BioEssays

www.bioessays-journal.com

Present-Day DNA Contamination in Ancient DNA Datasets

Stéphane Peyrégne and Kay Prüfer*

Present-day contamination can lead to false conclusions in ancient DNA studies. A number of methods are available to estimate contamination, which use a variety of signals and are appropriate for different types of data. Here an overview of currently available methods highlighting their strengths and weaknesses is provided, and a classification based on the signals used to estimate contamination is proposed. This overview aims at enabling researchers to choose the most appropriate methods for their dataset. Based on this classification, potential avenues for the further development of methods are discussed.

1. Introduction

Ancient DNA from historical or archaeological materials, such as bones, teeth, or hair, represents a valuable resource for studying the past. Ancient DNA has been isolated and sequenced from the remains of extinct humans[1-3] and other animals,[4,5] but also from herbaria,[6] dental calculus,[7,8] and environmental samples such as archaeological sediments.[9,10] These sequence data yielded insights into the evolutionary and population history of organisms.[11-16]

Yet, the degraded nature of ancient DNA complicates its analysis.[17,18] After the death of an organism, DNA is exposed to enzymes and chemical reactions that result in DNA fragmentation, the loss of bases, and base modifications.[19,20] Over time, DNA fragments will degrade and eventually become unrecoverable.[21] Even under favorable conditions, only a small amount of the original DNA remains from the organism under study (endogenous DNA). In addition, DNA from other organisms (exogenous DNA) contaminates most ancient specimens. This includes DNA from microbes that colonize decaying tissues,[22] and DNA from the environment that can seep into the specimen. Exogenous DNA can also be introduced by handling, lab equipment, and reagents.[23-25]

In most cases, researchers identify endogenous DNA sequences by aligning them to a closely related reference genome, thereby largely excluding sequences from distantly related organisms.[18,26-28] However, sequences from contaminating DNA that are similar to the reference genome can pass this filtering step. This is particularly problematic for the study of ancient human material, since human DNA is abundant in research environments and contamination by human DNA can lead to false signals of admixture or to underestimation of the divergence to present-day humans.[2,29,30]

Several precautions can guard against contamination.[17,31,32] Protective clothing during excavation[33,34] and lab work minimizes the introduction of contaminating DNA, and the irradiation of reagents, lab equipment, and clean room facilities are often used to degrade DNA from other potential sources.[35,36] Despite these efforts, a low level of contamination is unavoidable, and contamination is therefore closely monitored by using negative controls during DNA extraction and library preparation[37] and by the inclusion of unique combinations of DNA barcodes in each ancient DNA library.[38] However, these methods can only reveal contamination that is introduced during DNA extraction and library preparation or through cross-contamination between experiments. A measurement of contamination from all sources in the final sequencing dataset, which includes contamination introduced before DNA lab-work, remains crucial for downstream analyses.

In this review, we discuss methods that estimate the proportion of contamination in ancient DNA data. We focus on the case of ancient human samples with present-day human DNA contamination, since this is a particularly challenging problem (**Box 1**). However, many methods can also be applied to other organisms. We first describe the features of sequence data that can be used to quantify contamination. We then discuss specific approaches in more details, starting with methods for haploid loci, the mitochondria, and Y-chromosome, and then proceeding with methods for estimating contamination in diploid and recombining nuclear DNA.

2. Classification of Signals Used to Estimate Contamination

Three main signals are informative about the presence of contamination in ancient DNA datasets: sequence differences

Dr. S. Peyrégne, Dr. K. Prüfer
Department of Evolutionary Genetics
Max Planck Institute for Evolutionary Anthropology
Leipzig 04103, Germany
E-mail: stephane_peyregne@eva.mpg.de
Dr. K. Prüfer
Department of Archaeogenetics
Max Planck Institute for the Science of Human History
Jena 07745, Germany

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/bies.202000081

DOI: 10.1002/bies.202000081

BOX 1. Ancient DNA at the extremes

Poorly preserved ancient DNA samples can pose a particular challenge for analysis. Often this poor preservation is due to old age, although the oldest sequenced DNA to date, a permafrost sample from an ancient horse, is comparatively well preserved.[95] The poor DNA preservation in these samples means that the total concentration of ancient DNA is low and that contamination can constitute a large proportion of aligning sequences. Fortunately, old samples exhibit a high rate of C-to-T substitutions due to ancient DNA damage.

By using the presence of damage-associated substitutions to enrich for sequences that stem from ancient molecules, Meyer et al.[96] reconstructed the mitochondrial genome from the highly contaminated sequences of a Neanderthal ancestor found at Sima de los Huesos in Spain and dated to over 400 000 years ago. Although this approach only considered C-to-T substitutions at the ends of sequences, Skoglund et al.[44] developed a scoring system that considers all substitutions throughout the sequences. By filtering sequences based on these scores, they were able to reconstruct the mitochondrial genome from a Neanderthal found in Okladnikov Cave, Russia, with 10% present-day human DNA contamination.

Filtering sequences based on C-to-T substitutions can help to reduce contamination. However, if contamination accumulated some C-to-T substitutions, then these procedures may fail. In the previous two examples, mitochondrial genomes were reconstructed from multiple-fold coverage of sequences. Due to the high coverage, the correct endogenous mitochondrial genome will be reconstructed as long as contaminating sequences constitute at every site the minority among sequences with C-to-T substitutions. Unfortunately, such an approach is not possible for nuclear genomes, where the generated sequence coverage is typically far below onefold for highly degraded samples and informative sites are not always available. To deal with this issue and analyze nuclear sequences from Sima de los Huesos hominins, Meyer et al.[54] counted sequences in support of an assignment to the Neanderthal, Denisovan or modern human lineages. The inference of a closer relationship to Neanderthals was robust to high levels of modern human contamination, since contamination should not exhibit a high proportion of Neanderthal-specific variants. As an alternative to this approach, contamination levels can also be quantified and included in the calculation of statistics of interest.[42,43] By considering contamination, the relationship of late Neanderthals to early Neanderthals, with up to 65% contamination, was resolved.[42] Also, admixture rates can be estimated in the presence of contamination.[43]

These results show that even highly contaminated samples can yield insights when modeling or estimating the effect of contamination as part of the analyses.

between the contaminant and endogenous DNA, deviations from the expected ploidy, and time-dependent characteristics of ancient DNA such as damage-induced substitutions (**Figure 1**). We briefly describe these signals in this section.

2.1. Differences in the DNA Sequence

Sites that differ between the genome of interest and likely contaminants can be identified when their genome sequences are known in advance. For instance, the mitochondrial genomes of Neanderthals differ at some sites from those of all present-day humans[39] and contamination can be estimated by measuring the proportion of sequences that show the present-day human allele at these sites.[40] Other measures of sequence differences, such as sequence divergence, can also be used to estimate contamination if it is possible to predict what their value would be in the absence of contamination.[41,42] All approaches in this class require some a priori knowledge of the relationship between contaminating and ancient individuals. Also, these approaches gain power with increasing divergence between the contaminating and ancient genome sequences. Yet, once sequence differences are known, a few sequences overlapping these positions can be sufficient to estimate contamination.

2.2. Deviation from the Expected Ploidy

Contamination can cause a sample to show unusual patterns of ploidy. For instance, heterozygous sites on the X or Y-chromosomes in males, or Y-chromosome sequences in females, are signs of contamination. This signal is not limited to the sex chromosomes; a higher proportion of sequences supporting one allele at a heterozygous site on the autosomes can, for instance, indicate contamination from an individual carrying this allele.[2,43] In contrast to the previous class, ploidy-based methods often require multiple-fold coverage. However, they have the advantage that prior knowledge of the relationship between the contaminant and ancient individual is not required.

