Paleogenomics

Peter D. Heintzman[*]*, André E. R. Soares*[*]*, Dan Chang, and Beth Shapiro*
Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, California 95064, USA

[*]These authors contributed equally to this study.

Keywords

Paleogenomics
The science of reconstructing and analyzing the genomes of organisms that are not alive in the present day.

Ancient DNA (aDNA)
Ancient DNA is DNA that is extracted and characterized from degraded biological specimens, including preserved bones, teeth, hair, seeds, or other tissues. Target organismal DNA is termed endogenous DNA, whereas coextracted non-target DNA is termed exogenous or contaminant DNA

Next-generation sequencing (NGS)
NGS refers to the use of various high-throughput sequencing platforms that have emerged since 2005. These technologies generate large quantities of often highly accurate DNA sequences at significantly lower costs than was possible using first-generation (Sanger) sequencing technologies. NGS technologies have been integral to the production of paleogenomes

DNA sequencing library
A sequencing library consists of fragments of DNA that have been modified, for example, by the addition of technology-specific "adapters" to each end of the DNA strand, to allow their sequencing via next-generation sequencing technologies

Polymerase chain reaction (PCR)
A technique used to create many identical copies of fragments of DNA. In paleogenomic approaches, PCR is used to increase the concentration of DNA through the amplification of targeted DNA fragments (template molecules) into multiple copies (amplicons)

Mitochondrial (mt)DNA
This genetic material is located within the mitochondria, which are organelles found in the cytoplasm of most eukaryotic cells. MtDNA is usually organized into a circular DNA molecule. MtDNA is maternally inherited in most species, and does not undergo recombination

Hominins
Hominins are the members of the subfamily Homininae, which includes the genus *Homo* and species such as modern humans (*Homo sapiens*), Neandertals (*Homo neanderthalensis*), and Denisovans

Black Death
"Black Death" is a colloquial name for the bubonic plague that devastated Europe in the mid-fourteenth century, killing an estimated 200 million people. Black Death was caused by the bacterium *Yersinia pestis*

Single-nucleotide polymorphisms (SNPs)
SNPs are positions along a DNA sequence where two haplotypes sampled from within the same species differ from each other. In protein-coding regions, a SNP can be defined as synonymous (the sequence difference does not change the encoded amino acid) or non-synonymous (the sequence difference changes the encoded amino acid)

Admixture
Admixture occurs when populations that have been reproductively isolated interbreed. Admixture can introduce new genetic variation into a population via gene flow

Paleogenomics is the science of reconstructing and analyzing the genomes of organisms that are not alive in the present day. Paleogenomic analyses can provide insights as to when and by what means traits evolved, and how extinct organisms are related to living species and populations. Paleogenomics is a relatively new field that has been made possible through advances in technologies to recover DNA sequences from preserved remains, to characterize these recovered sequences using next-generation sequencing approaches, and to assemble complete genomes from complex pools of often highly damaged, short fragments of DNA. In this chapter, an outline is provided of how paleogenomes are reconstructed, and some of the biological insights that can be gained from studying the genomes of dead organisms are showcased.

1 Reconstructing Paleogenomes

1.1 Ancient DNA

1.1.1 The Nature of Ancient DNA

Ancient DNA (aDNA) is the degraded genetic material from deceased organisms that is preserved either in the remains of these organisms or in the environment. Unlike DNA from living organisms, aDNA is often reduced to short fragments (typically fewer than 100 base pairs (bp)) and damaged by hydrolytic and oxidative processes (see Sect. 1.1.2). Due to ongoing degradation, aDNA is not expected to survive for more than one million years, even under ideal conditions [1]. Ideal conditions for long-term DNA survival include preservation in low-temperature and low-humidity environments. The majority of aDNA has been retrieved from permanently frozen (e.g., arctic soil) and cave environments, and the oldest aDNA recovered as yet dates to between 700 and 800 thousand years ago [2, 3].

In addition to becoming degraded over time, aDNA is often coextracted with other DNA, which complicates aDNA experiments. For example, DNA from the burial environment – including plant, fungal and microbial DNA – will colonize the remains post mortem, and this coextracted fraction of exogenous DNA can be very high compared to the endogenous (host) fraction. For example, the first Neandertal paleogenome was constructed from DNA with an endogenous content of less than 5% Neandertal DNA [4]. In exceptional

circumstances, however, the endogenous content can be more than 70% [5–7] and even as high as 95% [8]. Another source of non-endogenous DNA in aDNA experiments is contamination [9], which occurs when DNA molecules from the present-day environment are introduced into the ancient sample at some stage of the experimental process.

A key challenge in aDNA research is to distinguish endogenous DNA from both exogenous (and sometimes also ancient) DNA and contaminating modern DNA. Both contamination by present-day DNA and the coextraction of exogenous DNA can lead to an erroneous interpretation of results, the most notorious of which include claims of "antediluvian" DNA from fossilized bone and amber [10–12] that were later found to be the result of contamination, or simply were irreproducible [13–15]. As a consequence of these and other early and erroneous claims, strict protocols have been developed for working with aDNA, including performing research only in sterilized environments [9, 16].

1.1.2 The Common Effects of DNA Damage

Post-mortem DNA decay involves the accumulation of both physical and chemical damage to the DNA molecule, including single- and double-strand breaks, chemical alterations to the nucleotides that make up the DNA strand, and the formation of blocking and miscoding lesions [17–19]. Blocking lesions, such as positions along the DNA sequence at which a nucleotide is missing, inhibit amplification and therefore reduce the likelihood that damaged molecules are recovered from an ancient specimen. Miscoding lesions, alternately, can be amplified using the polymerase chain reaction (PCR), but will generate an incorrect sequence. The most abundant miscoding lesion in aDNA data is the hydrolytic deamination of cytosine to uracil [17, 20–22]. Deaminated cytosines are copied as thymine (C → T) during PCR, as the latter base is analogous to uracil. However, if the deaminated cytosine is on the opposite DNA strand, a complemented version of this change is observed, with guanine appearing as adenine (G → A). Deaminated cytosines are abundant at the ends of aDNA molecules due to single-stranded overhangs [20], and their presence has been used to preferentially select for endogenous aDNA from within pools of sequenced DNA fragments using bioinformatic approaches [23, 24]. Alternatively, damaged fragments may be removed from the pool of available extracted DNA molecules, for example, via the use of uracil-DNA glycosylase (UDG), which hydrolyzes uracils so that they cannot be copied by the polymerase in PCR.

1.2 The Recovery and Sequencing of Ancient DNA

In reconstructing a paleogenome, the first steps involve the isolation and sequencing of aDNA. This requires that aDNA be extracted, prepared for sequencing (library preparation), enriched (capture-based target enrichment) if necessary, and, finally, sequenced.

1.2.1 Ancient DNA Extraction

A variety of methods have been developed to extract aDNA [25]. Most of these require that samples are first broken down (lysed), which is achieved in different ways depending on what type of tissue is to be processed. The lysis of bone cells most often takes place in a strong solution of ethylenediaminetetraacetic acid (EDTA) and proteinase K, breaking down hydroxyapatite and collagen, respectively [26, 27]. The lysis of hair cells often takes

places in a stabilized buffer with sodium dodecyl sulfate (SDS), dithiothreitol (DTT), and proteinase K, which will disintegrate keratin [28, 29]. For plant tissues, a lysis buffer containing cetyltrimethylammonium bromide (CTAB) and polyvinylpyrrolidone (PVP) can be used to break down and remove PCR-inhibiting substances, such as polysaccharides and polyphenols, respectively [30].

In either lysing scenario, the resulting solution is then purified using phenol–chloroform-based phase separation [31, 32], silica spin columns [26, 27, 32, 33], or a combination of the two. Commonly, the extracted DNA is stored in a Tris–HCl and EDTA (TE) buffer, with the optional addition of the detergent Tween-20 to ensure long-term extract viability. The majority of published paleogenomes have been constructed using DNA extracted by the silica column-based method first described by Rohland *et al.* [27], although a more recent method developed by Dabney *et al.* [26] appears to be more efficient at retaining shorter DNA fragments (<50 bp) and is therefore well suited to extracting DNA from very degraded samples.

1.2.2 DNA Library Preparation

Before the fragments of aDNA that have been extracted can be sequenced, each fragment needs to be modified slightly so that they are recognized by the sequencing platform. Modification usually means the addition of platform-specific sequences (adapters) to the ends of each DNA molecule; this process is termed "library preparation". Although platform-specific kits can be purchased and used for library preparation, protocols have been developed that are tailored specifically to the types of challenges (e.g., damage) that are common with aDNA. The most common method for preparing aDNA for sequencing using the Illumina sequencing platform is that developed by Meyer and Kircher [34] (Fig. 1a). In this protocol, the ends of each fragment of DNA are first made double-stranded by the removal or fill-in of single-stranded overhangs; this process is termed blunt-ending. Next, short Y-shaped adapters are ligated onto the ends of the blunt-ended DNA molecules, after which the single-stranded ends of the adapters are filled in to make each fragment double-stranded. Finally, PCR is used to amplify the adapter-ligated DNA molecules, using a pair of single-stranded DNA sequences (oligonucleotides) that bind to the adapter sequence. This "indexing PCR" adds both the platform-specific adapter and a sample-specific index to each DNA fragment in the library (see Sect. 1.2.4). The sequence-specific index makes it possible for multiple libraries to be pooled for sequencing in the same sequencing run.

