Available online at www.sciencedirect.com
ScienceDirect
C. R. Palevol 7 (2008) 113–124
COMPTES RENDUS PALEVOL
http://france.elsevier.com/direct/PALEVO/

General palaeontology (Taphonomy and Fossilisation)

Palaeogenomics

Michael Hofreiter

MPI for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany

Received 27 June 2007; accepted after revision 10 December 2007
Available online 4 March 2008

Written on invitation of the Editorial Board

Abstract

In many ways, palaeogenomics began when the first ancient DNA sequence was reported. This first sequence was derived from a stuffed museum specimen of the quagga, an extinct mammal related to the zebra. Unspecified and unselected DNA was extracted from the quagga specimen, cloned into a bacterial library, and then sequenced. It took another 17 years and the development of PCR before two independent groups successfully sequenced the complete mitochondrial genomes from several extinct moa species. Only 4 years later, using the original approach of cloning nonspecific ancient DNA extract and shotgun sequencing, the first ancient nuclear DNA sequences were determined, this time from the extinct cave bear. Since these early successes, palaeogenomics has rapidly expanded, because of both technological development and increasing interest in ancient DNA research. New methods, developed since the cave bear sequence was reported, have produced nuclear DNA on a megabase scale from two extinct species, the mammoth and the Neanderthal, our closest relative. For both species, low-coverage genome-sequencing projects have been proposed. It is likely that these will be successful, given the rapid technical development in sequencing techniques. This review carefully examines both the promise and the current limitations of palaeogenomic analyses for both mitochondrial and nuclear DNA. ***To cite this article: M. Hofreiter, C. R. Palevol 7 (2008).***

Résumé

Paléogénomique. La publication de la première séquence d'ADN ancien peut être considérée comme le début de la paléogénomique. Cette première séquence a été obtenue à partir d'un spécimen, taxidermisé et conservé dans un musée, un quagga, mammifère éteint et apparenté au zèbre. L'ADN a été extrait à partir de ce spécimen de quagga, cloné dans une bibliothèque bactérienne et puis séquencé sans qu'aucune séquence n'ait été particulièrement sélectionnée. Il a fallu attendre 17 ans et des progrès de la technique de PCR avant que deux groupes réussissent à séquencer indépendamment le génome mitochondrial complet de plusieurs espèces de moa, éteintes elles aussi. À peine quatre ans plus tard, en utilisant l'approche originale de clonage sans sélection d'un extrait d'ADN ancien suivi par une étape de clonage aléatoire, les premières séquences nucléaires anciennes ont été déterminées, cette fois-ci à partir de l'ours de cavernes éteint. Depuis ces réussites initiales, la paléogénomique a rapidement pris son essor, conséquence, d'une part, d'une évolution technologique et, d'autre part, d'un intérêt accru pour la recherche sur l'ADN ancien. Des nouvelles méthodes développées depuis la publication de la séquence de l'ours des cavernes ont permis la production de séquences d'ADN nucléaire à l'échelle de la mégabase à partir de deux espèces éteintes, le mammouth et l'homme de Neandertal, notre parent le plus proche. Pour les deux espèces, des projets de séquençage de faible couverture ont été proposés. Il est probable que ces projets

E-mail address: hofreite@eva.mpg.de.

1631-0683/$ – see front matter © 2007 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.crpv.2007.12.005

aboutiront, compte tenu du rythme de développement rapide des techniques de séquençage. Cette revue analyse soigneusement à la fois les promesses et les limites actuelles des analyses paléogénomiques de l'ADN mitochondrial et nucléaire. ***Pour citer cet article : M. Hofreiter, C. R. Palevol 7 (2008).***

Keywords: Ancient DNA; Palaeogenetics; Palaeogenomics; Proboscidians; Neanderthal

Mots clés : ADN ancien ; Paléogénétique ; Paléogénomique ; Proboscidiens ; Neandertal

1. Historical background and techniques for ancient DNA and palaeogenomics

The study of ancient DNA has a very interesting history [35,49]. Since its start, results have been published in the most prestigious scientific journals (e.g., [13,33]). At the same time, however, the field has witnessed the publication of completely erroneous results, such as the report of dinosaur DNA sequences [50], which was discredited very soon after its publication [51]. The study of ancient DNA has even been declared dead more than once (e.g., [4]), despite its relatively short history. Although this verdict may be true for certain aspects of ancient DNA research, such as the amplification of DNA sequences from many-million-year-old amber specimens [1], recent years have seen a proliferation of high-quality ancient DNA studies and the field's expansion into a number of new areas, such as population genetics over time (e.g., [25,45]), palaeogenomics (e.g., [32,38], also [5,12]) and functional palaeogenetics [42], to name just a few. This review focuses on one of the most recent developments, the establishment of palaeogenomics (Fig. 1) as a research field in its own right.

Genomics generally describes the sequencing and analysis of a substantial part of a given organism's nuclear genome. Given this definition, palaeogenomics should focus solely on analyzing nuclear DNA from extinct organisms. Even though the nuclear genome comprises the vast majority of any eukaryotic organism's genome, one should not forget that animals also contain a second – albeit much smaller – genome, the mitochondrial one. Plants even have a third genome, contained in their plastids. Due to the much higher copy number (several hundreds to several thousands) of mitochondrial (mt) and plastid (cp) DNA contained in living cells, as compared to nuclear (nu) DNA's two copies, mtDNA and cpDNA sequences are much easier to analyze from ancient specimens. Consequently, most ancient DNA studies to date have concentrated on mtDNA or cpDNA sequences, although there were a number of earlier nuDNA studies (e.g., [3,11,19,21]). Since nu, mt, and cpDNA comprise independent genomes, this discussion will include ancient DNA studies that analyze the complete mitochondrial genome (mitogenomics) of extinct species (there are no such studies on plastid genomes to date). At the same time, with regard to the nuclear genome, only suitable techniques for such analyses will be examined, such as shotgun sequencing, since the largest study published to date analyzed less than 0.5% of an extinct species' nuclear genome [38].

