array} \\ \begin{pmatrix} (1-l_o)^{d_5} & 1-(1-l_o)^{d_5} & 0 & 0 \\ 0 & (1-l_{ds})^{d_5}\times(1-o_2(1-l_o)^{d_3}) & (1-(1-l_{ds})^{d_5})\times(1-o_2(1-l_o)^{d_3}) & o_2(1-l_o)^{d_3} \\ 0 & 1-(1-l_{ss})^{d_5} & (1-l_{ss})^{d_5} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

with d_5 representing the distance from the previous observation (set to the current position for the initial observation), d_3 the distance to the end of the sequence and o_2 the probability of a 3′ single-stranded overhang. The modelling of the state transitions between informative sites assumes that only one transition happens between two informative sites.

The chain ends with the following transition probabilities:

$$\begin{array}{c} \\ 5'o \\ ds \\ ss \\ 3'o \end{array} \begin{array}{c} \begin{array}{cccc} 5'o & ds & ss & 3'o \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & (1-l_{ds})^{d_3} & 0 & 1-(1-l_{ds})^{d_3} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

For the emission probabilities of the three possible observations mentioned above, we assumed that all single-stranded states (*5’o*, *3’o* and *ss*) have the same emissions:

$$\begin{array}{c} \mathrm{M} \\ \mathrm{D} \\ \mathrm{E} \end{array} \begin{pmatrix} (1-e)(1-r_{ss})+\frac{e}{3}r_{ss} \\ (1-e)r_{ss}+\frac{e}{3}(1-r_{ss}) \\ \frac{2e}{3} \end{pmatrix}.$$

Similarly, the emission vector for the double-stranded state is:

$$\begin{array}{c} \mathrm{M} \\ \mathrm{D} \\ \mathrm{E} \end{array} \begin{pmatrix} (1-e)(1-r_{ds})+\frac{e}{3}r_{ds} \\ (1-e)r_{ds}+\frac{e}{3}(1-r_{ds}) \\ \frac{2e}{3} \end{pmatrix}.$$

Here, r_{ss} and r_{ds} denote the deamination rates in single-stranded regions (including the single-stranded overhangs) and double-stranded regions, respectively. We model polymorphisms and errors using a single rate, *e*, as these are indistinguishable without prior knowledge of polymorphic sites in the genomes of the source populations/species. We also note that all types of substitutions are assumed to be equally likely, a simplification that has also been used in previous work [15, 18].

Model of present-day DNA contamination

To identify contamination, we contrast the aDNA model with a model for DNA without deamination. We assumed that any difference to the reference genome arose from a constant mismatch rate e along the sequence. Assuming independence between sites, the probability of the data is simply: $e^d(1 - e)^s$ where d and s are the number of mismatches and matches to the reference at informative positions, respectively. We used the same rate e for both endogenous and contaminating sequences.

Estimating present-day DNA contamination

We use a mixture model to estimate the overall contamination rate c. We denote the i-th sequence as X_i, assuming we have N sequences in total. Using the aDNA sequence model, for each sequence, we calculate $P_E\ (X_i \mid \Theta, e)$, the probability of the sequence given that the corresponding DNA fragment is endogenous, conditional on the HMM parameters Θ. Similarly, using the model of contaminating DNA outlined above, we calculate $P_C\ (X_i \mid e)$, the probability of the sequence given that it originates from a contaminating DNA fragment. Therefore, we have $P\ (X_i \mid \Theta,\ e) = c\ P_C\ (X_i|e) + (1 - c)\ P_E\ (X_i \mid \Theta,\ e)$. Further assuming sequences are independent, the complete likelihood is $P(X|\Theta, e) = \prod_{i=1}^{N} (P_E, (X_i|\Theta, e), (1 - c), +, P_C, (X_i|e), c)$. We obtain a maximum likelihood estimate of c using the L-BFGS-B algorithm (tolerance: 10^{-10}) implemented in scipy.optimize (version 1.3.1) and estimate standard errors using the Hessian matrix of the likelihood function, generated using the numdifftools library (version 0.9.39). Assuming that the maximum likelihood estimates are normally distributed, the 95% confidence intervals (CI) are approximated as ± 1.96 standard errors.

Evaluation of AuthentiCT

We implemented this model in a program called AuthentiCT. In this section, we evaluate how well AuthentiCT is able to estimate the proportion of contaminating sequences and to separate aDNA from present-day DNA sequences.

Inference of present-day DNA contamination rates

Assessing the accuracy with artificial mixtures of ancient and present-day DNA

To test if AuthentiCT can estimate the proportion of present-day DNA contamination, we created artificial mixtures of aDNA and present-day DNA sequences in varying proportions, from 5 to 95%, in steps of 5%. As present-day contaminant, we used sequences generated from present-day human DNA previously sheared to short fragments and treated like aDNA (mimicking the treatment of a genuine present-day DNA contamination [9]; these sequences are available in Additional file 2). As aDNA sequences, we used sequences from archaic datasets generated from single-stranded libraries that exhibit minimal amounts of present-day human DNA contamination [58, 60].

For each dataset, we then compared the contamination rate estimates from AuthentiCT to the estimates from contDeam [17] and from the conditional substitution analysis ([22], as computed in [23]) (Fig. 3). The conditional substitution analysis underestimates the contamination proportion (average bias, − 6.73%; root mean square error (RMSE), 0.0741; based on 100,000 sequences), and contDeam overestimates it (average bias, 2.36–1.19%; RMSE, 0.0396–0.0320; based on 10,000 and 100,000 sequences, respectively). In contrast, our method yields more accurate estimates (average bias, 0.05–0.42%; RMSE, 0.0194–0.0191; based on 10,000 and 100,000 sequences, respectively). We note some variability in the results depending on the dataset used, which may reflect different properties that are not under our control (e.g. additional contamination in the Neandertal datasets or differences in error rates between the present-day and ancient DNA sequences).