One disadvantage of the Meyer and Kircher method is that, while many fragments of surviving aDNA may be single-stranded, only double-stranded fragments of DNA become part of the sequencing library. However, a recently developed library preparation method circumvents this limitation by incorporating both single-stranded and double-stranded aDNA into the resulting libraries [35] (Fig. 1b). In this new method, heat is first applied to denature any double-stranded DNA molecules so that they become single-stranded. A biotinylated single-stranded adapter is then joined to one end of each DNA molecule, which becomes anchored to streptavidin-coated beads via the now-attached biotin. An oligonucleotide is then attached to the single-stranded adapter, after which the single-stranded DNA molecule is filled in and becomes double-stranded. A short Y-shaped adapter

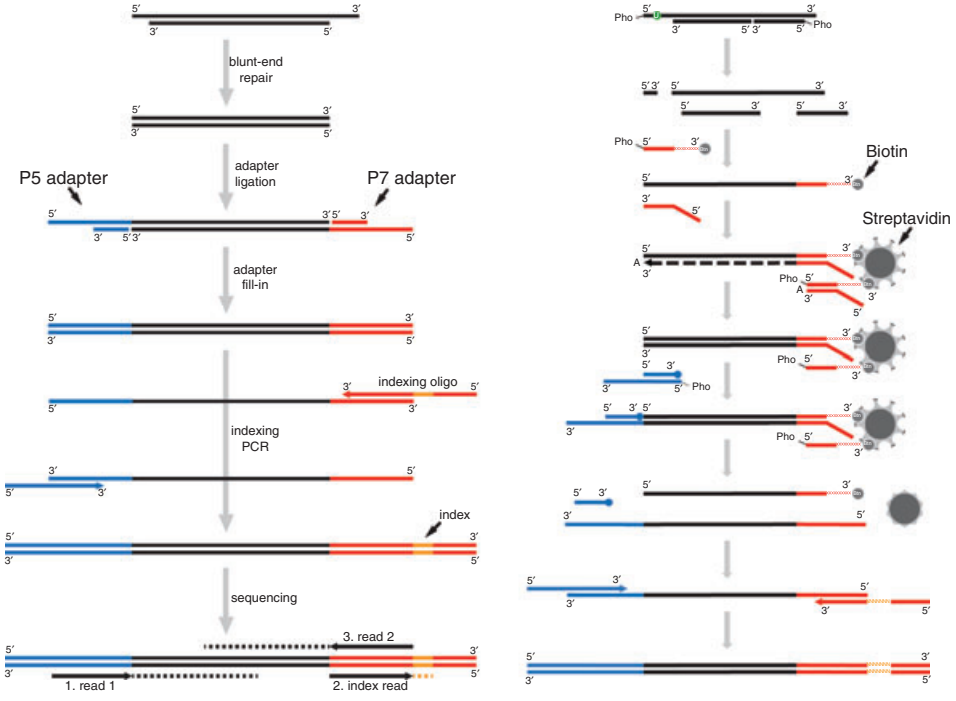

Fig. 1 Extracted DNA is prepared into a sequencing library. (a) Double-stranded and b) single-stranded library preparation, as described by Meyer & Kircher [34] and Gansauge & Meyer [35], respectively. (b) was adapted by permission from Macmillan Publishers Ltd: Nature Protocols [35], copyright 2013.

is then joined to the opposite end of the DNA molecule. Heat is used to release the anchored biotin from the streptavidin and also produce single-stranded DNA through denaturation. This single-stranded DNA, with adapters attached to both ends, can then undergo indexing PCR as described above. The single-stranded library preparation method has been shown to increase the amount of DNA that can be sequenced by an order of magnitude compared to the double-stranded method, and was integral to the production of the first high-coverage paleo-hominin genomes [36, 37].

1.2.3 Capture-Based Target Enrichment

In some instances, for example, when sequencing complete paleogenomes or targeting specific loci for population-level analysis, it may be useful to increase the proportion of endogenous DNA in an aDNA library prior to sequencing, which can reduce the amount – and therefore cost – of sequencing required. Fortunately, capture-based target enrichment methods have been developed to enrich either for total endogenous DNA (whole-genome enrichment; WGE) or for specific genomic regions of interest [38–42]. Target capture involves the isolation of

DNA molecules through binding to target-specific single-stranded DNA or RNA bait molecules (hybridization). Once the target DNA molecules are bound to the bait, any remaining unbound off-target DNA molecules are removed. Bait molecules may either be anchored to an array (array-based), or biotinylated (solution-based). In the case of biotinylated baits, the latter are bound to streptavidin-coated magnetic beads, which can then be immobilized in the presence of a magnet. The bait molecules, with the DNA of interest attached, are then washed to remove the unbound off-target DNA molecules and therefore enrich for DNA sequences of interest. The enriched DNA sequences are PCR-amplified to a concentration suitable for sequencing.

Companies such as MYcroarray (Ann Arbor, MI, USA) offer WGE kits specifically designed for aDNA, which include the target-specific bait molecules. If performing a large number of target capture reactions, however, it may be economically viable to self-manufacture bait molecules in the laboratory, as illustrated by the whole genome in-solution capture (WISC) method of Carpenter *et al.* [40] (Fig. 2). Using WISC, Carpenter *et al.* reported an up to 159-fold increase in the endogenous DNA content of 12 ancient human samples [40].

1.2.4 Quantification of aDNA and Sequencing

To ensure a successful sequencing run and/or capture-based target enrichment experiment, it is useful to assess the quality of aDNA libraries and the quantity of DNA recovered prior to sequencing. This involves assessing the distribution of the length of recovered sequence fragments, ensuring that adapter dimers are kept to a minimum, and calculating the concentration of DNA in the library. The first two indicators can be evaluated using gel-based systems such as a Bioanalyzer (Agilent, CA, USA), whereas library concentration may be calculated via quantitative PCR (qPCR) or Qubit (Life Technologies, CA, USA). The high sensitivity and accuracy of these methods make them ideal for quantifying aDNA. If the presence of adapter dimers is problematic, these can be selectively removed, for example, by using a MagNA bead cleanup protocol [43].

The distribution of the length of recovered DNA fragments in a sequencing library is important to determine how many nucleotides per molecule to sequence (sequencing cycles), as well as approximating DNA survival (shorter molecules being indicative of greater degradation). For example, if determining the number of sequencing cycles for a DNA library with a mean insert length within the adapters of 60 bp, a 75-cycle run would sequence the entire length of the DNA insert and the first 15 nucleotides of the adapter (Fig. 1a). This situation is acceptable, as the adapter sequence can be removed bioinformatically during data processing (see Sect. 1.3.1). However, were a 150-cycle run to be conducted, the DNA insert and full adapter length ($\sim$ 65 bp) would be sequenced and the final $\sim$ 25 bp would comprise noise. In this scenario, there is the possibility that the sequencing run may fail completely.

To reduce costs, it is often desirable to combine multiple DNA libraries into a single pool to be sequenced simultaneously. The introduction of an index (also referred to as a barcode) during library preparation (see Sect. 1.2.2) allows pooled library molecules to be identified and assigned to their correct sample (see also Sect. 1.3.1). A DNA barcode is a unique sequence that consists of six to eight nucleotides and is built into one of the adapters (Fig. 1). An increased accuracy of index-based assignment can be achieved through double-indexing, in which an additional