Within these parameters, we return to the very beginning of ancient DNA analyses, the sequencing of short mtDNA fragments from the extinct quagga [13], a subpopulation of the plains zebra [23]. These first ancient DNA results were produced by extracting DNA from the tissues of a museum specimen, cloning this DNA into phages and then sequencing two clones which were identified by hybridization with mountain zebra mtDNA [13]. The two clones were markedly similar to horse mtDNA sequences and were thus identified as endogenous quagga sequences. This result was – and remains – remarkable in a number of ways. First, it showed that DNA survives for considerable time after the death of an organism. The last quagga died at the Amsterdam zoo in 1883, and although it could not be determined which one of the three individuals from the Museum of Natural History in Mainz was used in the above study (R. Rau, personal communication), the animal died at least 140 years before the ancient DNA analysis was performed. Second, DNA not only survived, it did so in considerable amounts, as the technique applied is rather insensitive and requires comparatively large quantities of DNA. Third, it is likely that many of the clones contained quagga DNA fragments. Although between hundreds to thousands of copies of the mitochondrial genome exist for every copy of the nuclear genome in a living cell, given that the nuclear genome is 200,000 times longer than the mitochondrial one, one would expect between 40 and 1,000 clones containing nuDNA for every clone containing mtDNA (see also below). Considering that Higuchi and colleagues screened about 25,000 phage plaques and found two clones containing mtDNA, a substantial number of the remaining clones (between 80 and 2,000 at least, or between 0.3 and 10%) probably contained nuclear quagga sequences. This corresponds to a

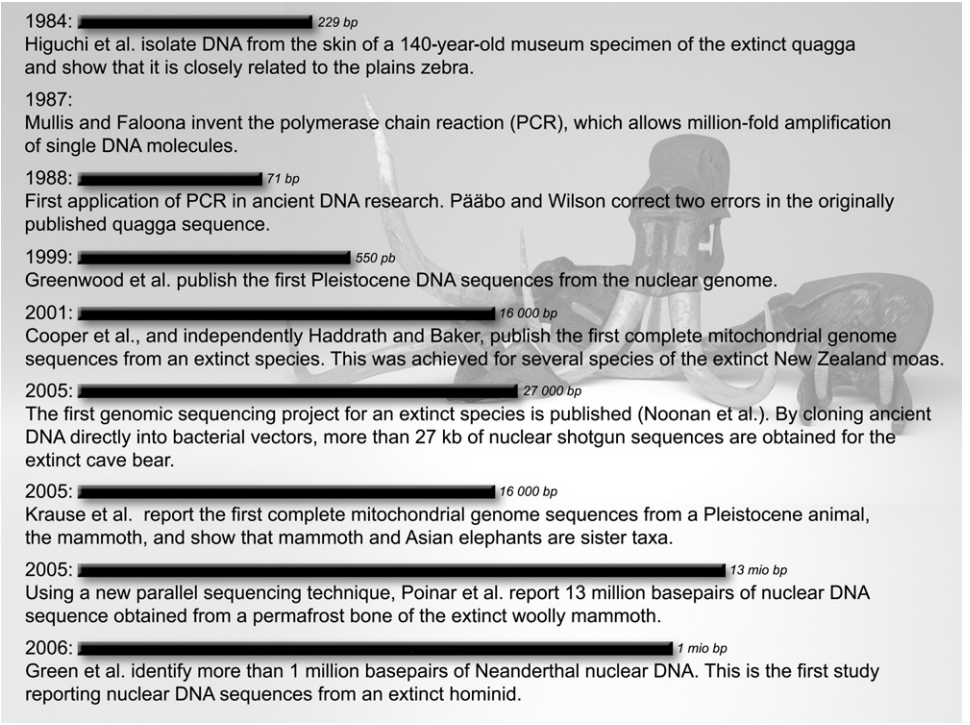


Fig. 1. Major achievements in ancient DNA research that have influenced palaeogenomics, including the sequence length obtained in the different studies. Sequence lengths are shown on a log-scale.
Fig. 1. Principaux résultats de la recherche dans le domaine de l'ADN ancien ayant contribué à la paléogénomique, et longueur des séquences obtenues dans le cadre de chacune de ces études (en échelle logarithmique).

ratio of endogenous DNA between 0.3% and 10%, similar to the ratios found in modern palaeogenomic studies [10,31,32,38]. In fact, by screening the 25,000-phage plaques using horse satellite DNA, Higuchi and colleagues also identified clones containing nuclear quagga DNA. Thus, palaeogenomics, in a way, began at the same time as research on ancient DNA.