Exploring the limits of the method with simulations

To further evaluate AuthentiCT in scenarios where we have full control over the parameters, we simulated aDNA and present-day DNA sequences using the model described above, varying deamination rates, error rates, sequence lengths, GC contents, and numbers of sequences. Unless stated otherwise, each dataset contained 100,000 sequences with lengths following a shifted geometric distribution with minimum and mean lengths of 35 and 45 bp, respectively. By default, we use a GC content of 40%, a

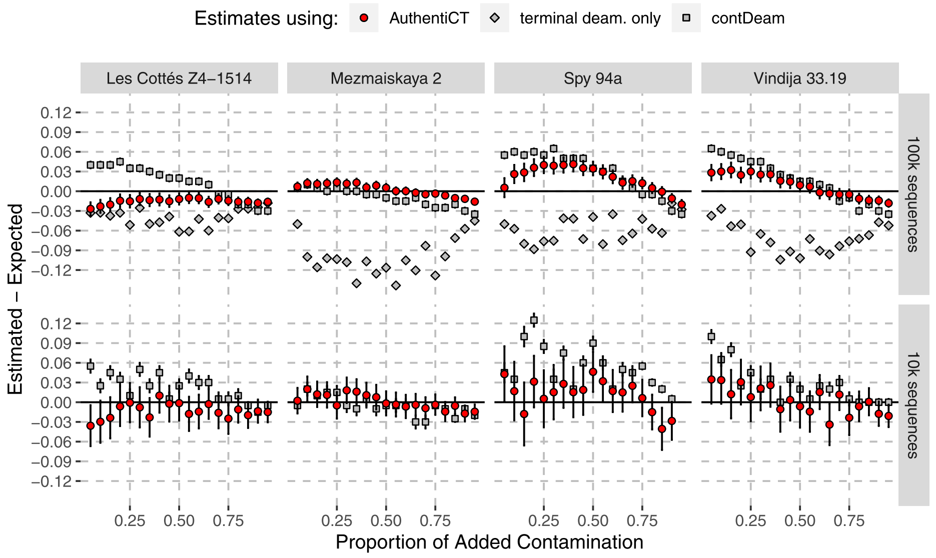

Fig. 3 Contamination estimates on artificial mixtures of Neandertal and present-day human DNA sequences. Estimates from AuthentiCT are shown with red dots, whereas the estimates from two other methods are shown in grey [17, 22]. Each point represents the difference between the estimated proportion of contamination and the proportion of present-day human DNA sequences introduced in the corresponding Neandertal dataset (different panels). Added contamination (*x*-axis) ranges from 5 to 95% in steps of 5% (see Additional file 1: Figure S3 for a narrower range from 1 to 10% in steps of 1%). Error bars correspond to 95% CI. There were not enough observations with 10,000 sequences for the method relying only on terminal deamination [3]. We note that the Neandertal datasets, particularly Spy 94a, may contain some present-day human DNA contamination in addition to the human contamination we introduced (see Additional file 1: Table S1 for the libraries used and the associated contamination estimates)

terminal deamination rate of 0.5 and an error/divergence rate of 0.001, and set the HMM parameters to $o = 0.5$, $o_2 = 0.5$, $l_o = 0.34$, $l_{ss} = 0.20$ and $l_{ds} = 0.003$.

We found that AuthentiCT performs well for datasets of 10,000 or more sequences (RMSE, 0.010 and 0.009; average bias, 0.006 and 0.007; Fig. 4a), and its performance is consistent over a wide range of deamination rates (from 0.03 to 0.5 in Fig. 4b; RMSE between 0.005 and 0.013), albeit with larger confidence intervals for lower deamination rates. The least reliable estimates were obtained for very small datasets (1000 sequences) with low contamination rates (below 10%; RMSE, 0.028; average bias, 0.010; Fig. 4a). We note that AuthentiCT overestimates contamination for low contamination rates (average bias of 0.020 and 0.009 for contamination estimates below 0.25 with a terminal deamination rate of 0.15 and 0.5, respectively). This likely represents overfitting to short sequences with few informative sites, as the bias decreases with longer sequences or higher GC contents (Additional file 1: Figure S1).

Another variable that may affect contamination estimates is the rate of C-to-T substitutions to the reference genome that is not induced by deamination. Although the sequencing error rate would typically not exceed 1% on sequencing platforms commonly used for ancient DNA sequencing [61, 62], divergence to the reference genome may be an issue. We therefore tested our method on simulated sequences with varying substitution rates, assuming the same rate for endogenous and contaminating sequences (see Additional file 1: Figure S2 for results with different rates). As expected, an increase of the substitution rate leads to a decrease in accuracy of the contamination estimates (RMSE of 0.006, 0.010 and 0.011 for substitution rates of 0.001, 0.01 and 0.02, respectively; average bias of 0.005, 0.009