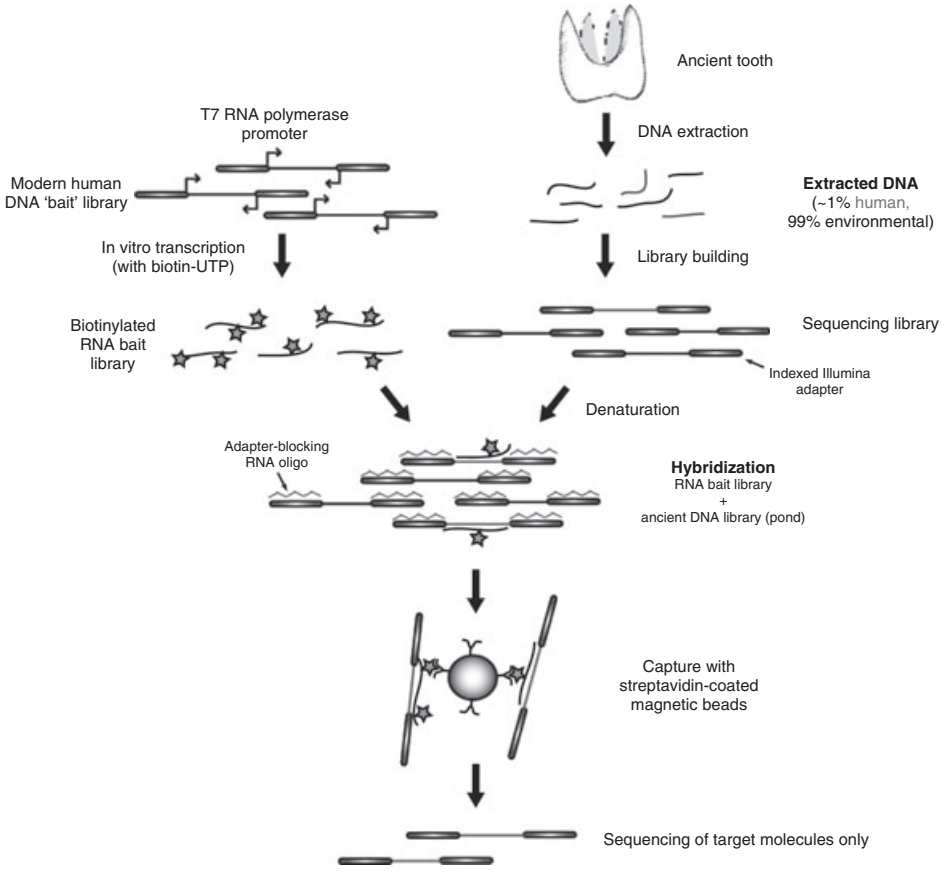

Fig. 2 A pipeline for generating ancient DNA molecules for sequencing and paleogenomic analysis. DNA is extracted, converted into a library, and enriched using self-manufactured bait molecules. Reprinted with permission from Ref. [40]; © 2013, Elsevier, The American Society of Human Genetics.

index is incorporated into the second adapter [44].

Sequencing can either be single- or paired-end. Single-end sequencing refers to sequencing the DNA in only one direction, reading the molecule from one end towards the middle. In paired-end sequencing, the DNA fragment is sequenced in both directions, starting from each end independently, and this results in two reads per fragment (Fig. 1a). Paired-end sequencing is often used when sequencing aDNA, because the short length of the fragment means that the entire read is often sequenced twice – once in each direction – providing higher-quality reads in the overlapping portion of the sequence.

1.3 Assembly and Analysis of a Paleogenome

The assembly of a paleogenome requires that sequence data be prepared for mapping, be mapped to a reference genome, and undergo a series of quality checks.

Tab. 1 An example bioinformatic pipeline for the assembly and analysis of a paleogenome from Illumina paired-end read data.

Major step	*Specific step*	*Example programs available*
Filtering	Paired end read merging	SeqPrep[a]; FLASH [50][b]
Filtering	Adapter trimming	SeqPrep[a]; Trimmomatic [51]; FASTX clipper[b] [c]
Filtering	Barcode splitting	FASTX barcode splitter[c]
Filtering	Remove low complexity reads	PRINSEQ [52]
Filtering	Initial duplicate removal	PRINSEQ [52]; FASTX collapser[c]; FastUniq [53]
Mapping	Alignment of reads	BWA [54]; Bowtie2 [55][b]
Mapping	Quality filtering	SAMtools [56]
Mapping	Further duplicate removal	SAMtools [56]; Picard[b] [d]
QC	Depth of coverage	GATK [57]; BEDTools [58]
QC	DNA damage	mapDamage2.0 [59][b]
QC	Contamination estimate	N/A
Analysis	Demographic history	PSMC [60]; Eigensoft[e]
Analysis	Admixture	ADMIXTURE [61]; TreeMix [62]; ADMIXTOOLS [63]
Analysis	Phylogenetic inference	BEAST [64]; RAxML [65][b]
Analysis	Evolutionary rate calibration	BEAST [64]
Analysis	Selection scan	N/A
Analysis	Phenotypic inference	IGV [66]; VCFtools [67]; SNPedia [68]

The list of programs available is not exhaustive and further experiment-specific steps may be required.
N/A: Not applicable – specialist scripts are generally used for these analyses. QC: quality control, FLASH: Fast Length Adjustment of SHort reads, BWA: Burrows-Wheeler Alignment tool, PSMC: Pairwise Sequentially Markovian Coalescent, GATK: Genome Analysis Toolkit, BEAST: Bayesian Evolutionary Analysis Sampling Trees, and IGV: Integrative Genomics Viewer.

a) Available from *https://github.com/jstjohn/SeqPrep*
b) Steps included in the PALEOMIX pipeline [46].
c) Available from *http://hannonlab.cshl.edu/fastx_toolkit/*
d) Available from *http://broadinstitute.github.io/picard/*
e) Available from *http://www.hsph.harvard.edu/alkes-price/software/*

1.3.1 Initial Data Processing

A typical pipeline for preparing aDNA sequence data for mapping to a reference genome consists of merging paired-end reads, trimming adapter sequences, removing low-complexity and short reads, and subdividing the pooled data into sample-specific data sets based on index [45, 46] (Table 1). Paired-end reads can be merged if the paired reads are sufficiently long so that they overlap each other (see Sect. 1.2.4). Merging reads reduces sequencing errors, allows for a more robust detection (and removal) of adapters, and also improves the mappability of the reads by extending the read length [45]. If the fragments of aDNA are shorter than the length of the sequencing read, adapters may be sequenced on both ends of the aDNA fragment. These adapter sequences need to be removed prior to mapping, but this can be automated on some sequencing machines. As with any sequencing experiment, some of the resulting data will comprise low-complexity reads, which occur both as sequencing artifacts and when low-complexity regions of the genome are sequenced. Low-complexity reads should be removed from the data set, as they have an increased likelihood of being erroneously mapped. Commonly,

reads that are shorter than ~30–35 bp long are also removed from aDNA data sets prior to mapping, as these may not be possible to map with high confidence to a unique region of the genome [36, 47]. If multiple libraries are pooled for sequencing (see Sect. 1.2.4), these can be disambiguated according to their index sequence prior to mapping [45] (see Sect. 1.2.2).

1.3.2 Mapping and Quality Control

Because aDNA tends to be highly fragmented, paleogenomes are assembled by mapping short reads to evolutionarily close and previously assembled reference genomes, rather than using overlapping short reads to create an assembly in the absence of a reference genome (*de novo* assembly). Software has been developed specifically for the reference-guided assembly of aDNA, for example, by considering explicitly the expectation that deaminated bases are likely to occur near the ends of sequence reads [47]. Parameters to run the available software can be optimized for aDNA; for example, disabling seed-based mapping strategies can compensate somewhat for challenges of mapping very short fragments and often highly damaged fragments to a reference genome [48]. Because most aDNA extracts contain only small amounts of DNA, aDNA sequencing libraries are often amplified via PCR prior to sequencing so as to generate sufficient amounts of DNA to sequence. This results in high rates of sequence duplication, which needs to be corrected by the removal of duplicates after read mapping [45, 46].

Mismatches or indels (insertions–deletions), caused by evolutionary divergence from the reference genome, and DNA damage can reduce read mappability. For this reason, it is desirable to use a reference genome that evolutionarily is as close as possible to the ancient organisms being sequenced. This challenge is an important limitation in paleogenomics (Table 2), but one that is lessening as the taxonomic diversity of published genomes increases and bioinformatic approaches improve.

Tab. 2 The diversity of published paleogenomes, as of September 2014.[a)]

Organism	*Reference(s)*
Ancient humans/hominins	
Aboriginal Australian	[69]
Native American (North, South)	[40, 70–73]
Egyptian	[73]
European	[6, 40, 74–78]
Siberian	[79]
Neandertal	[4, 37]
Denisovan	[7, 36]
Epigenomes	[80, 81]
Human pathogens	
Tuberculosis	[82]
Leprosy	[83]
Bubonic plague	[84]
Animals	
Horse	[2]
Pig	[85]
Myotragulus	[86]
Bear (polar, brown)	[5, 87]
Mammoth	[88, 89]
Plants and plant pathogens	
Cotton	[8]
Barley stripe mosaic virus	[90]
Irish potato famine	[91, 92]

[a)] List is expanded from Table 1 in Ref. [49].

Because of the myriad challenges of assembling paleogenomes, it is important that assembled paleogenomes undergo basic quality control prior to analysis. This is particularly important when assembling human or pathogen paleogenomes, as these organisms represent common sources of contamination. The quality control of paleogenomes includes calculating average coverage across the genome and estimating rates of DNA damage and contamination [36, 69–71, 74, 79].

