Reconstructing ancient genomes and epigenomes

Ludovic Orlando1,2, M. Thomas P. Gilbert1,3 and Eske Willerslev1

Research involving ancient DNA (aDNA) has experienced a true technological revolution in recent years through advances in the recovery of aDNA and, particularly, through applications of high-throughput sequencing. Formerly restricted to the analysis of only limited amounts of genetic information, aDNA studies have now progressed to whole-genome sequencing for an increasing number of ancient individuals and extinct species, as well as to epigenomic characterization. Such advances have enabled the sequencing of specimens of up to 1 million years old, which, owing to their extensive DNA damage and contamination, were previously not amenable to genetic analyses. In this Review, we discuss these varied technical challenges and solutions for sequencing ancient genomes and epigenomes.

The 1984 publication of a short mitochondrial DNA sequence from the quagga, a zebra-like equid that has been extinct since the 1880s, initiated the field of ancient DNA (aDNA) research1. Following concomitant development of PCR and realization that DNA survived in osseous materials2, the future of aDNA research looked bright. However, the degraded nature of aDNA coupled with the sensitivity of PCR to contamination — whether derived from environmental microorganisms or human handling, and thus embedded in the samples, or in the form of laboratory and/or reagent contamination — contributed to a series of publications based on false-positive results. Given that these problems seriously undermined the field’s broader scientific interest and reliability until the mid-2000s, few would have expected that, by the field’s twenty-fifth birthday, the genome of an ancient human3 and draft genomes of the extinct mammoth1,2 and Neanderthals3 would have been sequenced. Today, many tens of ancient genomes, ranging from microbial pathogens4-13 to vertebrate genomes14-29 (including the quagga30), have been sequenced.

Paleogenomics is driven by high-throughput sequencing (HTS) platforms, some of which generate data from billions of short DNA fragments per run30. In most paleogenomic studies, DNA libraries are generated by ligating the genomic extract to generic adaptors, amplified using PCR and then subjected to HTS using so-called second-generation sequencing platforms. This contrasts with traditional PCR-based approaches, in which loci are individually targeted and sub-amplicon-sized DNA is unexploitable. In addition to enabling whole-genome sequencing, HTS revealed how a diverse range of fossil specimens that were previously ignored owing to an inability to yield PCR amplicons nevertheless contained ultrashort aDNA fragments (~30–50 bp). Combining HTS with extraction methods tailored to the short, damaged aDNA molecules increased the time window for aDNA sequencing by an order of magnitude to at least 1 million years in permafrozen regions31 and 500,000 years in temperate caves32. Beyond genomes, the profiling of the epigenetic landscape (that is, epigenomes) of these ancient samples has recently become feasible33,34, conferring the potential to characterize regulatory changes throughout evolutionary timescales. However, there are also difficulties in paleogenomic studies. Indeed, HTS has enhanced some of the challenges, including data authentication and contaminant identification, as well as accounting for inflated error rates caused by damaged nucleotides.

In this Review, we discuss key technological developments underpinning the paleogenomic revolution (FIG. 1) and describe post-mortem damage types common to aDNA and how they can be accounted for (and even exploited). Furthermore, we discuss how aDNA targets can be enriched relative to other DNA, how the resulting sequences can be analysed, and recent progress in characterizing ancient epigenomes. Throughout, we highlight current limitations and provide perspectives for future developments. As most advances relate to human calcified tissues (bones and teeth), we principally focus on these. Some of the key findings addressing long-standing debates in our own global population history...
are summarized in BOX 1 to illustrate the diversity of information that can be gathered, and recent literature describing other key evolutionary insights revealed by ancient genomics have been reviewed elsewhere\textsuperscript{35–38}.

**aDNA damage and tailored extractions**

**aDNA damage.** aDNA damage accumulates over time and was originally characterized using enzymatic reactions to reveal the presence of particular types of DNA damage (such as abasic sites and crosslinks)\textsuperscript{3} or gas chromatography experiments coupled with mass spectrometry\textsuperscript{39}. Later approaches inferred damage types on the basis of mutational patterns in sequence data\textsuperscript{40–43}. Specifically, an excess of C→T mutations, and their significant reduction following treatment with uracil DNA glycosylase\textsuperscript{44}, revealed cytosine deamination to uracil (a thymine analogue) as the most prominent base modification.

HTS data subsequently refined our understanding of such damage, demonstrating that deamination increases towards read termini\textsuperscript{45}, consistent with expectations of faster rates in the overhanging single strands at the fragment termini\textsuperscript{46,47} (FIG. 2). HTS data also revealed that depurination drives post-mortem DNA fragmentation, as genomic positions preceding read starts (corresponding to breaks or abasic sites in aDNA molecules) often consist of purines\textsuperscript{44}. This bias appears towards adenines for younger samples but guanines for older samples, possibly reflecting differences in fragmentation dynamics\textsuperscript{48} and/or base-specific resonance structures\textsuperscript{47}. Statistical models exploiting nucleotide misincorporation patterns in HTS data sets revealed single-strand breaks in aDNA\textsuperscript{44,46}, most likely at nicks or abasic sites. Finally, whereas indirect detection methods indicated that polymerase-blocking lesions such as interstrand crosslinks could be prominent in aDNA\textsuperscript{49}, direct experimental assays based on HTS data suggested a more minor contribution\textsuperscript{50}. Therefore, their general importance may be context dependent.

**Targeting ultrashort fragments.** Extensive aDNA fragmentation was documented early in the field’s history, with later quantitative PCR assays revealing up to 100-fold decreases in the abundance of PCR templates for each doubling of target size\textsuperscript{51}. As HTS generally allows most aDNA molecules to be sequenced over their full length, the resulting distribution represents a size-decay curve\textsuperscript{52} that enables direct quantitative comparisons of fragmentation across specimens through space, time and environmental conditions\textsuperscript{53}. Although random DNA fragmentation should decrease molecule numbers exponentially as size increases, aDNA templates often peak at 40–80 bp before this decay is observed. The exact median length observed reflects the overall fragmentation levels experienced after death, which generally increase with the depositional temperature\textsuperscript{54,55}. However, the deviation from the expected exponential decay curve for ultrashort sizes suggests that common extraction protocols do not recover, and thus do not optimally exploit, this fraction of molecules.

This challenge was met by introducing improved silica-based extraction protocols that modify volume and environmental conditions\textsuperscript{53}. Although random DNA fragmentation should decrease molecule numbers exponentially as size increases, aDNA templates often peak at 40–80 bp before this decay is observed. The exact median length observed reflects the overall fragmentation levels experienced after death, which generally increase with the depositional temperature\textsuperscript{54,55}. However, the deviation from the expected exponential decay curve for ultrashort sizes suggests that common extraction protocols do not recover, and thus do not optimally exploit, this fraction of molecules.