2.3. Ancient DNA Degradation Patterns

The degradation of DNA leaves characteristic patterns that can be used to distinguish ancient DNA sequences from those from present-day DNA contamination.[44,45] The most common damage in ancient DNA originates from cytosine-deamination[46] and occurs more often at the ends of DNA molecules,[47] likely because of single-stranded overhangs that degrade faster than the mostly double-stranded interior.[48,49] Cytosine-deamination turns cytosines (C) into uracils that are then misread as thymines

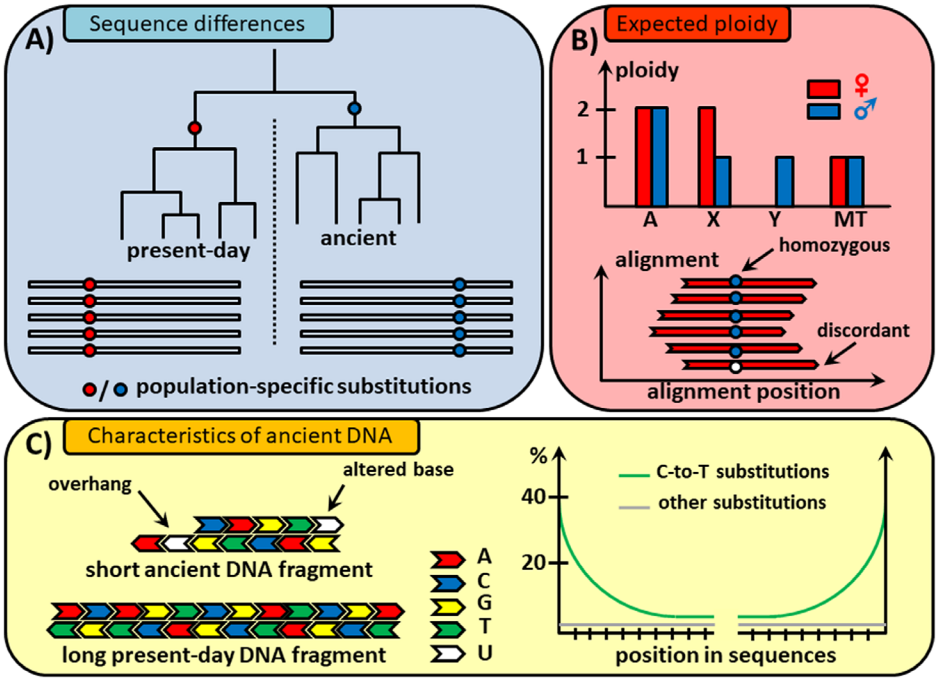

Figure 1. Classification of the signals used to estimate contamination. Each box illustrates one of the three signals. Box A shows the genealogical relationship of present-day and ancient individuals, including two derived variants that are informative for either group. The presence of a red variant indicates a contaminant sequence, whereas a blue variant indicates endogenous sequences. Box B shows the expected ploidy for the autosomes (A), the X- and Y-chromosomes (X, Y), and the mitochondrial genome (MT) for females (red) and males (blue). Deviations from these expectations indicate contamination from the opposite sex. The illustration below shows sequences aligned to a reference genome. These sequences carry two different alleles represented by dots. However, one allele (white) is rare compared to the other allele (blue). This observation is not compatible with the 50:50 ratio expected for a heterozygous site; therefore the discordant allele may originate from contamination. Box C illustrates ancient DNA damage. Left: ancient DNA fragments often contain uracils caused by ancient DNA damage whereas uracils are typically absent from present-day DNA fragments. Right: When no repair enzymes are used, uracils will be misread as thymines and their presence will result in high rates of C-to-T substitutions that occur primarily at the ends of sequences. Note that this signal depends on the library preparation protocol, and that high rates of G-to-A toward the 3′-end, instead of C-to-T exchanges, are also a possible signal. Neither pattern is expected for present-day DNA sequences.

(T) by the DNA polymerases used during DNA library preparation. This leads to erroneous C-to-T substitutions in the sequence data (and additional G-to-A substitutions, depending on the specifics of the library preparation protocol). The prevalence of deamination-induced C-to-T substitutions increases with the age of the sample,[50] although other factors, such as climate, have also a substantial effect on the rate of cytosine deamination.[51]The frequency of C-to-T substitutions can be used to classify sequences as likely ancient,[44,52] and to quantify contamination from undamaged present-day DNA.[53,54] Although not diagnostic, other features such as the length of ancient DNA sequences can also be used. Methods that are based solely on these ancient DNA degradation patterns require comparatively few sequences and no prior knowledge of genetic relationships.

3. Methods to Estimate Contamination

Many methods have been developed over the years to estimate contamination in ancient samples. Often, these methods rely on more than one of the signals described above (**Figure 2**). Thus, it is more helpful to consider the type of data to be analyzed when choosing a method. Here, we discuss methods grouped by the

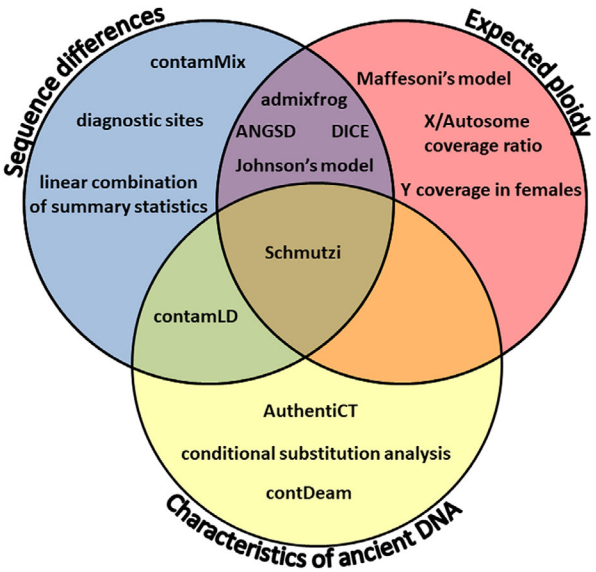

Figure 2. Classification of methods to estimate contamination. Methods are classified by the signals they used: sequence differences (blue), expected ploidy (red), or characteristics of ancient DNA (yellow). Some methods rely on multiple signals.

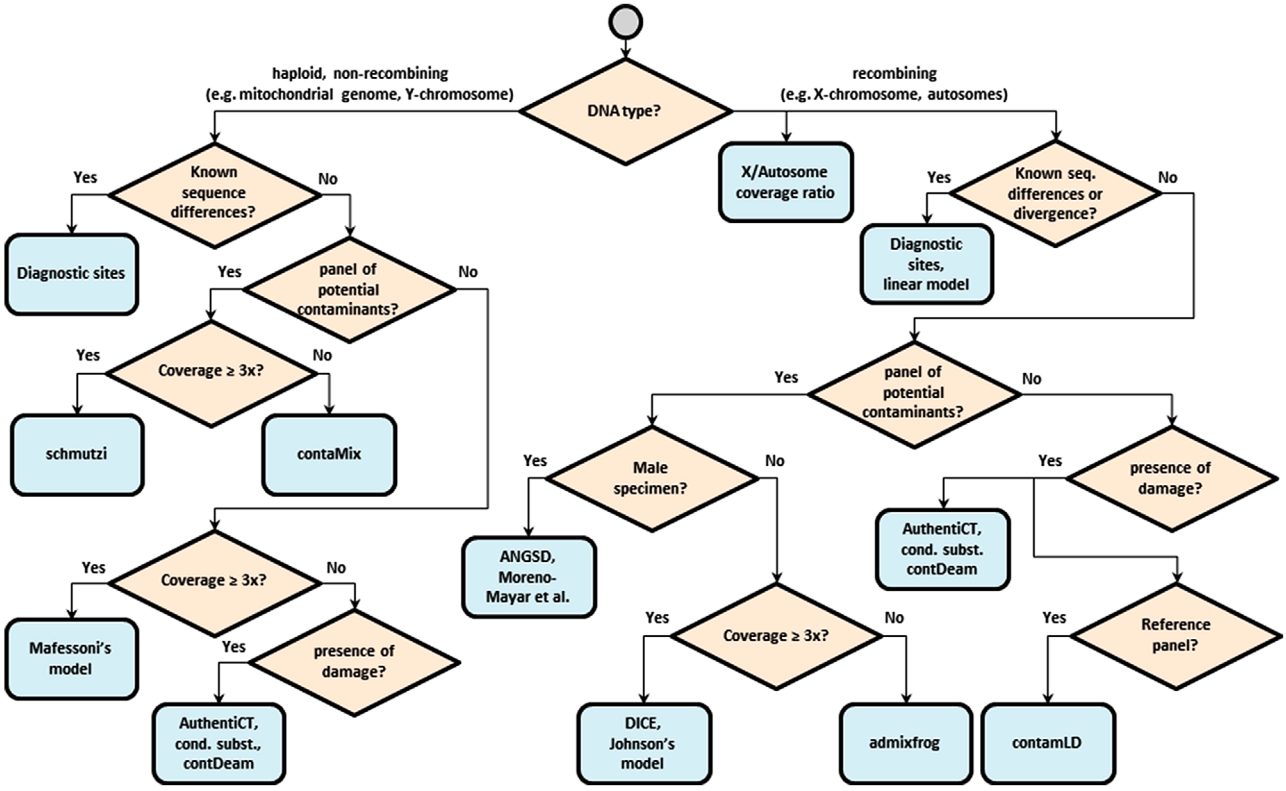

Figure 3. Flow chart describing the choice of methods depending on characteristics of the data available. Blue squares correspond to the methods, whereas peach diamonds illustrate the decisive questions. The gray circle represents the start of the flow chart. Questions that are answered with "No" do not indicate the requirement of an absence and further suitable methods can be identified by following these paths. The reference panel required for contamLD differs from the panel of potential contaminants used in some other methods in that it should be closely related to the ancient population of interest.

