Title: Toward high-resolution population genomics using archaeological samples

Running title: Progress in genetic analysis of ancient DNA

Irina Morozova[1,2#], Pavel Flegontov[3,4], Alexander S. Mikheyev[5], Hosseinali Asgharian[6], Petr Ponomarenko[6,7], Vladimir Klyuchnikov[8], GaneshPrasad ArunKumar[9], Sergey Bruskin[2], Egor Prokhortchouk[10,11], Yuriy Gankin[12], Evgeny Rogaev[2,13], Yuri Nikolsky[2,14,15], Ancha Baranova[15,16], Eran Elhaik[17], Tatiana V. Tatarinova[6,4,7#]

#corresponding authors: TT tatiana.tatarinova@usc.edu and IM irina.morozova@iem.uzh.ch

1. Institute of Evolutionary Medicine, University of Zurich, Switzerland
2. Vavilov Institute of General Genetics RAS, Moscow, Russia
3. Department of Biology and Ecology, Faculty of Science, University of Ostrava, Czech Republic
4. A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russian Federation
5. Ecology and Evolution Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan
6. University of Southern California, Los Angeles, CA, USA
7. Children's Hospital Los Angeles, Los Angeles, CA, USA
8. Donskaya Archeologia, Rostov, Russia
9. School of Chemical and Biotechnology, SASTRA University, Tanjore, India
10. Research Center of Biotechnology RAS, Moscow, Russia
11. Department of Biology, Lomonosov Moscow State University, Russia
12. EPAM Systems, Newtown, PA, USA
13. University of Massachusetts Medical School, Worcester MA USA
14. F1 Genomics, San Diego CA, USA
15. School of Systems Biology, George Mason University, VA, USA
16. Research Centre for Medical Genetics, Moscow, Russia
17. University of Sheffield, Sheffield, South Yorkshire, UK

1

2

Abstract

The term "ancient DNA" (aDNA) is coming of age, with over 1200 hits in the PubMed database, beginning in the early 1980s with the studies of "molecular paleontology." Rooted in cloning and limited sequencing of DNA from ancient remains during the pre-PCR era, the field has made incredible progress since the introduction of PCR and next-generation sequencing. Over the last decade, aDNA analysis ushered in a new era in genomics and became the method of choice for reconstructing the history of organisms, their biogeography, and migration routes, with applications in evolutionary biology, population genetics, archaeo-genetics, paleo-epidemiology, and many other areas. This change was brought by development of new strategies for coping with the challenges in studying aDNA due to damage and fragmentation, scarce samples, significant historical gaps, and limited applicability of population genetics methods. In this review, we describe the state-of-the-art achievements in aDNA studies, with particular focus on human evolution and demographic history. We present the current experimental and theoretical procedures for handling and analyzing highly degraded aDNA. We also review the challenges in the rapidly growing field of ancient epigenomics. Advancement of ancient DNA tools and methods signifies a new era in population genetics and evolutionary medicine research.

Keywords: ancient DNA, bioinformatics, epigenetics, population genetics, next generation sequencing

Ancient DNA as an indispensable source of information

The passion to unravel and reconstruct the history of life on Earth has always stimulated research in evolutionary biology. Although inferences of past events such as the states of ancestral organisms (e.g., ancestral sequences), evolutionary episodes (e.g., speciation)[1], and the dynamics governing change (e.g., mutation models) can be obtained through computational phylogenetic and coalescent approaches using contemporary data, naturalists have always valued direct observation above all other methods. Ancient DNA (aDNA) is thus expected to revolutionize evolutionary genetics in the same manner that the fossil record revolutionized paleontology: it is a direct window into the past —a "time capsule". Ancient DNA has already been invaluable in addressing many key questions in evolutionary biology[1-14], frequently providing the only available evidence.

Studies over the past several decades have demonstrated that aDNA can survive and be extracted from ancient and historical material (e.g., bones, teeth, eggshells; mummified, frozen, or artificially preserved tissues; plant remains). The first attempts to extract and analyze aDNA were performed before the PCR era. In a pioneer study in 1984, Higuchi et al.[15] managed to recover DNA using bacterial cloning from dried quagga muscle, an extinct subspecies of plains zebra (*Equus quagga*). However, due to extremely poor DNA preservation, analyses of aDNA were limited until an effective technology for DNA amplification, like PCR, made very small amounts of DNA accessible for study. In addition, next-generation sequencing (NGS) technologies and the plummeting cost of DNA sequencing have provided an unprecedented opportunity to perform millions of sequencing reactions in parallel. These advances enabled the first report on ancient sequences retrieved by NGS in 2005[16].

Human evolution and demographic history

Over the past decade, genomic techniques have been reshaping our fundamental understanding of human history and origins[17,18]. Until recently, much of what was known about history came from the study of archeological sites and anthropological investigations, piecing together patterns of human migration and admixture from physical features, pottery,

weapons, ornaments, art production, traditional customs, and studies of modern DNA[19-26]. Other sources of information included linguistic classifications and ancient texts. Although undeniably powerful, these approaches often yielded more questions than answers, and their resolution required incorporation of additional data. Analysis of ancient human remains can reveal migration patterns[10-13,27], address questions of kinship and family structure[1], and provide insight into physiological or morphological characteristics such as blood group, skin color, hair type[28-31], and climatic adaptation[1]. When combined with other evidence, sequencing ancient genomes could help settle important debates within archaeology or linguistics. This approach, although not infallible, is particularly valuable now when ancient genetics is considered to be highly robust tool and has significantly impacted many fields such as forensics and history[32].

Sequencing of the genomes from archaic hominids have illuminated earlier events in human evolution and suggested that early hominids had a richer evolutionary history than was previously appreciated. Analysis of Neanderthal genomes extracted from remains found in Europe and Western and Central Asia and dated 230–30 thousand years ago (kya) demonstrated that, contrary to previous suggestions, Neanderthals may have interbred with *Homo sapiens*[2,8,33-39]. The studies showed that Neanderthals share more genetic material with modern humans across Eurasia than those from sub-Saharan Africa, indicating that genetic flow from Neanderthals to Eurasian *Homo sapiens* likely occurred after the emergence of humans from Africa but before the divergence of Eurasian groups[2,8]. Additional gene flow events may have occurred later in Europe[38] and East Asia[35,40]. Mitochondrial DNA (mtDNA) sequences of morphologically ambiguous Neanderthal bones from Teshik-Tash cave in Uzbekistan and Okladnikov cave in Southern Siberia provided evidence that Neanderthals had an extensive range prior to their extinction[41]. Since the first description of the Neanderthal genome, a number of studies have suggested that various Neanderthal alleles have been preferentially retained in modern populations due to specific selective pressures[42].

Genetic analysis of mtDNA from a phalanx dated 48–30 kya recovered from the Denisova cave in Southern Siberia revealed another hominid, named Denisovan, which is genetically distinct from Neanderthals and modern humans[43]. Since then, only two more samples (molars) of Denisovans have been discovered[3]. Comparison of Neanderthal and Denisovan genomes suggested that their population histories were independent of each other. The Denisovan mtDNA represents a deep branch, with the Neanderthal mtDNA closer to that of modern humans[3]. Comparative analysis of the Denisovan and modern human genomes revealed that the genetic contribution from Denisovans to modern humans may have been restricted to Melanesia and Australia with hybridization events taking place mostly on the Southeast Asian mainland, although they may have permeated to Oceania[3,4,43-47] as recently suggested by the existence of a widespread, low-level signal of Denisovan ancestry across Eastern Eurasian and Native American populations[48]. However, the methodology used in these studies has been criticized, thus calling into question the ancient introgression of Denisovan DNA on the Asian mainland[49,50].

NGS analysis of a nearly complete mitochondrial genome of a hominid found in Sima de los Huesos cave in Atapuerca, Spain and dated to more than 300 kya[51] suggested existence of another branch in the human evolutionary tree. Surprisingly, the Sima de los Huesos mtDNA forms a clade with the mitochondrial genome of Denisovans rather than that of Neanderthals, demonstrating an unexpected link between Denisovans and Middle Pleistocene European hominids. Recently, approximately three million bases of nuclear sequences were obtained from a Sima de los Huesos femur fragment, an incisor, and a molar[52]. In contrast to the mtDNA, the nuclear genomic sequences of Sima de los Huesos are significantly

more similar to Neanderthals than to Denisovans[52]. These results agree with previous morphological analyses[53,54] but present an archaeological puzzle.

Studies of aDNA have also delineated human migration routes around the world, particularly in Europe. Analysis of human genomes from Europe and Siberia dated 24–5 kya[7,9,10,14,28,29] revealed at least three different sources of the population diversity of modern Europeans, i.e., West European hunter-gatherers, ancient North Eurasians with high similarity to Upper Paleolithic Siberians, and early European farmers originating mainly in the Near East but with limited West European hunter-gatherer ancestry[7].

Further aDNA studies allowed mapping migration in Europe in greater detail. A recent study of 69 European individuals who lived 8–3 kya[12] demonstrated that between 8–5 kya populations of Western and Eastern Europe were genetically distinct. Groups of early farmers of Middle Eastern origin arrived to Western Europe and mixed with local hunter-gatherers, whereas Eastern Europe at that time was inhabited by a distant branch of ancient North Eurasian hunter-gatherers[10]. However, Eastern Europe did not remain a "hunter-gatherer's refuge" for too long. Around 6–5 kya, farming populations of West Anatolian ancestry appeared in Eastern Europe and mixed with local hunter-gatherers in the Pontic-Caspian region, giving rise to pastoralist people of the extremely successful Yamnaya archaeological culture. Such multiethnic melting pots were fertile ground for many innovations, such as horse domestication and wheeled vehicles from the Yamnaya culture[55], which probably enabled massive migration or invasions into Western and Northern Europe approximately 4.5 kya, introducing their ancestry, languages, and customs to central Europeans. Haak and colleagues reported that this steppe ancestry persisted in central Europeans from at least 3 kya, and it is ubiquitous in present-day Europeans[12]. At approximately the same time, similar migrations spread Yamnaya-related cultures into South Siberia and Central Asia, as revealed by another large-scale study of 101 genomes from Eurasian Bronze Age (5-3 kya) burial sites[11]. Such large-scale aDNA studies[7,11,12] have not only made technical breakthroughs but also had significant interdisciplinary effects: they influenced the decades-long debate in archaeology and linguistics about the origin of Indo-European language speakers and shed light on perennial questions about the prevalence of traits like skin color and lactose intolerance in modern Europeans.

With the progress in aDNA sequencing technology, considerable studies of remains from all over the world have begun to emerge[56-58], shedding light onto, for example, settlement of China and the Pacific islands. Recently, Liu and colleagues reported the sequencing of an 80,000-year-old man (the earliest modern human in southern China) raising major questions about the "*out of Africa"* canonical paradigm, since there is no evidence that humans entered Europe before 45 kya[59]. A recent study of a 4,500-year-old Ethiopian skeleton preserved in relatively cool mountainous conditions was the first example of successful aDNA analysis in Africa[60], giving hope to forthcoming studies of this incredibly interesting region.

Pending future advances in functional genomics, aDNA might prove an unrivaled source of information on the evolution of traits associated with cognitive phenotypes. For example, discovering the genetic variants responsible for language acquisition may allow researchers to pinpoint the origin of complex language in the human lineage, indubitably a cornerstone event in human evolution. Approximately a decade ago the Neanderthals were found to bear a modern human version of *FOXP2*[61] (gene responsible for the ability to speak[62]). The authors suggested that the modern variant of *FOXP2* were presented in common ancestor of Neanderthals and modern humans[61].

We can also expect aDNA genomic studies to provide direct evidence about human adaptation substantiating the genetic basis of selection. For example, a genome-wide scan of 230 West Eurasians who lived 6.5–1 kya and their comparison with modern human genomes identified significant signatures of selection in a range of loci related to diet (lactase persistence, fatty acid metabolism, vitamin D levels, and some diet-associated diseases), pathogen resistance, and externally visible phenotypes (skin and eye pigmentation, tooth morphology, hair thickness, and body height).[63] This work demonstrated the utility of aDNA data in human adaptive evolution studies. The currently available set of published human aDNA NGS data, including sample IDs, dating, archaeological cultures, site names and locations, haplogroups, references, and links to data repositories, is accessible online at our website (http://chcb.saban-chla.usc.edu/ANDA_WEB/Ancient.xlsx) and illustrated in Figure 1.