This challenge was met by introducing improved silica-based extraction protocols that modify volume and composition of the DNA-binding buffer\textsuperscript{51}. These methodological improvements increased recovery rates of 35–50-bp molecules by twofold to fivefold, and greatly contributed towards the sequencing of even very short molecules.
**Box 1 | Human evolution insights: one of the principal achievements of ancient genomics**

An area of great interest in the study of human evolution is clarifying the admixture history and the migration routes followed by our ancestors to create contemporary patterns of genetic variation\(^1\)\. Study of the historical hair of an Aboriginal Australian revealed the existence of a migration from Africa or the Middle East that reached Australia and that took place 20,000–30,000 years earlier than the migration that gave rise to present-day Europeans and Asians\(^5\)\. The 36,200-year-old bone remains from an Upper Paleolithic man from Kostenki, Russia, were also found to be genetically closer to contemporary Europeans than to contemporary Asians, suggesting an earlier date for the separation between these populations\(^5\)\. The 24,000-year-old remains of a child from Mal’ta, south-central Siberia, Russia, showed strong genetic affinities not only with Europeans but also with Native Americans, indicating a mixed population ancestry for the first Americans\(^5\)\. The Solotuken theory, which assumed a European origin across the Atlantic for the Paleo-Indian Clovis culture in North America, could be ruled out because the 12,600-year-old cranial remains of the Anzick individual belonging to this culture shows greater genetic affinities to Native Americans than to Europeans\(^5\)\. The peopling of Europe and the effect of the agricultural revolution have also received great attention\(^5\),\(^11\),\(^18\),\(^21\),\(^22\),\(^27\),\(^18\),\(^51\),\(^53\),\(^71\),\(^122\).\(^\text{1}\)

The main genetic components present in modern Europeans seem to have already differentiated by 36,200 years ago\(^2\)\(^7\),\(^1\),\(^27\),\(^122\), and their later dispersal involved several migration waves\(^5\),\(^12\). The expansion of the first Neolithic farmers resulted in mixing hunter-gatherer Mesolithic and near-eastern population backgrounds within western Europe ~7,500 years ago\(^1\),\(^25\),\(^60\),\(^18\),\(^123\). A later extensive migration took place ~4,500 years ago from the steppes and was associated with the spread of Indo-European languages into Europe\(^5\). The possibility to gather genome-wide data at population scales from ancient individuals now provides an opportunity for a fine reconstruction of population migration and admixture patterns from classical antiquity to modern times.

In some cases, ancient genomes have revealed direct genetic continuity across different archaeological cultures, questioning theories assuming that culture only changes through the migration of peoples and not simply though the spread of ideas. The first example is provided by the Paleo-Eskimos from the New World Arctic, who represent distinct cultural units but were found to represent a single population, first replaced by Inuit ~1,000 years ago\(^2\)\(^7\),\(^1\). Ancient genomic data have also enhanced our understanding of deeper human evolution, showing that modern humans admixed with Neanderthals\(^6\),\(^9\) ~50,000–60,000 years ago\(^7\),\(^12\),\(^27\),\(^122\) while expanding out of Africa. Current data suggest that more-recent admixture events might even have taken place\(^12\), but this needs further investigation. The existence of the Denisovans, who may represent a distinct population of Neanderthals from the Altai mountains that substantially contributed to the genomic diversity of present-day Melanesians, was revealed based on genomic data\(^14\),\(^16\),\(^71\),\(^12\). The sequencing of the mitochondrial genome of ~400,000-year-old hominins from Atapuera, Spain, also revealed genetic affinities between these individuals and the Denisovans, although genome-wide information is needed to understand the underlying population history\(^7\),\(^12\),\(^2\).

Finally, when the age of a specimen can be determined using radiocarbon dating, it can be used as a calibration point to help in the estimation of genome-wide mutation rates. When applied to the remains of a 45,000-year-old human from Siberia, this technique confirmed that the autosomal mutation rate was about half of that estimated from the human–chimpanzee divergence\(^12\),\(^27\), which is consistent with recent estimates from human pedigrees\(^12\),\(^2\). Similarly, when the age of a sample falls outside the range of radiocarbon dating, the deficit of mutations observed along the phylogenetic branch leading to the specimen can be used to provide an estimate for when it lived\(^18\),\(^5\).

**Pre-digestion**

*Exposure of ancient calcified materials to a short initial digestion aimed at removing substantial fractions of exogenous contaminants.*

454

The initial generation of GS-FLX sequencing platforms based on pyrosequencing, before their acquisition and renaming by Roche.

**DNA library construction and amplification**

**General recommendations.** Second-generation sequencing requires template molecule modification through adaptor ligation\(^4\)\(^5\)\(^6\)\(^7\). Both library construction and subsequent PCR amplification represent sources of error\(^6\),\(^8\),\(^9\). The parts of a genome sequenced can be affected by adaptor binding biases and/or the relative efficacy of PCR enzymes to amplify the constructs. Which and where nucleotide misincorporations occur during these amplifications also confer errors in resulting sequences\(^10\),\(^11\),\(^12\). For example, the Phusion polymerase, which was originally part of the Illumina library building procedure, preferentially amplifies short and relatively GC-rich templates\(^4\)\(^5\)\(^6\)\(^7\). The same is true for related polymerases, such as Phusion Hot Start I and II, even when high-fidelity buffers are used. This bias is reduced, or even disappears, when other polymerases are used, and Accuprime Pfu, Herculease II Fusion and Pfu Turbo Cx Hotstart currently seem to be better alternatives than the most commonly used polymerases, AmpliTag Gold and Platinum Taq High-Fidelity\(^13\). Increasing PCR cycle number often reduces the molecular complexity of DNA libraries\(^14\); thus, polymerases should be carefully selected, PCR amplification cycles minimized and/or independent PCR reactions undertaken in parallel to limit such biases. This has important consequences for authenticating aDNA data and quantifying post-mortem DNA damage, as expected misincorporation models require tailoring to the exact experimental procedure followed\(^15\),\(^16\),\(^17\).

**Double-stranded DNA libraries.** Different DNA library construction methods also show clear differences in efficiency. Early aDNA libraries were based around 454-compatible blunt-end approaches\(^18\),\(^19\)\(^20\) (Fig. 3a),
However, as adaptor ligation is random, a fraction of the constructs do not contain both of the different adaptors and thus cannot be sequenced using this method. Another possible limitation is adaptor dimer formation during ligation; if amplified and sequenced, these waste sequencing capacity. Illumina introduced T/A ligation to overcome this in their original library construction procedure, in which aDNA fragments have an overhanging adenine added (known as A-tailing) to facilitate ligation to T-tailed adaptors ([FIG. 3b](#)). However, this strategy seems to be suboptimal for aDNA, mostly because templates starting with thymines are less efficiently processed during ligation. Thus the (often substantial) fraction of templates containing deaminated cytosine residues (thymine analogues) at their termini fails to incorporate into libraries. TruSeq libraries, which also rely on T/A ligation, have also been shown to introduce significant amounts of palindromic artefacts, whereby short sequence segments at read starts are copied towards read ends.