following three categories: mitochondrial DNA, sex chromosomes, and autosomes. Methods in each category differ in their requirements regarding sequence coverage, ancient DNA damage patterns, and a priori knowledge of sequence divergence. **Figure 3** shows a flowchart for choosing methods based on the data at hand.

3.1. Mitochondrial DNA

Ancient DNA studies often proceed by first sequencing mitochondrial genomes (or other nonrecombining haploid sequences such as chloroplasts[55]). The high copy number of mitochondria per cell, their small genome, and the wider availability of enrichment methods for mitochondrial DNA[56] make it easier to sequence the mitochondrial genome to high coverage than the nuclear genome, so that mitochondrial sequences often provide first insights into the level of contamination in a sample.

3.1.1. Methods Based on Differences in the DNA Sequence

Mitochondrial variation is often well characterized. For instance, full mitochondrial genomes have been reconstructed for the extinct human groups of Neanderthals and Denisovans[40-42,57-62] in addition to thousands of present-day and ancient modern humans.[63] These data can be used to identify positions where contaminating and endogenous mitochondrial genomes differ. At these diagnostic sites, the proportion of sequences carrying the contaminating allele represents an estimate of contamination.[2,40,64] However, diagnostic positions are not always known in advance. In this case, and when multi-fold coverage is available, one can reconstruct the endogenous sequence by consensus calling and then identify positions where the consensus sequence shows a variant that is either absent or present at a low frequency in a reference panel of potential contaminating genomes.[40,65]

Sometimes it is not possible to identify diagnostic positions if, for instance, the endogenous genome falls within the variation of contaminating genome(s). An alternative strategy exploits the fact that the mitochondrial genome is nonrecombining, and models the data as a mixture of sequences from different mitochondrial genomes. This approach is implemented in contamMix, an approach that considers both the reconstructed consensus mitochondrial sequence of the ancient individual and a reference panel of potentially contaminating mitochondrial genomes.[66] The method assumes that each sequence originates from one of these genomes and identifies them based on the number of matching bases, but allows for additional differences from sequencing errors. The proportion of sequences assigned

to other genomes than the consensus genome corresponds to the contamination estimate.

3.1.2. Methods Based on Deviations from the Expected Ploidy

The reconstruction of the endogenous consensus genome can be challenging if the data are highly contaminated or exhibit high error rates because of DNA damage. In this context, another approach, Schmutzi, takes advantage of the haploidy of the mitochondrial genome and requires multiple-fold sequence coverage to jointly reconstruct the endogenous and contaminant consensus sequences.[53] It uses a detailed model of errors, including substitutions from cytosine-deamination, and distinguishes endogenous from contaminant sequences by considering damage patterns and fragment lengths. It then gives a contamination estimate based on diagnostic sites, which are identified by comparing both consensus sequences to a reference panel of likely contaminants.

Schmutzi and contamMix can only be applied if a reference panel is available, but another approach that is independent of such panel exists (Mafessoni's model in[10]). This approach requires high sequence coverage and estimates the proportion of distinct mitochondrial genomes from differences among mapped sequences. Although the method is not aimed at quantifying contamination, one could apply it for this purpose, if contaminating and endogenous genomes are sufficiently different and sequencing error rates are low.

Contamination estimates based on mitochondrial sequences are often used as a proxy for nuclear contamination. However, it has been noted that contamination estimates obtained from the mitochondria and the nuclear genome can differ.[30,67] This is because mitochondrial DNA may degrade at different rates than nuclear DNA[21,68] and, more importantly, because the ratio of nuclear to mitochondrial sequences differs between cell types. This can, for example, lead to an underestimate of contamination if the source of contamination is a cell type with a low ratio of mitochondrial to nuclear genomes.[30,67,69] While none of the methods takes nuclear mitochondrial insertions (NuMts) or heteroplasmies into account, these factors are unlikely to introduce large errors in contamination estimates.

3.2. Sex Chromosomes

Although sex chromosomes are less accessible than mitochondrial genomes, they are recognized as a useful tool to study sex-biased migration and admixture among populations.[70] Contamination estimates that rely on the haploid state of the sex chromosomes in males or the absence of a Y-chromosome in females have been specifically developed for sex chromosome data.

3.2.1. Methods Based on Differences in the DNA Sequence

Many methods to estimate mitochondrial DNA contamination are also applicable to the nonrecombining part of the Y-chromosome. In particular, diagnostic sites can be identified from the large datasets of Y-chromosome diversity in humans (e.g.,[71,72]). While useful for studying Y-chromosomes, the estimates are insensitive to contamination from females and cannot be used as a measure of autosomal contamination.

3.2.2. Methods Based on Deviations from the Expected Ploidy

The X-chromosome is also present in a haploid state in males. In contrast to the mitochondrial genome and the Y-chromosome, however, the X-chromosome recombines, and methods based on haplogroups cannot be applied. As an alternative, Rasmussen et al.[73,74] and Moreno-Mayar et al.[75] used known variants on the X-chromosome to detect sequences that disagree with the majority call. These alternative alleles are unexpected in males that carry only one copy of the X-chromosome and can be used, together with an estimate of sequencing error from neighboring sites, to estimate contamination. Note that this X-chromosome contamination estimate gives an upper limit on the rate of contamination in the autosomes, since contamination originating from females has twice the impact on the X-chromosome of males compared to their autosomes.

Another approach, also based on the expected ploidy, is to compare the sequence coverage between the X-chromosome and the autosomes.[41] Female contamination in a male individual will increase the X-to-autosome ratio, while male contamination in a female individual will decrease it. The deviation of this ratio from the expected value of 0.5 and 1 for males and females, respectively, can thus be used to estimate contamination by the opposite sex. The advantage of this method is that it can be applied to both sexes and is unaffected by sequencing errors, as it does not rely on genetic variants.

Similarly, it is possible to estimate male contamination in a female sample by dividing the number of Y-chromosome sequences by the expected number of sequences that would map to the Y-chromosome if it were a male.[30] Assuming that the alignment efficiency is uniform among chromosomes, this expected number of sequences is simply half the observed number of sequences that map to the autosomes multiplied by the fraction of the genome that is the Y-chromosome.

3.3. Autosomal DNA

Estimating contamination from autosomal DNA is challenging because autosomes are diploid and recombine, and sequence coverage is often low. However, autosomal data are indispensable for the study of population history and selection, and accurate estimates of contamination are crucial to ensure correct results.

3.3.1. Methods Based on Differences in the DNA Sequence

As for mitochondrial and sex-chromosome sequences, contamination rates in autosomal data can rely on diagnostic sites. A prerequisite for the existence of such sites is a sufficiently large divergence between the likely source of contamination and the studied genome. This approach has, for instance, been used to estimate modern human DNA contamination in Neanderthal data.[54,76] In this setting, at least thousands of positions exist where most

modern human genomes carry a derived variant that is absent from sequenced Neanderthal genomes. Note that the approach is conservative in that it may lead to an overestimate of contamination rates, as newly sequenced Neanderthal genomes can carry modern human alleles because of hitherto unknown variation instead of contamination. The comparatively small fraction of Neanderthal ancestry in present-day humans would only lead to a minor underestimate.[2,76]

A natural extension of using diagnostic positions is to rely on expectations for a statistic that describes the relationship between contaminating and endogenous genomes. If these expectations differ between the contaminating and endogenous genomes, the level of contamination can be gauged by modeling the observed value of the summary statistic as a linear combination of these two expectations. Statistics that have been used for this purpose are estimates of sequence divergence and the sharing of derived alleles.[41,42] However, further statistics, such as admixture proportions, may also be suitable. Depending on the summary statistic used, this approach can use more of the data than diagnostic positions. However, similar to diagnostic positions, the approach relies on assumptions about the relationship of the contaminant to the ancient individual that can influence later analyses. Yet, as both methods only require that a few hundred sequences overlap informative sites, they are particularly useful for low-coverage data.