Historic patterns in the spread of infectious diseases

Some devastating pandemics, like the Black Death, remain infamous even centuries after these catastrophes. Ancient DNA enables discovery of the origin and spread of disease-carrying alleles to aid modern epidemiology. Such analyses are possible when genotypes of ancient humans are recovered along with the genomes of their pathogens. For example, Rasmussen and colleagues sequenced DNA extracted from ancient human teeth and found that *Yersinia pestis*, the etiological agent of plague, infected humans in Bronze Age Eurasia as early as 5 kya, three millennia before the first historical records of plague[64]. The authors concluded that the bacterium became the highly virulent, flea-borne bubonic plague strain only about 3 kya by acquiring specific genetic changes[64].

Analysis of *Mycobacterium tuberculosis* genomes from remains of ancient humans and animals helped in deciphering the origin and dispersal of tuberculosis in human populations. Ancient DNA studies provided support for the hypothesis that the appearance of tuberculosis in humans was not connected to animal domestication as it was suggested before. On the contrary, *M. tuberculosis* strain in humans is the most ancient one and other tuberculosis strains causing animal diseases evolved from the human strain[65]. Tuberculosis spread with humans and evolved in local conditions[66,67]. The most ancient, so far, human *M. tuberculosis* strain was discovered in 9000-Year-Old Pre-Pottery Neolithic Settlement in the Eastern Mediterranean[66] where, in spite of the presence of quantities of bovine bones, no signs of bovine strain, *M. bovis*, were found. Discovery of *M. bovis* strain in human remains from the Iron Age (as well as animal-like *Mycobacterium* strains in Pre-Columbian humans) showed that back infection from animals took place[6,64] at a later time.

Studying dental plaque of Europeans from different periods (Mesolithic, Neolithic, Bronze Age, Early Medieval, Late Medieval, and present time) demonstrated important shifts in human oral microbiota during recent evolution. The first shift took place in the early Neolithic period with the introduction of farming when more caries- and periodontal disease-associated taxa were detected. The oral microbiota composition remained stable between the Neolithic period and modern times. Recently, possibly during the Industrial Revolution in the 19th century, cariogenic bacteria became dominant, likely due to consumption of industrially processed flour and sugar. Consequently, the genetic diversity of the oral microbiotic ecosystem was impinged, which contributed to the spread of chronic oral and other diseases in countries with postindustrial lifestyles[68].

One of the most remarkable achievements in the field is the study of historical RNA. In 1997, Taubenberger and colleagues[69] extracted and analyzed RNA from the virus that caused the “Spanish flu” pandemic that killed at least 20

million people in 1918–1919[69-74]. Reconstruction of the viral genome helped to reveal its origin and discover the mechanism of its exceptional virulence. In contrast to modern influenza viruses, which require an exogenous protease for their replication, the 1918 pandemic virus could replicate without exogenous trypsin. The "Spanish flu" viral genome contained a constellation of genes, essential for optimal virulence, which contributed to the strain's ultra-high virulence[73]. This knowledge enabled epidemiologists to develop a vaccination strategy against another potential Spanish flu virus[72].

Adaptation of experimental and computational methods to the specific biochemistry of aDNA

After the death of an organism, all of its biomolecules are degraded either by host enzymes released from their proper compartments or by saprobic microorganisms. Therefore, compared to modern DNA, aDNA has lower concentration; it is fragmented, contaminated, and chemically modified[16,75,76]. Ancient DNA is also commonly damaged by strand breaks and cross-linking in addition to oxidative and hydrolytic degradation of bases or sugar residues. Relative preservation of DNA in old samples depends on environmental circumstances, such as temperature, humidity, pH, or oxygen, rather than the absolute age of the sample. For instance, DNA samples extracted from frozen remains dated thousands or even hundreds of thousands years can be of better quality than much more recent samples[5,77-79]. Recent studies showed that the age of "readable" (by current methods) aDNA products is restricted to about 1–1.5 million years[11,73]. At present, the 560–780 thousand years old Middle Pleistocene horse is the most ancient organism from which reliable aDNA data have been procured[5]. Below we describe methods for overcoming difficulties caused by each one of the special aDNA features.

Degradation

Early success with aDNA extraction and sequencing raised hopes that museum specimens, ancient samples, and archaeological finds would provide a plethora of aDNA, but such hopes faded when it became clear that these old samples did not yield any usable DNA[77]. Unfortunately it is not uncommon for aDNA projects to be disbanded due to low or undetectable DNA content[80-82]. In many other projects, the aDNA concentration is so low that it demands destructive sampling to yield adequate sequencing coverage. That, in turn, results in low genomic coverage (percentage of the length of the reference genome that is covered by mapped reads from the sample) and less reliable genotype calls. In their analysis of Neanderthal DNA, Green and colleagues reported GC content to be positively correlated ($r = 0.49$) with retrieval success of sequence fragments[2,83-85], likely due to the faster denaturation of AT-rich regions. They also found G and T overrepresented at the 5′ and 3′ ends of break points and suggested depurination as a significant cause of strand breaks[85].

Some of the difficulties in working with aDNA were resolved by technological breakthroughs. Improvement of extraction protocols can substantially increase the quantity and quality of aDNA. Thus, modern protocols[4,109] enable extraction and analysis of very short fragments (less than 50–60 bp, which constitutes the vast majority of aDNA). DNA fragmentation posed difficulties for conventional PCR, which requires amplification of a large number of overlapping fragments to cover a relatively long fragment of DNA, and it is impossible to sequence very short fragments (50–70 bp) using Sanger sequencing. However, NGS technologies generate short reads for any DNA. The average retrieved sequence length in most aDNA projects is 50–100 bp, which is the same order of magnitude as the length of reads produced by many current NGS instruments.

Fragmentation and decay of DNA is a natural occurrence not only *postmortem* but also *in vivo*. Spontaneous DNA degradation caused by damaging and mutagenic factors is prevented by DNA repair mechanisms that are not present after

death. However, controlled DNA degradation in living organism is implemented during programmed cell death (apoptosis) and differentiation of certain cell types (i.e., erythroid and lens cells, hair cortical cells). A large family of DNase enzymes performs the DNA degradation vital for proper development and functioning of living tissues. Apoptotic processes leading to these changes and DNA degeneration explain the average length of DNA fragments of 140–160 bp and under extracted from ancient mammoth hairs[110,111]. Many processes leading to DNA degradation, including those that accompany cell and tissue senescence (telomere shortening, error accumulation during DNA synthesis), occur naturally *in vivo*. Apoptosis finds its continuation in *postmortem* tissues, leading to further fragmentation of DNA even in favorable conditions for specimen preservation. The detailed biochemistry of processes occurring after death still requires further evaluation, and elucidation of their contribution to aDNA quality might be a promising area for research.

Contamination

Even after successful DNA extraction, results must always be checked for authenticity. Ancient DNA is often contaminated with some level of exogenous DNA (e.g., DNA from ancient or modern saprotrophic bacteria or fungi), *postmortem* juxtaposition of organisms, or modern human DNA from the researchers themselves. Naturally, low amounts of aDNA (or its complete absence) in the sample might facilitate the domination of PCR products by exogenous DNA, resulting in recovery of irrelevant sequences. Indeed, in the 1990s a large number of papers were published reporting DNA sequences from extremely ancient remains such as Miocene plant fossils[86,87], amber-entombed organisms[88,89], 250-million-year-old bacteria in salt crystal[90], and dinosaur bones and eggs[91,92]. In one such case, researchers reported successful extraction and amplification of mtDNA cytochrome *b* fragment from a Cretaceous Period dinosaur[91]. The sequences differed from all modern cytochrome *b* sequences. This led the authors to believe that they had sequenced authentic DNA from 80-million-year-old bones. It was later discovered that those mtDNA sequences were not close to avian and reptilian mtDNAs, as would be expected from their phylogenetic history, but rather to mammalian (including human) mtDNAs. It was thereby suggested that the alleged "dinosaur" DNA was contaminated, presumably by modern human DNA[93-96]. A similar course of events occurred in the study of ancient bacterial DNA supposedly preserved in 250-million-year-old salt crystals, which turned out to be modern bacterial DNA. In addition to these examples, several other aDNA projects have been impeded by contamination of ancient samples[97-101].

To prevent contamination, the experiment must be properly managed, including special requirements for sample collection, sterilization of the working area, DNA authentication, and independent reproducibility[112,113]. These protocols are constantly being refined and improved. For example, in addition to mechanical removal of the upper layer and UV and/or bleach treatment of the sample, a brief pre-digestion step was recently suggested[114], consisting of short-term sample incubation in an extraction buffer and its subsequent removal. According to the authors, this step alone increases the fraction of endogenous DNA several fold.

In shotgun sequencing of vertebrate samples, a substantial fraction of the reads comes from contamination with environmental DNA from bacteria and fungi[1,115,116]. Microbial sequences are often remarkably different from target species sequences and thus should be easily flagged by a standard BLAST search against the NCBI non-redundant nucleotide database. This strategy, however, fails to discover most of the microbial sequences that have yet to be sequenced. Therefore, it is not surprising that a large fraction of reads in many aDNA libraries is labeled as "unknown" or "unclassified," mainly due to the unidentified microbial content[84]. Frequently, mapping the shotgun sequencing reads onto the reference genome of

the target species (or the closest genome at hand) and discarding all reads below a certain level of similarity is preferred[117] alongside choosing tissues with less microbial DNA. For instance, it has been suggested that hair shafts or avian eggshells contain less microbial DNA than bone[110,118], but these tissues are not available for most ancient samples. Alternatively, recovery of bacterial or fungal sequences is not very likely for PCR-based capture methods, as primers are designed based on known sequences from the sample's own species or its close relatives.

The intricacy and method of detecting and removing modern human contamination depends on the distance of the target species from humans. Expectedly, it is much easier to handle distantly related species such as mammoths, penguins, or cave bears than archaic hominids, like the Denisovans and Neanderthals, and particularly ancient modern humans. Moreover, archeological material in Europe is usually excavated and later handled, extracted, and sequenced by Europeans, sometimes from the same region. The same is generally true for other territories around the world. When a limited number of loci are sequenced from PCR or cloning products, it is possible to examine alignments visually and to inspect individual polymorphic positions to determine which differences are genuine and which are likely artifacts or contamination[119,120]; however, with reads from shotgun sequencing technologies, automated methods are typically required.

Analyzing sequence reads in a phylogenetic framework along with sequences from ancient and extant relatives and outgroups is one of the initial steps to ensure that ancient sequences fit within the acceptable phylogeny and flag probable contamination. For instance, sequences from the mammoth were compared to those of the elephant, its closest kin, and to outgroups, such as humans and dogs, to ascertain phylogenetic correctness[116]. Filtering reads that were mapped onto the elephant genome with a high score and matched the elephant genome better than that of human, dog, or other species helped to remove human and microbial contamination. Neanderthal samples were phylogenetically examined to see if they fall outside the range of modern human variation[41].

Initially, a number of human and non-human studies filtered out samples with long sequence fragments considered evidence of contamination since authentic aDNA is supposed to be fragmented[80,120]. However, it has become clear that the average aDNA fragment length can vary substantially between samples and can overlap with contaminant fragment lengths; therefore, more elegant approaches are needed to develop authentication criteria based on length. In a study of Neanderthal DNA, estimates of human-Neanderthal sequence divergence and the percentage of C→T and G→A (equivalent events) misincorporations did not vary significantly with alignment length[108]. Existence of substantial modern DNA contamination would have produced two types of fragments: authentic ancient ones which were short and had high numbers of mismatches (showed high divergence vs. modern human reference), and modern contaminant ones which were long and showed few mismatches (showed low divergence vs. modern human reference). Noonan et al.[108] remarked that the absence of an inverse relationship between alignment length and divergence from the human reference meant that the level of contamination with modern DNA was negligible in their dataset; however, they did not provide a quantitative estimate. The problem with this approach is that even among authentic ancient fragments, short fragments presumably represent higher rates of base modification and consequently may produce upward-biased divergence estimates[83].