A similar sequencing approach as that taken by Higuchi and colleagues was independently applied to ancient human DNA [33]. Pääbo also extracted DNA from an organism that had died long ago and selectively cloned the longer DNA fragment into bacterial colonies. However, instead of an animal, he used a human mummy from Egypt, some 2400 years old. Moreover, instead of searching for mtDNA as in the quagga study, he screened about 1000 of the colonies obtained for a repetitive sequence from the human nuclear genome, the ALU element family. In this way, he managed to identify a 3.4-kb piece of nuDNA, which contained two copies of the ALU element. He sequenced about 900 bp of the insert, including one of the ALU elements. Given the generally degraded nature of ancient DNA, this result was quite remarkable. Four years after its publication, questions were raised about whether the resulting DNA was actually modern human contamination, rather than from the mummy itself [7]. The current knowledge about ancient DNA degradation [8,26,34,38,39] and the pervasive problems with modern human DNA contamination (e.g., [17,27,44,48]) suggest that this may indeed have been the case. The result, however, could also have been authentic, due to either exceptional DNA preservation or the artificial creation of the 3.4-kb insert via ligation of several smaller pieces during vector construction.

The general problem with these early studies lies in the fact that for technical reasons, the results could not be reproduced [18]. Thus, although correct in a broad sense, the quagga results also contained some errors. The originally published DNA sequences differed at two positions from sequences of the plains zebra; both differences resulted in amino acid substitutions at positions conserved across all vertebrate sequences known at that time [14]. Unfortunately, it was impossible to replicate the obtained sequences then, since genomic shotgun sequences only represent a random sampling of the genome. Therefore, substantial sequencing efforts would be necessary to obtain the same sequence again. For example, in a recent study that produced 13 million basepairs of mammoth DNA sequences using the shot-

gun method, the mitochondrial genome was not even completely sequenced once, despite its relatively small size and high copy number [38]. This situation changed in 1987, when the polymerase chain reaction (PCR, [29]) was developed, which allows for the targeted sequencing of any DNA sequence. Using this technique, Pääbo and Wilson [36] were able to show that both sequence substitutions from the original quagga sequences were artefacts, most likely due to damage to the ancient DNA template. Interestingly, the two substitutions were a C to T and a G to A substitution, both changes that can be caused by cytosine deamination, a common type of DNA damage in living organisms, and one of the most rapidly occurring types of in-vitro DNA damage [26]. Moreover, this type of damage had been suspected early on [34] and had later been confirmed [15] to also affect ancient DNA (Fig. 3). As discussed below, this poses a substantial problem for palaeogenomic analyses using shotgun-sequencing techniques.

2. Amplification of complete mtDNA genomes from extinct species

Given the advantages of PCR – high sensitivity, targeted sequence recovery – genomic library screening was no longer used for studying ancient DNA and PCR became the exclusive method used. Consequently, the first mitochondrial genomic sequences were also recovered using PCR – 17 years after the publication of the first ancient DNA sequences. These mitochondrial genomes came from a number of moa species [5,12], extinct flightless birds from New Zealand, related to the living ratites (basal flightless birds), i.e. the emu, rheas, kiwis, cassowaries, ostrich, and the extinct Malagasy elephant bird. Moas were among the first extinct species from which mtDNA sequences were determined and their phylogenetic relationship to their living relatives investigated [6]. At that time, it had already been discovered that their DNA was quite well preserved, with fragments of up to 438 bp being amplifiable. However, the authors of this initial study only determined about 1000 bp of DNA sequences, and with this limited amount of data, they were not able to resolve confidently the phylogenetic position of the moas. This only became possible almost ten years later with the analyses of almost complete mtDNA genome sequences by two independent research groups [5,12]. The resulting genomic sequences enabled the confident placement of moas within the ratite phylogenetic tree, as well as allowing various divergence events to be dated. For the first time, inferences about both the time and mode of population separation could be drawn from ancient genomic sequences. Thus, the divergence dates for the different ratite lineages allowed for re-evaluating the timing of the Gondwana continent's gradual separation into the modern southern continents, which, according to these data, started about 90 million years ago [5,12].

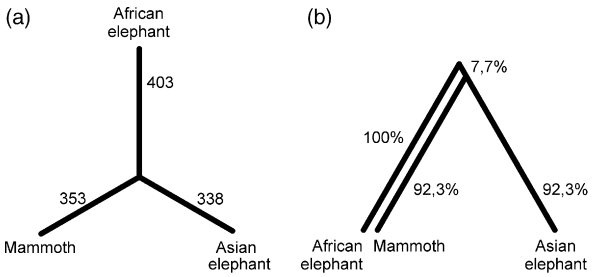


Fig. 2. Phylogenetic reconstructions for elephantidae (mammoth, African and Asian elephant), after [22], but with corrected mammoth sequence (see [41]). (**a**) An unresolved trifurcating tree based on complete mitochondrial genomes. The numbers indicate the absolute number of substitutions on each branch as inferred by parsimony. (**b**) The same tree as in (**a**), but resolved by midpoint rooting, which supports the basal position of the African elephant with more than 99% statistical support. Numbers show the relative lengths of external and internal branches in percentage.

Fig. 2. Relations phylogénétiques au sein des Elephantidae (mammouth, éléphants d'Afrique et d'Asie), d'après [22], en corrigeant cependant la séquence du mammouth (voir [41]). (**a**) Arbre à trois branches non résolu, fondé sur le génome mitochondrial complet. Les chiffres indiquent le nombre absolu de substitutions sur chacune des branches, tel que l'indique le calcul de parcimonie. (**b**) Même arbre qu'en (**a**), mais cette fois résolu par un point médian d'enracinement ; il soutient, avec 99 % de vraisemblance statistique, la position de base de l'éléphant d'Afrique ; les chiffres indiquent les longueurs relatives des branches externes et internes, en pourcentage.