Average coverage is usually expressed as "$N\times$", where N refers to the average number of times any particular base in the genome appears in an aligned read. Average coverage is calculated by summing the total length of all reads in the alignment and dividing this sum by the total length of the reference genome. For example, if the total length of all reads in an alignment is 9×10^9 bp and the reference genome is 3×10^9 bp, then each base is covered three times on average, and this would be expressed as a coverage depth of 3×. Coverage statistics are useful as a quality control measure as they allow the recognition of regions of the genome that are either dramatically over- or under-covered, which in turn can alert research teams to problematic (misassembled or low-complexity) regions of the genome.

Rates of DNA damage are calculated by observing base mismatches between reads and the reference genome. If the sequences are of ancient origin, there should be a noticeable increase in the rate of C → T and G → A transitions at the 5′ and 3′ ends of reads, respectively (see Sect. 1.1.2). If this pattern is not observed, the data may not be authentically ancient. Programs have been specifically designed to detect and quantify these damage patterns in aDNA [59]. Contamination rates can be estimated by calculating the rate of heterozygosity at homozygous regions of the genome, including the mitochondrion and the X and Y chromosomes (if the individual is male). In female individuals, the rate of contamination from males can be inferred from reads that align to the Y chromosome [36]. Rates of DNA damage and contamination are useful measures of the authenticity of the recovered data, and therefore the reliability of the assembled paleogenome.

1.3.3 Paleogenomic Analysis

Paleogenomes can be used to ask a wide variety of biological questions, including questions regarding demography [5, 36, 37], admixture (interbreeding) and gene flow [36, 37, 70, 79, 85], the relationships between ancient and present-day individuals [36, 37, 70, 75, 76, 84, 85, 90], evolutionary rates [2], the direction of natural selection [2, 4, 37], how genotype translates to phenotype [4, 6, 36, 37, 71, 74], and how genes are regulated [80, 81] (Table 1). Some of these topics are explored in detail in the next section.

2 Case Studies

2.1 Hominin Paleogenomics

2.1.1 The Origin of Hominin Paleogenetics

One of the first topics to be addressed using paleogenomic techniques was the evolution of the human species, *Homo sapiens*, and its closest evolutionary relatives. This research has improved significantly what is known about humans and about other members of the human family tree. Paleogenomes isolated from the remains of a variety of ancient hominins have revealed patterns of demographic change over time [36], elucidated phylogenetic relationships between humans and their closest evolutionary relatives [38, 93], and revealed ancient admixture events between hominin lineages [4, 7, 37].

In many ways, research that has been focused on clarifying human evolutionary history has driven the key developments in aDNA technologies. In 1997, for example, Krings *et al.* [93] performed one of the first aDNA experiments whereby DNA was extracted from a Neandertal-type

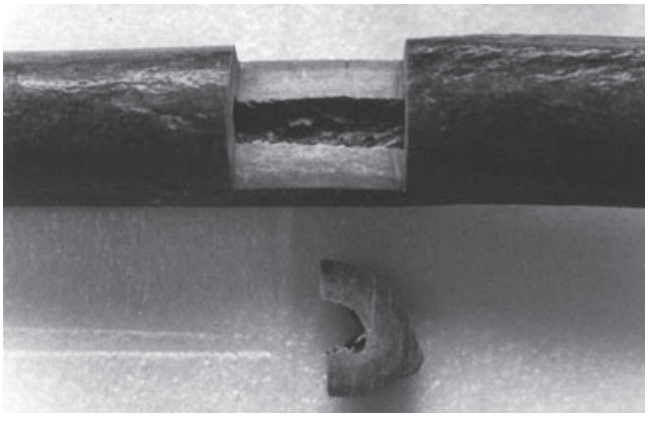

Fig. 3 The right humerus of the Neandertal type specimen used by Krings *et al.* [93] for ancient DNA extraction. Reprinted with permission from Ref. [93]; © 1997, Elsevier.

specimen, a fossil excavated from Feldhofer Cave, Germany (Fig. 3). Using two sets of oligonucleotides, Krings and coworkers PCR-amplified a 105 bp section of the mitochondrial genome, of which 30 amplicons were cloned and sequenced. Of these 30 sequences, 27 fell outside of the mitochondrial variation of modern humans, and the authors concluded that these were fragments of Neandertal mitochondrial DNA. Although this dataset was small, the results established the notion that Neandertals did not contribute genes to modern humans, and also opened the door to the direct observation of genetic differences between extinct hominins and modern humans.

The approach used by Krings and colleagues in this early aDNA study (DNA extraction, PCR amplification, molecular cloning, sequencing) was straightforward, and had the advantage of targeting specific regions of interest. Within these target regions, known diagnostic single-nucleotide polymorphisms (SNPs) can be used to distinguish between authentic endogenous DNA and contaminating DNA. However, one drawback of this approach is that it requires the targeting of specific genomic regions (usually via PCR), and this may not be feasible when the organism in question is highly diverged from any living organism. Further, aDNA molecules are often degraded into very short fragments, many of which are too short to be amplifiable via PCR. The latter process is also time-consuming and expensive, particularly when large segments of the genome are targeted for analysis.

2.1.2 Entering the Paleogenomics Era

During the years following the pioneering studies of Krings *et al.* [93], aDNA sequences were obtained from a variety of Neandertal individuals [94–100]. However, all of these investigations had been conducted using targeted amplification via PCR, and the insights gained had remained limited to a few small regions of the genome. In 2006, Green *et al.* [101] moved the field of aDNA into the era of paleogenomics with the announcement that they had generated more than one million base-pairs of Neandertal DNA. This achievement was made possible due to the development of what is now known as next-generation sequencing (NGS) technologies.

NGS dramatically increased the amount of data that could be recovered from ancient specimens. For example, shortly after Green *et al.* reported their one million base-pairs of Neandertal DNA, the first complete mitochondrial genome of a Neandertal was assembled [47]. The bone from which this genome was recovered, as well as the bones from which five other complete mitochondrial genomes were soon recovered [38], contained very small fractions of endogenous DNA, with between 0.01% and 1.5% of extracted DNA of Neandertal origin. However, because of the much higher volume of sequence data obtained through NGS, this low endogenous content was no longer a barrier. Analyses of complete mitochondrial genomes revealed, for example, that Neandertal populations had only one-third of the mitochondrial genetic diversity of modern humans, most likely due to a smaller population size [36].

Neandertals were not the only hominin remains to be analyzed during the early days of NGS. At about the same time that the first complete Neandertal mitochondrial genome were published, a second research group described a complete mitochondrial genome from an archaic modern human. In the latter case, Gilbert *et al.* [102] isolated the mitochondrial genome from the hair of a Saqqaq paleo-eskimo individual from Greenland that had been frozen for about 4000 years. Remarkably, this sample contained approximately 80% endogenous DNA [79], which may have been attributable to its relatively young age, preservation at very low temperature (see Sect. 1.1.1), or that the sample was hair rather than bone (keratinous samples are known to be less likely contaminated and more easily cleaned of potential surface contaminants than other sources of DNA) [29, 103, 104]. On analyzing this mitochondrial genome, Gilbert *et al.* concluded that this individual belonged to the mitochondrial haplogroup D2a1, which is observed in the modern Siberian Yuit people but not in modern Inuits or Native Americans. This result suggested that the Saqqaq people were descended from human populations that had moved to Greenland from the Bering Sea, and were not related to the Inuit populations that currently inhabit Greenland. These insights helped to elucidate the origins of the earliest human expansion into the most northern regions of the New World.

2.1.3 The First Complete Human Paleogenome

In 2010, Rasmussen *et al.* [71] published details of the first human paleogenome, which they had isolated from the remarkably well-preserved Saqqaq paleo-eskimo hair sample described above. With an average coverage depth of 20×, these authors were able to recover 79% of the genome, identifying 350 000 SNPs. Analyses of these SNPs allowed Rasmussen *et al.* to corroborate the mitochondrial relationship of the Saqqaq individual to modern human populations, to identify his blood type (A+), and to estimate his eye color, based on a variant of the HERC2-OCA2 locus that is strongly associated with brown eyes. They also identified 12 SNPs that have been associated with adaptation to a cold climate by influencing metabolism and body mass index.

2.1.4 Human and Neandertal Paleogenomics

Since this first human paleogenome, several additional complete genomes from both archaic hominins and the close evolutionary relatives of present-day humans have been reported [36, 37, 69, 74, 76, 77, 79], some of which have been sequenced to very

high coverage [36, 37, 77]. Possibly the most surprising conclusion of these studies has been the discovery that all living humans with ancestry outside of sub-Saharan Africa have a small portion of their nuclear genomes that is derived from ancient admixture with Neandertals [37, 105–108]. This conclusion is counter to the results of mitochondrial DNA analyses, which show that Neandertal sequences cluster outside of the mitochondrial diversity of modern humans [93, 97] (see Sect. 2.1.1).