Through accumulation of ancient sequences over time, positions at which the target sequence (e.g., Neanderthal or Denisovan) have been invariably different from the likely contaminant (e.g., modern humans) can be used to estimate modern DNA contamination[85,121,122]. Here, mtDNA is the marker of choice because of its high copy number, leading to

greater sequencing depth; however, the validity of extrapolating mtDNA contamination estimates to nuclear sequences has been questioned based on possible differences in the conservation properties of mtDNA and nuclear DNA[83]. As base modification and misincorporations in aDNA often involve C to U (T) and A to G transitions, contamination with external DNA can be more reliably estimated using transversion or indel counts[85]. Even when sufficient prior data on sequence variation in the archaic hominid population is available, the fraction of reads that deviate from consensus base calls at haploid loci, e.g., those on mtDNA or the Y-chromosome, can provide an estimate of exogenous DNA—assuming that authentic aDNA is more abundant than contamination and that correct sequence reads are more likely than errors. This method is especially applicable to positions at which the modern human population is fixed for the derived base while the archaic consensus base is ancestral[3,46].

Ancient modern humans are not expected to carry informative (fixed) substitutions compared to contemporary humans or necessarily to fall outside the range of modern human phylogeny, although they might do so. A first step in the QC of sequences from ancient modern human samples is to ascertain that all sequence reads come from a single individual. This can be done by estimating X-linked heterozygosity in male samples, Y-linked heterozygosity in male samples, Y-linked presence in female samples, or mtDNA heterozygosity for either gender[1,46,83,123]. Next, it is necessary to show that each specimen in the dataset carries unique sequences (e.g., mtDNA or Y-chromosome haplotypes) that are different from sequences of other specimens and from the researchers[81,107,124].

As most of the ancient sequences during the pre-NGS era or shortly thereafter were limited to mitochondrial markers[85,125,126], it was crucial to distinguish them from nuclear inserts of mtDNA (NUMTs). Generally, a higher alignment score to the mitochondrial sequence than to the nuclear sequences is the authentication criterion. For extinct species without a reference, where sequence reads must be mapped to the genome of another species, this becomes more complicated because divergence of orthologous sequences must be considered in addition to differences between NUMTs and their mitochondrial counterparts[85]. Considering the low likelihood of heteroplasmy, observing more than one allele with non-negligible frequencies at each position would indicate either external contamination or sequencing of NUMTs.

Postmortem base modification

Postmortem DNA modifications through hydrolysis and oxidation pose another substantial difficulty for studying aDNA. The most significant alteration is nucleotide deamination, which leads to false transitions during PCR: cytosine to uracil, 5-methyl-cytosine to thymine (both causing incorporation of T instead of C), and, more rarely, adenine to hypoxanthine (causing incorporation of G instead of A)[80,102-105]. Chemical modification of nucleotides can lead to reduced sequencing coverage because they prevent mapping of many authentic reads due to an overestimated number of mismatches compared to the reference. They can also result in erroneous base and genotype calls and false estimates of genomic parameters such as heterozygosity, nucleotide diversity, GC content, or divergence times. Base modifications are often observed in the 5–7 final bases of DNA fragments and are thought to occur more readily in terminal, single-stranded overhangs[106]. These terminal misincorporations are even more problematic because local sequence alignment methods used for mapping the NGS reads onto the reference genome rely heavily on matching initial bases to the reference[84].

To overcome problems with chemical modification, several approaches have been developed. Treatment with uracil-N-glycosylase (UNG) removes uracil residues, thus preventing replication of fragments with deaminated cytosine[102,127];

however, the resulting abasic sites prevent replication by DNA polymerases, which excludes all the fragments with uracil from the reaction. This can be crucial for valuable ancient samples already having low DNA concentrations. A simple modification was suggested recently to overcome this problem[128]: follow-up treatment with endonuclease VIII after UNG repairs most of the abasic sites and enables subsequent analysis of these fragments. This procedure, however, does not resolve the problem of false A→G transitions. Using DNA polymerases such as Phusion, which does not amplify uracil, also avoids false C→T (but not A→G) transitions but excludes all uracil-containing fragments from amplification, which further decreases the DNA template in the reaction. Also, since these enzymes can work with methylated, deaminated cytosine (i.e., 5-methyluracil, thymine), the problem remains for methylated aDNA. Single primer extension PCR (SP-PCR) enables analysis of separate DNA strands, which makes it possible both to distinguish real mutations from *postmortem* modifications and to evaluate the level of these modifications[102-105,129]. SP-PCR is performed in two steps: first, PCR with only one primer is carried out to accumulate only one DNA strand, and then the second primer is added to the reaction and PCR continues with a normal protocol. The resulting PCR product derives mainly from one of the DNA strands. Analysis of these products can identify in which DNA strand *postmortem* modification occurred. This method requires very thorough selection of PCR primers and annealing temperatures, otherwise non-specific annealing or formation of primer dimers is highly possible.

One estimation strategy for base modification compares the percentage of T and A calls at ultra-conserved C and G positions, respectively. These genomic positions are expected to have retained their ancestral state in the ancient sample, so transitions exclusively observed in the ancient sample can be attributed to base misincorporations[4,111]. Another method compares the frequencies of different types of transitions and transversions in ancient-modern and modern-modern sequence alignments of closely related species (e.g., Neanderthal-human, Neanderthal-chimp, and human-chimp). An excess of C→T (and G→A, respectively) transitions in modern-ancient alignments provides an estimate of base modification[84]. A third strategy takes advantage of the direction of transition induced by base modification. In a 2006 study, the C→U modification in mammoth DNA caused the apparent rate of (mammoth T) → (elephant C) transitions to be 1.9-fold larger that of (mammoth C) → (elephant T) transitions[116]. Recently developed experimental protocols, such as pre-treatment of aDNA with UNG, have reduced the magnitude of this problem.

If the level of base modification is non-negligible, steps must be taken to eliminate or lessen its effect on the output of downstream population genetic analyses. Sometimes, C→T / G→A or all transitions are simply left out of the analyses, and only transversions and indels are included in the calculation of divergence or reconstruction of phylogeny[2]. Alternatively, it is possible to polarize polymorphisms into ancestral and derived states using an outgroup (e.g., chimp for human-Neanderthal comparisons) to place the mutation events on the corresponding branches of the phylogenetic tree using a parsimony approach (of which the branch leading to the ancient sample will probably contain disproportionately high numbers) and to calculate divergence times using information from branches leading to modern samples only[84]. Another strategy takes advantage of the observation that most of the base modifications occur at the 5′ and 3′ ends of fragments and trims a few (5–7) bases off either end of each sequence read to lower the chance of including a misincorporated base[1,123].

Base and genotype calling

Once the sources of contamination or base misincorporation are detected and removed from the aDNA sequences, it is possible to use the consensus base for further analyses. However, genotypes can be inferred more accurately by combining

observed read bases with various estimators of contamination, base modification, sequencing error, and read alignment quality combined via a single maximum likelihood (ML) calculation[130]. The ML model can be designed in haploid mode for mtDNA, or X- and Y-chromosomes in males, or diploid mode for autosomal markers or X-linked markers in females[1]. Depending on the specific design, ML models can use these estimators to output the genotype or co-estimate all of these parameters simultaneously.

Although the problems associated with aDNA anomalies are not insurmountable using available experimental and bioinformatics technologies, drastic variations in the type and magnitude of damage among ancient remains make it impossible to develop a universally successful protocol for aDNA extraction and sequencing. For instance, the fraction of authentic Neanderthal mtDNA among six examined ancient samples varied from ~1% to ~99%[84], and the level of contamination in five well-preserved human bone specimens dated 800–1600 CE varied from 0% to 100%[107]. The Neanderthal mitochondrial genome and partial nuclear genome were retrieved using data from several sequencing attempts.[41,84,108] This compendium was crucial in determining design parameters for assembling the full Neanderthal nuclear genome[8]. The contaminating sequences in an ancient maize microsatellite genotyping project were found to be of different natures across samples: some exhibited mainly microbial contamination, whereas others contained copies of transposable elements[80]. Therefore, an initial round of extraction and sequencing is recommended to estimate quality parameters for each sample (e.g., yield, chemical modification, % contamination, % uniquely mapped reads, and % genome covered) to inform appropriate experimental and data preparation strategies. It is also important to remember that both experimental and computational methods of overcoming of ancient DNA problems have advantages and limitations. Therefore, to achieve the most reliable results, it is of great importance that researchers use both these approaches to examine aDNA.

Analysis of aDNA data

Software tools for preprocessing of aDNA NGS data

In order to implement the computational approaches needed to address the special features of aDNA (i.e. obstacles due to degradation, base damage, and contamination) in addition to regular NGS data quality control and preprocessing steps, application of highly specialized tools is required. Current NGS analyses of ancient DNA are performed with well-established but non-specialized computational tools as novel customized tools for aDNA analysis have not yet been widely accepted, and custom scripts have to be written to adjust for aDNA specifics. Base calling is frequently performed with Illumina's standard base-caller Bustard BayesCall[131] (flexible model-based tool) and freeIbis[132] (utilizing a multiclass Support Vector Machine algorithm). FastQC[133] is typically used for preliminary quality control of reads. AdapterRemoval[134], CutAdapt[135], and SeqPrep[136] are currently the most common tools in the aDNA world for de-multiplexing, adapter trimming, low-quality call trimming, and paired-end merging. BWA and Bowtie are commonly used for mapping of aDNA reads.[137] Since BWA and Bowtie were developed for high quality modern DNA reads, parameter adjustments to reflect properties of ancient DNA must be done. For example, it may be advisable to trim likely-damaged positions, disable the seed, adjust gap openings and penalties, and permit indels at read ends.[137] Genome Analysis Toolkit (GATK)[138,139] or SAMtools[140] are then used for variant calling. Methods should be optimized for shorter (17–35 nucleotide) reads with possible adapters on both ends of the read and a large overlap between paired reads, and the call corresponding to the highest quality score should be selected at each position. In addition, due to low quantities of endogenous DNA, the high number of short reads, and high levels of contamination, masking repeat regions may improve read mapping. Several pipelines (e.g., aLib[141] and PALEOMIX[142]) incorporating all these changes were designed for aDNA analysis. Sometimes

selected elements of such pipelines are combined to achieve optimal performance. For instance, the leeHom module of aLib is used to preprocess reads[143] while Anfo[2], MIA[121], and BWA-PSSM[144] are used for subsequent read mapping.

After read pretreatment and alignment, the text file containing the sequence alignment data (termed the SAM file or BAM for binary files) can be used to estimated contamination and degradation levels using tools such as mapDamage[145] and mapDamage2.0[146]. PMDtools[147] (identification of those DNA fragments that are unlikely to come from modern sources) and Schmutzi[148] (maximum *a posteriori* estimator of mitochondrial contamination for ancient samples) are utilized to select reads for re-analysis that have higher chances of coming from aDNA. Depending on the situation, analysis can be done in a fully automated cycle for the entire genome or only for mtDNA. Such algorithms report the probabilities of different types of *postmortem* DNA degradation, which allows for better statistical modeling at the variant calling stage, employing SNPest[149] or custom scripts. A typical NGS pipeline for aDNA analysis is shown in Figure 2.

The amount of extracted endogenous DNA may allow satisfactory coverage of aDNA sequences (as high as 10–20× coverage at a subset of regions for a few samples), comparable to modern DNA studies. Nevertheless, it is very common for ancient samples to have ~1× average coverage. In such cases, population genetics analysis can still be done using ADMIXTURE and other standard tools by choosing the variant with the highest number of supporting reads (or with the highest quality) instead of trying to make heterozygous/homozygous calls at each autosomal position (which is tricky when there are <5 reads covering a given position resulting in 3–4 conflicting variants). This method can be used when the amount of contaminating modern human DNA is much lower than the amount of endogenous DNA.

Special care needs to be exercised when combining SNP data from ancient and modern samples. Recently published analysis of the first ancient African genome[150] presented an erroneous conclusion that genomes of individuals throughout Africa contain DNA inherited from Eurasian immigrants.[151] The error was noticed, and the authors published an erratum stating that it had been necessary to convert the input produced by SAMtools to be compatible with PLINK, but this step was omitted causing removal of many positions homozygous to the human reference genome[151]. This example illustrates importance of using validated pipelines for aDNA analysis.

NGS data analysis of aDNA

Below, we discuss the analytical methods for biologically relevant interpretation of aDNA data. Various population genetics analytical methods have been applied to infer past demographic events of populations based on data obtained from aDNA studies. One of the basic methods for identifying ancient haplotypes is scanning present-day populations for variants identified in the aDNA. This simple approach provides an estimate of populations/regions that harbor such ancient genetic signatures and has been successfully applied to identify modern European populations with mtDNA mutations that were found in aDNA samples[152,153]. Analysis of single-nucleotide polymorphisms (SNPs) in prehistoric samples can shed light on ancestral phenotypes, including pigmentation of skin, hair, and eyes[31], and the sex of the sample can be computed as the ratio of reads mapping to the Y- and X-chromosomes[154]. In the case of uni-parental markers such as mtDNA variants and Y-chromosome markers, the mutational distance between the ancient and modern haplotypes is visualized using phylogenetic network analysis programs[153,155]. Network analysis of haplotype data reveals genetic distance, mutation rate, and regions of haplotype spread.