Mitogenomics was again picked up several years later with the publication of mitochondrial genomes from the extinct woolly mammoth, again by two independent groups [22,41]. The first of these studies used an innovative type of PCR, a two-step multiplex approach. In this approach, a large number of primer pairs are mixed together, and the targeted fragments are pre-amplified in the first step, using a low – about 30 – number of PCR cycles. The reaction is then diluted, and each fragment is amplified individually using another 30–40 cycles with either the same primers as in the first step or nested primers [22,40]. In this way, it became possible to amplify and replicate the 16,700 bp of the mitochondrial genome using a fossil extract containing as little as 200 mg bone powder. The second study used conventional PCR from an exceptionally well-preserved mammoth sample and which yielded a very similar sequence [41]. Both studies used the resulting sequences to clarify the hotly debated relationship between the mammoth and the living Asian and

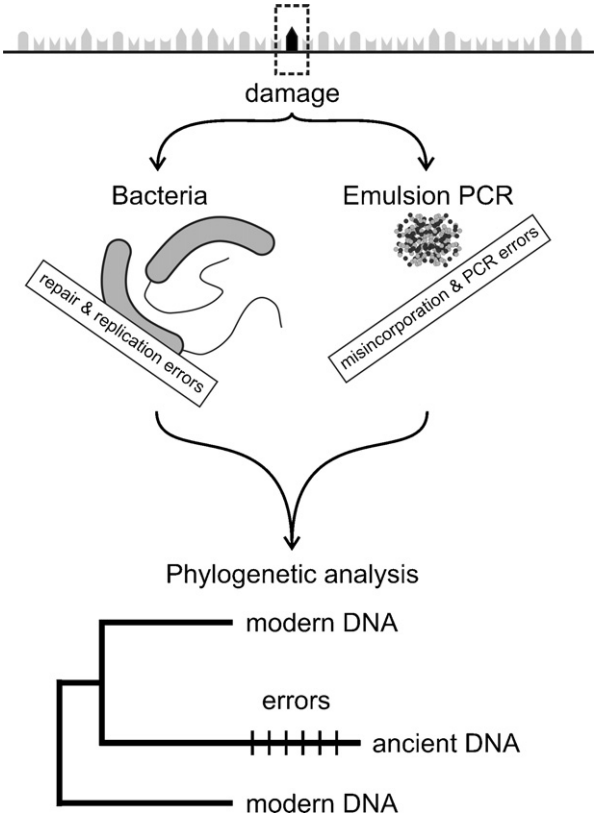


Fig. 3. Mechanisms that introduce errors into non-replicated ancient DNA sequences obtained either by reconstruction of a bacterial library (left column) or 454 sequencing (right column), and the effect on phylogenetic trees reconstructed from such sequences. Many ancient DNA templates contain damaged nucleotides such as uracil, a deamination product of cytosine, a regular base in DNA. Upon cloning into bacteria, such errors may become repaired to restore the original condition, but sometimes the modified nucleotide may be used as template for replication without prior repair, leading to errors in the sequence. Bacteria may insert further differences in the original DNA sequence due to errors during DNA replication (left column). During emulsion PCR, damaged nucleotides can result in misincorporations, resulting in changed DNA sequences. For example, uracil pairs with adenine instead of guanine, the partner of the original cytosine, the precursor of uracil prior to deamination. Replication errors during PCR may introduce further changes (right column).