Nakatsuka et al.[77] recently presented another approach that relies on linkage between pairs of sites. As the contaminant and endogenous genomes often carry different haplotypes, this approach tests for a reduction of linkage between sites compared to the expectations derived from a panel of reference genomes. As a reduction in linkage may also be due to divergence to this reference panel, the method uses deaminated sequences to correct the linkage that is expected without contamination. This approach is applicable to ancient genomes with little divergence to the contaminant(s), which makes it valuable for the study of nuclear sequences from modern humans.

3.3.2. Methods Based on Deviations from the Expected Ploidy

In the previous section, the methods required a priori knowledge about the endogenous genome to infer contamination rates. Often, it is easier to make assumptions about the contaminant rather than the endogenous sequences. Assuming that some divergence exists between contaminating genomes and the endogenous genome, Philip L. F. Johnson[2,78] uses sites where likely sources of contamination are all or nearly all derived. Although the endogenous genome can carry any allele at these positions, we expect contamination to contribute sequences with derived alleles. Thus, the method infers contamination rates as an excess of sequences with derived alleles compared to the expectation of 0% at homozygous ancestral positions or 50% at heterozygous positions in the endogenous genome.

An extension of this approach, implemented in the software DICE,[43] increases the set of informative sites to also include those that are derived at a lower frequency in the contaminating source population. For this, Racimo et al.[43] model the relationship of the ancient sample to a set of known background populations to infer the probability of each genotype in the endogenous genome. Contamination then corresponds to the excess of sequences with either ancestral or derived alleles compared to expectations derived from the most likely genotype (i.e., the absence of derived and ancestral alleles at homozygous ancestral and homozygous derived sites, respectively, or an equal proportion of ancestral and derived alleles at heterozygous sites).

In contrast to the methods described in the previous section, both methods require multiple-fold coverage at informative sites since the inference of contamination relies on deviations from the expected ploidy. Requiring higher coverage also ensures that contamination can be estimated when contaminating and endogenous genomes show little divergence. However, we note that a recently introduced method, admixfrog, which is similar to DICE in how it models contamination, can yield contamination estimates with low sequence coverage (0.1× for an archaic human genome;[79]). This is achieved by taking advantage of sequence differences among multiple panel populations and assuming that the ancestries of the ancient genome derive from some of these source populations.

Another approach based on deviations from the expected ploidy is to take advantage of regions in the genome that are homozygous because of inbreeding. Contamination introduces alternative alleles randomly along the genome, including in these homozygous regions where only one allele is expected at any given position. An implementation based on this idea used homozygous regions in the genome of a Neanderthal woman whose parents were related at the level of half-siblings to jointly estimate error rates and the proportion of contamination.[80]

3.3.3. Methods Based on Patterns of Ancient DNA Damage

Substitutions associated with DNA damage have long been used as a signal to determine whether ancient sequences are preserved in a sample and to distinguish these sequences from contaminating sequences.[44,54,81-84] Contamination can be estimated from deamination patterns under the assumption that such patterns are absent in contaminating sequences. This is achieved by contrasting the true rate of damage-associated substitutions for the endogenous sequences to the observed rate of such substitutions for all sequences. To estimate the frequency of damage at the terminal positions of endogenous sequences, Meyer et al.[54] conditioned on the presence of a damage-associated substitution on one end of a sequence to enrich for genuine ancient sequences. The opposite ends of these sequences are then used to infer the frequency of substitutions in this ancient fraction, assuming that damage at both ends is independent. Because several biases can influence the frequency of substitutions at the ends of sequences (e.g., alignment bias against sequences with many substitutions), this method has not been used to quantify contamination. However, Meyer et al. could identify samples with substantial contamination using this approach.

As a prior for the mitochondrial contamination estimates of Schmutzi, Renaud et al.[53] implemented a method, contDeam, that solely uses patterns of ancient DNA damage. Contamination is estimated as the mixture proportion between two models of substitutions, one with and another without ancient DNA

BOX 2. Contamination per sequence or contamination per base?

Some contamination estimates give the proportion of bases that originate from contamination, while others give the proportion of contaminating sequences. These estimates can differ when the sequence length of contaminating and endogenous sequences differ. For instance, if contaminating sequences are on average twice as long as endogenous sequences, then a given informative site is twice as likely to be covered by a contaminating sequence compared to an endogenous sequence. A method based on informative sites would thus give an estimate per base, but would in this example overestimate the proportion of contaminating sequences.

Although methods based on informative sites naturally yield estimates per base, methods relying on ancient DNA damage produce estimates per sequence. However, methods based on ploidy or coverage can be formulated as either per base or per sequence.

Downstream analyses often benefit from contamination estimates per base. To convert estimates per sequence to estimates per base, it is possible to either weight sequences proportionally to their length or subsample sequences so that longer sequences are represented more often. For instance, restricting the estimation of contamination to the subset of sequences overlapping specific sites will automatically correct for sequence length.

damage. Compared to the previous approach, this method is not limited to the terminal bases of sequences, but instead infers site-specific deamination frequencies from sequences that exhibit a C-to-T substitution at one end.

By taking into account the dependence between C-to-T substitutions along ancient DNA sequences, a more recent method models the observed frequency of damage-associated substitutions in single- and double-stranded parts of the original DNA fragments.[85] Each sequence originates from either contamination, which does not contain damage, or an ancient molecule that contains damage according to the explicit model of the structure of ancient DNA fragments. Like the previous method, the approach estimates contamination as the mixture proportion of sequences fitting to one of these two models. Note that both methods provide estimates that correspond to the proportion of contaminant sequences, while the estimates for most other methods correspond to the proportion of contaminant bases (**Box 2**).

Contamination estimates based on DNA damage have the advantage that they are independent of sequence differences between endogenous and contaminating genomes, and that estimates can be obtained even for very low-coverage samples (10 000 sequences can be sufficient to estimate contamination). However, current methods assume that the contaminant is devoid of deamination, which several studies have shown is not true in all cases.[25,41,42] Other factors, such as heterogeneity in preservation within the sample or biases in extraction or library preparation, may limit the validity of the deamination models.[86] With further knowledge about the structure of ancient DNA fragments and the reduction of bias from protocols,[87] the use of these methods may increase.

4. Perspectives

Although many methods to estimate contamination are now available, these methods are not applicable in all circumstances. In addition, more accurate estimates of contamination may help to infer more details about the evolutionary and population history. Here we ask: What future development may we expect?

4.1. Methods Based on Differences in the DNA Sequence

The knowledge about the relationships among human groups increased substantially in recent years.[12] This knowledge includes the timing of population movements that resulted in large-scale admixtures. Some of these admixture signals may represent a useful source of information to estimate contamination. For instance, contamination from a present-day admixed population could be quantified in populations that pre-date these admixture events by measuring the admixture proportion.

Admixture between populations also results in large uninterrupted segments of different ancestries within an individual's genome. Contamination from other individuals will often carry a different ancestry at the same locations, so that contaminating sequences yield a reduction in linkage within ancestry segments. This information can in principle be used to quantify contamination.

4.2. Methods Based on Deviations from the Expected Ploidy

Sex-chromosomes are often used to estimate contamination levels in ancient samples, since even for shallow sequence data a substantial difference in coverage is expected for these regions of the genome. Large-scale insertion/deletion differences or segmental duplications can similarly yield such expected differences in coverage. We are hopeful that a better knowledge of the frequency and location of these polymorphisms will lead to the development of new methods for quantifying contamination.