Recent population genomic studies have attempted to reveal the evolutionary consequences of this admixture. The publication of high-coverage paleogenomes from two archaic hominins – a Neandertal and a Denisovan – confirmed that admixture had taken place between archaic hominins and the ancestors of modern humans, most likely after the dispersal of modern non-African humans out of Africa [36, 37] (Fig. 4). Genomic admixture from Denisovans seems to be particularly prominent in Papuans and Micronesians [36], whereas admixture from Neandertals has been detected in modern Eurasians [4], with Asian populations generally exhibiting a greater amount of Neandertal ancestry than European populations [105]. Admixture introduced novel haplotypes into ancestral Eurasian human populations, some of which were driven to high frequencies. For example, haplotype frequencies of archaic hominin-derived HLA haplotypes are higher than expected in both modern Asian and African populations, given the genome-wide average percentage of introgression. This suggests that, following admixture into ancient Asian populations, the archaic HLA haplotypes were driven to high frequency in these populations and later spread to African populations [107]. Similarly, haplotype N of the *STAT2* gene is at a higher than expected frequency in non-Africans, and is evolutionarily similar to the Neandertal haplotype [109]. Another interesting example is the *EPAS1* gene, which is associated with adaptation to living at high altitudes [110]. The most commonly observed allele of the *EPAS1* gene in modern Tibetans exhibits a high sequence similarity with the allele isolated from the Denisovan paleogenome; this allele is not found in other human populations except for two Han Chinese individuals, and the

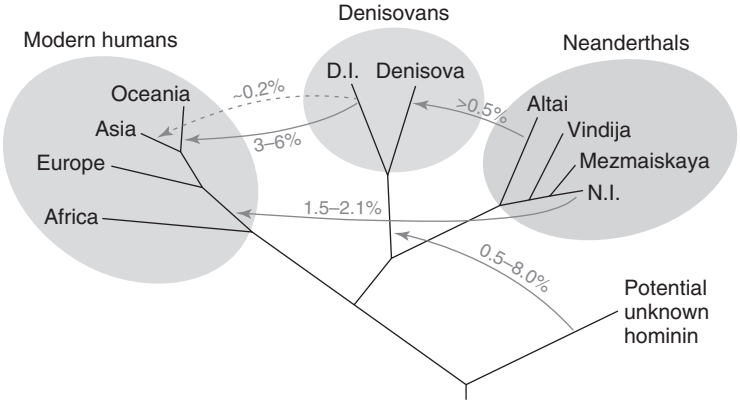

Fig. 4 A model of gene flow between archaic hominins and modern humans, as proposed by Prüfer *et al*. [37]. The arrows indicate the direction and magnitude of the gene flow events. Dashed lines indicate uncertainty. Reprinted with permission from Ref. [37]; © 2013, Macmillan Publishers Ltd.

high frequency of this allele in Tibetans is not likely to have resulted from *de novo* mutations. This suggests that, after admixture, the Denisovan allele of the *EPAS1* gene was driven to high frequency in Tibetans, perhaps as an adaptation to high-altitude hypoxia [110]. It is very likely that further analyses will discover other adaptations in modern human populations that are derived from admixture with humankind's archaic cousins.

2.2 The Black Death and Insights into Ancient Plagues

The desire to understand human evolution drove many of the technical advances during the early years of aDNA research. However, aDNA analyses have not been limited to studying human evolution, and aDNA techniques have been used to study the evolutionary histories of a wide variety of organisms over the past 30 years. One interesting and notable example is the series of studies that have focused on uncovering the causative agent and evolutionary history of the Black Death (bubonic plague), which caused the death of one-third of the European population during the mid-fourteenth century.

In 2010, Haensch *et al.* [111] identified *Yersinia pestis* DNA from human mass-grave remains in northern and southern Europe, resolving a longstanding debate about the causative agent of the Black Death. *Y. pestis* is an anaerobic bacterium that can infect both human and animals and is believed to have originated from *Y. pseudotuberculosis* between 1500 and 20 000 years ago [112]. Haensch *et al.* [111] sampled the teeth and bone from 76 plague victims that had been excavated from putative fourteenth- to seventeenth-century plague pits in England, France, Germany, Italy, and the Netherlands. The remains were prepared for analysis in a dedicated clean facility. For each excavated specimen, the outer surface was ultraviolet-irradiated and sandblasted to destroy and remove any surface contaminants. In order to identify any bacterial matter preserved within these samples, the team attempted to amplify (via PCR) a *Y. pestis*-specific gene, *pla*, which was located on the multicopy plasmid pPst. The resulting amplified fragments were shown to be highly similar to a modern strain (CO92) of *Y. pestis.*

Next, the team's aim was to identify which strain of *Y. pestis* was responsible for Black Death. Modern strains of *Y. pestis* have been divided into three biovars, *Medievalis*, *Orientalis*, and *Antiqua*, which can be identified based on their ability to ferment glycerol and/or reduce nitrate [113]. The biovars are also genetically distinct from one another, providing a means to classify the ancient specimens using recovered aDNA. Interestingly, typing of the ancient specimens led Haensch *et al.* to identify two new clades of *Y. pestis*, both of which were ancestral to the *Orientalis* and *Medievalis* biovars. The authors concluded that at least two strains of *Y. pestis* had spread through Europe during the epidemic between 1347 and 1750 AD (Fig. 5).

After identifying and genotyping *Y. pestis*, the next step towards an in-depth understanding of the dynamics of the Black Death epidemic was to reconstruct the *Y. pestis* paleogenome [84]. The sequencing of ancient diseases is historically and epidemiologically important, as it helps to shed light on the mechanisms of pathogen adaptation and infection dynamics, but has been a considerable challenge for the field of aDNA because ancient pathogens are preserved at very low abundance in ancient remains. In 2011, Bos *et al.* [84] extracted DNA from the dental pulp of four individuals that had

Fig. 5 Spread of the Black Death in Europe according to the infections routes per year. Reprinted with permission from Ref. [84]; © 2011, Macmillan Publishers Ltd.

been buried at the East Smithfield cemetery in London which, as historical documents indicate, was created for the burial of plague victims around 1348 AD. Interestingly, paleogenomic analyses indicated that the strains of *Y. pestis* recovered from these individuals were very similar, but not identical, to modern strains of *Y. pestis.* The authors concluded that the high virulence of the disease circulating during the Black Death epidemic may not have been due to differences in the bacterial genotype, but instead to some environmental or host susceptibility factor [84].

2.3 Reconstruction and Selection of Ancient Phenotypes

Most paleogenomic studies have focused on estimating the evolutionary history of a population or species. However, paleogenomes can also be used to explore the link between genotype and phenotype, despite the fact that phenotypes are mostly not preserved in the fossil record. Genotyping samples that span a broad temporal range provides a means of inferring which genetic mutations may be associated with physical, physiological or behavioral changes, such as those that occur in response to natural or artificial selection.

One example of the resurrection and analysis of an ancient phenotype that has been made possible with paleogenomic technologies is the adaptation of woolly mammoths to an arctic environment. The common ancestor of elephants and woolly mammoths is believed to have been tropically adapted and, for this reason, it should be possible to detect adaptations to a cold climate by searching within the woolly mammoth (*Mammuthus primigenius*) genome. To this end, Campbell *et al.* [114] amplified and sequenced two

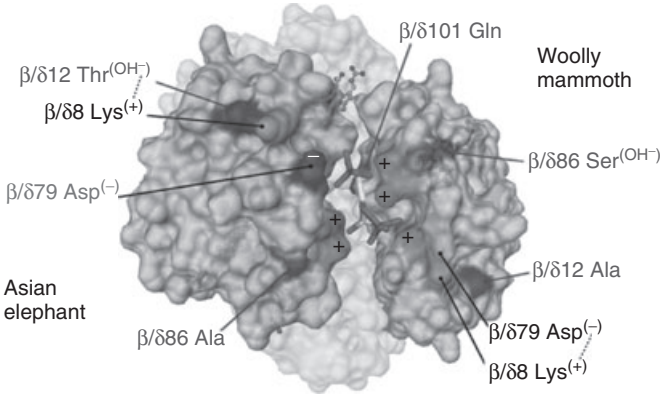

Fig. 6 Surface model of a deoxyhemoglobin molecule bound to 2,3-bisphosphoglycerate (BPG). The left side is modeled after the Asian elephant, and the right side after the woolly mammoth. Specific amino acid substitutions are highlighted. Reprinted with permission from Ref. [114]; © 2010, Macmillan Publishers Ltd.

hemoglobin genes, the *HBA-T2* gene of the α-chain globin and the *HBB/HBD* gene of the β/δ-chain fusion, from 43 000-year-old woolly mammoth remains. The team discovered three amino acid replacements that had occurred within the *HBB/HBD* gene along the lineage leading to woolly mammoths (Fig. 6). To test whether these changes would have been adaptive, the team synthesized mammoth-specific hemoglobin and compared it to a synthesized version of elephant hemoglobin in a functional assay. Subsequently, it was found that two of the changes reduced the effect of temperature variation on the oxygen-carrying potential of mammoth hemoglobin. The temperature-insensitivity of mammoth hemoglobin would likely have facilitated the transmission of oxygen to exposed extremities (such as limbs and digits) in cold climates, thereby providing an adaptive advantage to mammoths in their arctic environment.