**Single-stranded DNA libraries.** A subsequent development was library construction directly on single-stranded DNA (ssDNA) templates. In this method, DNA is denatured using heat into single strands and then ligated to a first adaptor, before extension with Bst polymerase generates the complementary strand. A second adaptor is ligated at the 3’ end of the complementary strand, and the full construct is then amplified by PCR ([FIG. 3c](#)). Inclusion of biotin in the first adaptor allows minimal DNA loss during purification using streptavidin-coated paramagnetic beads. The development of this method enabled characterization of the Denisovan genome at ~30x coverage using DNA extracts generated from 40 mg of bone material. Although the method is sometimes beneficial on highly degraded osseous materials (as both strands and every single-strand break of endogenous DNA molecules have 3’ termini that are compatible with their incorporation into libraries), its benefit on less-degraded and non-osseous materials remains unverified.

**Enriching for aDNA**

* aDNA extracts are metagenomic mixtures. The endogenous DNA within most ancient specimens is usually embedded within high levels of environmental microbial DNA. Although there are notable exceptions (including some keratinized materials, particularly dense bones such as the petrosal bone, and intentionally preserved materials from museums or herbaria), it is unusual for the endogenous DNA content in most calcified remains to account for more than a few percent of the total DNA content. DNA preservation and environmental microbial contamination levels can show extreme variation within a single bone. For example, extracts and libraries constructed from a single 36,000-year-old European human bone yielded 0.1–8.0% of human DNA, and even greater variation (0.5–27.8%) was seen using the early Native American ‘Anzick’ cranial bone.

High microbial contaminant DNA levels render shotgun sequencing of genomes uneconomical. Thus, several methods have been developed that improve accessibility to endogenous aDNA. These enrichment strategies are used either during library construction, by preferentially incorporating damaged aDNA fragments, or after library construction, by separating

---

**Figure 2 | Typical ancient DNA molecules.** A diverse range of degradation reactions affect DNA post-mortem and result in extensive fragmentation (preferentially at purine nucleotides) and base modifications. The most common base modification identified in high-throughput sequencing data sets is deamination of cytosines into uracils (red), or thymines (blue) when cytosines were methylated (°C). Such deaminations occur much faster at overhanging ends. Other modifications include abasic sites (green) and single-strand breaks (vertical lines). The chemical structures of three damage by-products (uracils, thymines and abasic sites) are shown. R, purine; Y, pyrimidine.
endogenous and exogenous fractions through annealing to pre-defined sets of probes (in solution or on microarrays). Intended capture targets range from whole mitochondrial genomes (~16 kb) or ancient commensal and pathogenic bacterial genomes (~4 Mb) to large sets of single-nucleotide polymorphisms (SNPs) (~400,000 SNPs), whole exomes (~30 Mb), chromosomes (~30 Mb) and even whole nuclear genomes (~3 Gb). Other approaches that have been demonstrated, although not used in the most recent relevant studies, include targeted digestion of environmental microbial DNA using restriction enzymes and primer extension capture (PEC). Before discussing enrichment strategies further, we highlight that currently none is able to recover 100% of the target molecules, and thus they come at a cost of reduced

Figure 3 | Constructing ancient DNA libraries. The three most common types of ancient DNA (aDNA) libraries are shown. 5’-phosphate groups are indicated with black circles, single-strand DNA breaks are shown as vertical lines, biotinylated adaptor groups are shown in red, and streptavidin-coated beads are shown in grey. a | To construct a double-stranded DNA (dsDNA) library, aDNA is first end-repaired. It is then ligated to double-stranded adaptors (blue), and the resultant nicks are filled in to construct library templates devoid of single-strand breaks. b | To construct an A-tailed DNA library, aDNA is end-repaired and then A-tailed (that is, an adenine is added to the 3’ ends of the strands) to facilitate subsequent ligation to T-tailed adaptors while disfavouring ligation between adaptor pairs. The adaptors are typically Y-shaped (that is, they are complementary at the T-tailed end but have non-complementary arms at the other end). The use of such adaptors results in aDNA strands being flanked by distinct non-complementary adaptor sequences at each end to enable subsequent unidirectional sequencing through the aDNA fragment. Nicks resulting from ligation are filled-in through PCR post-ligation. c | To construct a single-stranded DNA (ssDNA) library, aDNA is first denatured into single strands using heat and then ligated to biotinylated single-stranded adaptors. The original DNA strand is then copied using DNA polymerase extension, and a second adaptor is ligated to enable further PCR amplification and sequencing. Purification steps are performed using streptavidin-coated paramagnetic beads. Part c adapted with permission from REF. 16, American Association for the Advancement of Science.
library complexity. Therefore, the upper threshold on the maximum sequencing depth attainable from a given library is reduced, and users must consider the end goal of their analyses before determining whether capture is a sensible strategy over direct shotgun sequencing. If the goal is to sequence to high coverage, highly complex libraries showing relatively high endogenous content can be shotgun-sequenced\(^ {15,16,18,22,28}\), but enrichment of multiple libraries is advisable in other cases\(^ {21,72}\).

**Damaged template enrichment.** One approach selectively targets damaged DNA molecules\(^ {46}\) during ssDNA library preparation\(^ {25,66}\). After the DNA strand complementary to the original template is generated, constructs are 5’-phosphorylated, which enables ligation to a non-phosphorylated adaptor (FIG. 4a). Following extension with Bst polymerase to fill the nick located 5’ of this adaptor, treatment with uracil DNA glycosylase and endonuclease VIII ( USER mix) is implemented to first replace deaminated cytosines with abasic sites and then to cleave out these abasic sites\(^ {41}\). The new 3’ end is then dephosphorylated and used for priming a new extension. Thus, all library strands that originally harboured deaminated cytosines are reconstructed over their full length and are available in the reaction supernatant for further amplification and sequencing. The undamaged DNA template fraction remains attached to streptavidin-coated paramagnetic beads and can be retained for other uses. This method has shown great specificity when applied to samples from Late Pleistocene Neanderthals showing extreme levels of deamination\(^ {48}\). Importantly, in all extracts tested, the relative contamination from modern human DNA decreased by ~1.6-fold following selective enrichment, suggesting that undamaged templates resulting from recent manipulations of the specimen could readily be filtered. Furthermore, the endogenous content of one sample increased by 3.7–5-fold, which markedly reduced the genome sequencing cost. Future experiments will no doubt explore the wider potential of this method. For now, users should bear in mind that any endogenous undamaged molecules will not be retained and will thus be lost, making the method only appropriate for the most damaged samples. Additionally, any DNA carrying damage will be enriched, potentially providing access to the genomes of associated ancient microorganisms (although these can show reduced DNA damage levels compared to their human hosts\(^ {7}\)).