4.3. Methods Based on Characteristics of Ancient DNA

Future method development could include additional features of ancient DNA such as the propensity of ancient DNA sequences to align to positions that are adjacent to purines.[47] This feature may be explained by a process called depurination, the loss of purine bases that can lead to a break of the DNA backbone. If

BOX 3. Contamination and metagenomics

This review focused on the analyses of a single species' genome from an ancient sample. However, the study of the species composition in ancient metagenomic samples gained attention in recent years, for instance, to reconstruct the diet of ancient populations[97] or to study the evolution of pathogens.[98] Often, these samples yield very few sequences from the genomes of interest. This means that low levels of contamination pose a challenge. We note that contamination in the context of metagenomics datasets can also stem from misassignments of sequences from closely related endogenous species, an issue that goes beyond the scope of this review.

The presence of ancient DNA damage-associated C-to-T substitutions can authenticate ancient sequences. Weiss et al.[82] devised a method to detect substitution patterns typical of ancient DNA when only a few hundred sequences are available. Even though this method does not exclude the presence of contamination, it can provide a positive indication that at least some sequences are ancient.

When authenticating sequences assigned to a single species, one can also use the distribution of edit distances (number of substitutions to the reference genome per sequence) to further increase confidence in the species assignment.[99,100] This is because truly related sequences will tend to carry a lower number of mismatches to the reference compared to spurious alignments. Competitive mapping is another way to identify and exclude contamination.[101,102] After mapping sequences to the genomes of the organism of interest and its closest known relatives, only sequences that map to the genome of interest are retained, thereby increasing the confidence that the sequences originated from this genome. More recently, this idea was also applied to faunal datasets to exclude human DNA contamination.[103]

Finally, the availability of high-throughput profiling tools[104-106] and large databases of microbial sequences[107-109] makes it possible to characterize the microbial content of metagenomics datasets. This information can be used to authenticate the origin of the sequences.[110] For instance, oral and soil microbiomes differ sufficiently to rule out substantial environmental contamination when studying dental calculus.[111] For more in-depth reviews about authenticating ancient microbial sequences, see for instance refs. [112] and [99].

present-day contamination does not exhibit this pattern, which is generated over a long time, then future contamination tests could include this signal.

DNA molecules invariably break into shorter fragments over time[20] and the length of ancient DNA sequences has been used as a signal to indicate the presence of ancient DNA.[88] However, some studies have found a poor correlation between age and fragment length over different archaeological sites.[50,51] This poor correlation may be explained by different preservation conditions, but length distributions are also influenced by extraction and library preparation methods.[89] These issues make it problematic to distinguish comparatively recent contaminating molecules, which are in principle subject to similar processes, from those that are truly old.

Several other features of ancient DNA have been discussed and could perhaps be used to quantify contamination. Regulatory signals in DNA are partially preserved in the substitution patterns and breakage of ancient DNA molecules. For instance, a nucleosome map has been reconstructed from the periodicity of fragment length.[90] Such periodicity could serve as a diagnostic feature of the tissue of origin, for example bone for an ancient bone sample versus skin for contamination from handling. Similarly, one could leverage signals of methylation that also differ between tissues. It was indeed possible to reconstruct partial methylation maps from C-to-T substitutions in CpG context from ancient DNA libraries that have been treated with repair enzymes or amplified with a proofreading polymerase.[90-92]

5. Conclusion

Ancient DNA is a rapidly growing field that has yielded unique insights into the past. However, from the start, the development of the field was hindered by the issue of contamination, which remains a major concern. The recently expanding field of palaeoproteomics is facing similar challenges.[93] Here, we have surveyed methods to quantify contamination and introduced a classification scheme of the signals they currently use.

We focused on methods applied to human samples, for which contamination is particularly problematic. However, we note that contamination is also a pervasive issue for the analysis of other organisms[24] (**Box 3**). For instance, it may be difficult to study reliably ancient samples from agricultural products, as these are often common in our environment.[82] Similarly, residual microbial sequences on lab-ware can influence metagenomic studies.[94]

Our classification puts the different approaches into perspective and may help to identify further avenues of method development. Methods that provide accurate estimates of contamination in low-coverage data are particularly needed. Such methods will help researchers to make better decisions during data production by avoiding wasting resources and preserving rare samples.

Contamination tests ensure the validity of insights drawn from ancient DNA studies. We hope that the continued development of methods to estimate contamination increases the confidence in ancient DNA results and helps to further push the limits of this field.

Acknowledgements

The authors thank Janet Kelso and Benjamin M. Peter for helpful comments on the manuscript, as well as Franziska Honigschnabel and Linda Schymanski from the multimedia team of the MPI-EVA for preparing the graphical abstract. The authors are also grateful for the constructive feedback from Hendrik Poinar and one other anonymous reviewer.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

ancient DNA, contamination, paleogenomics

Received: April 14, 2020
Revised: May 20, 2020
Published online:

[1] M. Krings, A. Stone, R. W. Schmitz, H. Krainitzki, M. Stoneking, S. Pääbo, *Cell* **1997**, *90*, 19.
[2] R. E. Green, J. Krause, A. W. Briggs, T. Maričić, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Fritz, N. F. Hansen, E. Y. Durand, A. S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, et al., *Science* **2010**, *328*, 710.
[3] D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. Johnson, T. Maričić, J. M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. E. Eichler, M. Stoneking, M. Richards, S. Talamo, M. V. Shunkov, A. P. Derevianko, J. J. Hublin, J. Kelso, M. Slatkin, S. Pääbo, *Nature* **2010**, *468*, 1053.
[4] R. Higuchi, B. Bowman, M. Freiberger, O. A. Ryder, A. C. Wilson, *Nature* **1984**, *312*, 282.
[5] L. Orlando, *BioEssays* **2020**, *42*, 1900164.
[6] C. Pont, S. Wagner, A. Kremer, L. Orlando, C. Plomion, J. Salse, *Genome Biol.* **2019**, *20*, 29.
[7] C. Warinner, C. Speller, M. J. Collins, *Philos. Trans. R. Soc. B* **2015**, *370*, 20130376.
[8] L. S. Weyrich, K. Dobney, A. Cooper, *J. Hum. Evol.* **2015**, *79*, 119.
[9] M. W. Pedersen, S. Overballe-Petersen, L. Ermini, C. D. Sarkissian, J. Haile, M. Hellstrom, J. Spens, P. F. Thomsen, K. Bohmann, E. Cappellini, I. B. Schnell, N. A. Wales, C. Carøe, P. F. Campos, A. M. Schmidt, M. T. Gilbert, A. J. Hansen, L. Orlando, E. Willerslev, *Philos. Trans. R. Soc. B* **2015**, *370*, 20130383.
[10] V. Slon, C. Hopfe, C. L. Weiss, F. Mafessoni, M. de la Rasilla, C. Lalueza-Fox, A. Rosas, M. Soressi, M. V. Knul, R. Miller, J. R. Stewart, A. P. Derevianko, Z. Jacobs, B. Li, R. G. Roberts, M. V. Shunkov, H. de Lumley, C. Perrenoud, I. Gušić, Ž. Kućan, P. Rudan, A. Aximu-Petri, E. Essel, S. Nagel, B. Nickel, A. Schmidt, K. Prüfer, J. Kelso, H. A. Burbano, S. Pääbo, M. Meyer, *Science* **2017**, *356*, 605.
[11] M. Slatkin, F. Racimo, *Proc. Natl. Acad. Sci. U.S.A* **2016**, *113*, 6380.
[12] P. Skoglund, I. Mathieson, *Annu. Rev. Genomics Hum. Genet.* **2018**, *19*, 381.
[13] M. A. Yang, Q. Fu, *Trends Genet.* **2018**, *34*, 184.
[14] D. E. MacHugh, G. Larson, L. Orlando, *Annu. Rev. Anim. Biosci.* **2017**, *5*, 329.
[15] G. P. McHugo, M. J. Dover, D. E. MacHugh, *BMC Biol.* **2019**, *17*, 98.
[16] S. Marciniak, G. H. Perry, *Nat. Rev. Genet.* **2017**, *18*, 659.
[17] S. Pääbo, H. Poinar, D. Serre, V. Jaenicke-Despres, J. Hebler, N. Rohland, M. Kuch, J. Krause, L. Vigilant, M. Hofreiter, *Annu. Rev. Genet.* **2004**, *38*, 645.
[18] K. Prüfer, U. Stenzel, M. Hofreiter, S. Pääbo, J. Kelso, R. E. Green, *Genome Biol.* **2010**, *11*, R47.
[19] J. Dabney, M. Meyer, S. Pääbo, *Cold Spring Harbor Perspect. Biol.* **2013**, *5*, a012567.
[20] S. Pääbo, *Proc. Natl. Acad. Sci. U.S.A* **1989**, *86*, 1939.
[21] M. E. Allentoft, M. Collins, D. Harker, J. Haile, C. L. Oskam, M. L. Hale, P. F. Campos, J. A. Samaniego, M. T. Gilbert, E. Willerslev, G. Zhang, R. P. Scofield, R. N. Holdaway, M. Bunce, *Proc. R. Soc. B* **2012**, *279*, 4724.
[22] C. Der Sarkissian, L. Ermini, H. Jonsson, A. N. Alekseev, E. Crubezy, B. Shapiro, L. Orlando, *Mol. Ecol.* **2014**, *23*, 1780.
[23] B. Llamas, G. Valverde, L. Fehren-Schmitz, L. S. Weyrich, A. Cooper, W. Haak, *Sci. Technol. Archaeol. Res.* **2017**, *3*, 1.
[24] H. Malmström, J. Stora, L. Dalén, G. Holmlund, A. Götherström, *Mol. Biol. Evol.* **2005**, *22*, 2040.
[25] M. L. Sampietro, M. T. Gilbert, O. Lao, D. Caramelli, M. Lari, J. Bertranpetit, C. Lalueza-Fox, *Mol. Biol. Evol.* **2006**, *23*, 1801.
[26] M. Kircher, *Methods Mol Biol* **2012**, *840*, 197.
[27] C. de Filippo, M. Meyer, K. Prüfer, *BMC Biol.* **2018**, *16*, 121.
[28] M. Schubert, A. Ginolhac, S. Lindgreen, J. F. Thompson, K. A. Al-Rasheid, E. Willerslev, A. Krogh, L. Orlando, *BMC Genomics* **2012**, *13*, 178.
[29] J. D. Wall, S. K. Kim, *PLoS Genet.* **2007**, *3*, 1862.
[30] R. E. Green, A. W. Briggs, J. Krause, K. Prüfer, H. A. Burbano, M. Siebauer, M. Lachmann, S. Pääbo, *EMBO J.* **2009**, *28*, 2494.
[31] M. Hofreiter, D. Serre, H. N. Poinar, M. Kuch, S. Pääbo, *Nat. Rev. Genet.* **2001**, *2*, 353.
[32] A. Cooper, H. N. Poinar, *Science* **2000**, *289*, 1139b.
[33] E. Pilli, A. Modi, C. Serpico, A. Achilli, H. Lancioni, B. Lippi, F. Bertoldi, S. Gelichi, M. Lari, D. Caramelli, *PLoS One* **2013**, *8*, e52524.
[34] D. Y. Yang, K. Watt, *J. Archaeol. Sci.* **2005**, *32*, 331.
[35] S. Champlot, C. Berthelot, M. Pruvost, E. A. Bennett, T. Grange, E. M. Geigl, *PLoS One* **2010**, *5*, e13042.
[36] M. Knapp, A. C. Clarke, K. A. Horsburgh, E. A. Matisoo-Smith, *Ann. Anat.* **2012**, *194*, 3.
[37] N. Rohland, I. Glocke, A. Aximu-Petri, M. Meyer, *Nat. Protoc.* **2018**, *13*, 2447.
[38] M. Kircher, S. Sawyer, M. Meyer, *Nucleic Acids Res.* **2012**, *40*, e3.
[39] D. Serre, A. Langaney, M. Chech, M. Teschler-Nicola, M. Paunovic, P. Mennecier, M. Hofreiter, G. Possnert, S. Pääbo, *PLoS Biol.* **2004**, *2*, e57.
[40] R. E. Green, A. S. Malaspinas, J. Krause, A. W. Briggs, P. L. Johnson, C. Uhler, M. Meyer, J. M. Good, T. Maričić, U. Stenzel, K. Prüfer, M. Siebauer, H. A. Burbano, M. Ronan, J. M. Rothberg, M. Egholm, P. Rudan, D. Brajković, Z. Kućan, I. Gusić, M. Wikström, L. Laakkonen, J. Kelso, M. Slatkin, S. Pääbo, *Cell* **2008**, *134*, 416.
[41] S. Sawyer, G. Renaud, B. Viola, J. J. Hublin, M. T. Gansauge, M. V. Shunkov, A. P. Derevianko, K. Prüfer, J. Kelso, S. Pääbo, *Proc. Natl. Acad. Sci. U.S.A* **2015**, *112*, 15696.
[42] S. Peyrégne, V. Slon, F. Mafessoni, C. de Filippo, M. Hajdinjak, S. Nagel, B. Nickel, E. Essel, A. Le Cabec, K. Wehrberger, N. J. Conard, C. J. Kind, C. Posth, J. Krause, G. Abrams, D. Bonjean, K. Di Modica, M. Toussaint, J. Kelso, M. Meyer, S. Pääbo, K. Prüfer, *Sci. Adv.* **2019**, *5*, eaaw5873.
[43] F. Racimo, G. Renaud, M. Slatkin, *PLoS Genet.* **2016**, *12*, e1005972.
[44] P. Skoglund, B. H. Northoff, M. V. Shunkov, A. P. Derevianko, S. Pääbo, J. Krause, M. Jakobsson, *Proc. Natl. Acad. Sci. U.S.A* **2014**, *111*, 2229.
[45] J. Krause, A. W. Briggs, M. Kircher, T. Maričić, N. Zwyns, A. Derevianko, S. Pääbo, *Curr. Biol.* **2010**, *20*, 231.