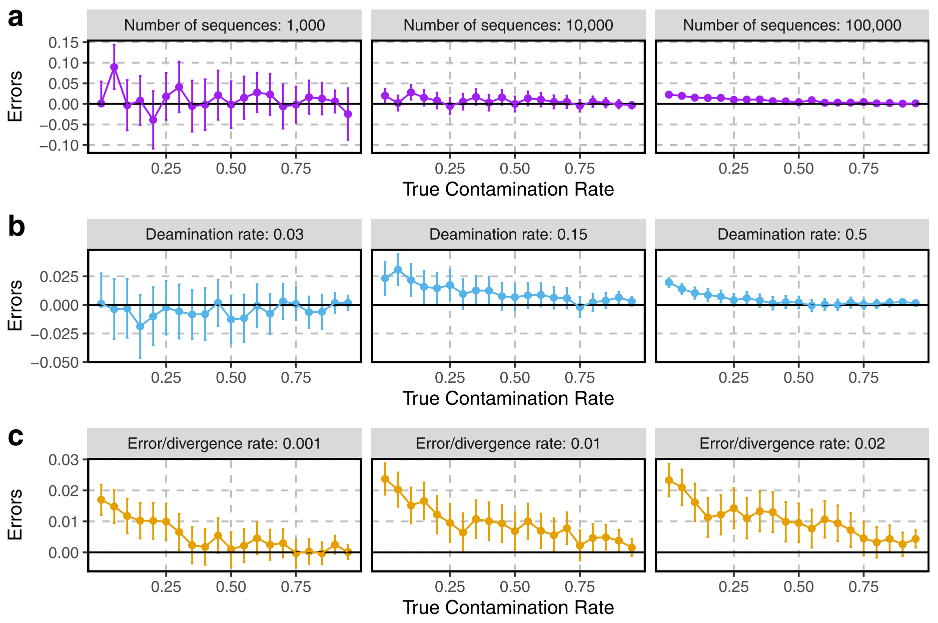

Fig. 4 Contamination estimates on simulated datasets. The datasets differ in the number of sequences (**a**), deamination rates (**b**) and sequencing error/divergence rates (**c**). Each point corresponds to a set of simulated sequences with varying proportions of contamination (*x*-axis) from 0 to 95% in steps of 5%. The errors (*y*-axis) correspond to the difference between the estimated and true contamination rates. The bars correspond to the 95% CI

and 0.01 for the same substitution rates; Fig. 4c). We note that divergence can also lead to alignment issues that may further introduce biases in the damage patterns, as mismatches to the reference may lead to the selective loss of sequences with additional C-to-T substitutions.

Application to real ancient DNA datasets

To further validate the method, we next applied AuthentiCT to published sequences from archaic human specimens [22, 23, 58, 60] without introducing sequences from present-day DNA. We compared our results with a previous approach [23] that relies solely on the divergence to a present-day African genome (HGDP00456, [45]). This independent method works well for Neandertals (as the contaminating modern human DNA will be much more similar to the African genome than the Neandertal genome), but will not translate to samples genetically close to their contaminant (i.e. early modern humans). Here, we use the $F(A|B)$ statistic as a measure of divergence, as it varies little between individuals from the same population [63]. The value of this statistic for a contaminated sample is simply a linear combination of the values for the endogenous and contaminating genomes, i.e. $F_{\text{observed}} = c \times F_{\text{contaminant}} + (1 - c) \times F_{\text{endogenous}}$ where c is the contamination rate. We set $F_{\text{contaminant}}$ to 0.275 (computed from the genotype calls of a present-day European genome, HGDP00521 [45]) and $F_{\text{endogenous}}$ to 0.176 (Table S20 in [60]).

We note, however, that the two approaches measure slightly different contamination proportions. While AuthentiCT relies on every sequence, $F(A|B)$ relies only on sequences that overlap informative positions. Therefore, AuthentiCT yields contamination estimates corresponding to the proportion of contaminating sequences, whereas the approach based on $F(A|B)$ provides estimates corresponding to the proportion of bases that come from contaminating sequences. These estimates can differ if the length distributions of endogenous and contaminating sequences differ, as is the case for some datasets (Additional file 1: Supplementary Note 3). To get comparable estimates of contamination, we ran AuthentiCT on the subset of sequences that overlap the informative sites used for quantifying contamination based on the $F(A|B)$ statistic (Fig. 5a; see Additional file 1: Table S2 for the contamination estimates per sequences). Using the same sequences may also account for potential differences in the proportion of contamination along the genome. The contamination estimates from both methods are highly correlated (Pearson's coefficient, 0.98; p value $< 10^{-15}$; Fig. 5a).

Finally, we applied AuthentiCT to ancient modern human DNA sequences generated by hybridization capture and with minimal amounts of present-day DNA contamination [48, 56]. We identified 19 DNA libraries derived from male specimens for which we could estimate contamination based on the proportion of observed alternative alleles on X-chromosome sequences (using contaminationX [18]), with a minimal coverage depth of 2 sequences, at least 100 informative positions and the HapMap CEU population as reference panel [64], excluding variants with a minor allele frequency lower than 5%. Both methods yielded similar low contamination estimates (Fig. 5b), which demonstrate that AuthentiCT can be applied to datasets generated by hybridization capture and with low levels of contamination (for estimates with ancient DNA from domesticated species, see also Additional file 1: Figure S4; these sequences are available in Additional file 2).