**Extension-free target enrichment in solution.** Target enrichment approaches based on target–probe hybridization are currently widely used. These require heat denaturation of DNA libraries to enable annealing of library inserts to overlapping tiled probes along target regions. Probes can be economically generated using long-range PCR, if fresh DNA material from closely related species can be extracted\(^ {18}\), through PCR amplification and then ligation to a biotinylated adaptor. This probe library can be amplified (with biotinylated primers) and used in an unlimited number of enrichment reactions. Following annealing at stringencies that can be adapted depending on the phylogenetic distance between targets and probes, streptavidin-coated beads are washed to eliminate library constructs with inserts showing no genetic proximity to the targeted regions, and the final fraction is amplified and sequenced.

This strategy has predominantly been used for sequencing mitochondrial genomes\(^ {72,74,75,82–85}\), bacterial plasmids\(^ {86}\) and short nuclear loci\(^ {87}\). Hybridization is even successful when probes diverge from targets by 10–13%\(^ {82}\), which is useful if no close living relative and/or reference genome is available. This can also be exploited to detect probe carry-over post-sequencing if the DNA from a distantly related organism was used for preparing probes (for example, if DNA from a European bison was used when enriching for aDNA from aurochs\(^ {88}\)). Alternatively, potential probe carry-over can be eliminated before sequencing using dedicated molecular tools. For example, replacement of deoxythymidine triphosphate (dTTP) by deoxyuridine triphosphate (dUTP) in probes enables subsequent digestion with uracil DNA glycosylase before amplification and sequencing\(^ {89}\).

Biotinylated probes can also be custom designed and synthesized, which enables specific probe tiling and *in silico* assessment for secondary structures, homogeneous GC content and annealing temperatures. Different manufacturers can now deliver such probes, with related procedures apparently achieving similar efficiency\(^ {90}\). Depending on the overall size of the genomic regions targeted, multiple libraries can, in theory, be enriched as pools to achieve faster hands-on times. However, owing to the probable formation of chimeric DNA libraries during post-capture PCRs, pooling of libraries before capture should ideally be avoided, or if pooling is used then the constituent libraries should at least be double-indexed DNA libraries\(^ {91}\) to enable chimera identification and elimination from subsequent analyses. Increasing probe tiling densities (11 bp versus 24 bp) did not consistently improve enrichment for ~670 nuclear loci in archaeological maize, suggesting that even relatively reduced probe densities can be used to efficiently recover the full molecular complexity of DNA libraries\(^ {88}\).

In general, custom-synthesized biotinylated probes are most economical when targeting fairly small regions (hundreds of kilobases to a few megabases) owing to probe synthesis costs. However, microarrays can achieve extremely high probe numbers (approximately 1 million each) and, if manufacturers consent, can be chemically treated to cleave the probes from the microarray surface, thus recovering large sets of probes at relatively reasonable costs\(^ {92,76}\). Synthetic DNA probes are built into biotinylated probe libraries using biotinylated adaptors of minimal size (~20 bp) to limit interference during probe–target annealing. The known adaptor sequence allows further amplification, thereby immortalizing the probe set at low cost. In this way, Fu *et al.*\(^ {72}\) used 8.7 million probes to recover most of the non-repetitive fraction of chromosome 21 from a 40,000-year-old human specimen from Tianyuan cave, China. In addition, they targeted ~3,500 200-bp-long regions around positions...
**Figure 4 | Enriching DNA libraries for ancient inserts.**

**a** | Selective uracil enrichment is shown. 5′-phosphate groups are indicated with black circles, single-strand DNA breaks are shown as vertical lines, biotinylated adaptor groups are shown in red, and streptavidin-coated beads are shown in grey. A single-stranded DNA (ssDNA) library is built until the polymerase extension step. DNA is then phosphorylated to enable the ligation of the second adaptor. This contrasts with the ssDNA library procedure, in which the ligation occurs between the 5′ end of the second adaptor and the 3′ end of the newly synthesized strand (FIG. 3c). DNA is then treated with uracil DNA glycosylase and endonuclease VIII (USER mix) to generate and then cleave out abasic sites at cytosines that were deaminated into uracils post-mortem. The 3′-phosphate groups at these new termini are then removed (not shown). The resulting 3′-OH ends now serve to prime an extension with a DNA polymerase, which copies throughout the whole length of the strand complementary to where the damage was. As a result, the supernatant now contains double-stranded DNA (dsDNA) library templates corresponding to the original deaminated strands. Other library templates remain unaffected and can be separated, as they remain bound to streptavidin-coated paramagnetic beads.

**b** | In whole-genome in-solution capture (WISC), ssDNA templates from an ancient DNA (aDNA) library are prepared. The target, endogenous aDNA is shown as thin black lines, whereas the exogenous contaminating DNA is shown as thin green lines; adaptors are shown as thick blue lines. In parallel, a probe DNA library is prepared from fresh modern DNA extracts (thin red lines) and used to generate biotinylated RNA probes through *in vitro* transcription. T7 adaptors to enable *in vitro* transcription are shown in thick purple lines. The aDNA library is annealed to the RNA probes, low-complexity DNA and adaptor blockers (the latter two are not shown for simplicity). The library fraction of interest is then recovered following elution from streptavidin-coated paramagnetic beads.

Part **a** adapted with permission from REF. 68, Cold Spring Harbor Laboratory Press. Part **b** adapted with permission from REF. 78, The American Society of Human Genetics.
known to carry allelic variants in archaic and modern humans, thereby enabling direct estimates of archaic hominin ancestry within the Tianyuan specimen. The method was also used to obtain the exome sequence of two Neanderthals from Spain and Croatia and, more recently, sequence data from ~400,000 loci within a single reaction. This target enrichment procedure reduced the genotyping costs by at least 45-fold per ancient specimen and enabled genome-wide analyses of ancient individuals at population scales. In this analysis, two 52-nucleotide-long probes were selected to be located on each side of a polymorphic site, and two were centred on the polymorphic site, each representing one of the two possible alleles.

**Solid-phase target enrichment.** Direct application of microarrays can also enrich large sets of targets, using approaches originally described for modern DNA. First used in the aDNA context to characterize exome sequences from a 49,000-year-old Neanderthal specimen, microarrays have subsequently enabled whole-genome sequencing from bacterial strains responsible for major historical epidemiological outbreaks, including the Black Death. Microarrays also provide interesting alternatives to real-time PCR and shotgun sequencing for parallel screening of >100 pathogens. This is particularly appropriate for identifying ancient pathogens, which often leave no physical skeletal evidence and are generally found only as trace material. Possible drawbacks are poor detection of the most divergent genomic regions and omission of regions with important genomic rearrangements (such as insertions) or unknown additional plasmids that do not segregate in modern strains.