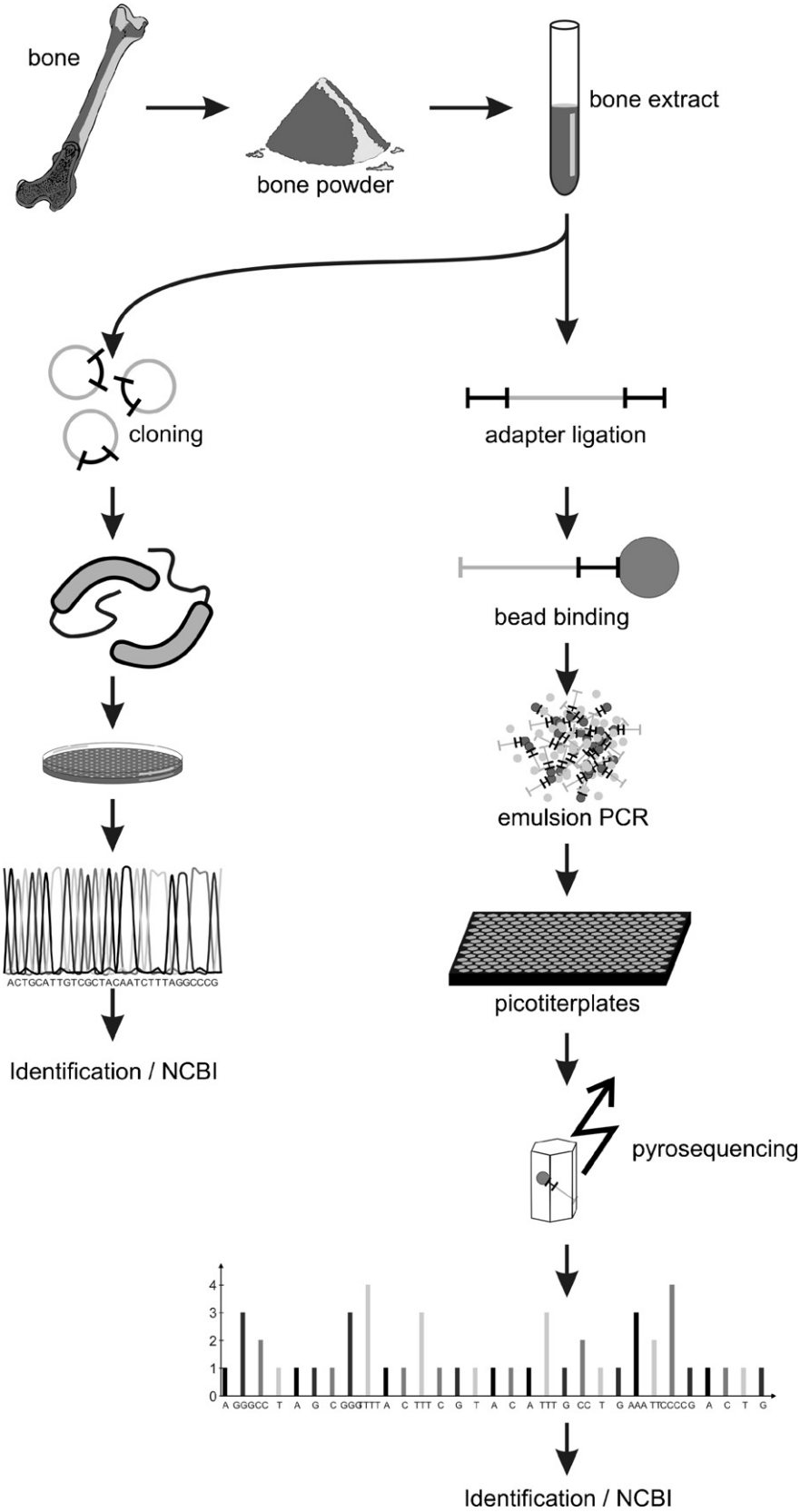
Fig. 3. Mécanismes responsables des erreurs de non-réplication des séquences d'ADN ancien intervenant au cours du clonage bactérien (colonne de gauche) ou du séquençage 454 (colonne de droite), et leurs effets sur les arbres phylogénétiques correspondants. Beaucoup d'extraits d'ADN ancien contiennent des nucléotides endommagés, tels que l'uracile, qui résulte de la déamination de la cytosine, base naturelle de l'ADN. Lorsqu'on procède par clonage bactrien, de telles erreurs peuvent être réparées, mais parfois le nucléotide modifié peut être utilisé comme base de réplication sans avoir été réparé, ce qui conduit à des erreurs de séquence. Les bactéries peuvent aussi induire d'autres différences par rapport à la séquence d'ADN originelle, en raison d'erreur de réplication (colonne de gauche). Au cours de l'amplification par PCR, les nucléotides endommagés peuvent provoquer des erreurs d'incorporation, entraînant elles-mêmes des modifications de séquences. C'est par exemple le cas lorsque l'uracile s'apparie avec l'adénine au lieu de le faire avec la guanine, comme l'aurait fait une cytosine avant de subir la déamination qui l'a transformée en uracile. Les erreurs de réplication au cours de la PCR peuvent aussi induire d'autres modifications de séquence (colonne de droite).

African elephants. Until the publication of these studies, many molecular studies using short DNA fragments had argued for a sister-group relationship between mammoth and African elephants, contrary to most morphological studies, which have argued that the Asian elephant is more closely related to the mammoth. The mitogenomic sequences eventually resolved this controversy in favour of the morphological view [22,41].

Together, these studies also vividly illustrate the importance of determining each sequence position at least twice from independent primary PCRs [15,17,35]. As pointed out by Rogaev and colleagues, the sequence from Krause and colleagues contained an unusually large number of amino acid substitutions in part of the ND2 gene. Re-sequencing the original PCR products for this fragment only, together with sequencing additional amplification products from the same region, showed that in the initial study, the same PCR product had inadvertently been sequenced twice, thus giving the false impression of replication. The additional sequences showed that this error had led to as many as eight consistent substitutions (Krause et al., unpublished data), seven of them being C to T substitutions, most likely due to cytosine deamination [15]. Correction of this sequence fragment resulted in an even more symmetrical phylogenetic tree than in the original publication (Fig. 2). The correction also provided greater support for the sister-group relationship between mammoth and Asian elephant, increasing the results of the maximum likelihood test from 98.8% to 99.9%. Moreover, the internal branch length, representing the time span between the two divergence events relative to the total length of the tree, increased slightly from 7.3% to 7.7%. These results show that, at least with regard to the mitochondrial genome, mammoths clearly shared a more recent common ancestor with Asian elephants than either species did with African elephants.