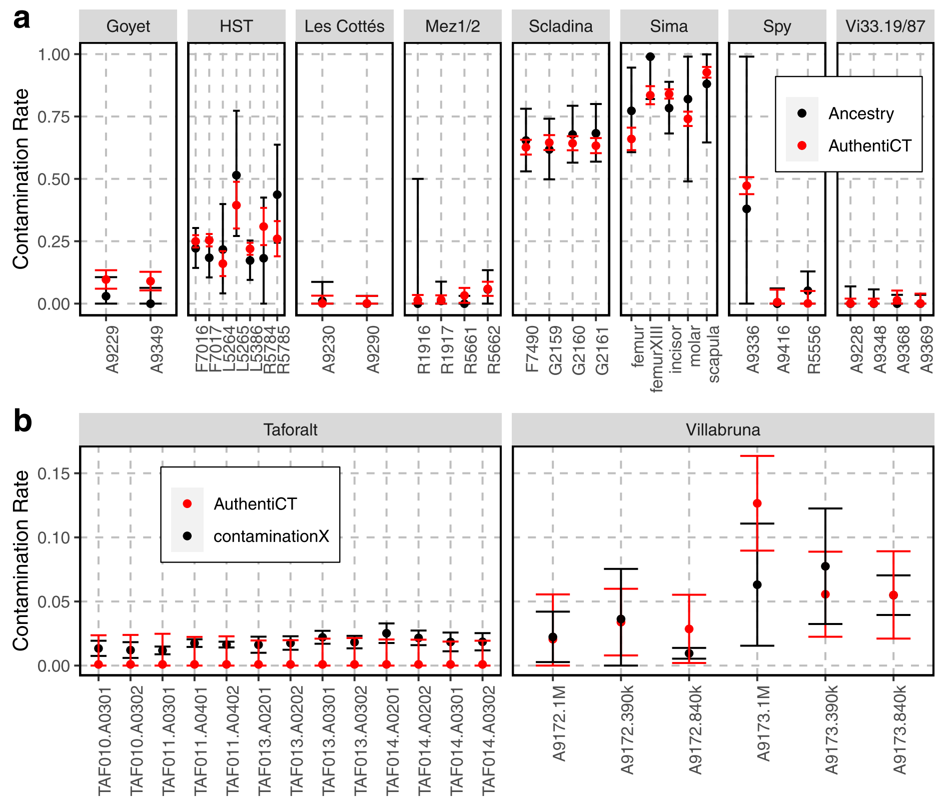

Fig. 5 Contamination estimates in archaic and modern human DNA datasets. **a** Estimates from AuthentiCT (in red) and a method based on the proportion of shared derived alleles with a present-day human genome (in black; see Material and Methods). When available, we used 10,000 sequences from public archaic human DNA datasets generated by shotgun sequencing of single-stranded DNA libraries [22, 23, 58, 60]. For more details about the libraries (identifiers on the *x*-axis), see Additional file 1: Table S2. Bars represent 95% CI. **b** Estimates from AuthentiCT (in red; using 10,000 sequences) and contaminationX (in black [18];) for public DNA datasets generated by hybridization capture of single-stranded DNA libraries prepared from ancient modern human samples [48, 56]. For the Villabruna samples, the numbers after the identifiers correspond to the number of target sites in the corresponding hybridization capture experiment. Bars represent 95% CI

Separating ancient DNA and present-day DNA sequences

As most downstream applications assume the absence of contamination, it is often necessary to identify endogenous sequences in contaminated aDNA samples. A common way to achieve this is to compare models for each sequence whether it is endogenous or a contaminant, as done by PMDtools [41] using a likelihood-ratio test. Here, we investigate how the likelihoods of AuthentiCT contrasts with the ones from PMDtools (using recommended parameters) on the same artificial mixtures of Neandertal and present-day human DNA sequences described above. For computational convenience, we fitted the parameter values of AuthentiCT to datasets of 10,000 sequences. Note that once fitted on a subset of sequences, AuthentiCT can then be applied to larger datasets with million sequences akin to PMDtools.

A strict filter for sequences with at least one C-to-T substitution within the first or last three positions leads to about 30% of endogenous sequences being detected, with 0.4% of false positives (Fig. 6a; see Additional file 1: Figure S5 for comparison with a filter based on the number of C-to-T substitutions anywhere along the sequences). At the same false-positive rate, both PMDtools and AuthentiCT detect