[46] M. Hofreiter, V. Jaenicke, D. Serre, A. von Haeseler, S. Pääbo, *Nucleic Acids Res.* **2001**, *29*, 4793.
[47] A. W. Briggs, U. Stenzel, P. L. Johnson, R. E. Green, J. Kelso, K. Prüfer, M. Meyer, J. Krause, M. T. Ronan, M. Lachmann, S. Pääbo, *Proc. Natl. Acad. Sci. U.S.A* **2007**, *104*, 14616.
[48] T. Lindahl, *Nature* **1993**, *362*, 709.
[49] T. Lindahl, B. Nyberg, *Biochemistry* **1974**, *13*, 3405.
[50] S. Sawyer, J. Krause, K. Guschanski, V. Savolainen, S. Pääbo, *PLoS One* **2012**, *7*, e34131.
[51] L. Kistler, R. Ware, O. Smith, M. Collins, R. G. Allaby, *Nucleic Acids Res.* **2017**, *45*, 6310.
[52] A. Helgason, S. Palsson, C. Lalueza-Fox, S. Ghosh, S. Sigurdardóttir, A. Baker, B. Hrafnkelsson, L. Arnadóttir, U. Thorsteinsdóttir, K. Stefánsson, *J. Mol. Evol.* **2007**, *65*, 92.
[53] G. Renaud, V. Slon, A. T. Duggan, J. Kelso, *Genome Biol.* **2015**, *16*, 224.
[54] M. Meyer, J. L. Arsuaga, C. de Filippo, S. Nagel, A. Aximu-Petri, B. Nickel, I. Martínez, A. Gracia, J. M. Bermúdez de Castro, E. Carbonell, B. Viola, J. Kelso, K. Prüfer, S. Pääbo, *Nature* **2016**, *531*, 504.
[55] S. Wagner, F. Lagane, A. Seguin-Orlando, M. Schubert, T. Leroy, E. Guichoux, E. Chancerel, I. Bech-Hebelstrup, V. Bernard, C. Billard, Y. Billaud, M. Bolliger, C. Croutsch, K. Čufar, F. Eynaud, K. U. Heussner, J. Köninger, F. Langenegger, F. Leroy, C. Lima, N. Martinelli, G. Momber, A. Billamboz, O. Nelle, A. Palomo, R. Piqué, M. Ramstein, R. Schweichel, H. Stäuble, W. Tegel, X. Terradas, F. Verdin, C. Plomion, A. Kremer, L. Orlando, *Mol. Ecol.* **2018**, *27*, 1138.
[56] T. Maričić, M. Whitten, S. Pääbo, *PLoS One* **2010**, *5*, e14004.
[57] A. W. Briggs, J. M. Good, R. E. Green, J. Krause, T. Maričić, U. Stenzel, C. Lalueza-Fox, P. Rudan, D. Brajkovic, Z. Kucan, I. Gušić, R. Schmitz, V. B. Doronichev, L. V. Golovanova, M. de la Rasilla, J. Fortea, A. Rosas, S. Pääbo, *Science* **2009**, *325*, 318.
[58] M. Hajdinjak, Q. Fu, A. Hübner, M. Petr, F. Mafessoni, S. Grote, P. Skoglund, V. Narasimham, H. Rougier, I. Crevecoeur, P. Semal, M. Soressi, S. Talamo, J. J. Hublin, I. Gušić, K. Ž, P. Rudan, L. V. Golovanova, V. B. Doronichev, C. Posth, J. Krause, P. Korlević, S. Nagel, B. Nickel, M. Slatkin, N. Patterson, D. Reich, K. Prüfer, M. Meyer, S. Pääbo, J. Kelso, *Nature* **2018**, *555*, 652.
[59] V. Slon, B. Viola, G. Renaud, M. T. Gansauge, S. Benazzi, S. Sawyer, J. J. Hublin, M. V. Shunkov, A. P. Derevianko, J. Kelso, K. Prüfer, M. Meyer, S. Pääbo, *Sci. Adv.* **2017**, *3*, e1700186.
[60] J. Krause, Q. Fu, J. M. Good, B. Viola, M. V. Shunkov, A. P. Derevianko, S. Pääbo, *Nature* **2010**, *464*, 894.
[61] C. Posth, C. Wissing, K. Kitagawa, L. Pagani, L. van Holstein, F. Racimo, K. Wehrberger, N. J. Conard, C. J. Kind, H. Bocherens, J. Krause, *Nat. Commun.* **2017**, *8*, 16046.
[62] K. Douka, V. Slon, Z. Jacobs, C. B. Ramsey, M. V. Shunkov, A. P. Derevianko, F. Mafessoni, M. B. Kozlikin, B. Li, R. Grün, D. Comeskey, T. Devièse, S. Brown, B. Viola, L. Kinsley, M. Buckley, M. Meyer, R. G. Roberts, S. Pääbo, J. Kelso, T. Higham, *Nature* **2019**, *565*, 640.
[63] E. Ehler, J. Novotny, A. Juras, M. Chylenski, O. Moravcík, J. Paces, *Nucleic Acids Res.* **2019**, *47*, D29.
[64] R. E. Green, J. Krause, S. E. Ptak, A. W. Briggs, M. T. Ronan, J. F. Simons, L. Du, M. Egholm, J. M. Rothberg, M. Paunovic, S. Pääbo, *Nature* **2006**, *444*, 330.
[65] I. Olalde, M. E. Allentoft, F. Sanchez-Quinto, G. Santpere, C. W. Chiang, M. DeGiorgio, J. Prado-Martinez, J. A. Rodríguez, S. Rasmussen, J. Quilez, O. Ramírez, U. M. Marigorta, M. Fernández-Callejo, M. E. Prada, J. M. Encinas, R. Nielsen, M. G. Netea, J. Novembre, R. A. Sturm, P. Sabeti, T. Marquès-Bonet, A. Navarro, E. Willerslev, C. Lalueza-Fox, *Nature* **2014**, *507*, 225.
[66] Q. Fu, A. Mittnik, P. L. F. Johnson, K. Bos, M. Lari, R. Bollongino, C. Sun, L. Giemsch, R. Schmitz, J. Burger, A. M. Ronchitelli, F. Martini, R. G. Cremonesi, J. Svoboda, P. Bauer, D. Caramelli, S. Castellano, D. Reich, S. Pääbo, J. Krause, *Curr. Biol.* **2013**, *23*, 553.
[67] A. Furtwängler, E. Reiter, G. U. Neumann, I. Siebke, N. Steuri, A. Hafner, S. Lösch, N. Anthes, V. J. Schuenemann, J. Krause, *Sci. Rep.* **2018**, *8*, 14075.
[68] D. Higgins, A. B. Rohrlach, J. Kaidonis, G. Townsend, J. J. Austin, *PLoS One* **2015**, *10*, e0126935.
[69] C. Schwarz, R. Debruyne, M. Kuch, E. McNally, H. Schwarcz, A. D. Aubrey, J. Bada, H. Poinar, *Nucleic Acids Res.* **2009**, *37*, 3215.
[70] T. Kivisild, *Hum. Genet.* **2017**, *136*, 529.
[71] M. Karmin, L. Saag, M. Vicente, M. A. Wilson Sayres, M. Järve, U. G. Talas, S. Rootsi, A. M. Ilumäe, R. Mägi, M. Mitt, L. Pagani, T. Puurand, Z. Faltyskova, F. Clemente, A. Cardona, E. Metspalu, H. Sahakyan, B. Yunusbayev, G. Hudjashov, M. DeGiorgio, E. L. Loogväli, C. Eichstaedt, M. Eelmets, G. Chaubey, K. Tambets, S. Litvinov, M. Mormina, Y. Xue, Q. Ayub, G. Zoraqi, et al., *Genome Res.* **2015**, *25*, 459.
[72] G. D. Poznik, Y. Xue, F. L. Mendez, T. F. Willems, A. Massaia, M. A. Wilson Sayres, Q. Ayub, S. A. McCarthy, A. Narechania, S. Kashin, Y. Chen, R. Banerjee, J. L. Rodriguez-Flores, M. Cerezo, H. Shao, M. Gymrek, A. Malhotra, S. Louzada, R. Desalle, G. R. Ritchie, E. Cerveira, T. W. Fitzgerald, E. Garrison, A. Marcketta, D. Mittelman, M. Romanovitch, C. Zhang, X. Zheng-Bradley, G. R. Abecasis, S. A. McCarroll, et al., *Nat. Genet.* **2016**, *48*, 593.
[73] M. Rasmussen, X. Guo, Y. Wang, K. E. Lohmueller, S. Rasmussen, A. Albrechtsen, L. Skotte, S. Lindgreen, M. Metspalu, T. Jombart, T. Kivisild, W. Zhai, A. Eriksson, A. Manica, L. Orlando, F. M. De La Vega, S. Tridico, E. Metspalu, K. Nielsen, M. C. Ávila-Arcos, J. V. Moreno-Mayar, C. Muller, J. Dortch, M. T. Gilbert, O. Lund, A. Wesolowska, M. Karmin, L. A. Weinert, B. Wang, J. Li, et al., *Science* **2011**, *334*, 94.
[74] M. Rasmussen, M. Sikora, A. Albrechtsen, T. S. Korneliussen, J. V. Moreno-Mayar, G. D. Poznik, C. P. E. Zollikofer, M. P. de León, M. E. Allentoft, I. Moltke, H. Jónsson, C. Valdiosera, R. S. Malhi, L. Orlando, C. D. Bustamante, T. W. Stafford Jr, D. J. Meltzer, R. Nielsen, E. Willerslev, *Nature* **2015**, *523*, 455.
[75] J. V. Moreno-Mayar, T. S. Korneliussen, J. Dalal, G. Renaud, A. Albrechtsen, R. Nielsen, A. S. Malaspinas, *Bioinformatics* **2019**, *36*, 828.
[76] K. Prüfer, C. de Filippo, S. Grote, F. Mafessoni, P. Korlević, M. Hajdinjak, B. Vernot, L. Skov, P. Hsieh, S. Peyrégne, D. Reher, C. Hopfe, S. Nagel, T. Maričić, Q. Fu, C. Theunert, R. Rogers, P. Skoglund, M. Chintalapati, M. Dannemann, B. J. Nelson, F. M. Key, P. Rudan, Ž. Kućan, I. Gušić, L. V. Golovanova, V. B. Doronichev, N. Patterson, D. Reich, E. E. Eichler, et al., *Science* **2017**, *358*, 655.
[77] N. Nakatsuka, E. Harvey, S. Mallick, M. Mah, , N. Patterson, D. Reich, *bioRxiv* **2020**, https://doi.org/10.1101/2020.02.06.938126.
[78] M. Meyer, M. Kircher, M. T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, et al., *Science* **2012**, *338*, 222.
[79] B. M. Peter, *bioRxiv* **2020**, https://doi.org/10.1101/2020.03.13.990523.
[80] K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, *Nature* **2014**, *505*, 43.

