METHODS OF aDNA ANALYSIS

Methods of aDNA analysis are usually essentially the same as of contemporary DNA. One has only to remember that aDNA is usually highly degraded. Its degradation depends on its age and conditions (temperature, humidity) prevailing at the site where fossil materials were found. Nevertheless, aDNA molecules are usually severely fragmented and bear chemical modifications due to a number of post-mortem processes, such as enzymatic activity or spontaneous chemical decay (Fig. 2). Signal from such degraded genetic material can be easily outcompeted by just a few contaminating molecules of long-chain intact modern contaminant DNA. Therefore, researchers dealing with aDNA should take several precautions against contamination of the sample. Usually, samples subjected to aDNA analyses have much lower content of the actual aDNA than of the contaminating modern DNA coming from the environment as well as from the person (persons) who handled the sample. It is therefore important that the extraction, amplification and all other manipulations in the case of aDNA should be performed in a dedicated aDNA facility. Such aDNA laboratory is usually a space physically separated from laboratories in which modern DNA work is performed. The space is often UV irradiated, which destroys surface and airborne DNA molecules, hence minimising the risk of contamination. Access to such laboratory is strictly limited to people performing the experiment. They should wear sterile lab coats, mouth masks and gloves (Fig. 3). The access to the laboratory should be possible only through the double doors, separated by the space where the sterile laboratory coats, face masks and shoes are kept.

Material used for aDNA analysis usually consists of teeth and bones. However, aDNA has been successfully extracted from such sources as hair,

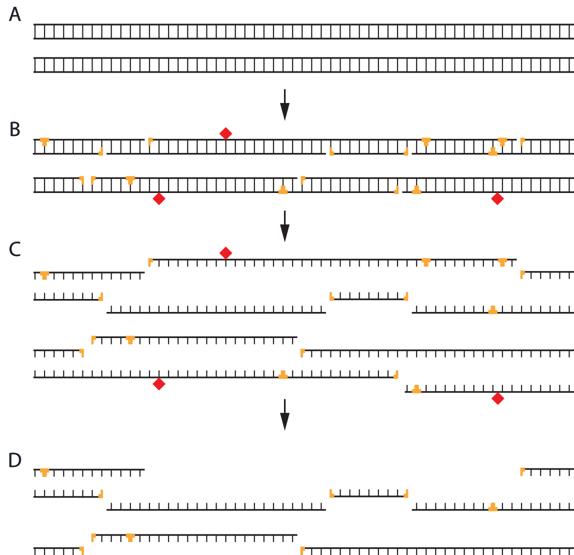

Fig. 2. An overview of the DNA damage processes. A) Long intact DNA molecules, after the death of their host cells, are subjected to extensive decay. B) The molecules that do not get digested by the release of lytic enzymes during cell death or by microorganisms preying on the decomposing body over time accumulate strand breaks (represented by gaps in the illustrated DNA molecule strands) and chemical lesions occurring spontaneously (red diamonds and yellow markings). C) DNA fragments available for DNA library building are fragmented and contain chemical modifications, some of which prevent further analysis by, e.g., blocking the enzymes used during DNA library building or DNA sequence reading (such lesions represented here by red diamonds). D) Only a portion of DNA molecules are sequenced. These are usually short and carry chemical modifications resulting in incorrect read of a nucleotide during DNA sequencing (such lesions represented here by yellow markings). Such lesions more frequently occur near the strand breaks and usually result in a spurious T (thymine) read instead of an original C (cytosine).

Created by Martyna Molak

mummified soft tissue, feathers, eggshells, soil or even coprolites (subfossilized faeces). By the nature of their abundance, bones are the most common source of ancient DNA. Successful extraction of nucleic material was reported even from bone over 700,000 years old (Orlando et al., 2013). One always has to take into consideration the fact that in bones the levels of contaminating microbial DNA are high with most bones having over 99% exogenous DNA. However, the analytical power obtained by increasing the

Fig. 3. Ancient DNA laboratory. The researcher preparing bone samples for DNA extraction is wearing protective gear to minimize the risk of contamination of the sample with modern DNA. The work is performed in a laminar flow cabinet situated in a dedicated separated aDNA facility.