**Whole-genome enrichment.** There is a growing interest in characterizing the entire genome sequence of ancient individuals at population scales. However, none of the methods presented above is appropriate for pulling down whole human genomes, as this requires synthesizing gigabases of probes. Whole-genome in-solution capture (WISC) and a commercial alternative with similar performance fill this niche, enabling economical whole-genome enrichment. WISC starts with the preparation of a genome-wide RNA probe library from a species with a genome that is closely related to the target genome in the aDNA sample. These RNA probes are generated from a genomic DNA library flanked by adaptors containing T7 promoters that enable a relatively inexpensive reaction, in vitro transcription. This in vitro transcription step is carried out in the presence of biotin 16–UTP, so that the resultant RNA probes are biotinylated. The biotinylated RNA probes are annealed to the ssDNA of a heat-denatured aDNA library, while low-complexity DNA and adaptor-blocking RNA oligonucleotides improve stringency and reduce enrichment for highly repetitive regions. Non-hybridized DNA is washed away, whereas the bound, enriched library fraction is finally released following RNase treatment (which precludes probe carry-over) and amplified before sequencing.

WISC-like approaches consistently improve the proportion of sequences that can be mapped to the human reference genome compared to shotgun sequencing (6–159-fold), at least when based on double-stranded DNA libraries. As hybridization efficiency increases with target length, its efficacy may be reduced when analysing libraries built using single-strand methods, which routinely exhibit smaller mean target molecule sizes. The fraction of reads that align to repetitive regions also generally increases with WISC, despite the use of an excess of low-complexity DNA. Unsurprisingly, WISC-enriched libraries show reduced complexity, so that almost every unique insert can be sequenced with minimal sequencing efforts. As an example, 5–10 million sequencing reads generated using WISC-enriched libraries of a Bronze Age Danish human hair sample and a pre-Columbian Peruvian human bone were found to cover 7,000–21,000 ancestry-informative markers, which proved to be sufficient for inferring the continental groups that are the closest to these ancient individuals.

**Analysing aDNA**

**From reads to genome alignments.** Most available paleogenomes were generated using Illumina technologies, although there are exceptions. Analysis of the underlying sequence data mainly relies on computational approaches developed for handling HTS data from modern DNA material, with some additional particularities. Most procedures are implemented within the open-source PALEOMIX package, in which reads are trimmed of adaptor sequences using AdapterRemoval and collapsed when mate pairs are available and overlap significantly, filtered for a minimal size of 25–30 bp and aligned against reference genomes of interest using Burrows–Wheeler Aligner (BWA) or Bowtie 2 (REF. 94). Alignments showing low-quality scores and PCR duplicates are further removed using the MarkDuplicates program from Picard tools, and reads are locally realigned around small insertions and deletions (indels) to improve overall genome quality using the IndelRealigner tool from the Genome Analysis Toolkit (GATK). PALEOMIX can also quantify DNA damage levels using mapDamage2 and perform phylogenomic and metagenomic analyses using modules mostly based on inferences deriving from ExaML (Exascale Maximum Likelihood) and MetaPhiAn (Metagenomic Phylogenetic Analysis), respectively.

Unlike sequences in other re-sequencing genome projects, in which mismatches relative to the reference genome generally are derived from sequencing errors and polymorphisms, aDNA sequences exhibit substantial fractions of nucleotide misincorporations that result from sequencing damaged bases. As these misincorporations cluster towards read termini, seeding approaches, whereby only the most upstream part of the sequence is used for speeding up identification of possible alignments along the genome, should be avoided. Parameters controlling acceptance thresholds for read-to-reference edit distance should be adapted to the phylogenetic distance to the reference genome, as overly conservative procedures will under-represent the
most polymorphic regions and under-estimate heterozygosity levels. Conversely, overly permissive procedures will inflate the alignment false-positive rate, resulting in regions with many reads from different organisms, which is a particular challenge for aDNA data, given its complex mixture of endogenous and exogenous reads\textsuperscript{52,57}.

Owing to the accumulation of nucleotide misincorporation towards read ends, probabilistic aligners based on position-scoring matrices have been developed to embed aDNA features from the aligning step. Available aligners include Mapping Iterative Assembler (MIA)\textsuperscript{46}, ANFO Short Read Aligner/Mappe\textsuperscript{4} and BWA-PSSM (position-specific scoring matrix)\textsuperscript{99}, and these generally show good performance for short reads and/or low-quality data, although some show running times that are compatible only with alignments against relatively small reference genomes (for example, mitochondrial genomes). Importantly, such probabilistic approaches handle platform-specific error profiles in a sound statistical framework.

**Authenticating aDNA data.** Following read alignment, analyses often focus on authenticating whether sequencing data are ancient. Software such as mapDamage\textsuperscript{46,100} or pmdtools\textsuperscript{49} can test the presence of typical nucleotide misincorporation patterns that result from inflated cytosine deamination rates at overhangs. Such patterns can be first obtained by preparing libraries on an aliquot of the DNA extract, while saving the remaining fraction for preparing almost damage-free libraries following USER treatment\textsuperscript{41}. This will limit nucleotide misincorporation effects on downstream analyses. Alternatively, mild USER treatment, which removes most, but not all, of the damage signature, has been proposed to enable sequence authentication and population analyses using the same sample aliquot\textsuperscript{72}.

Nucleotide misincorporation patterns can be exploited to fit statistical models of post-mortem DNA damage and estimate cytosine deamination rates and nick frequencies\textsuperscript{44,48}. Even though deamination rates at overhangs were reported to increase linearly with time across a wide range of archaeological sites and preservation conditions\textsuperscript{86}, this pattern has not been confirmed within archaeological sites in permafrost\textsuperscript{17} or temperate environments\textsuperscript{72}. Additionally, different remainders from the same specimen and/or extracts from the same remain can show variable levels of DNA damage\textsuperscript{23,42}. This suggests complex relationships in which both global conditions, as reflected in the thermal age of a given specimen, and microenvironmental factors (within and between remains) drive the amount of DNA damage ultimately measured. In our opinion, these complex relationships, and the dependency of damage quantification on the library preparation and amplification procedures, preclude the use of strict minimal thresholds of expected DNA damage levels as authentication criteria. Thus, quantitative comparison with the levels observed for samples excavated at the same or similar archaeological sites, and processed with the same experimental tools, is recommended.