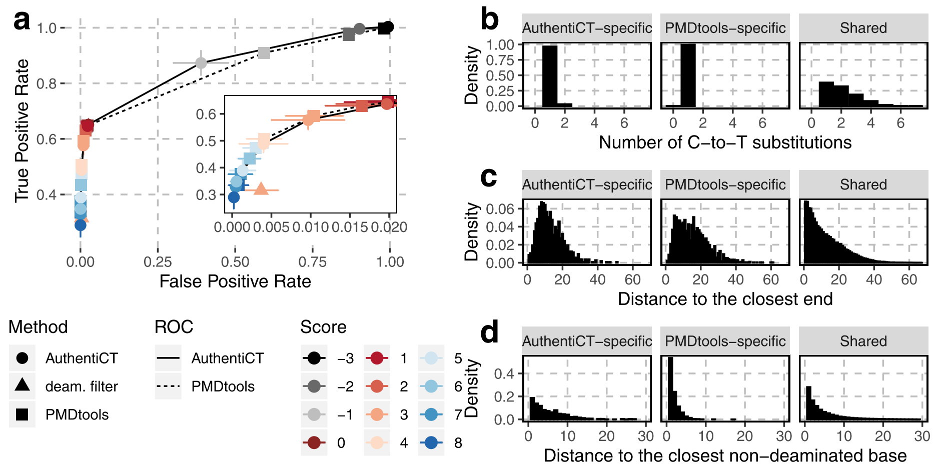

Fig. 6 Classification of aDNA and present-day DNA sequences. **a** The receiver operating characteristic (ROC) curves illustrate the performance of AuthentiCT (solid) and PMDtools (dashed) to identify aDNA sequences. A sequence is considered ancient if the log-likelihood ratio (score) of an ancient versus present-day origin is equal to or higher than a threshold (different colours). Each point represents the average performance over 19 datasets (of 10,000 sequences each) with varying proportions of ancient and present-day DNA sequences (5 to 95% in steps of 5%) for AuthentiCT (circles), PMDtools (squares) and a filter for sequences exhibiting at least one C-to-T substitution within the first or last three positions (deam. filter; triangles). Sequences are from Mezmaiskaya 2 (libraries A9180, A9288, A9289 and R1917 [58]) and the present-day human control. The bars correspond to two standard errors. **b**–**d** The distributions illustrate the number of C-to-T substitutions per sequences (**b**), the distance between a C-to-T substitution and the closest end of the sequence (**c**) or the closest non-deaminated base (**d**) in sequences classified as ancient only by AuthentiCT (left), only by PMDtools (middle) or both (right), using a score threshold of 3 as recommended for PMDtools [41]

around 50% of ancient sequences, and the likelihood models allow fine-tuning of precision with recall (see Additional file 1: Figure S6 for the performance on other datasets). The performance of AuthentiCT and PMDtools is similar at low false-positive rates (< 0.02), which are most important for classifying sequences. However, AuthentiCT performs better for more ambiguous sequences and yields higher likelihoods for a contaminant origin for present-day DNA sequences (Additional file 1: Figure S7), which explains why it performs well for estimating contamination (Fig. 3). The two methods mostly differ in their classification of sequences that exhibit only one C-to-T substitution (Fig. 6b) in the internal part of sequences (Fig. 6c). Compared to AuthentiCT's classification, there is an excess of sequences classified as ancient by PMDtools that exhibit non-deaminated bases adjacent to the C-to-T substitution (Fig. 6d). Some of these are therefore likely to represent polymorphisms or sequencing errors rather than deamination, indicating that AuthentiCT is more conservative when classifying these sequences as endogenous.

Discussion

Estimating present-day DNA contamination in aDNA datasets remains a difficult task, particularly if the contaminating DNA is closely related to the DNA of the studied organism. Most approaches rely on genetic differences between the endogenous and contaminating genomes, which are often unknown a priori. Besides, the dependence on the same information used in downstream analyses is not desirable as contamination may confound signals of interest (e.g. modern human ancestry in a Neandertal genome may look like present-day human DNA contamination). Here, we developed an alternative method to estimate the proportion of present-day DNA contamination based solely on a model of aDNA damage.

AuthentiCT overcomes several shortcomings of previous methods for estimating contamination that are based on aDNA damage. Most notably, it uses every position that is potentially informative, irrespective of whether it is close to the end of a sequence or not, and accounts for clusters of C-to-T substitutions in the internal part of aDNA sequences. The latter represent a feature of aDNA deamination patterns that, to our knowledge, has not been described or exploited before. We demonstrated that this more detailed modelling of the distribution of C-to-T substitutions along aDNA sequences leads to more accurate estimates of the proportion of present-day DNA contamination than previous approaches. However, the performance of AuthentiCT and PMDtools to classify sequences with multiple C-to-T substitutions were almost identical, and filtering for aDNA sequences will yield very similar results for both methods. Thus, the improvement stems largely from a more confident detection of contaminant sequences. In contrast to PMDtools, a sequence with many non-deaminated bases is considered to be more likely a contaminant under the AuthentiCT model (Additional file 1: Figure S6).

It is important to note potential caveats. First, we assume the absence of significant levels of deamination in the contaminating DNA. This assumption does not always hold true (e.g. [13, 23, 59]) and would lead to an underestimate of the proportion of contamination (Additional file 1: Figure S8). One could test this by first identifying sequences that carry contaminant alleles at diagnostic positions in the mitochondrial genome and then checking for the presence of C-to-T substitutions [65]. Second, there may be populations of DNA fragments with different rates of damaged bases, even within a single sample, because of microstructural differences in DNA preservation, or because of different treatments [66]. These differences may lead to an overestimate of the proportion of present-day DNA contamination, as it would lead to an excess of sequences without C-to-T substitutions. Yet, amongst the 50 libraries that we tested (from 19 specimens), the confidence intervals of the contamination estimates from AuthentiCT and another independent method diverge only twice (library A9349 from the Goyet Q56-1 Neandertal, Fig. 5a, and library A0201 of Taforalt 14, Fig. 5b). As AuthentiCT can run on as few as 10,000 sequences (in 3–10 min, see Additional file 1: Figures S9 and S10 for evaluations of the run time), one could split the data by sequencing run, sequence length or other covariates to obtain stratified contamination estimates. Finally, AuthentiCT is not applicable to libraries generated after treatments that alter deamination patterns, e.g. uracil selection or treatment with a uracil-DNA-glycosylase (UDG) [67, 68].

In the future, it will be useful to extend its application to data where patterns of DNA damage differ, such as in double-stranded aDNA libraries [46]. In addition, contamination estimates might be further improved if additional features typical of aDNA are incorporated into the model, such as fragment length or the frequency of purines in the reference genome at positions flanking the sequence alignments [39].