Photo credit: Martyna Molak private collection

number of sequence reads to millions of independent DNA sequences, significantly reduced problems with contamination of ancient samples with modern DNA. The field improved further with the realization that hair constitutes a remarkable source of high-quality aDNA that could be subjected to efficient decontamination procedures inapplicable to bones. DNA can be also extracted from eggshells. aDNA recovered from eggshells is relatively rich in endogenous DNA with bacterial DNA contamination much lower than in bones. Other sources of aDNA are coprolites and soil but its extraction is difficult and the aDNA is usually highly modified and degraded.

To extract the minute amounts of endogenous DNA, the surface of the samples is usually either physically removed by sanding or at least decontaminated with UV light and hypochloride solution which destroy the

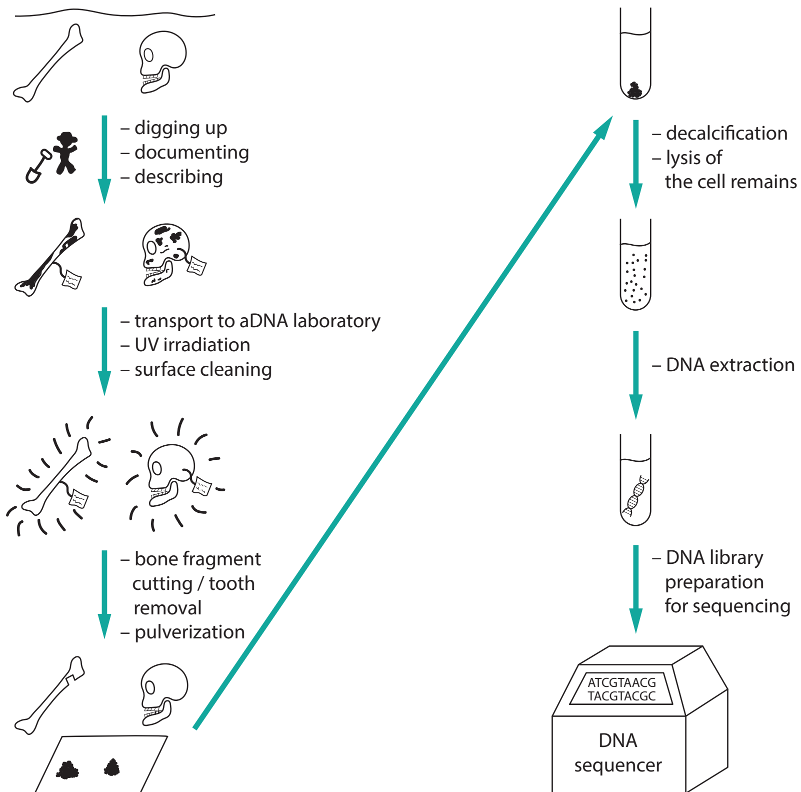

Fig. 4. The streamline of aDNA analysis. The organic material for the aDNA analysis often comes from archaeological or paleontological excavations. After the archaeological, paleontological and/or anthropological examination and description, the chosen pieces of the material get transported to the aDNA laboratory for genetic analyses. The surface of the material usually gets washed with hypochloride (bleach) and/or UV-irradiated to maximally remove any surface modern DNA contamination. A target fragment of the sample – usually the dense bone section which prevents post-mortem DNA damage the most – gets pulverized. The powder is then treated with chemical reagents that release DNA molecules from the remains of cells and tissue structures. The released DNA is extracted using a method selected from a range of available methods depending on the type and condition of the material. The purified DNA extract is further processed to obtain DNA sequences for genetic analyses. Usually this process involves turning DNA molecules into a DNA library, which allows amplification of the whole DNA content of the extract, which making the molecules available for detection, and sequencing (reading the DNA nucleotide sequence in individual molecules) using an Illumina platform.