3. Nuclear palaeogenomics

The ability to resolve difficult phylogenetic questions, such as the positions of the moas or mammoth, illustrates the power of mitogenomic analyses in ancient DNA research. However, the data do not always need to be obtained via PCR. Just two days after the first


bone
bone powder
bone extract
cloning
adapter ligation
bead binding
emulsion PCR
picotiterplates
pyrosequencing
ACTGCATTGTCGCTACAATCTTTAGGCCCG
Identification / NCBI
Identification / NCBI

mammoth mitochondrial genome was published online [22], another almost complete mammoth mitochondrial genome was published, also online [38]. Unlike the other two, it was not produced by sequencing PCR products, but by a recently developed shotgun sequencing technique ([28], see below). However, this was not even the first shotgun genomic sequencing study to be published. Even before the first mtDNA genome of a Pleistocene species was published, Noonan and colleagues [32] identified almost 27,000 bp of nuclear DNA sequences from two 40,000-year-old cave bear fossils. They had returned to the original approach used for sequencing ancient DNA; they directly cloned the ancient DNA into a bacterial vector and constructed a genomic library (Fig. 4). In their study, they used DNA extracts from two cave bear fossils, a standard extract from a cave bear tooth of about 200 mg and a large-scale extract, concentrated from about 40 g of bone. Although both extracts yielded cave bear sequences, a closer look at the results readily illustrates several drawbacks of this technique. First, to obtain 389 clones that contained cave bear DNA, they had to sequence about 10,000 clones from the first library and 5000 clones from the second, as only about 1% and 6%, respectively, of the clones contained inserts with cave bear DNA sequences. The remaining clones contained bacterial, fungal and unknown sequences, which is common for such metagenomic approaches [47]. However, these results show that almost impossibly large numbers of clones would be necessary to produce just 1x coverage of an extinct genome. Second, as with any shotgun technique, the fragments obtained represent random pieces of the genome. Therefore, it is impossible to sequence specific regions, for example to isolate certain genes that may have a phenotypic effect. Moreover, if two individuals from the same or closely related species are sequenced using this technique, no overlap between the sequence data sets exists. Thus, comparable data sets for two or more individuals cannot be obtained. Third, every fragment is sequenced only once, resulting in incorrect sequence positions. This effect became immediately obvious when Noonan and colleagues reconstructed a phylogenetic tree using the cave bear data and corresponding sequences from several modern bear species. The branch leading to the cave bear showed a significantly longer branch length as well as an increased ratio of C to T and G to A substitutions, in comparison to all other substitutions. Both effects are expected when ancient DNA templates contain miscoding lesions due to cytosine deamination, but are absent in replicated data sets obtained using PCR [10,15]. Interestingly, the effect disappeared when two sequence fragments that contained three and four C to T changes, respectively, were removed from the data set, and the remaining clone sequences were then compared to the corresponding brown bear sequences. Similar problems were also observed in larger palaeogenomic data sets (see below).

While sequencing 27 kb of nuclear DNA from a Pleistocene species was no doubt a major achievement, it pales in comparison to the 13 million bp of mammoth DNA sequence published only half a year later [38]. This amount of sequence data was obtained using a radically different technique, developed to rapidly analyze extremely large amounts of shotgun sequences, e.g. for the sequencing of complete bacterial genomes [28]. Similar to the construction of a genomic library, the ancient DNA extract is used directly without prior PCR amplification (Fig. 4). Using several steps, two different oligonucleotide linkers are added to the 5′ and 3′ ends of the template DNA, respectively, and only one strand of each template fragment is processed further. This makes it possible to also obtain very detailed information on miscoding lesions, as complementary substitutions such as C to T versus G to A substitutions can be differentiated, which is not possible with PCR amplified data

Fig. 4. Schematic diagram showing the process of obtaining genomic ancient DNA sequences using genomic libraries (left) and direct 454 sequencing (right), respectively. For both processes, ancient DNA extract obtained from bone (or any other source) is used. For library construction, DNA fragments are blunt end repaired, ligated into plasmids, and a bacterial culture is transformed. Many thousand individual bacteria are grown to colonies, the plasmids are isolated from each colony, and the sequences of the inserts are determined, usually using Sanger sequencing. For 454 sequencing, double stranded DNA adaptors are ligated to both ends of the ancient DNA fragments, the fragments are bound individually to small beads, and each DNA portion bound to a single bead is amplified separately in water-in-oil droplets in an emulsion PCR. Following that, the beads are loaded onto a picotiterplate and the sequence for each fragment is determined using pyrosequencing. For more information, see main text.

Fig. 4. Schéma montrant le processus d'obtention de séquences génomiques d'ADN ancien à partir de bibliothèques génomiques (à gauche) et à partir d'une séquençage 454 direct (à droite). Dans les deux cas, on utilise un extrait d'ADN obtenu à partir d'un os (ou de toute autre source). Pour le clonage bactérien, les extrémités endommagées des fragments d'ADN sont réparées, puis chimiquement liées dans des plasmides, tandis qu'une culture bactérienne est réalisée. Plusieurs milliers de bactéries sont multipliées en colonies, les plasmides sont isolés de chaque colonie et les séquences des inserts sont déterminées, en général par séquençage Sanger. Lors d'un séquençage 454, des amorces d'ADN double brin sont liées chimiquement aux deux extrémités des fragments d'ADN ancien, les fragments sont attachés individuellement à de petites billes et chaque portion d'ADN ainsi accrochée à chaque microbille est amplifiée séparément par PCR dans une émulsion de gouttelettes d'huile en milieu aqueux. Ensuite, les billes sont chargées sur une plaque comportant des puits de dimension picométrique, et la séquence de chaque fragment est déterminée par pyroséquençage. Pour plus d'information, voir le texte.