Conclusions

AuthentiCT is useful for estimating contamination in small datasets, e.g. when screening ancient specimens with shallow sequencing depth or when the samples are badly preserved. The independence of the contamination estimates from genetic differences between the contaminating and endogenous genomes makes this method both particularly valuable for the study of ancient human samples and broadly applicable to other species.

Specimen	Archive	Accession number
Denisova 2	European Nucleotide Archive	PRJEB20653
Goyet Q56–1	European Nucleotide Archive	PRJEB21870
Hohlenstein-Stadel	European Nucleotide Archive	PRJEB29475
Les Cottés Z4–1514	European Nucleotide Archive	PRJEB21875
Mezmaiskaya 1	European Nucleotide Archive	PRJEB21195
Mezmaiskaya 2	European Nucleotide Archive	PRJEB21881
Sima de los Huesos specimens	European Nucleotide Archive	PRJEB10597
Scladina I-4A	European Nucleotide Archive	PRJEB29475
Spy 94a	European Nucleotide Archive	PRJEB21883
Taforalt specimens	Sequence Read Archive	SRP132033
Villabruna	European Nucleotide Archive	PRJEB13123
Vindija 33.19	European Nucleotide Archive	PRJEB21157
Vindija 87	European Nucleotide Archive	PRJEB21882

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02123-y.

Additional file 1: Supplementary Figures, Notes and Tables. Figure S1: Contamination estimates on simulated datasets with different GC contents (A) and average sequence lengths (B). **Figure S2**: Effect of different rates of errors (including polymorphisms) between the endogenous and contaminating sequences. **Figure S3**: Contamination estimates on artifical mixtures of Neandertal and 1 to 10% (in steps of 1%) present-day human DNA sequences. **Figure S4**: Contamination estimates for ancient DNA datasets generated from domesticated species. **Figure S5**: Comparison of AuthentiCT, PMDtools and simple filters (based on the observed number of C-to-T substitutions) to identify ancient DNA sequences. **Figure S6**: Comparison of AuthentiCT and PMDtools to classify ancient and present-day DNA sequences for multiple datasets (when false positive rates are below 0.03). **Figure S7**: Comparison of the distributions of likelihood ratios («Score») computed with AuthentiCT (solid lines) and PMDtools (dashed lines) for different mixtures of ancient and present-day DNA sequences. **Figure S8**: Example of a Denisovan sample (Denisova 2, [5]) with a likely deaminated contaminant and the associated contamination underestimate from AuthentiCT. **Figure S9**: Runtime (wall clock) to estimate contamination depending on the number of sequences. **Figure S10**: Runtime (wall clock) to estimate contamination depending on the number of sequences when the model parameters values are provided in a configuration file (e.g. fitted a priori with a subset of the sequences). **Table S1**: Contamination estimates in the Neandertal datasets used in Figure 3. **Table S2**: Summary of the libraries used to compare the contamination estimates from AuthentiCT and a method based on the proportion of shared derived alleles with a present-day human genome («Ancestry-based» estimates). **Supplementary Note 1**: Excess of adjacent C-to-T substitutions inside ancient DNA sequences. **Supplementary Note 2**: Inference of the structure of ancient DNA fragments. **Supplementary Note 3**: Differences in contamination estimates per sequence and per base.

Additional file 2. Archive containing additional sequences used to test AuthentiCT. (GZ 12727 kb)

Additional file 3. Review history.

Acknowledgements

We thank Johann Visagie for valuable input in the implementation of AuthentiCT, Alba Bossom Mesa, Cesare de Filippo, Mateja Hajdinjak, Janet Kelso, Matthias Meyer and Elena Zavala for the constructive comments on the manuscript and Svante Pääbo, Kay Prüfer and Viviane Slon for the helpful discussions. We also thank Elena Essel, Mateja Hajdinjak, Petra Korlević, Cosimo Posth, Federico Sánchez-Quinto and Hannes Schröder for providing ancient DNA datasets to test AuthentiCT.

Review history

The review history is available as Additional file 3.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

SP conceived the study. SP and BMP designed the experiments. SP implemented the software. SP performed the analyses. SP and BMP wrote the manuscript. Both authors read and approved the final manuscript.

Funding

This work was supported by the Max Planck Society and the European Research Council [694707]. Open access funding provided by Projekt DEAL.

Availability of data and materials

An open-source implementation of AuthentiCT in python and a test dataset are available on GitHub https://github.com/StephanePeyregne/AuthentiCT, under the GPLv3 [69]. The version of the source code used in this manuscript has also been deposited in Zenodo and referenced as https://doi.org/10.5281/zenodo.3948256 [70]. The analysed datasets were generated in previous studies, and the corresponding accession numbers are listed below. Other sequences used in this study are available in Additional file 2.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 12 March 2020 Accepted: 27 July 2020
Published online: 15 September 2020