Created by Martyna Molak

predominantly contaminant surface DNA. Next, to enable the reagents to interact with the material, the sample needs to be pulverized or fragmented into small pieces at least. For bone and teeth material, special mills are often used to crush the hard tissue into powder, which is then usually decalcified using chelating solution (e.g. EDTA). The remnants of cell membranes are disrupted by surfactants and proteins get digested by proteolytic enzymes. DNA is extracted from such lysate using one of the range of methods based i.a. on DNA solubility in different media (phenol:chloroform method), on hydrophobic bonding between DNA and silica (silica solution and silica column methods) or on DNA binding to magnetic surfaces under certain conditions (magnetic beads extraction). The resulting pure DNA extract is then subjected to PCR amplification of target genetic markers or to DNA library preparation, which is curation of the DNA molecules in a way that enables amplification and sequencing of whole, yet predominantly short, DNA molecules in the extract (also called DNA immortalization). See Figure 4 for illustration of the streamline of obtaining genetic information from ancient materials.

Genetic markers

To estimate the range of genetic differences or similarities between individuals or populations, we rely on genetic markers. As markers, various types of genetic polymorphisms are used, including single nucleotide polymorphisms (SNPs), simple sequence length polymorphisms (SSLPs), and restriction fragment length polymorphisms (RFLPs). SSLPs include repeat sequences, variations known as minisatellites (variable number of tandem repeats, or VNTRs), microsatellites (simple tandem repeats, STRs) and insertions/deletions (indels). In the human genome, the most common types of markers are SNPs, STRs and indels. SNPs affect only one of the bases in a fragment of DNA. STRs are markers in which a piece of sequence is repeated several times in a row, and the number of repeats is variable within and across individuals. Indels are polymorphisms in which a piece of DNA sequence exists in some individuals within population (insertion allele), while it is absent in others (deletion allele).

Until the rapid progress in the DNA sequencing techniques occurred, viruses and bacterial plasmids were the major object of DNA studies.

This was simply because of their small size and the ease of their isolation. In case of higher organisms, mitochondrial DNA (mtDNA) and Y chromosome DNA were intensively studied for the same reasons. Each cell contains hundreds to thousands of mitochondria, which are located in the cytoplasm. In humans, mitochondrial DNA is composed of 16,569 bp and contains 37 genes, all of which are essential for normal mitochondrial function. Thirteen of these genes code for enzymes involved in oxidative phosphorylation. The remaining genes code for transfer RNA (tRNA) and ribosomal RNA (rRNA), the molecules involved in protein synthesis.

Recently, the study of aDNA has been greatly enhanced by the development of second-generation DNA sequencing technologies. The HTS (high-throughput sequencing) technique allows for sequencing of huge aDNA libraries on platforms such as the Illumina NovaSeq. These techniques are affordable even for small laboratories, although many labs decide to send their libraries to be sequenced by specialized companies like Beijing Genomics Institute. It is possible to use the HTS technique for sequencing of only one chosen locus in the whole population. This approach is named "metabarcoding" and it is applied for sequencing samples taken from lake water, sediments etc. The main advantage of the metabarcoding is the possibility to simultaneously analyze large numbers of samples at a limited cost.

According to Gnerre et al. (2011), massively parallel DNA sequencing technologies are revolutionizing genomics by making it possible to generate billions of relatively short (~100-base) sequence reads at very low cost. They presented an algorithm (ALLPATHS-LG) for genome assembly of human and mouse genomes, generated on the Illumina platform. The authors claim that combination of improved sequencing technology and improved computational methods makes it possible to increase dramatically the *de novo* sequencing of large genomes. This might be particularly useful for organism for which no close relatives have their genome sequenced, which i.a. is relevant to many extinct species studied using aDNA.