With increased numbers of recovered ancient and historic DNA samples and steady improvements in aDNA sequencing technology, scientists can study the distribution of ancient human genetic variation and compare it to that of modern populations[156] or gain a deeper-level understanding of the distribution of genetic variation within populations by applying admixture-based tools for joint analysis of modern and ancient samples at a population level. Tools and approaches, such as ADMIXTURE[157], ADMIXTOOLS[158], GPS[159], or reAdmix[160], which were initially developed for population analysis of contemporary individuals, can be applied in combination with anthropological data and historic records to reconstruct migration patterns, provenances, and local and global ancestries of extinct populations. ADMIXTURE is a computational tool for ML estimation of individual ancestries from multi-locus SNP genotype datasets. Recently, Allentoft et al.[11] inferred the ancestral components from modern samples and then projected the ancient samples onto the inferred components using the ancestral allele frequencies inferred by ADMIXTURE. Comparison of admixture profiles of ancient and modern populations within a given region informs generation of hypotheses about population migrations that can be validated with independent sources and methods of analysis. NGSADMIX uses genotype likelihoods instead of called genotypes to resolve ancestry, which is particularly useful considering the myriad sources of uncertainty in aDNA NGS data[161].

The geographical origins of samples can be inferred with the Geographic Population Structure (GPS) algorithm developed by Elhaik et al.[159], which analyzes genome-wide ancestry-informative markers. First, the algorithm converts the file with an individual's genotype into a *K*-dimensional vector using ADMIXTURE in a supervised mode. The resulting vector is compared to reference vectors corresponding to worldwide populations to identify the population with the smallest Euclidean distance to the individual. Finally, the genetic distances are converted to geographic distances based on the linear relationships between genetics and geography, which yields the position of the test sample. The same framework can be applied to analysis of ancient samples to identify the geographical origins of ancient populations in the case of migratory individuals. GPS can also identify the closest populations to the individual of interest. To infer multiple such populations, the reAdmix algorithm developed by Kozlov et al.[160] is a more suitable approach. reAdmix infers the complex provenances of highly mixed individuals and represents a tested individual as a weighted sum of reference populations. In addition, reAdmix can work in a conditional mode that limits the search space based on archaeological data.

To infer the geographical origin of a haplotype, it is essential to partition the genome into haplotypes with distinct ancestries that may have been inherited from multiple populations. Such haplotypes can be obtained using identical-by-descent tools such as BEAGLE[162], HAPMIX[163], or SABER[164], which allow inference of "local ancestry" instead of the "global ancestry" that can be inferred with PCA, GPS, and reAdmix, and their usage depends on the complexity of the dataset, the expected mixture levels, and the available phenotypic data. For instance, if the phenotype is associated with a particular trait, a "local ancestry" tool is preferred, whereas a "global ancestry" tool should be used when the phenotype is a complex trait involving multiple unknown loci.

When several individuals of the ancient population are available, certain population genetic parameters can be estimated. For example, by examining a number of microsatellite loci in 160–200-year-old Daphnia samples[165], researchers were able to calculate the heterozygosity, gene diversity, deviation from Hardy-Weinberg Equilibrium, and linkage disequilibrium between pairs of markers. An analysis of 21 samples from a graveyard in Germany dated ~6 kya allowed analysis of the mtDNA and Y-chromosomal haplotype diversity as well as the selection forces inferred from Tajima's D[81]. Jaenicke-

Despres et al. discovered allelic variants of three genes that differentiate modern maize and teosinte from 11 maize cobs dating to 660–4400 years ago, opening a window to the genetic chronology of maize domestication[166].

Even when only a single member of an ancient population can be recovered, a number of genomic and evolutionary inferences can be made about its taxon or population. Based on remarkably low levels of dN/dS (non- synonymous to synonymous substitution ratio), it was concluded that mitochondrial proteins were under strong purifying selection in Denisovans[43]. Conversely, the higher dN/dS ratio calculated from Neanderthal genomes was attributed to smaller effective population size and inefficient purifying selection[85,121]. The heterozygosity of the TAS2R38 locus in a single Neanderthal individual was used to infer that he was a bitter-taster and, further, that this trait varied among Neanderthals[167]. Using zygosity correlation along the genome, it was suggested that the Denisovans experienced a 30-fold decrease in effective population size compared to African humans[168].

Beyond DNA sequence: ancient epigenomics

Physical characteristics, predisposition to diseases, and psychological features are the result of expression of genes through complex and poorly understood interactions of so-called "genetic determinants" and external environmental signals. Epigenetic mechanisms including DNA methylation, histone modifications, and a spectrum of noncoding RNAs often affect chromatic structure and modulate gene expression in response to signals from the external or intra-organismal environment. Until recently, reconstructing the gene expression profile of specific *postmortem* samples using only DNA was deemed impossible. A common feature of all aDNA samples is their highly degraded state and the presence of deaminated cytosines. Several groups analyzing the methylation maps (methylome) of Neanderthals proposed that patterns of CpG methylation were preserved in the DNA[128,169]. Later, Llamas et al. applied bisulfite allelic sequencing of loci to late Pleistocene *Bison priscus* remains and demonstrated that methylation patterns are indeed preserved in ancient DNA[170], although *postmortem* deamination of methylated cytosine to thymine prevents accurate quantification of methylated cytosine levels.

In 2014, several groups proposed a reliable approach toward genome-wide methylation studies in ancient samples[169,171,172]. Using high-throughput sequencing in combination with a variety of computational approaches, researchers assembled genome-wide methylomes of human and mouse genomes in cells of different types and origins (e.g., somatic, stem, germline, and cancer)[173]. In bisulfite sequencing, unmethylated cytosines are chemically converted into uracils, which are then read by some polymerases, such as *Taq*, as thymines (T). These C→T mutations can be located in vertebrate genomes with single-nucleotide resolution. Similar chemical conversions occur naturally in *postmortem* samples, mainly through the hydrolytic deamination of cytosines located in single-stranded overhangs[102]. With *Taq* DNA polymerases that can replicate through uracils this results in increasing C→T misincorporation rates at sequence starts. Such misincorporations should not be observed with high-fidelity DNA polymerases, like Phusion, that cannot bypass uracils, except at 5mC where deamination leaves T instead of U (See Fig. 3 for the detailed interpretation). Thus, it became possible to detect methylated cytosines in ancient genomes by their elevated C→T mismatch rates at read starts compared to non-methylated cytosines.

Since deamination is a stochastic process, it is impossible to detect cytosine methylation with single-nucleotide resolution. The methylated CpG sites that did not experience *postmortem* deamination events are copied as regular CpG's and leave no

methylation footprint in the genome. Methylation may then be detected with single-nucleotide resolution upon a significant increase in coverage depth or by recording CpG→TpG mismatches within full genomic regions that include many CpG's. This strategy was applied to analyze aDNA from Neanderthal (50 kya), Denisovan (40 kya), and a relatively recent Paleo-Eskimo individual (4 kya). DNA methylation patterns in ancient human bones or hairs were almost indistinguishable from those in modern humans. However, by examining differentially methylated regions, Gokhman et al.[169] found that some key regulators of limb development, like *HOXD9* and *HOXD10*, had methylated promoters and gene bodies, respectively, in Neanderthal and Denisovan samples, whereas these regions are hypo-methylated in bone samples from present-day humans. Because deregulation of HOXD cluster genes in general, and *HOXD9* and *HOXD10* in particular, results in morphological changes in mice[174] that reflect Neanderthal-modern human differences, it can be inferred that epigenetic changes in HOXD clusters might have played a key role in the recent evolution of human limbs.

Differentially methylated regions were also found within the *MEIS1* gene, which encodes a protein that controls the activity of the HOXD cluster[174]. Elucidation of the ancient methylome from aggregated C→T mismatch information over large genomic regions allows analysis of whether extended regions with altered DNA methylation were present in ancestral bone and hair samples. These include not only hypermethylated CpG islands but also 1) large (several 100 kb to several Mb) partially methylated gene-poor domains that co-localize with lamina-associated domains[175,176]; 2) DNA methylation valleys extending over several kb of DNA, which are strongly hypomethylated in most tissues, enriched in transcription factors and developmental genes[177,178]; 3) undermethylated canyons (up to dozens of kb) that were recently identified in hematopoietic stem cells[179]; and 4) epigenetic programs associated with intestinal inflammation and characterized by hypermethylation of DNA methylation valleys with low CpG density and active chromatin marks[180].

Methylation analysis usually focuses on genomic regions that span several kb or even mb[178-180]; the C→T mismatch aggregation strategy, applied in a recent aDNA epigenomic study[171], could yield new perspectives on adaptation signals or disease markers in the rare case that tissues normally unpreserved in anthropological samples (brain, intestine, muscle, and blood) are found and may only be applicable to remains found in permafrost soil (such as mammoths and other extinct Pleistocene species). In exceptional cases, such as when the epigenome map was deconvoluted from a single hair of a Paleo-Eskimo man,[171] analysis of epigenetic patterns also allows estimation of the age at death. A recent forensic study generated data indicating a correlation between the methylation state of specific CpGs and the age of an individual[181]; however, such calculations assume that environmental signals 6 kya produced the same genomic methylation response observed today to estimate the age of ancient humans using modern databases. Using this assumption, Pedersen et al. calculated that the Saqqaq individual was probably in his late thirties when he died[171].

Because methylated CpG's are found almost exclusively in vertebrate somatic cells, whereas bacterial genomes feature methylated cytosines and adenines but almost never in a CpG context, CpG methylation of vertebrate genomes could be used to enrich the endogenous content of a human aDNA sample and separate it from bacterial contaminants[182]. Methyl DNA binding domain (MBD) affinity chromatography, which allows separation of methylated DNA probes containing a single methylated CpG, has become a routine method for establishing the methylomes of genomes of different origins[183]. Applying this method to aDNA not only promises both to facilitate characterization of ancient methylomes and to separate vertebrate and microbial fractions of aDNA extracts. Using the remains of the Saqqaq Paleo-Eskimo individual, woolly mammoths, polar bears, and two equine species, methylation marks were shown to survive in a variety of tissues and

environmental contexts and over a large temporal span (>45–4 kya). Additionally, MBD enrichment allows characterization of the microbiome of ancient samples and potentially ancient pathogen genomes.

Although DNA methylation may serve as an indicator of gene silencing, epigenetic analysis alone is insufficient to determine whether the gene was destined for transcription or silencing. Additional data, such as histone modification marks, chromatin structure, and transcription factor binding information, are essential for gene activity prediction. Although research on ancient proteins is in its nascent stage, shotgun sequencing of aDNA has provided a surprisingly rich source of epigenetic information. Pedersen et al. observed unexpected periodicity in the density of covered nucleotides along the Saqqaq genome[171] and hypothesized that these periodic patterns could stem from protection of DNA by nucleosome binding with preferential degradation of linker regions between nucleosomes, either by DNases that enter the nucleus at cell death or as a result of *postmortem* stand breaks, instead of from alignment or sequencing artifacts. In such a scenario, the observed read depth would reflect the nucleosome occupancy. Fourier transformation of the function that links sequencing coverage of a single nucleotide with its genomic coordinate revealed a strong signal at 200 bp from transcription start sites (TSSs), where nucleosomes are strongly positioned, and downstream. Moreover, a phasogram from Fourier transform also revealed short-range (10 bp) periodicity, reflecting preferential shifts in nucleosome positioning every 10 bp and/or preferential cleavage of the DNA backbone facing away from nucleosome protection. Strongly positioned nucleosomes in an ancient sample were also found within the vicinity (4 kb) of CTCF binding sites, and their order was negatively correlated with uncovered DNA methylation. Since DNase I-hypersensitive sites (DHSs) within the TSS are reliable predictive markers for gene transcription[184], regions within open chromatin structures may be more susceptible to *postmortem* or apoptosis-induced DNase cleavage, in which case the density of NGS reads within the TSS of active genes would be lower than at silent genes. Based on read density at known TSSs and DHSs from the ENCODE project and using *de novo* methods of TSS prediction (e.g., NPEST[185] or TSSer[186]), it is possible to sort TSSs according to the transcriptional activity of the corresponding genes. In the near future, it may be possible to quantitatively reconstruct gene expression patterns of ancient samples by combining nucleosome positioning, the presence of DHSs at TSSs, and DNA methylation. Therefore, the discovery of preserved brains, like those from bog bodies[187], is of particular interest. The recent discovery of a woolly mammoth that retained brain structures of very high quality[188] raised hopes that exciting discoveries are at the horizon that would allow us to test whether the higher nervous system activity in modern humans differs from that of ancient humans, mainly at the epigenetic level[189,190].