sets ([46]; see below). After the linkers have been added to the template DNA, the fragments are amplified in an emulsion PCR. In this technique, a water-in-oil emulsion is produced and the water droplets are used as micro reactors, ideally, each of which will amplify only one template DNA fragment using primers complementary to the linkers. This prevents the different templates from competing during the amplification, which would be the case with a regular PCR in solution. Moreover, the PCR is performed in a way that causes one strand of the products to be covalently linked to micro beads, which is necessary for the sequencing step. Thus, after the emulsion PCR, hundreds of thousands of beads are obtained, each coated with single-stranded DNA originating from a single template molecule. These beads are then put onto a picotiterplate that contains more than one million wells. The DNA fragments are sequenced using pyrosequencing, thus yielding about a million sequence reads. For technical reasons, 'only' about 200,000–300,000 wells per run produce high-quality sequences of about 100-bp length each. Pyrosequencing uses the fact that during DNA synthesis, one pyrophosphate molecule is produced per added nucleotide. Using several enzymes, including luciferase, the released pyrophosphate eventually leads to the emission of light, which can be measured. As in pyrosequencing only one of the four nucleotides is added each time, light will only be emitted from the wells in which the specific nucleotide can be added to the template. In the next sequencing cycle, a different nucleotide is added. Thus, after four cycles, all four nucleotides have been added and the process starts from the beginning. This way, with one sequence run, about 20–30 million bp of DNA sequences can be obtained. In fact, the new generation of machines allows the sequencing of almost 100 million bp, as they have both a longer read length (about 200 bp) and allow denser loading of the picotiterplate (see www.454.com).

Poinar and colleagues used this technique to analyze an unusually well preserved mammoth bone. Previous quantitative PCR analyses had shown that this sample contained about 10 million copies of mitochondrial DNA fragments of about 100 bp in length per gram of bone, as well as measurable quantities of mtDNA fragments up to 900 bp in length [38]. Moreover, the bone originated from a permafrost environment, and it had been stored in an ice cave after excavation. It was thawed only immediately before DNA extraction, giving microorganisms little chance to contaminate the bone with their DNA. Consequently, when the authors compared the obtained sequences to the publicly available African elephant genome, they found close matches for more than 40% of the reads. Although the elephant genome is incomplete, they could extrapolate that more than 50% of their sequences were most likely derived from mammoth DNA [38], a remarkable ratio compared to the 1% and 6% obtained in the earlier cave bear study. Nevertheless, their technique suffers from the same shortcomings as sequencing a genomic library. For example, despite 13 million bp of mammoth DNA sequence, Poinar and colleagues were unable to reconstruct a phylogenetic tree for mammoth and the two living elephant species, since the corresponding sequences for the Asian elephant are not available. Substantial efforts would need to be undertaken to amplify sufficient DNA fragments from the Asian elephant in order to develop a meaningful phylogeny for the three species. Moreover, the sequence data are likely to contain a significant number of sequence errors due to template damage. The severity of this problem was illustrated by an analysis of a smaller mammoth DNA sequence data set produced using the same technique [46], as well as by comparison of mtDNA sequences obtained from additional sequence runs performed on the Poinar samples [9]. Both groups found a significant excess of C to T substitutions compared to all other types of substitutions and also – albeit of a smaller magnitude– an excess of G to A substitutions. While the former type of substitutions can be explained chemically by cytosine deamination, a conclusive biochemical explanation for the observed G to A substitutions is currently lacking. However, Stiller and colleagues found, under conditions similar to those employed in emulsion PCR, that guanine to xanthine deamination might account at least for some of the G to A substitutions, since a thymine is incorporated opposite to a xanthine at least in some of the cases [46]. Whether this can account for the comparatively large number of G to A substitutions in the data is another – yet unresolved – question.

4. Problems and extensions of palaeogenomics studies

Regardless of the reasons for the template DNA damage, it is critical to recognize that shotgun sequence data contain a substantial number of errors, large enough to affect adversely data analyses. This problem is probably best illustrated by one of the most recent palaeogenomic studies, the analysis of about 1 million bp of Neanderthal DNA sequences [10]. These data were aligned and compared to the genome sequences of human and chimpanzee. When the number of substitutions on the human and Neanderthal lineages was compared, it was found that since the split of humans and Neanderthals, apparently ten times more substitutions had occurred on the Neanderthal lineage. As a 10-fold acceleration of the

substitution rate on the Neanderthal lineage compared to the human lineage is rather unlikely, the best explanation for this result is a high error rate in the Neanderthal sequences due to both template damage and sequencing errors. Thus, if we assume that the evolutionary rates on the human and the Neanderthal lineage are the same, then nine out of ten substitutions observed on the Neanderthal lineage are inaccurate, due to either damaged template DNA or sequencing errors. Although it is possible to correct for the overall number of substitutions, as done by Green and colleagues, assessing the reliability of individual substitutions is almost impossible [10]. Therefore, studies that aim to investigate substitutions specific to extinct species must rely on verification via PCR. This can be done, as shown by Green and colleagues for mitochondrial DNA data ([10], see below), as well as by Römpler and colleagues for a complete nuclear gene from an extinct species [42].

The principal feasibility of this approach had already been demonstrated in 1999, when short nuclear DNA sequences from Pleistocene samples were obtained for the first time ([11], Fig. 1). Although these early sequences were uninformative with regard to their function or phenotypic effect, in 2003, sequences obtained from several-thousand-year-old maize samples revealed evidence of positive selection at certain sites linked to domestication [21]. As recently shown, it is even possible to sequence complete nuclear genes from Pleistocene specimens, several tens of thousands of years old. This was achieved for the complete MC1R gene from the mammoth [42]. To prevent DNA damage from influencing the results, the variable sites detected were subsequently confirmed by replication via both multiplex PCR and SNP typing in additional individuals. Expression of the detected alleles even allowed functional analyses of the mammoth gene variants, despite the mammoth's extinction at least 4,000 years ago.