References

1. Lindahl T. Instability and decay of the primary structure of DNA. Nature. 1993;362:709–15.
2. Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, Campos PF, Samaniego JA, Gilbert MT, Willerslev E, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proc Biol Sci. 2012;279:4724–33.
3. Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, Arsuaga JL, Martinez I, Gracia A, de Castro JM, Carbonell E, Pääbo S. A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature. 2014;505:403–6.
4. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature. 2013;499:74–8.
5. Der Sarkissian C, Ermini L, Jonsson H, Alekseev AN, Crubezy E, Shapiro B, Orlando L. Shotgun microbial profiling of fossil remains. Mol Ecol. 2014;23:1780–98.
6. Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, Rabeder G, Krause J, Detter JC, Pääbo S, Rubin EM. Genomic sequencing of Pleistocene cave bears. Science. 2005;309:597–9.
7. Prüfer K, Stenzel U, Hofreiter M, Pääbo S, Kelso J, Green RE. Computational challenges in the analysis of ancient DNA. Genome Biol. 2010;11:R47.
8. Kircher M. Analysis of high-throughput ancient DNA sequencing data. Methods Mol Biol. 2012;840:197–228.
9. de Filippo C, Meyer M, Prüfer K. Quantifying and reducing spurious alignments for the analysis of ultra-short ancient DNA sequences. BMC Biol. 2018;16:121.
10. Llamas B, Valverde G, Fehren-Schmitz L, Weyrich LS, Cooper A, Haak W. From the field to the laboratory: controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. STAR. 2017;3:1–14.
11. Krause J, Briggs AW, Kircher M, Maričić T, Zwyns N, Derevianko A, Pääbo S. A complete mtDNA genome of an early modern human from Kostenki, Russia. Curr Biol. 2010;20:231–6.
12. Richards MB, Sykes BC, Hedges REM. Authenticating DNA extracted from ancient skeletal remains. J Archaeol Sci. 1995; 22:291–9.
13. Sampietro ML, Gilbert MT, Lao O, Caramelli D, Lari M, Bertranpetit J, Lalueza-Fox C. Tracking down human contamination in ancient human teeth. Mol Biol Evol. 2006;23:1801–7.
14. Malmstrom H, Stora J, Dalen L, Holmlund G, Gotherstrom A. Extensive human DNA contamination in extracts from ancient dog bones and teeth. Mol Biol Evol. 2005;22:2040–7.
15. Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J, et al. A revised timescale for human evolution based on ancient mitochondrial genomes. Curr Biol. 2013;23:553–9.
16. Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, Good JM, Maričić T, Stenzel U, et al. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell. 2008;134:416–26.
17. Renaud G, Slon V, Duggan AT, Kelso J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biol. 2015;16:224.
18. Moreno-Mayar JV, Korneliussen TS, Dalal J, Renaud G, Albrechtsen A, Nielsen R, Malaspinas AS: A likelihood method for estimating present-day human contamination in ancient male samples using low-depth X-chromosome data. Bioinformatics 2019;36(3):828–41.

19. Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T, et al. An Aboriginal Australian genome reveals separate human dispersals into Asia. Science. 2011;334:94–8.
20. Green RE, Krause J, Briggs AW, Maričić T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, et al. A draft sequence of the Neandertal genome. Science. 2010;328:710–22.
21. Racimo F, Renaud G, Slatkin M. Joint estimation of contamination, error and demography for nuclear DNA from ancient humans. PLoS Genet. 2016;12:e1005972.
22. Meyer M, Arsuaga JL, de Filippo C, Nagel S, Aximu-Petri A, Nickel B, Martinez I, Gracia A, Bermudez de Castro JM, Carbonell E, et al. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature. 2016;531: 504–7.
23. Peyrégne S, Slon V, Mafessoni F, de Filippo C, Hajdinjak M, Nagel S, Nickel B, Essel E, Le Cabec A, Wehrberger K, et al. Nuclear DNA from two early Neandertals reveals 80,000 years of genetic continuity in Europe. Sci Adv. 2019;5:eaaw5873.
24. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522:207–11.
25. Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, Korneliussen TS, Gronnow B, Appelt M, Gullov HC, Friesen TM, et al. The genetic prehistory of the New World Arctic. Science. 2014;345:1255832.
26. Castellano S, Parra G, Sanchez-Quinto FA, Racimo F, Kuhlwilm M, Kircher M, Sawyer S, Fu Q, Heinze A, Nickel B, et al. Patterns of coding variation in the complete exomes of three Neandertals. Proc Natl Acad Sci U S A. 2014;111:6666–71.
27. Schwarz C, Debruyne R, Kuch M, McNally E, Schwarcz H, Aubrey AD, Bada J, Poinar H. New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. Nucleic Acids Res. 2009;37:3215–29.
28. Higgins D, Rohrlach AB, Kaidonis J, Townsend G, Austin JJ. Differential nuclear and mitochondrial DNA preservation in post-mortem teeth with implications for forensic and ancient DNA studies. PLoS One. 2015;10:e0126935.
29. Furtwängler A, Reiter E, Neumann GU, Siebke I, Steuri N, Hafner A, Losch S, Anthes N, Schuenemann VJ, Krause J. Ratio of mitochondrial to nuclear DNA affects contamination estimates in ancient DNA analysis. Sci Rep. 2018;8:14075.
30. Green RE, Briggs AW, Krause J, Prüfer K, Burbano HA, Siebauer M, Lachmann M, Pääbo S. The Neandertal genome and ancient DNA authenticity. EMBO J. 2009;28:2494–502.
31. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014;15:356.
32. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. PLoS One. 2012;7:e34131.
33. Kistler L, Ware R, Smith O, Collins M, Allaby RG. A new model for ancient DNA decay based on paleogenomic meta-analysis. Nucleic Acids Res. 2017;45:6310–20.
34. Dabney J, Meyer M, Pääbo S. Ancient DNA damage. Cold Spring Harb Perspect Biol. 2013;5(7):a012567. https://doi.org/10.1101/cshperspect.a012567. Published 2013 Jul 1.
35. Weiss CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, Stinchcombe JR, Krause J, Burbano HA. Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. R Soc Open Sci. 2016;3: 160239.
36. Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Pääbo S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Res. 2001;29:4793–9.
37. Hansen A, Willerslev E, Wiuf C, Mourier T, Arctander P. Statistical evidence for miscoding lesions in ancient DNA templates. Mol Biol Evol. 2001;18:262–5.
38. Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, Austin J, Cooper A. Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Res. 2007;35:5717–28.
39. Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S. Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci U S A. 2007;104:14616–21.
40. Lindahl T, Nyberg B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. Biochemistry. 1974;13: 3405–10.
41. Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Pääbo S, Krause J, Jakobsson M. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc Natl Acad Sci U S A. 2014;111:2229–34.
42. Jonsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013;29:1682–4.
43. Ginolhac A, Rasmussen M, Gilbert MT, Willerslev E, Orlando L. mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics. 2011;27:2153–5.
44. Al-Asadi H, Dey KK, Novembre J, Stephens M. Inference and visualization of DNA damage patterns using a grade of membership model. Bioinformatics. 2019;35:1292–8.
45. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338:222–6.
46. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376–80.
47. Gansauge MT, Meyer M. A method for single-stranded ancient DNA library preparation. Methods Mol Biol. 1963;2019: 75–83.
48. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwängler A, Haak W, Meyer M, Mittnik A, et al. The genetic history of Ice Age Europe. Nature. 2016;534:200–5.
49. Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, Rohland N, Nagele K, Adamski N, Bertolini E, et al. Reconstructing the deep population history of Central and South America. Cell. 2018;175:1185–97 e1122.
50. Olalde I, Mallick S, Patterson N, Rohland N, Villalba-Mouco V, Silva M, Dulias K, Edwards CJ, Gandini F, Pala M, et al. The genomic history of the Iberian Peninsula over the past 8000 years. Science. 2019;363:1230–4.
51. Thomas JE, Carvalho GR, Haile J, Rawlence NJ, Martin MD, Ho SY, Sigfusson A, Josefsson VA, Frederiksen M, Linnebjerg JF, et al. Demographic reconstruction from ancient DNA supports rapid extinction of the great auk. Elife. 2019;8:e47509.
52. Barlow A, Cahill JA, Hartmann S, Theunert C, Xenikoudakis G, Fortes GG, Paijmans JLA, Rabeder G, Frischauf C, Grandal-d'Anglade A, et al. Partial genomic survival of cave bears in living brown bears. Nat Ecol Evol. 2018;2:1563–70.