Conclusions

Ancient DNA research has revolutionized a multitude of scientific disciplines. Representing the most direct route to address a large number of questions in evolution, medicine, anthropology, and even history, aDNA became an indispensable tool in archaeo-genetics, population genetics, paleo-epidemiology, and related fields. Analysis of aDNA has made tremendous progress since its humble beginning in the early 1990s, when contamination with modern DNA sources was commonplace, and only limited analysis was possible due to DNA fragmentation and sparse sampling. In this review, we attempted to provide a detailed overview of recent innovations aimed at coping with these limitations, both through experimental procedures and bioinformatics algorithms. We also considered challenges regarding aDNA biochemistry and degradation, particular bioinformatics tools compensating for short reads and gaps in sequencing coverage, and advances in population genetics to handle sparse sampling. Lastly, we described the particularities of aDNA epigenetics and functional interpretation of deduced activities of particular genes and pathways.

In envisioning future progress in aDNA studies, we would like to note that not every advance in genomics or experimental biology may affect the field. Recent breakthroughs in genomic technologies drastically increased the amount of information obtained from aDNA, and new inventions, e.g., progress in targeted enrichment methods and single-molecule sequencing, would likely allow investigation of previously intractable samples from hot climates and more distant eras. However, experimental approaches will always be limited by the quantity and quality of aDNA in ancient remains. Thus, development of computational methods to cope with aDNA-specific biases and extract meaningful information from low-coverage aDNA data is critical. Studies of aDNA will hugely benefit from further improvement of sophisticated bioinformatics tools coupled with the rapid accumulation of content (reference genomes and variant databases) from both ancient samples and freshly sequenced modern human populations. Regarding the latter, one can hardly overestimate the effect of international projects on systematic sequencing of small and remote human populations, such as those undertaken by National Geographic and others[13,20,155,159,191-194]. Parallel improvement of experimental and computational methods will enable studies of ancient populations instead of just a few individuals. The utility of aDNA data will increase with further progress in genotype-to-phenotype mapping of humans. For the first time, we can anticipate the direct study of evolution for traits that are not associated with the fossil record, such as metabolic and behavioral details. Ancient DNA will provide an important source of information on the origins of cells that harbored DNA thousands years ago, the age of samples at the time of death, and the environmental influences. Altogether, analysis of ancient DNA will help us to better understand our world and our role in it.

Acknowledgements

HA was supported by NIH grants GM098741 and MH091561. EP was supported by Russian Science Foundation grant 14-14-01202. PF was supported by the Moravian-Silesian region projects MSK2013-DT1, MSK2013-DT2, and MSK2014-DT1 and by the Institution Development Program of the University of Ostrava. TVT, EE, and PP were supported by NSF Division of Environmental Biology award 1456634. EE was supported by The Royal Society International Exchanges Award (IE140020) and MRC Confidence in Concept Scheme award 2014-University of Sheffield (Ref: MC_PC_14115). ER was supported by Russian Science Foundation grant 14-50-00029. We are grateful to Dr. Lana Grinberg and Dr. Steven D. Aird for proofreading the manuscript. The idea of the manuscript was conceptualized during the Ancient DNA Symposium sponsored by the Okinawa Institute of Science and Technology.

References

1. Rasmussen, M., Li, Y., Lindgreen, S., et al. 2010, Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, **463**, 757-762.
2. Green, R. E., Krause, J., Briggs, A. W., et al. 2010, A draft sequence of the Neandertal genome. *Science*, **328**, 710-722.
3. Reich, D., Green, R. E., Kircher, M., et al. 2010, Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468**, 1053-1060.
4. Meyer, M., Kircher, M., Gansauge, M.-T., et al. 2012, A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*.
5. Orlando, L., Ginolhac, A., Zhang, G., et al. 2013, Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, **499**, 74-78.
6. Bos, K. I., Harkins, K. M., Herbig, A., et al. 2014, Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, **514**, 494-497.

7. Lazaridis, I., Patterson, N., Mittnik, A., et al. 2014, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, **513**, 409-413.
8. Prufer, K., Racimo, F., Patterson, N., et al. 2014, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, **505**, 43-49.
9. Raghavan, M., DeGiorgio, M., Albrechtsen, A., et al. 2014, The genetic prehistory of the New World Arctic. *Science*, **345**, 1255832.
10. Raghavan, M., Skoglund, P., Graf, K. E., et al. 2014, Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, **505**, 87-91.
11. Allentoft, M. E., Sikora, M., Sjogren, K. G., et al. 2015, Population genomics of Bronze Age Eurasia. *Nature*, **522**, 167-172.
12. Haak, W., Lazaridis, I., Patterson, N., et al. 2015, Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*.
13. Raghavan, M., Steinrucken, M., Harris, K., et al. 2015, POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*, **349**, aab3884.
14. Skoglund, P., Malmstrom, H., Raghavan, M., et al. 2012, Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science*, **336**, 466-469.
15. Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A. and Wilson, A. C. 1984, DNA sequences from the quagga, an extinct member of the horse family. *Nature*, **312**, 282-284.
16. Pääbo, S., Poinar, H., Serre, D., et al. 2004, Genetic analyses from ancient DNA. *Annual review of genetics*, **38**, 645-679.
17. Pickrell, J. K. and Reich, D. 2014, Toward a new history and geography of human genes informed by ancient DNA. *Trends in genetics : TIG*, **30**, 377-389.
18. Veeramah, K. R. and Hammer, M. F. 2014, The impact of whole-genome sequencing on the reconstruction of human population history. *Nature reviews. Genetics*, **15**, 149-162.
19. Elhaik, E. 2013, The Missing Link of Jewish European Ancestry: Contrasting the Rhineland and the Khazarian Hypotheses. *Genome Biology and Evolution*, **5**, 61-74.
20. Elhaik, E., Greenspan, E., Staats, S., et al. 2013, The GenoChip: a new tool for genetic anthropology. *Genome Biol Evol*, **5**, 1021-1031.
21. Elhaik, E., Tatarinova, T. V., Klyosov, A. A. and Graur, D. 2014, The 'extremely ancient' chromosome that isn't: a forensic bioinformatic investigation of Albert Perry's X-degenerate portion of the Y chromosome. *Eur. J. Hum. Genet.*
22. Finkelstein, I. and Silberman, N. A. 2002, *The Bible unearthed : archaeology's new vision of ancient Israel and the origin of its sacred texts*. Simon and Schuster: New York, NY.
23. Elhaik, E., Tatarinova, T. V., Klyosov, A. A. and Graur, D. 2014, The 'extremely ancient' chromosome that isn't: a forensic bioinformatic investigation of Albert Perry's X-degenerate portion of the Y chromosome. *Eur J Hum Genet*, **22**, 1111-1116.
24. Marean, C. W., Anderson, R. J., Bar-Matthews, M., et al. 2015, A new research strategy for integrating studies of paleoclimate, paleoenvironment, and paleoanthropology. *Evol Anthropol*, **24**, 62-72.
25. Wang, X., Kellner, A. W., Cheng, X., et al. 2015, Eggshell and Histology Provide Insight on the Life History of a Pterosaur with Two Functional Ovaries. *An Acad Bras Cienc*, **87**, 1599-1609.
26. Merkhofer, L., Wilf, P., Haas, M. T., et al. 2015, Resolving Australian analogs for an Eocene Patagonian paleorainforest using leaf size and floristics. *Am J Bot*, **102**, 1160-1173.
27. Meyer, M., Kircher, M., Gansauge, M. T., et al. 2012, A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, **338**, 222-226.
28. Keller, A., Graefen, A., Ball, M., et al. 2012, New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature communications*, **3**, 698.
29. Olalde, I., Allentoft, M. E., Sanchez-Quinto, F., et al. 2014, Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature*, **507**, 225-228.
30. Olalde, I., Sanchez-Quinto, F., Datta, D., et al. 2014, Genomic analysis of the blood attributed to Louis XVI (1754-1793), king of France. *Scientific reports*, **4**, 4666.

31. Cerqueira, C. C., Paixao-Cortes, V. R., Zambra, F. M., Salzano, F. M., Hunemeier, T. and Bortolini, M. C. 2012, Predicting Homo pigmentation phenotype through genomic data: from Neanderthal to James Watson. *American journal of human biology : the official journal of the Human Biology Council*, **24**, 705-709.
32. Cale, C. M. 2015, Forensic DNA evidence is not infallible. *Nature*, **526**, 611.
33. Lohse, K. and Frantz, L. A. 2014, Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. *Genetics*, **196**, 1241-1251.
34. Sankararaman, S., Mallick, S., Dannemann, M., et al. 2014, The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, **507**, 354-357.
35. Vernot, B. and Akey, J. M. 2015, Complex history of admixture between modern humans and Neandertals. *American journal of human genetics*, **96**, 448-453.
36. Vernot, B. and Akey, J. M. 2014, Human evolution: genomic gifts from archaic hominins. *Current biology : CB*, **24**, R845-848.
37. Vernot, B. and Akey, J. M. 2014, Resurrecting surviving Neandertal lineages from modern human genomes. *Science*, **343**, 1017-1021.
38. Fu, Q., Hajdinjak, M., Moldovan, O. T., et al. 2015, An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, **524**, 216-219.
39. Sanchez-Quinto, F. and Lalueza-Fox, C. 2015, Almost 20 years of Neanderthal palaeogenetics: adaptation, admixture, diversity, demography and extinction. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **370**, 20130374.
40. Wall, J. D., Yang, M. A., Jay, F., et al. 2013, Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics*, **194**, 199-209.
41. Krause, J., Orlando, L., Serre, D., et al. 2007, Neanderthals in central Asia and Siberia. *Nature*, **449**, 902-904.
42. Racimo, F., Sankararaman, S., Nielsen, R. and Huerta-Sanchez, E. 2015, Evidence for archaic adaptive introgression in humans. *Nature reviews. Genetics*, **16**, 359-371.
43. Krause, J., Fu, Q., Good, J. M., et al. 2010, The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, **464**, 894-897.
44. Reich, D., Patterson, N., Kircher, M., et al. 2011, Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American journal of human genetics*, **89**, 516-528.
45. Cooper, A. and Stringer, C. B. 2013, Paleontology. Did the Denisovans cross Wallace's Line? *Science*, **342**, 321-323.
46. Krause, J., Briggs, A. W., Kircher, M., et al. 2010, A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia. *Current Biology*, **20**, 231-236.
47. Gibbons, A. 2011, Paleoanthropology Society & American Assoc. of Physical Anthropologists. Ancient footprints tell tales of travel. *Science*, **332**, 534-535.
48. Qin, P. and Stoneking, M. 2015, Denisovan Ancestry in East Eurasian and Native American Populations. *Molecular biology and evolution*.
49. Eriksson, A. and Manica, A. 2012, Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 13956-13960.
50. Cooper, A. and Stringer, C. B. 2013, Did the Denisovans Cross Wallace's Line? *Science*, **342**, 321-323.
51. Meyer, M., Fu, Q., Aximu-Petri, A., et al. 2014, A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature*, **505**, 403-406.
52. Meyer, M., Arsuaga, J. L., Nagel, S., et al. 2015, Nuclear DNA sequences from the hominin remains of Sima de los Huesos, Atapuerca, Spain. *5TH ANNUAL MEETING OF THE European Society for the study of Human Evolution* ESHE, LONDON / UK.
53. Arsuaga, J. L., Martinez, I., Arnold, L. J., et al. 2014, Neandertal roots: Cranial and chronological evidence from Sima de los Huesos. *Science*, **344**, 1358-1363.
54. Arsuaga, J. L., Carretero, J.-M., Lorenzo, C., et al. 2015, Postcranial morphology of the middle Pleistocene humans from Sima de los Huesos, Spain. *PNAS*, **112**, 11524-11529.