Statistical damage models also allow correction of base quality scores depending on their probability of being the result of nucleotide misincorporations at damage sites\textsuperscript{49}, thus limiting their possible effect on downstream analyses. However, we emphasize that for low-coverage data — in which mismatches are observed on a few reads at best and penalized when close to read termini — this procedure can potentially inflate the genetic proximity to the reference genome. SNP calling can also benefit from genotype callers, such as SNPeSt\textsuperscript{42,102}, that explicitly model post-mortem DNA damage as a possible source of error. Furthermore, nucleotide misincorporation patterns can be used by computational tools to sort the fraction of reads that show evidence of post-mortem damage\textsuperscript{41}, which is useful when there is substantial modern DNA contamination. Although extremely conservative and not cost-effective (as not all aDNA molecules carry post-mortem DNA damage and many true aDNA reads will be discarded), damage-based filtering approaches have shown great success in characterizing whole-mitochondrial sequences from extensively contaminated Neanderthal specimens\textsuperscript{101} and an ~400,000-year-old hominin\textsuperscript{73}. Finally, comparing analytical outcomes when considering the full population of reads or only the most damaged fraction (and disregarding mutations, such as transitions, that derive from post-mortem damage\textsuperscript{40–44}) can provide evidence that the results are not driven by damage and contamination artefacts\textsuperscript{101}.

In addition to revealing nucleotide misincorporation patterns, mapDamage also delivers the base composition of the genomic regions directly flanking DNA inserts and therefore tests depurination as the main driver for DNA fragmentation\textsuperscript{44,46,100}, which can also help authentication. This pattern is substantially affected following USER treatment, which mainly cleaves DNA downstream of unmethylated cytosine residues, therefore resulting in an excess of cytosines at genomic positions just preceding read starts\textsuperscript{46,72}.

**Estimating contamination levels.** Nucleotide misincorporation and base compositional patterns can be detected in even substantially contaminated samples. This can happen when treating the outer sample surface with bleach before DNA extraction, which can help to remove a fraction of fresh DNA contaminants but also introduces signatures of DNA damage within the remaining contaminants\textsuperscript{104}. This can also happen when a mixture of highly degraded aDNA templates and undamaged DNA contaminants is incorporated into libraries. A suite of tools has thus been developed for further authenticating aDNA data (in particular for human aDNA). The current methods available exploit the sequence information at sites and/or haplotypes with known variation across species and/or populations. For example, modern human contamination in Neanderthal HTS data has been estimated using the relative proportion of derived alleles and ancestral alleles observed at mitochondrial sites showing nearly fixed derived alleles in modern humans\textsuperscript{4}. A similar rationale was used to estimate the possible contribution of different human population backgrounds\textsuperscript{105}.
or species to final mitochondrial consensus sequences. A statistically more powerful contamination estimator for mitochondrial reads that uses linkage information at the read level has been developed.

As the cellular mitochondrial number is variable across cell types and tissues, contamination estimates based on mitochondrial sequence data do not directly reflect the true contamination levels of the nuclear genome. Heterozygosity levels observed on male X chromosomes can be used as a nuclear contamination proxy. As males are haploid for most X chromosome loci, base discordance between overlapping reads should result only from sequencing errors and should be distributed randomly along the chromosome. However, if modern human DNA contamination is present, discordance rates should inflate at sites that are polymorphic within contemporary populations. For archaic hominin specimens, nuclear contamination rates can be calculated from fixed alleles that are derived in modern humans. For female ancient human samples, the presence of sequences that are known to be unique to the Y chromosome can also reflect the presence of contamination from male-derived sources. Triallelic sites at autosomes could potentially be used in the future to estimate levels of nuclear contamination with modern human DNA, irrespective of the sample gender.

**Genome completion and error rates.** Reliable contamination estimates can generally be recovered from the data aligning to the X chromosome using even low-depth information, as long as each single genomic position is covered once on average (that is, ~1x coverage). Ultimately, the exact fraction of the genome that is covered depends on the sequencing effort and the sequence length. For aDNA sequence reads of 60 nucleotides, ~87% of the human genome is non-repetitive, and therefore reads of similar size (or shorter) cannot be uniquely aligned to the remaining ~13% of the genome. For example, the genome of a Paleo-Eskimo Greenlander of the Saqqaq culture was sequenced to ~16x coverage, and a high-quality genome from a living conspecific closely related outgroup (for example, the chimpanzee) three ing errors will have an impact on downstream analyses. Another parameter that is worth considering, as sequencing variation along the genome (see below), can also lead to specific patterns of depth variation. For aDNA sequence reads of 60 nucleotides, alignment to the remaining ~13% of the genome should result only from sequencing errors and should be distributed randomly along the chromosome. However, if modern human DNA contamination is present, discordance rates should inflate at sites that are polymorphic within contemporary populations. For archaic hominin specimens, nuclear contamination rates can be calculated from fixed alleles that are derived in modern humans. For female ancient human samples, the presence of sequences that are known to be unique to the Y chromosome can also reflect the presence of contamination from male-derived sources. Triallelic sites at autosomes could potentially be used in the future to estimate levels of nuclear contamination with modern human DNA, irrespective of the sample gender.

**Genome-wide epigenetic maps.** Cytosine methylation, which predominantly occurs in vertebrates at CpG dinucleotides, is the most commonly studied epigenetic modification. Classically, methylomes (that is, genome-wide maps of methylated cytosines) are reconstructed using bisulfite sequencing that converts unmethylated cytosines into uracils, which are sequenced as thymines, leading to CpG→TpG mutations. Methylated cytosines are not converted and are sequenced as regular CpG sites. Although it was successfully used to reveal fine-scale methylation patterns at four nuclear loci on the DNA extracted from a Late Pleistocene bison bone, this approach generally requires large amounts of DNA, as it inflicts extensive damage to DNA and is therefore generally inappropriate for aDNA. However, similar CpG→UpG conversions naturally occur post-mortem. In contrast to bisulfite treatment, methylated epialleles (mCpGs) are deaminated into TpGs, whereas unmethylated epialleles (CpGs) are deaminated into UpGs. As cytosine deamination is elevated in methylated contexts, a substantial fraction of mCpGs is expected to be converted into TpGs. Molecular tools that prevent UpG sequencing can therefore reveal methylated epialleles by tracking CpG→TpG mutations.