53. Froese D, Stiller M, Heintzman PD, Reyes AV, Zazula GD, Soares AE, Meyer M, Hall E, Jensen BJ, Arnold LJ, et al. Fossil and genomic evidence constrains the timing of bison arrival in North America. Proc Natl Acad Sci U S A. 2017;114:3457–62.
54. Schroeder H, Avila-Arcos MC, Malaspinas AS, Poznik GD, Sandoval-Velasco M, Carpenter ML, Moreno-Mayar JV, Sikora M, Johnson PL, Allentoft ME, et al. Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean. Proc Natl Acad Sci U S A. 2015;112:3669–73.
55. Sanchez-Quinto F, Malmstrom H, Fraser M, Girdland-Flink L, Svensson EM, Simoes LG, George R, Hollfelder N, Burenhult G, Noble G, et al. Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society. Proc Natl Acad Sci U S A. 2019;116:9469–74.
56. van de Loosdrecht M, Bouzouggar A, Humphrey L, Posth C, Barton N, Aximu-Petri A, Nickel B, Nagel S, Talbi EH, El Hajraoui MA, et al. Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. Science. 2018;360:548–52.
57. Glocke I, Meyer M. Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. Genome Res. 2017;27:1230–7.
58. Hajdinjak M, Fu Q, Hubner A, Petr M, Mafessoni F, Grote S, Skoglund P, Narasimham V, Rougier H, Crevecoeur I, et al. Reconstructing the genetic history of late Neanderthals. Nature. 2018;555:652–6.
59. Slon V, Viola B, Renaud G, Gansauge MT, Benazzi S, Sawyer S, Hublin JJ, Shunkov MV, Derevianko AP, Kelso J, et al. A fourth Denisovan individual. Sci Adv. 2017;3:e1700186.
60. Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017;358:655–8.
61. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour. 2011;11:759–69.
62. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. 2016;17:125.
63. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505:43–9.
64. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467:52–8.
65. Welker F, Hajdinjak M, Talamo S, Jaouen K, Dannemann M, David F, Julien M, Meyer M, Kelso J, Barnes I, et al. Palaeoproteomic evidence identifies archaic hominins associated with the Chatelperronian at the Grotte du Renne. Proc Natl Acad Sci U S A. 2016;113:11162–7.
66. Korlević P, Gerber T, Gansauge MT, Hajdinjak M, Nagel S, Aximu-Petri A, Meyer M. Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. Biotechniques. 2015;59:87–93.
67. Bokelmann L, Hajdinjak M, Peyrégne S, Brace S, Essel E, de Filippo C, Glocke I, Grote S, Mafessoni F, Nagel S, et al. A genetic analysis of the Gibraltar Neanderthals. Proc Natl Acad Sci U S A. 2019;116:15610–5.
68. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 2010;38:e87.
69. Peyrégne S. AuthentiCT. GitHub. 2020; https://github.com/StephanePeyregne/AuthentiCT. Accessed 16 July 2020.
70. Peyrégne S. AuthentiCT. Zenodo. 2020; https://doi.org/10.5281/zenodo.3948256. Accessed 16 July 2020.

Publisher's Note