55. Anthony, D. W. 2007, *The horse, the wheel, and language : how Bronze-Age riders from the Eurasian steppes shaped the modern world*. Princeton University Press: Princeton, N.J. ; Woodstock.
56. Matisoo-Smith, E. 2015, Ancient DNA and the human settlement of the Pacific: a review. *Journal of human evolution*, **79**, 93-104.
57. Gao, S. Z., Zhang, Y., Wei, D., et al. 2015, Ancient DNA reveals a migration of the ancient Di-qiang populations into Xinjiang as early as the early Bronze Age. *American journal of physical anthropology*, **157**, 71-80.
58. Zhao, Y. B., Li, H. J., Cai, D. W., et al. 2010, Ancient DNA from nomads in 2500-year-old archeological sites of Pengyang, China. *J Hum Genet*, **55**, 215-218.
59. Liu, W., Martinon-Torres, M., Cai, Y. J., et al. 2015, The earliest unequivocally modern humans in southern China. *Nature*, **526**, 696-699.
60. Llorente, M. G., Jones, E. R., Eriksson, A., et al. 2015, Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science*.
61. Krause, J., Lalueza-Fox, C., Orlando, L., et al. 2007, The derived FOXP2 variant of modern humans was shared with Neandertals. *Current biology : CB*, **17**, 1908-1912.
62. Lai, C. S., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F. and Monaco, A. P. 2001, A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*, **413**, 519-523.
63. Mathieson, I., Lazaridis, I., Rohland, N., et al. 2015, Eight thousand years of natural selection in Europe.
64. Rasmussen, S., Allentoft, M. E., Nielsen, K., et al. 2015, Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago. *Cell*, **163**, 571-582.
65. Brosch, R., Gordon, S. V., Marmiesse, M., et al. 2002, A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 3684-3689.
66. Nerlich, A. G. and Losch, S. 2009, Paleopathology of human tuberculosis and the potential role of climate. *Interdiscip Perspect Infect Dis*, **2009**, 437187.
67. Darling, M. I. and Donoghue, H. D. 2014, Insights from paleomicrobiology into the indigenous peoples of pre-colonial America - a review. *Mem Inst Oswaldo Cruz*, **109**, 131-139.
68. Adler, C. J., Dobney, K., Weyrich, L. S., et al. 2013, Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nat Genet*, **45**, 450-455, 455e451.
69. Taubenberger, J. K., Reid, A. H., Krafft, A. E., Bijwaard, K. E. and Fanning, T. G. 1997, Initial genetic characterization of the 1918 "Spanish" influenza virus. *Science*, **275**, 1793-1796.
70. Reid, A. H., Fanning, T. G., Hultin, J. V. and Taubenberger, J. K. 1999, Origin and evolution of the 1918 "Spanish" influenza virus hemagglutinin gene. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 1651-1656.
71. Taubenberger, J. K., Hultin, J. V. and Morens, D. M. 2007, Discovery and characterization of the 1918 pandemic influenza virus in historical context. *Antivir Ther*, **12**, 581-591.
72. Taubenberger, J. K., Morens, D. M. and Fauci, A. S. 2007, The next influenza pandemic: can it be predicted? *JAMA*, **297**, 2025-2027.
73. Tumpey, T. M., Basler, C. F., Aguilar, P. V., et al. 2005, Characterization of the reconstructed 1918 Spanish influenza pandemic virus. *Science*, **310**, 77-80.
74. Tumpey, T. M., Garcia-Sastre, A., Taubenberger, J. K., Palese, P., Swayne, D. E. and Basler, C. F. 2004, Pathogenicity and immunogenicity of influenza viruses with genes from the 1918 pandemic virus. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 3166-3171.
75. Millar, C. D., Huynen, L., Subramanian, S., Mohandesan, E. and Lambert, D. M. 2008, New developments in ancient genomics. *Trends in ecology & evolution*, **23**, 386-393.
76. Parks, M., Subramanian, S., Baroni, C., et al. 2015, Ancient population genomics and the study of evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **370**, 20130381.

77. Haile, J., Froese, D. G., Macphee, R. D., et al. 2009, Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 22352-22357.
78. Handt, O., Hoss, M., Krings, M. and Paabo, S. 1994, Ancient DNA: methodological challenges. *Experientia*, **50**, 524-529.
79. Handt, O., Richards, M., Trommsdorff, M., et al. 1994, Molecular genetic analyses of the Tyrolean Ice Man. *Science*, **264**, 1775-1778.
80. Lia, V. V., Confalonieri, V. a., Ratto, N., et al. 2007, Microsatellite typing of ancient maize: insights into the history of agriculture in southern South America. *Proceedings. Biological sciences / The Royal Society*, **274**, 545-554.
81. Haak, W., Balanovsky, O., Sanchez, J. J., et al. 2010, Ancient DNA from European early Neolithic farmers reveals their near eastern affinities. *PLoS Biology*, **8**.
82. Malaspinas, A. S., Lao, O., Schroeder, H., et al. 2014, Two ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil. *Current biology : CB*, **24**, R1035-1037.
83. Green, R. E., Briggs, A. W., Krause, J., et al. 2009, The Neandertal genome and ancient DNA authenticity. *The EMBO journal*, **28**, 2494-2502.
84. Green, R. E., Krause, J., Ptak, S. E., et al. 2006, Analysis of one million base pairs of Neanderthal DNA. *Nature*, **444**, 330-336.
85. Green, R. E., Malaspinas, A. S., Krause, J., et al. 2008, A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. *Cell*, **134**, 416-426.
86. Golenberg, E. M. 1991, Amplification and analysis of Miocene plant fossil DNA. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **333**, 419-426; discussion 426-417.
87. Golenberg, E. M., Giannasi, D. E., Clegg, M. T., et al. 1990, Chloroplast DNA sequence from a miocene Magnolia species. *Nature*, **344**, 656-658.
88. Cano, R. J. and Poinar, H. N. 1993, Rapid isolation of DNA from fossil and museum specimens suitable for PCR. *BioTechniques*, **15**, 432-434, 436.
89. Cano, R. J., Poinar, H. N., Pieniazek, N. J., Acra, A. and Poinar, G. O., Jr. 1993, Amplification and sequencing of DNA from a 120-135-million-year-old weevil. *Nature*, **363**, 536-538.
90. Vreeland, R. H., Rosenzweig, W. D. and Powers, D. W. 2000, Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature*, **407**, 897-900.
91. Woodward, S. R., Weyand, N. J. and Bunnell, M. 1994, DNA sequence from Cretaceous period bone fragments. *Science*, **266**, 1229-1232.
92. Li, Y., An, C.-C. and Zhu, Y.-X. 1995, DNA isolation and sequence analysis of dinosaur DNA from Cretaceous dinosaur egg in Xixia Henan, China. *Acta Sci. Nat. Univ. Pekinensi* **31**, 148–152.
93. Allard, M. W., Young, D. and Huyen, Y. 1995, Detecting dinosaur DNA. *Science*, **268**, 1192; author reply 1194.
94. Hedges, S. B. and Schweitzer, M. H. 1995, Detecting dinosaur DNA. *Science*, **268**, 1191-1192; author reply 1194.
95. Henikoff, S. 1995, Detecting dinosaur DNA. *Science*, **268**, 1192; author reply 1194.
96. Zischler, H., Hoss, M., Handt, O., von Haeseler, A., van der Kuyl, A. C. and Goudsmit, J. 1995, Detecting dinosaur DNA. *Science*, **268**, 1192-1193; author reply 1194.
97. Graur, D. and Pupko, T. 2001, The Permian bacterium that isn't. *Molecular biology and evolution*, **18**, 1143-1146.
98. Austin, J. J., Ross, A. J., Smith, A. B., Fortey, R. A. and Thomas, R. H. 1997, Problems of reproducibility--does geologically ancient DNA survive in amber-preserved insects? *Proc Biol Sci*, **264**, 467-474.
99. Austin, J. J., Smith, A. B. and Thomas, R. H. 1997, Palaeontology in a molecular world: the search for authentic ancient DNA. *Trends in ecology & evolution*, **12**, 303-306.
100. Gutierrez, G. and Marin, A. 1998, The most ancient DNA recovered from an amber-preserved specimen may not be as ancient as it seems. *Molecular biology and evolution*, **15**, 926-929.

101. Nicholls, H. 2005, Ancient DNA comes of age. *PLoS Biol*, **3**, e56.
102. Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A. and Paabo, S. 2001, DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res*, **29**, 4793-4799.
103. Gilbert, M. T., Hansen, A. J., Willerslev, E., et al. 2003, Characterization of genetic miscoding lesions caused by postmortem damage. *American journal of human genetics*, **72**, 48-61.
104. Gilbert, M. T., Willerslev, E., Hansen, A. J., et al. 2003, Distribution patterns of postmortem damage in human mitochondrial DNA. *American journal of human genetics*, **72**, 32-47.
105. Hansen, A., Willerslev, E., Wiuf, C., Mourier, T. and Arctander, P. 2001, Statistical evidence for miscoding lesions in ancient DNA templates. *Molecular biology and evolution*, **18**, 262-265.
106. Briggs, A. W., Stenzel, U., Johnson, P. L. F., et al. 2007, Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 14616-14621.
107. Kolman, C. J. and Tuross, N. 2000, Ancient DNA analysis of human populations. *American journal of physical anthropology*, **111**, 5-23.
108. Noonan, J. P., Coop, G., Kudaravalli, S., et al. 2006, Sequencing and Analysis of Neanderthal Genomic DNA. *Science*, **314**, 1113-1118.
109. Dabney, J., Knapp, M., Glocke, I., et al. 2013, Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 15758-15763.
110. Gilbert, M. T. P., Tomsho, L. P., Rendulic, S., et al. 2007, Whole-Genome Shotgun Sequencing of Mitochondria from Ancient Hair Shafts. *Science*, **317**, 1927-1930.
111. Miller, W., Drautz, D. I., Ratan, A., et al. 2008, Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, **456**, 387-390.
112. Cooper, A. and Poinar, H. N. 2000, Ancient DNA: do it right or not at all. *Science*, **289**, 1139.
113. Knapp, M., Lalueza-Fox, C. and Hofreiter, M. 2015, Re-inventing ancient human DNA. *Investigative genetics*, **6**, 4.
114. Damgaard, P. B., Margaryan, A., Schroeder, H., Orlando, L., Willerslev, E. and Allentoft, M. E. 2015, Improving access to endogenous DNA in ancient bones and teeth. *Scientific reports*, **5**, 11184.
115. Burbano, H. a., Hodges, E., Green, R. E., et al. 2010, Targeted Investigation of the Neanderthal Genome by Array-Based Sequence Capture. *Science*, **328**, 723-725.
116. Poinar, H. N., Schwarz, C., Qi, J., et al. 2006, Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA. *Science*, **311**, 392-394.
117. Prüfer, K., Stenzel, U., Hofreiter, M., Pääbo, S., Kelso, J. and Green, R. E. 2010, Computational challenges in the analysis of ancient DNA. *Genome biology*, **11**, R47-R47.
118. Oskam, C. L., Haile, J., McLay, E., et al. 2010, Fossil avian eggshell preserves ancient DNA. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 1991-2000.
119. Poinar, H., Kuch, M., McDonald, G., Martin, P. and Pääbo, S. 2003, Nuclear gene sequences from a late Pleistocene sloth coprolite. *Current biology*, **13**, 1150-1152.
120. Adcock, G. J., Dennis, E. S., Easteal, S., et al. 2001, Mitochondrial DNA sequences in ancient Australians: Implications for modern human origins. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 537-542.
121. Briggs, A. W., Good, J. M., Green, R. E., et al. 2009, Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*, **325**, 318-321.
122. Lalueza-Fox, C., Gigli, E., de la Rasilla, M., et al. 2008, Genetic characterization of the ABO blood group in Neandertals. *BMC evolutionary biology*, **8**, 342-342.
123. Rasmussen, M., Guo, X., Wang, Y., et al. 2011, An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science*, **334**, 94-98.
124. Lacan, M., Keyser, C., Ricaut, F.-X., et al. 2011, Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 9788-9791.