In another example of reconstructing an ancient phenotype using paleogenomic technologies, D'Costa *et al.* identified the presence of genes associated with antibiotic-resistance (AB-R) in naturally occurring, 30 000-year-old bacteria [115]. These results indicated that AB-R had persisted in the environment for many thousands of years prior to the widespread use of antibiotics. In order to demonstrate the link between genotype and phenotype, D'Costa *et al.* reconstructed the proteins encoded by the recovered sequences and confirmed their AB-R potential [115]. The results showed that AB-R was not likely to have originated from the widespread use of antibiotics within the past century, although the overuse of antibiotics during this time – in both humans and animals – has led to a proliferation in naturally occurring AB-R, causing the current problems associated with AB-R [116].

Phenotypic changes associated with domestication can also be assessed using paleogenomics. In animals, for example, coat color is often a target for artificial selection during domestication, and genes encoding coat color variation have been revealed by several comparative genomic studies of wild and domesticated animals [117–119]. Ludwig *et al.* [120] genotyped 89

ancient horses, representing pre- and post-domestication horses from the Eurasian steppe and Iberian Peninsula, for six genes that control coat color and spotting. The authors examined changes in the allelic composition of these genes through time and among locations, and concluded that selection for coat color variation, transitioning from bay and black in pre-domestic horses to a greater variety of coat coloration and spotting in domesticated horses, had occurred prior to the end of the Copper Age in Siberia and East Europe [120].

One of the best-studied genes with known phenotypic consequences in paleogenomics is *MC1R*, which is associated with hair pigmentation. Analyses of the *MC1R* gene of wooly mammoths revealed that these animals were probably a mixture of dark and light colors, and that the difference between these phenotypes may have resulted from a single non-synonymous substitution [121, 122]. Reconstructions of *MC1R* in Neandertals identified mutations that impaired the activity of this gene, likely resulting in pale skin and red hair [123].

To date, most links between genotype and phenotype in paleogenomes have focused on single genes, or clusters of genes whose function has been established by the functional analysis of living organisms. As the number and taxonomic diversity of sequenced paleogenomes grows, this is an area of research that is likely to grow rapidly, providing new and more detailed discoveries regarding the relationship between genome sequence and the appearance, physiology and behavior of ancient organisms.

2.4 Epigenetics of Ancient Species

In addition to analyzing the sequence of paleogenomes, it was shown recently that it might be possible to infer an ancient epigenome directly from aDNA sequence data, thanks in part to the way that DNA degrades over time. Although the function, heritability and malleability of the epigenome is far from understood, it is clear that a variety of cellular mechanisms are regulated epigenetically, such as genomic imprinting and transposition [124, 125]. One of the main systems for epigenetic regulation is DNA methylation, whereby the epigenome modifies the genome by attaching a methyl group (CH_3) to a cytosine. When cytosine is degraded in aDNA, it loses an amine group and becomes uracil; however, when methylated cytosine bases are deaminated the interaction between the two chemical modifications converts the cytosine into thymine rather than uracil. An ancient epigenome can be reconstructed by distinguishing deaminated cytosine bases that become thymine from those that become uracil.

In 2014, Pederson *et al.* [80] took advantage of this observation to generate the first paleo-human nucleosome map (how DNA is packaged in a chromosome) for the Saqqaq individual (see Sect. 2.1). Gokhman *et al.* [81] expanded upon this study, aiming to fully reconstruct the DNA methylation maps of Neandertals and Denisovans. To achieve this, they designed an experiment whereby, using the different properties of polymerase enzymes, they could build sequencing libraries capable of differentiating between methylated and unmethylated cytosines. The results showed that the methylation patterns of Neandertals, Denisovans and modern humans do not differ significantly, with the exception of the *HOXD* gene cluster that is responsible for limb development [126]. Gokhman *et al.* suggested that these differences in methylation may partly explain

the morphological differences between Neandertals and humans.

Patterns of methylation in paleogenomes can also be used to infer how an ancient individual responded to its environment. In some plants, for example, a genomic methylation process mediated by short interfering RNA (siRNA) occurs as a response to environmental stress, such as viral infections. An increase in the proportion of methylated cytosines at certain loci has therefore been proposed as potential evidence for viral infection in these ancient plants [127]. In other studies, ancient RNA has itself been targeted, with results showing that it is not only retrievable but also has potential for further functional investigation [90, 128].

3 Conclusions

Paleogenomics is a relatively young field, but is expanding rapidly in terms of taxonomic breadth and research foci. Today, methodological advancements make it possible to extract DNA from a greater variety of substrates, and improvements in sequencing technologies continue to increase the amount of data that can be generated for the same cost. New methods aim to increase the efficiency of DNA recovery from specimens that are preserved under less-than-ideal conditions, which will allow the recovery of older paleogenomes from a geographically and taxonomically more diverse sample of organisms. Taken together, these technical advances have propelled the field of aDNA towards the generation and analysis of complete paleogenomes, rather than single genes. Such analyses will continue to provide new and more comprehensive insights into how species, populations and entire ecosystems evolve over time.

Acknowledgments

The authors thank Sam Vohr and James Cahill for suggestions that enhanced Table 1. P.D.H., D.C., and B.S. were supported by grants from the Packard Foundation and Gordon and Betty Moore Foundation. A.E.R.S. was supported by the Capes – Science without Borders Program.

References

1. Lindahl, T. (1993) Instability and decay of the primary structure of DNA. *Nature*, **362** (6422), 709–715.
2. Orlando, L., Ginolhac, A., Zhang, G., Froese, D. *et al.* (2013) Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, **499** (7456), 74–78.
3. Willerslev, E., Cappellini, E., Boomsma, W., Nielsen, R. *et al.* (2007) Ancient biomolecules from deep ice cores reveal a forested Southern Greenland. *Science*, **317** (5834), 111–114.
4. Green, R.E., Krause, J., Briggs, A.W., Maricic, T. *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328** (5979), 710–722.
5. Cahill, J.A., Green, R.E., Fulton, T.L., Stiller, M. *et al.* (2013) Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. *PLos Genet.*, **9** (3), e1003345.
6. Keller, A., Graefen, A., Ball, M., Matzas, M. *et al.* (2012) New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.*, **3**, 698.
7. Reich, D., Green, R.E., Kircher, M., Krause, J. *et al.* (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468** (7327), 1053–1060.
8. Palmer, S.A., Clapham, A.J., Rose, P., Freitas, F.O. *et al.* (2012) Archaeogenomic