[81] H. Jonsson, A. Ginolhac, M. Schubert, P. L. Johnson, L. Orlando, *Bioinformatics* **2013**, *29*, 1682.
[82] C. L. Weiss, M. Dannemann, K. Prüfer, H. A. Burbano, *eLife* **2015**, 4.
[83] H. Al-Asadi, K. K. Dey, J. Novembre, M. Stephens, *Bioinformatics* **2019**, *35*, 1292.
[84] A. Ginolhac, M. Rasmussen, M. T. Gilbert, E. Willerslev, L. Orlando, *Bioinformatics* **2011**, *27*, 2153.
[85] S. Peyrégne, B. M. Peter, *bioRxiv* **2020**, https://doi.org/10.1101/2020.03.13.991240.
[86] A. Seguin-Orlando, M. Schubert, J. Clary, J. Stagegaard, M. T. Alberdi, J. L. Prado, A. Prieto, E. Willerslev, L. Orlando, *PLoS One* **2013**, *8*, e78575.
[87] I. Glocke, M. Meyer, *Genome Res.* **2017**, *27*, 1230.
[88] O. Handt, M. Richards, M. Trommsdorff, C. Kilger, J. Simanainen, O. Georgiev, K. Bauer, A. Stone, R. Hedges, W. Schaffner, *Science* **1994**, *264*, 1775.
[89] J. Dabney, M. Knapp, I. Glocke, M. T. Gansauge, A. Weihmann, B. Nickel, C. Valdiosera, N. García, S. Pääbo, J. -. L. Arsuaga, M. Meyer, *Proc. Natl. Acad. Sci. U.S.A* **2013**, *110*, 15758.
[90] J. S. Pedersen, E. Valen, A. M. Velazquez, B. J. Parker, M. Rasmussen, S. Lindgreen, B. Lilje, D. J. Tobin, T. K. Kelly, S. Vang, R. Andersson, P. A. Jones, C. A. Hoover, A. Tikhonov, E. Prokhortchouk, E. M. Rubin, A. Sandelin, M. T. Gilbert, A. Krogh, E. Willerslev, L. Orlando, *Genome Res.* **2014**, *24*, 454.
[91] A. W. Briggs, U. Stenzel, M. Meyer, J. Krause, M. Kircher, S. Pääbo, *Nucleic Acids Res.* **2010**, *38*, e87.
[92] D. Gokhman, E. Lavi, K. Prüfer, M. F. Fraga, J. A. Riancho, J. Kelso, S. Pääbo, E. Meshorer, L. Carmel, *Science* **2014**, *344*, 523.
[93] J. Hendy, F. Welker, B. Demarchi, C. Speller, C. Warinner, M. J. Collins, *Nat. Ecol. Evol.* **2018**, *2*, 791.
[94] L. S. Weyrich, A. G. Farrer, R. Eisenhofer, L. A. Arriola, J. Young, C. A. Selway, M. Handsley-Davis, C. J. Adler, J. Breen, A. Cooper, *Mol. Ecol. Resour.* **2019**, *19*, 982.
[95] L. Orlando, A. Ginolhac, G. Zhang, D. Froese, A. Albrechtsen, M. Stiller, M. Schubert, E. Cappellini, B. Petersen, I. Moltke, P. L. Johnson, M. Fumagalli, J. T. Vilstrup, M. Raghavan, T. Korneliussen, A. S. Malaspinas, J. Vogt, D. Szklarczyk, C. D. Kelstrup, J. Vinther, A. Dolocan, J. Stenderup, A. M. Velazquez, J. Cahill, M. Rasmussen, X. Wang, J. Min, G. D. Zazula, A. Seguin-Orlando, C. Mortensen, et al., *Nature* **2013**, *499*, 74.
[96] M. Meyer, Q. Fu, A. Aximu-Petri, I. Glocke, B. Nickel, J. L. Arsuaga, I. Martínez, A. Gracia, J. M. de Castro, E. Carbonell, S. Pääbo, *Nature* **2014**, *505*, 403.
[97] S. L. Schnorr, K. Sankaranarayanan, C. M. Lewis, Jr., C. Warinner, *Curr. Opin. Genet. Dev.* **2016**, *41*, 14.
[98] M. A. Spyrou, K. I. Bos, A. Herbig, J. Krause, *Nat. Rev. Genet.* **2019**, *20*, 323.
[99] F. M. Key, C. Posth, J. Krause, A. Herbig, K. I. Bos, *Trends Genet.* **2017**, *33*, 508.
[100] R. Hübler, F. M. Key, C. Warinner, K. I. Bos, J. Krause, A. Herbig, *Genome Biol.* **2019**, *20*, 280.
[101] S. Rasmussen, M. E. Allentoft, K. Nielsen, L. Orlando, M. Sikora, K. G. Sjögren, A. G. Pedersen, M. Schubert, A. Van Dam, C. M. Kapel, H. B. Nielsen, S. Brunak, P. Avetisyan, A. Epimakhov, M. V. Khalyapin, A. Gnuni, A. Kriiska, I. Lasak, M. Metspalu, V. Moiseyev, A. Gromov, D. Pokutta, L. Saag, L. Varul, L. Yepiskoposyan, T. Sicheritz-Pontén, R. A. Foley, M. M. Lahr, R. Nielsen, K. Kristiansen, E. Willerslev, *Cell* **2015**, *163*, 571.
[102] A. Andrades Valtuena, A. Mittnik, F. M. Key, W. Haak, R. Allmäe, A. Belinskij, M. Daubaras, M. Feldman, R. Jankauskas, I. Janković, K. Massy, M. Novak, S. Pfrengle, S. Reinhold, M. Šlaus, M. A. Spyrou, A. Szécsényi-Nagy, M. Tõrv, S. Hansen, K. I. Bos, P. W. Stockhammer, A. Herbig, J. Krause, *Curr. Biol.* **2017**, *27*, 3683.
[103] T. R. Feuerborn, E. Palkopoulou, T. van der Valk, J. von Seth, A. Munters, P. Pečnerová, M. Dehasque, I. Ureña, E. Ersmark, V. K. Lagerholm, M. Krzewinska, R. Rodríguez-Varela, A. Götherström, L. Dalén, D. Díez-del-Molino, *bioRxiv* **2020**, https://doi.org/10.1101/2020.03.05.974907.
[104] A. Herbig, F. Maixner, K. I. Bos, A. Zink, J. Krause, D. H. Huson, *bioRxiv* **2016**, https://doi.org/10.1101/050559.
[105] G. Louvel, C. Der Sarkissian, K. Hanghoj, L. Orlando, *Mol. Ecol. Resour.* **2016**, *16*, 1415.
[106] D. E. Wood, J. Lu, B. Langmead, *Genome Biol.* **2019**, *20*, 257.
[107] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, et al., *Nucleic Acids Res.* **2016**, *44*, D733.
[108] Human Microbiome Project C, *Nature* **2012**, *486*, 215.
[109] J. A. Gilbert, J. K. Jansson, R. Knight, *BMC Biol.* **2014**, *12*, 69.
[110] D. Knights, J. Kuczynski, E. S. Charlson, J. Zaneveld, M. C. Mozer, R. G. Collman, F. D. Bushman, R. Knight, S. T. Kelley, *Nat. Methods* **2011**, *8*, 761.
[111] C. Warinner, J. F. Rodrigues, R. Vyas, C. Trachsel, N. Shved, J. Grossmann, A. Radini, Y. Hancock, R. Y. Tito, S. Fiddyment, C. Speller, J. Hendy, S. Charlton, H. U. Luder, D. C. Salazar-García, E. Eppler, R. Seiler, L. H. Hansen, J. A. Castruita, S. Barkow-Oesterreicher, K. Y. Teoh, C. D. Kelstrup, J. V. Olsen, P. Nanni, T. Kawai, E. Willerslev, C. von Mering, C. M. Lewis Jr, M. J. Collins, M. T. Gilbert, F. Rühli, E. Cappellini, *Nat. Genet.* **2014**, *46*, 336.
[112] C. Warinner, A. Herbig, A. Mann, J. A. Fellows Yates, C. L. Weiß, H. A. Burbano, L. Orlando, J. Krause, *Annu. Rev. Genomics Hum. Genet.* **2017**, *18*, 321.