Briggs et al. first exploited this feature of aDNA, using HTS data generated from library inserts in which uracil residues were removed following USER treatment. Gokhman et al. subsequently generated high-resolution aDNA methylation maps using high-quality Neanderthal and Denisovan genomes, identifying ~2,000 regions that show differential methylation patterns in the bones of modern versus archaic hominins. Homeobox D10 (HOXD10), which encodes a key regulator of limb development, is found in one such region, and its epigenetic reprogramming is proposed to have participated in the shaping of specific morpho-anatomical features along the human lineage. Pedersen et al. used sequence data from the high-coverage Paleo-Eskimo Saqqaq genome to also extract genome-wide methylation information, as DNA libraries were amplified using a polymerase that does not bypass uracils. Furthermore, CpG→TpG substitution rates were

---

**Epialleles**

Allelic variants showing identical genetic sequences but different epigenetic marks, such as methylation patterns.
Box 2 | Reconstructing population histories

One of the most common first steps in the analysis of genome-wide data from ancient humans is the characterization of their closest relatives among modern populations. Such inferences are generally based on principal component analysis (PCA) or statistical clustering, using software such as Admixture\(^1\). A benefit of statistical clustering is that it also enables documentation of contamination levels through determining whether the ancient samples exhibit a genetic contribution that could be derived from the research team\(^1\). With shotgun sequencing at low depth of coverage (for example, ≤8×), genotypes cannot be reliably determined, and analyses are generally performed using pseudo-haploid data in which sequence reads from many loci consist of a random sampling of only one of the two constituent alleles, and thus individuals are considered to be homozygous for the unique allele sampled at a given locus. The genomic regions covered across multiple individuals are then also limited, which reduces the number of orthologous loci overlapping known genetic variation in modern populations. In such cases, the ancestry of each ancient individual can be determined using multidimensional scaling (MDS), which exploits pairwise measures of genetic distances in a panel of individuals, calculated by normalizing the sum of all instances where two individuals show different alleles by the total number of loci with no missing data in each pair. This procedure is implemented in the bammds package\(^17\). Additionally, Procrustes transformation of individual PCA projections based on the particular vector of single-nucleotide polymorphisms covered in each specimen and the same reference panel can help to visualize the population affinities of a group of ancient individuals within a single analysis\(^18\).

However, PCA-based approaches reflect not only population ancestries but also the temporal sampling between ancient and modern individuals\(^18\). Thus, at best, MDS, PCA and clustering analyses should be viewed as formulating evolutionary hypotheses, which subsequently require testing using approaches such as model-based inference, as well as coalescence simulations\(^14\). D-statistics\(^19\) and population f-statistics\(^20\).

Population f-statistics methods, such as the \( f_r \) statistics, have been developed for detecting populations with mixed ancestries and identifying populations that are closest to ancient individuals\(^10\). D-statistics has received particular attention because it originally supported the theory that admixture occurred between Neandertals and non-African modern humans\(^5\). D-statistics is based on four-way alignments that include one outgroup (O) and three populations (H\(_1\), H\(_2\), and H\(_3\)), of which two (H\(_2\) and H\(_3\)) are more closely related. For example, in the case of Neandertals, with the following configuration (O = Chimpanzee, H\(_1\) = Neandertals; H\(_2\) = Eurasians, H\(_3\) = Africans), positive D-statistics indicate an excess of shared polymorphisms and possible admixture between Neandertals and Eurasians\(^5,16,21\). However, this observation is also compatible with gene flow into Africans from a currently unsampled and divergent ghost population\(^12\), as well as with population subdivision in Africa, with Neandertal and Eurasian ancestors leaving Africa from related population backgrounds\(^11\).

Admixture events can be further dated from the distribution of introgressive block lengths in modern and ancient individuals\(^22,12,13\), as recombination reduces their size over time. The resulting date seemed to be too recent to be compatible with a scenario involving population subdivision in Africa, which confirmed admixture with Neandertals outside Africa.

---

Ghost population

An unsampled population that exchanges migrants with other sampled populations and that can be identified based on admixture signatures left in descending populations.

Introgressive block lengths

Population admixture introduces a mosaic of ancestry blocks along the genome, the lengths of which decrease with each subsequent generation owing to recombination. Introgressive block lengths can therefore be exploited to determine the date of admixture events.

Genome-wide nucleosome maps

Library inserts derived from endogenous aDNA generally show unimodal size distributions that are typically centred around 40–80 bp. However, several aDNA sequence data sets exhibit striking 10-bp periodicity in their size distributions\(^2,21,25,26,34,79\). Pedersen et al\(^34\) proposed that this result from nucleosome protection, with DNA fragmentation preferentially occurring at nucleotides facing away from nucleosomes. Assuming that nucleosomes are strongly positioned and phased along DNA scaffolds, and recalling that the turn of the DNA helix is 10 bp long, only 1 nucleotide per 10 bp would be fully exposed to hydrolysis. If this is true, then nucleosome protection should also drive additional patterns. For example, DNA fragmentation should occur preferentially within spacers, which are nucleosome-free regions of ~50 bp separating successive ~150-bp DNA blocks covered by nucleosomes. Fewer endogenous reads should therefore map to spacer regions, leading to depth-of-coverage periodicities of ~200 bp, with peaks of coverage corresponding to nucleosome centres and correlating with both in silico predicted and experimentally derived nucleosome maps. These predicted periodicities were confirmed in the Saqqaq sample data, even following correction for base compositional effects, which can substantially affect depth-of-coverage variation during library amplification\(^8\). This finding, together with expected patterns of methylation and depth of coverage within CTCF regions and splicing sites, confirmed the nucleosome protection hypothesis.

Nucleosomes might protect DNA from cleavage that occurs during cellular apoptosis and/or post-mortem\(^34\). As similar periodicity patterns have been found not only in ancient hair follicles\(^13,14\), which have undergone extensive apoptosis, but also in other ancient tissues that are not particularly affected by apoptosis, such as teeth\(^21\) and bone\(^25,26,79\), we expect that ancient nucleosome maps could, in the future, be reconstructed across a wide range of samples. Recalling that such patterns are also absent from many of the samples analysed so far, further work is needed to understand which factors drive the preservation of signatures of nucleosome protection after death.

Assessing ancient gene expression levels

Post-mortem DNA damage enables the reconstruction of ancient methylome and nucleosome maps. Given the central role of epigenetic states in regulating chromatin accessibility to transcription factors, this information can be tentatively used to infer ancient gene expression levels. Encouragingly, methylation ratios between gene bodies and promoter regions (a proxy for gene expression) showed strong correlation with hair follicle expression levels measured using high-throughput RNA sequencing (RNA-seq)\(^34\). However, further work is needed to develop genuine proxies that accurately measure ancient gene expression levels. The epigenome of each cell type is complex, and ancient samples will necessarily span a range of tissues, with unbalanced contributions from different cell types, which will possibly result in variable validity of expression predictions across samples, age, sex and health conditions. As one example, genome hypermethylation is a known response to viral infection.
in plants, and methylation assays for ancient plant material can therefore be used to monitor viral exposure in ancient populations\textsuperscript{109}.