evidence of punctuated genome evolution in *Gossypium*. *Mol. Biol. Evol.*, **29** (8), 2031–2038.
9. Cooper, A. and Poinar, H.N. (2000) Ancient DNA: do it right or not at all. *Science*, **289** (5482), 1139.
10. Cano, R.J., Poinar, H.N., Pieniazek, N.J., Acra, A. *et al.* (1993) Amplification and sequencing of DNA from a 120–135-million-year-old weevil. *Nature*, **363** (6429), 536–538.
11. Desalle, R., Gatesy, J., Wheeler, W., and Grimaldi, D. (1992) DNA-sequences from a fossil termite in oligomiocene amber and their phylogenetic implications. *Science*, **257** (5078), 1933–1936.
12. Woodward, S.R., Weyand, N.J., and Bunnell, M. (1994) DNA sequence from Cretaceous period bone fragments. *Science*, **266** (5188), 1229–1232.
13. Austin, J.J., Ross, A.J., Smith, A.B., Fortey, R.A. *et al.* (1997) Problems of reproducibility – does geologically ancient DNA survive in amber-preserved insects? *Proc. R. Soc. London, Ser. B: Biol. Sci.*, **264** (1381), 467–474.
14. Gutierrez, G. and Marin, A. (1998) The most ancient DNA recovered from an amber-preserved specimen may not be as ancient as it seems. *Mol. Biol. Evol.*, **15** (7), 926–929.
15. Hebsgaard, M.B., Phillips, M.J., and Willerslev, E. (2005) Geologically ancient DNA: fact or artefact? *Trends Microbiol.*, **13** (5), 212–220.
16. Fulton, T.L. (2012) Setting up an Ancient DNA laboratory, in *Ancient DNA: Methods and Protocols*, Methods in Molecular Biology (eds B. Shapiro and M. Hofreiter), Humana Press, Springer, New York, pp. 1–11.
17. Dabney, J., Meyer, M., and Paabo, S. (2013) Ancient DNA damage. *Cold Spring Harbor Perspect. Biol.*, **5** (7), a012567.
18. Gilbert, M.T.P., Binladen, J., Miller, W., Wiuf, C. *et al.* (2007) Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis. *Nucleic Acids Res.*, **35** (1), 1–10.
19. Gilbert, M.T.P., Hansen, A.J., Willerslev, E., Rudbeck, L. *et al.* (2003) Characterization of genetic miscoding lesions caused by postmortem damage. *Am. J. Hum. Genet.*, **72** (1), 48–61.
20. Briggs, A.W., Stenzel, U., Johnson, P.L., Green, R.E. *et al.* (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. USA*, **104** (37), 14616–14621.
21. Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A. *et al.* (2001) DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.*, **29** (23), 4793–4799.
22. Stiller, M., Green, R.E., Ronan, M., Simons, J.F. *et al.* (2006) Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. *Proc. Natl Acad. Sci. USA*, **103** (37), 13578–13584.
23. Skoglund, P., Northoff, B.H., Shunkov, M.V., Derevianko, A.P. *et al.* (2014) Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl Acad. Sci. USA*, **111** (6), 2229–2234.
24. Meyer, M., Fu, Q., Aximu-Petri, A., Glocke, I. *et al.* (2014) A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature*, **505** (7483), 403–406.
25. Shapiro, B. and Hofreiter, M. (eds) (2012) *Ancient DNA: Methods and Protocols*, Methods in Molecular Biology, Humana Press, Springer, New York.
26. Dabney, J., Knapp, M., Glocke, I., Gansauge, M.T. *et al.* (2013) Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl Acad. Sci. USA*, **110** (39), 15758–15763.
27. Rohland, N., Siedel, H., and Hofreiter, M. (2010) A rapid column-based ancient DNA extraction method for increased sample throughput. *Mol. Ecol. Resour.*, **10** (4), 677–683.
28. Campos, P.F. and Gilbert, T.M.P. (2012) DNA extraction from keratin and chitin, in *Ancient DNA: Methods and Protocols*, Methods in Molecular Biology (eds B. Shapiro and M. Hofreiter), Humana Press, Springer, New York, pp. 43–49.
29. Gilbert, M.T.P., Wilson, A.S., Bunce, M., Hansen, A.J. *et al.* (2004) Ancient mitochondrial DNA from hair. *Curr. Biol.*, **14** (12), R463–R464.
30. Kistler, L. (2012) Ancient DNA extraction from plants, in *Ancient DNA: Methods and*

Protocols, Methods in Molecular Biology (eds B. Shapiro and M. Hofreiter), Humana Press, Springer, New York, pp. 71–79.
31. Barnett, R. and Larson, G. (2012) A phenol–chloroform protocol for extracting DNA from ancient samples, in *Ancient DNA: Methods and Protocols*, Methods in Molecular Biology (eds B. Shapiro and M. Hofreiter), Humana Press, Springer, New York, pp. 13–19.
32. Rohland, N. and Hofreiter, M. (2007) Comparison and optimization of ancient DNA extraction. *Biotechniques*, **42** (3), 343–352.
33. Yang, D.Y., Eng, B., Waye, J.S., Dudar, J.C. *et al.* (1998) Technical note: improved DNA extraction from ancient bones using silica-based spin columns. *Am. J. Phys. Anthropol.*, **105** (4), 539–543.
34. Meyer, M. and Kircher, M. (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protoc.*, **2010** (6), pdb prot 5448.
35. Gansauge, M.T. and Meyer, M. (2013) Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.*, **8** (4), 737–748.
36. Meyer, M., Kircher, M., Gansauge, M.T., Li, H. *et al.* (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science*, **338** (6104), 222–226.
37. Prufer, K., Racimo, F., Patterson, N., Jay, F. *et al.* (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, **505** (7481), 43–49.
38. Briggs, A.W., Good, J.M., Green, R.E., Krause, J. *et al.* (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*, **325** (5938), 318–321.
39. Burbano, H.A., Hodges, E., Green, R.E., Briggs, A.W. *et al.* (2010) Targeted investigation of the Neandertal genome by array-based sequence capture. *Science*, **328** (5979), 723–725.
40. Carpenter, M.L., Buenrostro, J.D., Valdiosera, C., Schroeder, H. *et al.* (2013) Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.*, **93** (5), 852–864.
41. Maricic, T., Whitten, M., and Paabo, S. (2010) Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One*, **5** (11), e14004.
42. Fu, Q., Meyer, M., Gao, X., Stenzel, U. *et al.* (2013) DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl Acad. Sci. USA*, **110** (6), 2223–2227.
43. Rohland, N. and Reich, D. (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.*, **22** (5), 939–946.
44. Kircher, M., Sawyer, S., and Meyer, M. (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.*, **40** (1), e3.
45. Kircher, M. (2012) Analysis of high-throughput ancient DNA sequencing data, in *Ancient DNA: Methods and Protocols*, Methods in Molecular Biology (eds B. Shapiro and M. Hofreiter), Humana Press, Springer, New York, pp. 197–228.
46. Schubert, M., Ermini, L., Der Sarkissian, C., Jonsson, H. *et al.* (2014) Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.*, **9** (5), 1056–1082.
47. Green, R.E., Malaspinas, A.S., Krause, J., Briggs, A.W. *et al.* (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, **134** (3), 416–426.
48. Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J.F. *et al.* (2012) Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, **13**, 178.
49. Shapiro, B. and Hofreiter, M. (2014) A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science*, **343** (6169), 1236573.
50. Magoc, T. and Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27** (21), 2957–2963.
51. Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30** (15), 2114–2120.
52. Schmieder, R. and Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27** (6), 863–864.
53. Xu, H., Luo, X., Qian, J., Pang, X. *et al.* (2012) FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One*, **7** (12), e52249.

54. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25** (14), 1754–1760.

55. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9** (4), 357–359.

56. Li, H., Handsaker, B., Wysoker, A., Fennell, T. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25** (16), 2078–2079.

57. McKenna, A., Hanna, M., Banks, E., Sivachenko, A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20** (9), 1297–1303.

58. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26** (6), 841–842.

59. Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P.L. *et al*. (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, **29** (13), 1682–1684.

60. Li, H. and Durbin, R. (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475** (7357), 493–496.

61. Alexander, D.H. and Lange, K. (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinf.*, **12**, 246.

62. Pickrell, J.K. and Pritchard, J.K. (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLos Genet.*, **8** (11), e1002967.

63. Patterson, N., Moorjani, P., Luo, Y., Mallick, S. *et al.* (2012) Ancient admixture in human history. *Genetics*, **192** (3), 1065–1093.

64. Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, **29** (8), 1969–1973.

65. Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30** (9), 1312–1313.

66. Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings Bioinf.*, **14** (2), 178–192.

67. Danecek, P., Auton, A., Abecasis, G., Albers, C.A. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27** (15), 2156–2158.

68. Cariaso, M., Lennon, G. (2012) SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res.*, **40** (Database issue), D1308–D1312.

69. Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E. *et al.* (2011) An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science*, **334** (6052), 94–98.

70. Rasmussen, M., Anzick, S.L., Waters, M.R., Skoglund, P. *et al.* (2014) The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*, **506** (7487), 225–229.

71. Rasmussen, M., Li, Y.R., Lindgreen, S., Pedersen, J.S. *et al.* (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, **463** (7282), 757–762.

72. Raghavan, M., DeGiorgio, M., Albrechtsen, A., Moltke, I. *et al.* (2014) The genetic prehistory of the New World Arctic. *Science*, **345** (6200), 1255832.

73. Khairat, R., Ball, M., Chang, C.C., Bianucci, R. *et al.* (2013) First insights into the metagenome of Egyptian mummies using next-generation sequencing. *J. Appl. Genet.*, **54** (3), 309–325.

74. Olalde, I., Allentoft, M.E., Sanchez-Quinto, F., Santpere, G. *et al.* (2014) Derived immune and ancestral pigmentation alleles in a 7000-year-old Mesolithic European. *Nature*, **507** (7491), 225–228.

75. Sanchez-Quinto, F., Schroeder, H., Ramirez, O., Avila-Arcos, M.C. *et al.* (2012) Genomic affinities of two 7000-year-old Iberian hunter-gatherers. *Curr. Biol.*, **22** (16), 1494–1499.

76. Skoglund, P., Malmstrom, H., Raghavan, M., Stora, J. *et al.* (2012) Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science*, **336** (6080), 466–469.

77. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G. *et al.* (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, **513** (7518), 409–413.

78. Skoglund, P., Malmstrom, H., Omrak, A., Raghavan, M. *et al.* (2014) Genomic

diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science*, **344** (6185), 747–750.

79. Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M. *et al.* (2014) Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, **505** (7481), 87–91.
80. Pedersen, J.S., Valen, E., Velazquez, A.M., Parker, B.J. *et al.* (2014) Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res.*, **24** (3), 454–466.
81. Gokhman, D., Lavi, E., Prufer, K., Fraga, M.F. *et al.* (2014) Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science*, **344** (6183), 523–527.
82. Bos, K.I., Harkins, K.M., Herbig, A., Coscolla, M. *et al.* (2014) Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, **514**, 494.
83. Schuenemann, V.J., Singh, P., Mendum, T.A., Krause-Kyora, B. *et al.* (2013) Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science*, **341** (6142), 179–183.
84. Bos, K.I., Schuenemann, V.J., Golding, G.B., Burbano, H.A. *et al.* (2011) A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature*, **478** (7370), 506–510.
85. Ramirez, O., Burgos-Paz, W., Casas, E., Ballester, M. *et al.* (2014) Genome data from a sixteenth century pig illuminate modern breed relationships. *Heredity*, **131**, 46–52.
86. Ramirez, O., Gigli, E., Bover, P., Alcover, J.A. *et al.* (2009) Paleogenomics in a temperate environment: shotgun sequencing from an extinct Mediterranean caprine. *PLoS One*, **4** (5), e5670.
87. Miller, W., Schuster, S.C., Welch, A.J., Ratan, A. *et al.* (2012) Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proc. Natl Acad. Sci. USA*, **109** (36), E2382–E2390.
88. Miller, W., Drautz, D.I., Ratan, A., Pusey, B. *et al.* (2008) Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, **456** (7220), 387–390.
89. Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B. *et al.* (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, **311** (5759), 392–394.
90. Smith, O., Clapham, A., Rose, P., Liu, Y. *et al.* (2014) A complete ancient RNA genome: identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. *Sci. Rep.*, **4**, 4003.
91. Martin, M.D., Cappellini, E., Samaniego, J.A., Zepeda, M.L. *et al.* (2013) Reconstructing genome evolution in historic samples of the Irish potato famine pathogen. *Nat. Commun.*, **4**, 2172.
92. Yoshida, K., Schuenemann, V.J., Cano, L.M., Pais, M. *et al.* (2013) The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *Elife*, **2**, e00731.
93. Krings, M., Stone, A., Schmitz, R.W., Krainitzki, H. *et al.* (1997) Neandertal DNA sequences and the origin of modern humans. *Cell*, **90** (1), 19–30.
94. Orlando, L., Darlu, P., Toussaint, M., Bonjean, D. *et al.* (2006) Revisiting Neandertal diversity with a 100 000 year old mtDNA sequence. *Curr. Biol.*, **16** (11), R400–R402.
95. Lalueza-Fox, C., Sampietro, M.L., Caramelli, D., Puder, Y. *et al.* (2005) Neandertal evolutionary genetics: mitochondrial DNA data from the Iberian Peninsula. *Mol. Biol. Evol.*, **22** (4), 1077–1081.
96. Ovchinnikov, I.V., Gotherstrom, A., Romanova, G.P., Kharitonov, V.M. *et al.* (2000) Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature*, **404** (6777), 490–493.
97. Serre, D., Langaney, A., Chech, M., Teschler-Nicola, M. *et al.* (2004) No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biol.*, **2** (3), e57.
98. Krings, M., Capelli, C., Tschentscher, F., Geisert, H. *et al.* (2000) A view of Neandertal genetic diversity. *Nat. Genet.*, **26** (2), 144–146.
99. Caramelli, D., Lalueza-Fox, C., Condemi, S., Longo, L. *et al.* (2006) A highly divergent mtDNA sequence in a Neandertal individual from Italy. *Curr. Biol.*, **16** (16), R630–R632.
100. Schmitz, R.W., Serre, D., Bonani, G., Feine, S. *et al.* (2002) The Neandertal type site revisited: interdisciplinary investigations of

skeletal remains from the Neander Valley, Germany. *Proc. Natl Acad. Sci. USA*, **99** (20), 13342–13347.

101. Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W. *et al.* (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature*, **444** (7117), 330–336.

102. Gilbert, M.T.P., Kivisild, T., Gronnow, B., Andersen, P.K. *et al.* (2008) Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland. *Science*, **320** (5884), 1787–1789.

103. Gilbert, M.T.P., Tomsho, L.P., Rendulic, S., Packard, M. *et al.* (2007) Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science*, **317** (5846), 1927–1930.

104. Willerslev, E., Gilbert, M.T.P., Binladen, J., Ho, S.Y. *et al.* (2009) Analysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. *BMC Evol. Biol.*, **9**, 95.

105. Wall, J.D., Yang, M.A., Jay, F., Kim, S.K. *et al.* (2013) Higher levels of Neanderthal ancestry in East Asians than in Europeans. *Genetics*, **194** (1), 199–209.

106. Plagnol, V. and Wall, J.D. (2006) Possible ancestral structure in human populations. *PLos Genet.*, **2** (7), e105.

107. Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A. *et al.* (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science*, **334** (6052), 89–94.

108. Lachance, J., Vernot, B., Elbers, C.C., Ferwerda, B. *et al.* (2012) Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell*, **150** (3), 457–469.

109. Mendez, F.L., Watkins, J.C., and Hammer, M.F. (2012) A haplotype at STAT2 introgressed from Neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am. J. Hum. Genet.*, **91** (2), 265–274.

110. Huerta-Sánchez, E., Jin, X., Asan Bianba, Z., Peter, B.M. *et al.* (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, **512**, 194–197.

111. Haensch, S., Bianucci, R., Signoli, M., Rajerison, M. *et al.* (2010) Distinct clones of *Yersinia pestis* caused the Black Death. *PLoS Pathog.*, **6** (10), e1001134.

112. Achtman, M., Zurth, K., Morelli, G., Torrea, G. *et al.* (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc. Natl Acad. Sci. USA*, **96** (24), 14043–14048.

113. Perry, R.D. and Fetherston, J.D. (1997) *Yersinia pestis* – etiologic agent of plague. *Clin. Microbiol. Rev.*, **10** (1), 35–66.

114. Campbell, K.L., Roberts, J.E., Watson, L.N., Stetefeld, J. *et al.* (2010) Substitutions in woolly mammoth hemoglobin confer biochemical properties adaptive for cold tolerance. *Nat. Genet.*, **42** (6), 536–540.

115. D'Costa, V.M., King, C.E., Kalan, L., Morar, M. *et al.* (2011) Antibiotic resistance is ancient. *Nature*, **477** (7365), 457–461.

116. Huttner, A., Harbarth, S., Carlet, J., Cosgrove, S. *et al.* (2013) Antimicrobial resistance: a global view from the 2013 World Healthcare-Associated Infections Forum. *Antimicrob. Resist. Infect. Control*, **2**, 31.

117. Qanbari, S., Pausch, H., Jansen, S., Somel, M. *et al.* (2014) Classic selective sweeps revealed by massive sequencing in cattle. *PLos Genet.*, **10** (2), e1004148.

118. Andersson, L. and Georges, M. (2004) Domestic-animal genomics: deciphering the genetics of complex traits. *Nat. Rev. Genet.*, **5** (3), 202–212.

119. Wiener, P. and Wilkinson, S. (2011) Deciphering the genetic basis of animal domestication. *Proc. R. Soc. B: Biol. Sci.*, **278** (1722), 3161–3170.

120. Ludwig, A., Pruvost, M., Reissmann, M., Benecke, N. *et al.* (2009) Coat color variation at the beginning of horse domestication. *Science*, **324** (5926), 485.

121. Workman, C., Dalén, L., Vartanyan, S., Shapiro, B. *et al.* (2011) Population-level genotyping of coat colour polymorphism in woolly mammoth (*Mammuthus primigenius*). *Quarternary Sci. Rev.*, **30** (17-18), 2304–2308.

122. Römpler, H., Rohland, N., Lalueza-Fox, C., Willerslev, E. *et al.* (2006) Nuclear gene indicates coat-color polymorphism in mammoths. *Science*, **313** (5783), 62.

123. Lalueza-Fox, C., Römpler, H., Caramelli, D., Stäubert, C. *et al.* (2007) A melanocortin 1 receptor allele suggests varying

pigmentation among Neanderthals. *Science*, **318** (5855), 1453–1455.

124. Hollister, J.D. and Gaut, B.S. (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.*, **19** (8), 1419–1428.

125. Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16** (1), 6–21.

126. Davis, A.P., Witte, D.P., Hsieh-Li, H.M., Potter, S.S. *et al*. (1995) Absence of radius and ulna in mice lacking hoxa-11 and hoxd-11. *Nature*, **375** (6534), 791–795.

127. Smith, O., Clapham, A.J., Rose, P., Liu, Y. *et al*. (2014) Genomic methylation patterns in archaeological barley show de-methylation as a time-dependent diagenetic process. *Sci. Rep.*, **4**, 5559.

128. Fordyce, S.L., Avila-Arcos, M.C., Rasmussen, M., Cappellini, E. *et al*. (2013) Deep sequencing of RNA from ancient maize kernels. *PLoS One*, **8** (1), e50961.