125. Shapiro, B., Drummond, A. J., Rambaut, A., et al. 2004, Rise and Fall of the Beringian Steppe Bison. *Science*, **306**, 1561-1565.
126. Barnes, I., Matheus, P., Shapiro, B., Jensen, D. and Cooper, A. 2002, Dynamics of Pleistocene Population Extinctions in Beringian Brown Bears. *Science*, **295**, 2267-2270.
127. Lindahl, T., Ljungquist, S., Siegert, W., Nyberg, B. and Sperens, B. 1977, DNA N-glycosidases: properties of uracil-DNA glycosidase from Escherichia coli. *The Journal of biological chemistry*, **252**, 3286-3294.
128. Briggs, A. W., Stenzel, U., Meyer, M., Krause, J., Kircher, M. and Paabo, S. 2010, Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res*, **38**, e87.
129. Willerslev, E., Davison, J., Moora, M., et al. 2014, Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature*, **506**, 47-51.
130. Burbano, H. a., Green, R. E., Maricic, T., et al. 2012, Analysis of human accelerated DNA regions using archaic hominin genomes. *PloS one*, **7**, 1-8.
131. Kao, W. C., Stevens, K. and Song, Y. S. 2009, BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. *Genome Res*, **19**, 1884-1895.
132. Renaud, G., Kircher, M., Stenzel, U. and Kelso, J. 2013, freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers. *Bioinformatics*, **29**, 1208-1209.
133. Andrews, S. 2015, Babraham Bioinformatics. Babraham Institute.
134. Lindgreen, S. 2012, AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC research notes*, **5**, 337.
135. Martin, M. 2015, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**.
136. St. John, J. 2015, SeqPrep.
137. Schubert, M., Ginolhac, A., Lindgreen, S., et al. 2012, Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, **13**, 178.
138. DePristo, M. A., Banks, E., Poplin, R., et al. 2011, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, **43**, 491-498.
139. McKenna, A., Hanna, M., Banks, E., et al. 2010, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297-1303.
140. Li, H., Handsaker, B., Wysoker, A., et al. 2009, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.
141. Renaud, G. 2015, aLib: a sequencing pipeline for ancient and modern DNA.
142. Schubert, M., Ermini, L., Der Sarkissian, C., et al. 2014, Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature protocols*, **9**, 1056-1082.
143. Renaud, G., Stenzel, U. and Kelso, J. 2014, leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Res*, **42**, e141.
144. Kerpedjiev, P., Frellsen, J., Lindgreen, S. and Krogh, A. 2014, Adaptable probabilistic mapping of short reads using position specific scoring matrices. *BMC bioinformatics*, **15**, 100.
145. Ginolhac, A., Rasmussen, M., Gilbert, M. T., Willerslev, E. and Orlando, L. 2011, mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics*, **27**, 2153-2155.
146. Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. and Orlando, L. 2013, mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, **29**, 1682-1684.
147. Jakobsson, M. 2015, PMDtools. http://www.ieg.uu.se/jakobsson/software/pmdtools/.
148. Renaud, G., Slon, V., Duggan, A. T. and Kelso, J. 2015, Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol*, **16**, 224.
149. Lindgreen, S., Krogh, A. and Pedersen, J. S. 2014, SNPest: a probabilistic graphical model for estimating genotypes. *BMC research notes*, **7**, 698.
150. Gallego Llorente, M., Jones, E. R., Eriksson, A., et al. 2015, Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science*, **350**, 820-822.
151. Callaway, E. 2016, Error found in study of first ancient African genome. *Nature*.

152. Rivollat, M., Mendisco, F., Pemonge, M. H., et al. 2015, When the waves of European neolithization met: first paleogenetic evidence from early farmers in the southern paris basin. *PloS one*, **10**, e0125521.
153. Der Sarkissian, C., Balanovsky, O., Brandt, G., et al. 2013, Ancient DNA reveals prehistoric gene-flow from siberia in the complex human population history of North East Europe. *PLoS Genet*, **9**, e1003296.
154. Skoglund, P., Malmstrom, H., Omrak, A., et al. 2014, Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science*, **344**, 747-750.
155. Brotherton, P., Haak, W., Templeton, J., et al. 2013, Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. *Nature communications*, **4**, 1764.
156. Elhaik, E. 2012, Empirical Distributions of F_{ST} from Large-Scale Human Polymorphism Data. *PloS one*, **7**, e49837.
157. Alexander, D. H., Novembre, J. and Lange, K. 2009, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, **19**, 1655-1664.
158. Patterson, N. J., Moorjani, P., Luo, Y., et al. 2012, Ancient Admixture in Human History. *Genetics*.
159. Elhaik, E., Tatarinova, T., Chebotarev, D., et al. 2014, Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nature communications*, **5**, 3513-3513.
160. Kozlov, K., Chebotarov, D., Hassan, M., et al. 2015, Differential Evolution Approach to Detect Recent Admixture. *BMC Genomics*, **16**, S9.
161. Skotte, L., Korneliussen, T. S. and Albrechtsen, A. 2013, Estimating individual admixture proportions from next generation sequencing data. *Genetics*, **195**, 693-702.
162. Browning, B. L. and Browning, S. R. 2013, Improving the Accuracy and Efficiency of Identity by Descent Detection in Population Data. *Genetics*.
163. Price, A. L., Tandon, A., Patterson, N., et al. 2009, Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, **5**, e1000519.
164. Tang, H., Coram, M., Wang, P., Zhu, X. and Risch, N. 2006, Reconstructing genetic ancestry blocks in admixed individuals. *American journal of human genetics*, **79**, 1-12.
165. Limburg, P. A. and Weider, L. J. 2002, 'Ancient' DNA in the resting egg bank of a microcrustacean can serve as a palaeolimnological database. *Proc Biol Sci*, **269**, 281-287.
166. Jaenicke-Despres, V., Buckler, E. S., Smith, B. D., et al. 2003, Early allelic selection in maize as revealed by ancient DNA. *Science*, **302**, 1206-1208.
167. Lalueza-Fox, C., Gigli, E., de la Rasilla, M., Fortea, J. and Rosas, A. 2009, Bitter taste perception in Neanderthals through the analysis of the TAS2R38 gene. *Biol Lett*, **5**, 809-811.
168. Lynch, M., Xu, S., Maruki, T., Jiang, X., Pfaffelhuber, P. and Haubold, B. 2014, Genome-wide linkage-disequilibrium profiles from single individuals. *Genetics*, **198**, 269-281.
169. Gokhman, D., Lavi, E., Prufer, K., et al. 2014, Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science*, **344**, 523-527.
170. Llamas, B., Holland, M. L., Chen, K., Cropley, J. E., Cooper, A. and Suter, C. M. 2012, High-resolution analysis of cytosine methylation in ancient DNA. *PloS one*, **7**, e30226.
171. Pedersen, J. S., Valen, E., Velazquez, A. M., et al. 2014, Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res*, **24**, 454-466.
172. Smith, O., Clapham, A. J., Rose, P., Liu, Y., Wang, J. and Allaby, R. G. 2014, Genomic methylation patterns in archaeological barley show de-methylation as a time-dependent diagenetic process. *Scientific reports*, **4**, 5559.
173. Gerstein, M. B., Kundaje, A., Hariharan, M., et al. 2012, Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91-100.
174. Zakany, J. and Duboule, D. 2007, The role of Hox genes during vertebrate limb development. *Current opinion in genetics & development*, **17**, 359-366.
175. Hansen, K. D., Timp, W., Bravo, H. C., et al. 2011, Increased methylation variation in epigenetic domains across cancer types. *Nat Genet*, **43**, 768-775.

176. Berman, B. P., Weisenberger, D. J., Aman, J. F., et al. 2012, Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet*, **44**, 40-46.
177. Xie, W., Schultz, M. D., Lister, R., et al. 2013, Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, **153**, 1134-1148.
178. Nakamura, R., Tsukahara, T., Qu, W., et al. 2014, Large hypomethylated domains serve as strong repressive machinery for key developmental genes in vertebrates. *Development*, **141**, 2568-2580.
179. Jeong, M., Sun, D., Luo, M., et al. 2014, Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat Genet*, **46**, 17-23.
180. Abu-Remaileh, M., Bender, S., Raddatz, G., et al. 2015, Chronic Inflammation Induces a Novel Epigenetic Program That Is Conserved in Intestinal Adenomas and in Colorectal Cancer. *Cancer research*.
181. Kader, F. and Ghai, M. 2015, DNA methylation and application in forensic sciences. *Forensic science international*, **249**, 255-265.
182. Seguin-Orlando, A., Korneliussen, T. S., Sikora, M., et al. 2014, Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. *Science*, **346**, 1113-1118.
183. Hendrich, B. and Bird, A. 1998, Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Molecular and cellular biology*, **18**, 6538-6547.
184. Crawford, G. E., Holt, I. E., Whittle, J., et al. 2006, Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res*, **16**, 123-131.
185. Tatarinova, T., Kryshchenko, A., Triska, M., et al. 2013, NPEST: a nonparametric method and a database for transcription start site prediction. *Quant Biol*, **1**, 261-271.
186. Troukhan, M., Tatarinova, T., Bouck, J., Flavell, R. B. and Alexandrov, N. N. 2009, Genome-wide discovery of cis-elements in promoter sequences using gene expression. *OMICS*, **13**, 139-151.
187. Menotti, F. and O'Sullivan, A. 2013, *The Oxford handbook of wetland archaeology*. Oxford University Press: Oxford, United Kingdom.
188. Kharlamova A, S. S., Kurtova A, Chernikov V, Protopopov A, Boeskorov G, Plotnikov V, Ushakov V, Maschenko E 2014, Preserved brain of the Woolly mammoth (Mammuthus primigenius (Blumenbach 1799)) from the Yakutian permafrost. *Quaternary International*.
189. Zeng, J., Konopka, G., Hunt, B. G., Preuss, T. M., Geschwind, D. and Yi, S. V. 2012, Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *American journal of human genetics*, **91**, 455-465.
190. Chopra, P., Papale, L. A., White, A. T., et al. 2014, Array-based assay detects genome-wide 5-mC and 5-hmC in the brains of humans, non-human primates, and mice. *BMC Genomics*, **15**, 131.
191. ArunKumar, G., Tatarinova, T. V., Duty, J., et al. 2015, Genome-wide signatures of male-mediated migration shaping the Indian gene pool. *J Hum Genet*.
192. Kushniarevich, A., Utevska, O., Chuhryaeva, M., et al. 2015, Genetic Heritage of the Balto-Slavic Speaking Populations: A Synthesis of Autosomal, Mitochondrial and Y-Chromosomal Data. *PloS one*, **10**, e0135820.
193. Yunusbayev, B., Metspalu, M., Metspalu, E., et al. 2015, The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet*, **11**, e1005068.
194. Flegontov, P., Changmai, P., Zidkova, A., et al. 2016, Genomic study of the Ket: a Paleo-Eskimo-related ethnic group with significant ancient North Eurasian ancestry. *Scientific reports*, **6**.

Tables

Table 1. Difficulties of working with ancient DNA and specialized methods developed to encounter them.

Problem	Experimental solutions	Bioinformatics solutions
Degradation	Improved extraction protocols Using NGS approach (*catches short DNA fractions*)	Algorithms based on genotype likelihoods rather than a single best genotype for low coverage genomic positions.
Base damage	Using DNA Polymerases which does not amplify through uracil (*remove uracil- containing fragments from the reaction*) Treatment with Uracil-N-glycosylase plus Endonuclease VIII (*removes uracil, then repairs abasic sites*) Single-primer Extension PCR (*analyses separate DNA strands*)	Trimming 5-7 bases from read ends Counting and excluding C→T and G→A mutations at ultra-conserved positions Apparently accelerated evolution on branches leading to ancient samples Comparing frequencies of different classes of transitions in modern-modern and modern-ancient alignments Estimation of contamination or divergence based on indels and transversions only, not transitions Exclusion of common ancestor-ancient sample branches from calculation of divergence
Contamination	Special protocols for sample collection, transport and storage Special pre-digestion steps (*including mechanical and chemical decontamination, short-time pre-incubation*) Independent replication in two labs PCR-capture with species-specific primers	Exclusion of extremely long reads or alignments (in case of 454 or Sanger sequencing) Phylogenetic correctness (exclusion of reads based on similarity with non-target species; inclusion of reads based on similarity with the target species or a close relative) Conformity to species- or ethnicity-specific variants or haplotypes Unique individual origin of reads from one specimen: homozygosity of X and Y positions in male specimens, absence of Y reads in female specimens, homozygosity of mtDNA positions Absence of haplotypes present mainly in unrelated specimens or research team members Distinguishing mtDNA sequences from NUMTs

The solutions aimed at one or more of the problems are not mutually exclusive and are often used in combination for better results. Also, various bioinformatics ideas for tackling contamination and base damage are sometimes integrated into a single Maximum Likelihood framework for base and genotype calling.