Mitochondrial DNA

For a long time, mitochondrial DNA was the primary target of the ancient DNA research. This is because mitochondrial genome is copied multiple

Fig. 5. The routes of dispersal of maternal lineages (as represented by mitochondrial genome haplogroups) during human Out-of-Africa expansion. Major haplogroups and the place of their dominance in the modern-day populations are indicated by haplogroup names (single letters, sometimes followed by a digit). Inferred routes and timing of particular expansions are indicated by arrows and their coloring. The most recent common ancestor of all modern-day mitochondrial genomes (named "mitochondrial Eve" belonging to macrohaplogroup L) lived probably in Southern Africa around 200 kya. From there and then the descendant lineages were spreading along with human migrations. Part of lineage L3 left Africa around 60–70 kya and populated the rest of the world, reaching South-East Asia and Australia around 50 kya, Central Asia only slightly after 50 kya, Western Europe around 40 kya and the Americas by 14 kya. The distribution of mitochondrial haplogroups worldwide mostly reflects the route of human migrations and thus learning one's haplogroup can the ancestral story of that person's maternal lineage.

Author: Maulucioni CC by 3.0 Source: https://commons.wikimedia.org/wiki/File:Human_migrations_and_mitochondrial_haplogroups.PNG

times in each cell and therefore targeting any genetic marker located on the mitochondrial genome increases the probability of successful amplification as compared to targeting nuclear markers. The non-coding fragment of mitochondrial genome – the D-loop – due to low evolutionary constraints on mutation rate is highly variable in most species and thus is often used in intraspecific genetic structure analyses. Other mitochondrial markers, such as cytochrome b, are more conserved and serve as useful markers for analyses involving multiple species. Moreover, mitochondrial DNA is generally strictly maternally inherited which on the one hand means that it bears only a part of the individual's genetic history but on the other hand makes it much easier to trace genealogies with and hence to interpret the results based on mitochondrial as compared to autosomal markers. The abovementioned properties of mitochondrial DNA made it a marker of choice not only for ancient DNA research but also for a large part of phylogenetic and population genetics research in general.

Early DNA-based human population genetics research, although pioneered by Luca Cavalli-Sforza's analyses of various genetic markers in the 1980s, has been predominantly based on mitochondrial data. Mitochondrial lineages have been arising along the course of human migrations which started probably somewhere in South Africa and resulted in peopling almost every piece of land on Earth and hence their uneven geographical distribution reflects the history of these migrations. The lineages have been grouped into haplogroups according to shared ancestry and assigned alphabet letter labels for easy identification. Thanks to mitochondrial marker analyses, the routes as well as timing of the expansion and diversification of particular lineages have been reconstructed (Fig. 5). Inferences made on mitochondrial data analyses remained a benchmark for further genomic analyses.

Y chromosome

The human Y chromosome is a structure comprising around 60 Mb (million base pairs), and it expresses 45 proteins, some associated with sex and fertility. Apart from these genes, the Y chromosome houses multiple repetitive sequences without a known function, which can be deleted without

apparent phenotypic consequence. Genes on the Y chromosome fall into two distinct categories: those expressed throughout the body in all cell types, and those expressed predominantly or solely in the testes most probably involved in testis development and/or spermatogenesis. One of these genes (*Sry*), codes for the regulatory protein, controlling expression of sex determining genes. The *Sry* gene is required to turn on *Sox9* expression to develop male gonads. XX females lack the *Sry* region, and hence fail to activate *Sox9*, and develop ovaries instead. Evolution and structure of Y chromosome was reviewed by Graves (2006).

In the majority of the aDNA studies from ancient fossils mtDNA was used. Progress in techniques of DNA extraction and sequencing made it possible to turn towards analysis of nuclear DNA. Here, the Y chromosome appeared to be utterly useful and thoroughly investigated particularly in studies of human populations and individuals. Kivisild (2017) provided an extensive review of the Y chromosome studies. The genetic diversity of the Y chromosome is highest within and between African populations. Analysis of this diversity allowed distinguishing several clades. Genetic variation outside Africa consists mainly of three clades which coalesce to a single founding lineage within 40,000–60,000 years ago, consistent with the Out of Africa (OOA) dispersal model (Stringer, 2002). In addition to these three clades, the populations of the Near East and Europe also show the presence of young sub-clades of another haplogroup, which probably reflects recent episodic gene flow from Africa within the last 25,000 years. The oldest ancient Y chromosomes that have been sequenced at sufficiently high coverage to allow for phylogenetic mapping are from northern parts of Eurasia. Nine old ancient Eurasian Y chromosome sequences obtained by Seguin-Orlando et al. (2014) can be assigned to one of the three major lineages that have been inferred from the analyses of the extant Y variation. Two of the oldest human Y chromosomes sequenced so far, Ust'Ishim Man (Fu et al., 2014) and the Oase Man (Fu et al., 2015), are both placed near the root of haplogroup K, which is the most frequent Y chromosome lineage present today. K is an ancestral group for a number of regional haplogroups, which today are frequent in Europe, East Asia, Oceania, and Americas (Fig. 6).