Conclusions

Recent technical developments have enhanced our understanding of the properties of aDNA molecules and how we should best proceed to maximize their retrieval. In some environments, this enables genomic characterization throughout much of the past million years\textsuperscript{12,31,32}. Ongoing research and the increasing wealth of sequencing data generated will undoubtedly further improve current approaches in the near future. DNA extraction represents an area with great potential for improvement, especially if tailored to the molecular structures, niches and microenvironmental parameters that best preserve DNA.

The discovery that post-mortem cytosine deamination preferentially occurs at overhangs was important for the development of authentication criteria\textsuperscript{41}. However, other base modifications, including pyrimidine derivatives, have been identified\textsuperscript{39}. Improved characterization of the chemical features of aDNA molecules, as well as their methylation and nucleosome protection patterns, could therefore open new avenues for data authentication. This will also improve our ability to correct sequence analyses from as-yet- unidentified biases and provide opportunities for targeting damaged templates before sequencing. The development of engineered DNA polymerases that can bypass specific DNA lesions introduced post-mortem\textsuperscript{116} could also facilitate library construction and amplification.

Importantly, although the approaches outlined here improve aDNA retrieval and analyses, the HTS technologies themselves had the greatest impact on the field. Although not originally designed for aDNA, their massive throughput coupled with their ability to sequence short molecules rendered them ideal for aDNA applications. Therefore, it is likely that future HTS platforms that directly sequence DNA bases and their modifications with minimal (if any) library preparation will drive the future of aDNA research. The results of the initial application of true single-molecule DNA sequencing are encouraging, having demonstrated substantial improvement in relative amounts of accessible endogenous sequences\textsuperscript{17,45,56}.

Although most paleogenomic studies have focused on a limited number of individuals, current approaches allow the characterization of genome-wide SNP variation at ancient population scales\textsuperscript{71,111}. Future studies can be expected to investigate genetic variation in large population samples on the high-density SNP or even whole-genome scale, thus improving our understanding of past demographic, adaptive and admixture trajectories with greater detail\textsuperscript{112}.

Besides delivering ancient genomes and epigenomes, new methodological developments have also provided access to ancient transcriptomes\textsuperscript{113,114} and proteomes\textsuperscript{115,116}. Owing to the biochemical processes inherent in animal cell death, animal tissues are unlikely to represent good reservoirs for long-term RNA survival. Materials still exist in other organisms that do not undergo autolysis. One example is plant seed, a tissue that requires RNA survival for germination and that has demonstrated ancient RNA survival going back hundreds to thousands of years\textsuperscript{113,114}. Such materials may contribute to our understanding of how gene expression pathways have been remodelled during domestication. Additionally, a wide range of ancient proteins have been sequenced from Late\textsuperscript{115} and Middle\textsuperscript{117} Pleistocene specimens. With half-lives exceeding that of DNA, ancient peptides might be the only way to retrieve genetic information from the early Pleistocene and even earlier time periods. Within a much more recent time range, namely the past few thousand years, studies of proteins have already delivered information that is not obtainable from DNA, such as whether milk products were already consumed in particular ancient societies\textsuperscript{118}. Molecular analyses of dental plaque, which offers a rich reservoir entrapping biomolecules derived not only from the host but also from its diet and the oral microbiome\textsuperscript{119,120}, may also hold great promises, especially now that computational approaches have been developed to compare the diversity of past and present microbiomes\textsuperscript{121}.

A final question worth considering is whether the technological breakthroughs in ancient genomics may offer pathways towards de-extinction\textsuperscript{122}. Bringing back lost species is of growing interest, and although it is a topic fraught with challenges ranging from the ethical to the technological, for many extinct species a key starting requisite will be a well-characterized reference genome. As new extraction and computational methods expand the age range and quality of specimens from which such data can reliably be obtained, so too will the range of species that could be considered as possible targets for de-extinction attempts.
This study takes advantage of both second-generation (high-throughput, and library- and amplification-independent) and third-generation (high-throughput, and library- and amplification-independent) sequencing technologies to present the sequence history of a ringed seal characterized: that of an ~7000-year-old horse.


35. This study exploits DNA degradation patterns in HTS data sets to characterize, for the first time, genome-wide nucleosome and methylation maps from an ancient human and infer ancient gene expression levels and the age at death of the individual.


44. This paper describes a novel method for constructing deNA libraries using ssDNA templates, which provides new insights into the characterization of the Denisovan genome at a quality rivaling that of modern genomes, starting from only minute amounts of DNA extracts.


51. This study reports the first genetic analysis of ancient specimens based on a HTS technology, paving the way for whole-genome sequencing from ancient specimens.

This paper presents a fully automated pipeline performing all sequence analyses associated with recent archaeological projects, genomic inference and metagenomic profiling. It is applicable to both modern and ancient sequence data sets.


**Acknowledgements**

This work was supported by the Danish Council for Independent Research, Natural Sciences (FRU, 4002-00152B and 0602-02383B), the Danish National Research Foundation (DNRF94), the Lundbeck Foundation (852-5062), a Marie Curie Career Integration Grant (FP7-CIG-295845), and the “Chaires d’Attraitivité 2014” IDEX, University of Toulouse, France.

**Competing interests statement**

The authors declare no competing interests.

**FURTHER INFORMATION**

[AdaptorRemoval](https://github.com/lindgreen/AdapterRemoval)

**Admixture**

[https://www.genetics.ucla.edu/software/admixture/Aufhor] (https://github.com/MikkelSchubert/paleomix)

**ANFO Short Read Aligner**

[https://bioinf.eva.mpg.de/anfo/aligner.html](https://github.com/MikkelSchubert/paleomix)

Bayesian reconstruction of ancient DNA fragments:

[https://github.com/crean/FreeHop](https://github.com/crean/FreeHop)

[https://www.broadinstitute.org/](https://www.broadinstitute.org/)

**Genome Analysis Toolkit**

[https://github.com/slindgreen/](https://github.com/slindgreen/)

**mapDamage and mapDamage2**

[https://www.broadinstitute.org/](https://www.broadinstitute.org/)

**Mapping Iterative Assembler**

[https://github.com/mikkel-schubert/paleomix](https://github.com/mikkel-schubert/paleomix)

**MetaPhAn**

[https://hutterereeseq.github.io/metaphan](https://hutterereeseq.github.io/metaphan)

**PALEOMIX**

[https://github.com/MikkelSchubert/paleomix](https://github.com/MikkelSchubert/paleomix)

**PicCARD**

[https://code.google.com/p/piccard/](https://code.google.com/p/piccard/)

**SNAPE**

[https://github.com/slindgreen/SNAPE](https://github.com/slindgreen/SNAPE)

**Thermal App Web Tool**

[https://github.com/slindgreen/ThermalApp](https://github.com/slindgreen/ThermalApp)

**ALL LINKS ARE ACTIVE IN THE ONLINE PDF**