1041 Table 2. Software packages, pipelines and tools suitable for aDNA analysis

Program name	Short description	Available at
BayesCall	Base calling algorithm for high-throughput short-read sequencing	http://bayescall.sourceforge.net/
AdapterRemoval	Searches and removes remnant adapter sequences from High-Throughput Sequencing data	https://github.com/MikkelSchubert/adapterremoval
ADMIXTOOLS	A software package that supports formal tests of whether admixture occurred, and makes it possible to infer admixture proportions and dates.	https://github.com/DReichLab/AdmixTools
ADMIXTURE	Software tool for maximum likelihood estimation of individual ancestries from multilocus SNP genotype datasets	https://www.genetics.ucla.edu/software/admixture/
aLib	Sequencing pipeline for ancient and modern DNA	https://github.com/grenaud/aLib
aLib/ leeHom	Bayesian reconstruction of ancient DNA fragments	https://github.com/grenaud/leehom
Anfo	Reads mapper similar to Soap/Maq/Bowtie	https://bioinf.eva.mpg.de/anfo/
BEAGLE	Software package that performs genotype calling, genotype phasing, imputation of ungenotyped markers, and identity-by-descent segment detection	https://faculty.washington.edu/browning/beagle/beagle.html
BWA-PSSM	Probabilistic short read aligner	http://bwa-pssm.binf.ku.dk/
CutAdapt	Tool to find and remove adapter sequences	https://cutadapt.readthedocs.org
FastQC	Quality control tool for high throughput sequence data	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
FreeIbis	An efficient basecaller for Illumina sequencers with calibrated quality scores	https://github.com/grenaud/freeIbis
GPS	Sample Provenance Predictor	http://chcb.saban-chla.usc.edu/gps/
HAPMIX	Software for identifying ancestry segments in admixed individuals	http://www.stats.ox.ac.uk/~myers/software.html
MIA	Consensus calling (or "reference assisted assembly"), chiefly of ancient mitochondria	https://github.com/mpieva/mapping-iterative-assembler
NGSADMIX	Tool for finding admixture proportions from NGS data	http://www.popgen.dk/software/index.php/NgsAdmix
NPEST	de-novo TSS prediction	http://link.springer.com/article/10.1007%2Fs40484-013-0022-2
PALEOMIX	User-friendly package for largely automates the analyses related to whole genome re-sequencing.	https://github.com/MikkelSchubert/paleomix
reAdmix	Prediction of provenance for individuals of recently admixed origin	http://chcb.saban-chla.usc.edu/reAdmix/
SABER	Reconstructing genetic ancestry blocks in admixed individuals	http://med.stanford.edu/tanglab/software/saber.html
SeqPrep	Stripping adaptors and/or merging paired reads with overlap into single reads	https://github.com/jstjohn/SeqPrep
TSSer	de-novo TSS prediction	http://online.liebertpub.com/doi/abs/10.1089/omi.2008.0034
TREXML	A maximum likelihood method for inferring phylogenetic trees	No longer distributed
PUZZLE	A maximum likelihood method for inferring phylogenetic trees	http://www.tree-puzzle.de/
NETWORK	Parsimony method to construct phylogenetic trees and networks	http://www.fluxus-engineering.com/sharenet.htm
TFPGA	Tools for population genetic analysis of allozymes and other molecular markers	http://www.marksgeneticsoftware.net/tfpga.htm
Arlequin	A tool for population genetic data analysis	http://cmpg.unibe.ch/software/arlequin35/
MEGA	A tool for molecular evolutionary genetic analyses (phylogenetic trees, distances, substitution models, etc.)	http://www.megasoftware.net/
GenePop	A tool for population genetic data analysis	http://genepop.curtin.edu.au/
MUSCLE	Multiple sequence alignment	http://www.ebi.ac.uk/Tools/msa/muscle/
RAXML	A method for maximum likelihood phylogenetic tree reconstruction	http://sco.h-its.org/exelixis/web/software/raxml/index.html

PAUP	Tools for phylogenetic reconstruction including distance methods, parsimony and maximum Likelihood	http://paup.csit.fsu.edu/
MRBAYES	Bayesian inference of phylogeny	http://mrbayes.sourceforge.net/
PAML	A tool for maximum likelihood phylogenetic tree reconstruction	http://abacus.gene.ucl.ac.uk/software/paml.html
DnaSP	Tools for DNA sequence polymorphism data within and between populations plus tests of neutrality	http://www.ub.edu/dnasp/
BayeSSC	Tool for Bayesian inference based on coalescent simulations	http://web.stanford.edu/group/hadlylab/ssc/index.html

042

Figure legends

Fig. 1 Location of aDNA finds

Fig. 2 Schema of bioinformatics pipeline for aDNA analysis from NGS sequencing

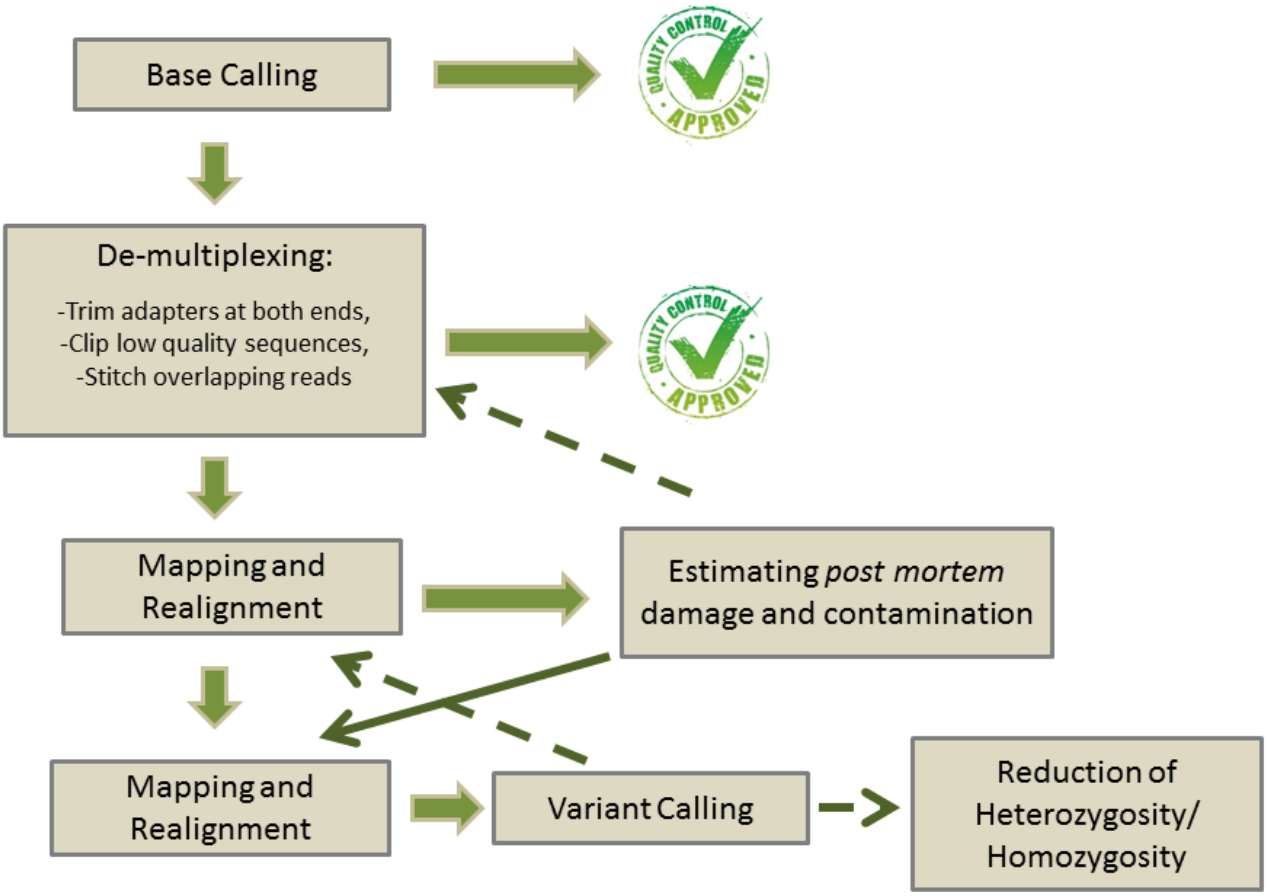
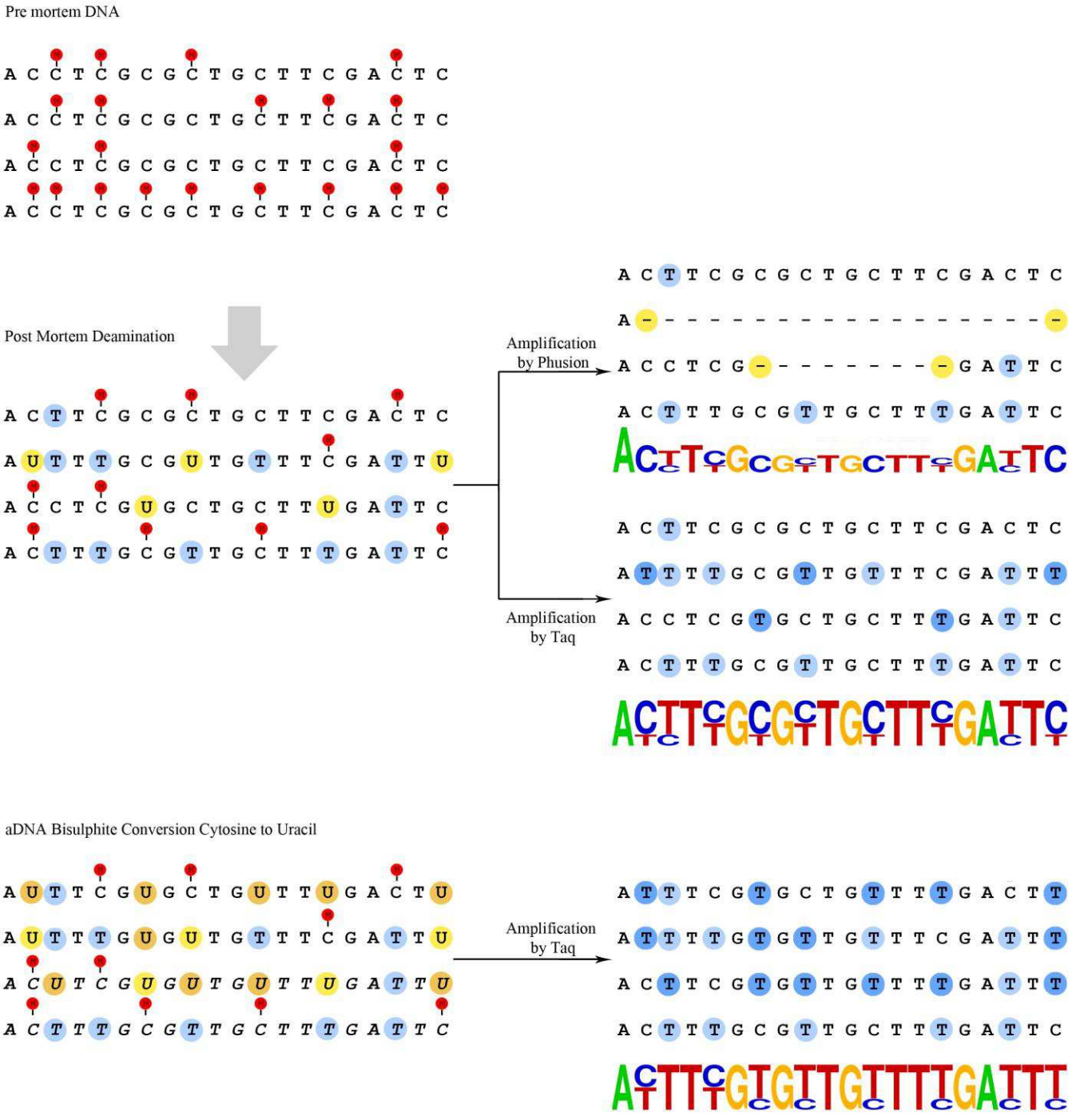


Fig. 3. ???


Pre mortem DNA
Post Mortem Deamination
Amplification by Phusion
Amplification by Taq
aDNA Bisulphite Conversion Cytosine to Uracil
Amplification by Taq