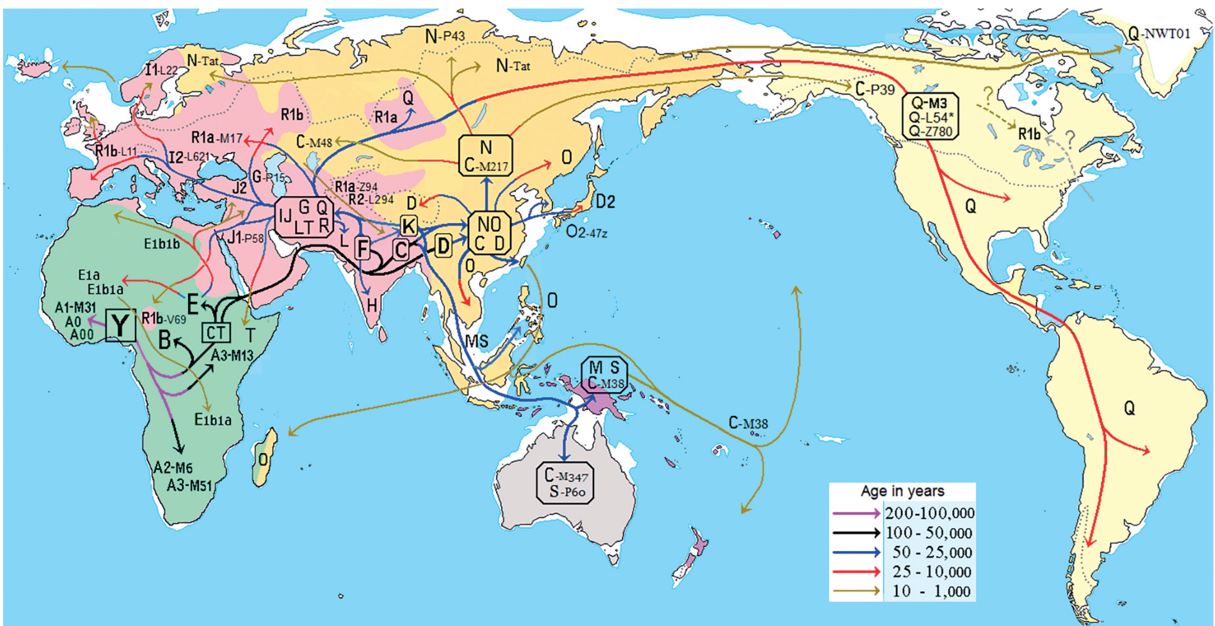

Fig. 6. The routes of dispersal of paternal lineages (as represented by Y chromosome haplogroups) during human Out-of-Africa expansion. Major haplogroups and the place of their dominance in the modern-day populations are indicated by haplogroup names (alphanumeric, sometimes accompanied by a name of an SNP determining the lineage). Inferred routes and timing of particular expansions are indicated by arrows and their coloring. The most recent common ancestor of all modern-day Y chromosomes (named "Y-chromosomal Adam", labelled here as "Y") lived probably in Central-Northwest Africa around 200 kya. The routes of dispersal are mostly concordant with mitochondrial DNA (Fig. 5) at a large scale, although some details differ noticeably. The distribution of Y-chromosomal haplogroups worldwide mostly reflects the route of human migrations and thus learning one's haplogroup can the ancestral story of that person's paternal lineage.

Author: Maulucioni CC by 3.0

Source: https://commons.wikimedia.org/wiki/File:Migraciones_humanas_en_haplogrupos_de_ADN-Y.PNG)