Due to the fragmented nature of ancient DNA, SNP typing is probably the best method for verifying positions of potential interest detected in palaeogenomic studies, although it also suffers from some limitations. Thus, if very short fragments are amplified, the sequences are no longer informative about species origin. A piece of 10 bp may well be sufficiently preserved across all mammalian species to make it impossible to assign species origin. Since PCR primers accept some mismatches, it is possible to amplify at least some conserved sequence fragments from any mammalian species using the same primer pairs. Thus, contamination of mammoth or Neanderthal bones or of PCR reagents with cattle or pig DNA [24] could easily lead to false inferences, if SNP typing is used as the sole method of analyses.

However, in their Neanderthal study, Green and colleagues were not so much interested in positions specific to Neanderthals, but rather in estimating the divergence time between humans and Neanderthals. They reported a surprisingly young age of just 500,000 years. As sequence divergence always predates population divergence [30], human and Neanderthal populations must have diverged even more recently. A similar result was obtained in a second study analyzing a smaller data set of Neanderthal genomic sequences. Using another aliquot of the extract Green and colleagues sequenced, Noonan and colleagues constructed a genomic library as in the cave bear study [31]. This approach produced about 60,000 bp of Neanderthal sequences. Their reconstructed human–Neanderthal sequence divergence date centred around 700,000 years, but the confidence intervals of the two studies widely overlap. They also calculated a population divergence date, at 370,000 years B.P. Taking the two studies together, it is clear that the population divergence obtained from these data is more recent than palaeoanthropological studies suggest. Whether this is true or an artefact of either sample contamination – although unlikely since both studies tested for contamination in the extract – or gene flow from humans to Neanderthals, as speculated by Green and colleagues, will most likely be revealed further in later studies.

Finally, it should be noted that the estimates from both studies have very large confidence intervals. This is a general problem of phylogenetic and population genetic studies, as many of the parameters used for obtaining molecular dates can only be estimated with limited precision. Both studies used the divergence of humans and chimps for calibration of the molecular clock, which in itself has a large confidence interval, and varying substitution rates across the genome or selection on parts of it may further influence the results of such estimates. However, while this is a problem shared with any study on phylogenetics or population genetics, increasing amounts of data should ameliorate at least some of the problems in reconstructing the Neanderthal–modern human divergence in the future.

Green and colleagues also obtained about 1/5 of the Neanderthal mitochondrial genome. Interestingly, when using the raw shotgun data together with a large number of human sequences and chimpanzee as an outgroup, the reconstructed phylogenetic tree had a much longer branch leading to Neanderthals than the one leading to humans, similar to the nuclear data [10]. To investigate this effect further, they multiplex amplified all positions unique to the Neanderthal lineage or shared only between Neanderthals and chimpanzees to verify or disprove their status. Strikingly, they could verify only 7

out of 20 Neanderthal specific positions, whereas they could verify 13 of 14 positions shared between Neanderthals and chimpanzees. Using these replicated data, the branch lengths of the human and Neanderthal lineages are no longer different, which again stresses the importance of sequence replication with ancient DNA. Moreover, this result shows that although it is difficult to obtain reliable Neanderthal specific positions using low-coverage shotgun sequences, such data can be used to identify positions that have changed on the human lineage after humans and Neanderthals separated. These human specific positions can then be further investigated to determine whether they may have contributed to human specific phenotypes [10].

It is worthwhile to note, however, that for such analyses two closely related reference genomes of high quality are necessary. Ideally, one of these genomes comes from a sister taxon of the extinct species (such as humans to Neanderthals), whereas the second reference genome should come from a basal species common to the other two (such as chimpanzee to the genus *Homo*). Only then is it possible to identify gene positions that are identical between Neanderthals and chimpanzees, but different in humans, and thus infer changes on the human lineage following their separation from Neanderthals. Conversely, this information is also crucial for tentatively identifying positions that changed on the Neanderthal lineage. For example, it is currently not possible to assign genetic changes occurring on the mammoth lineage that are possibly associated with its adaptation to an arctic climate, because only a draft sequence from the African elephant is available. Consequently, it is only possible to identify positions that differ between the two species, but it is impossible to determine which of these changed on the mammoth lineage. However, given the decreasing costs of genome sequencing, sequences from basal species, such as sea cow or the more closely related Asian elephant may become available within the next couple of years. Although the problems of template damage and sequencing errors will always limit the information that can be deduced about extinct species from low-coverage genomic sequences, such sequences have the potential to teach us a lot about closely related living species, which in the case of Neanderthals is about ourselves.

5. Conclusions and perspectives

Although having its roots back to 1984, palaeogenomics in its own right is a very young research field. The analysis of complete mitochondrial genomes has shown great potential for resolving the phylogenetic position of extinct species, including dating critical divergence events using molecular clock approaches [5,12,22,41]. Hopefully, future studies will help to clarify the phylogenetic relationships between other interesting groups, such as the extinct ground sloths or the enigmatic South American endemic Litopterans. Its potential and utility for population genetic analyses as demonstrated with modern humans [20] has yet to be evaluated.

However, mtDNA comprises just a tiny fraction of a species' complete genome; the vast majority of DNA is encoded in the nuclear genome. Moreover, mtDNA is inherited strictly maternally in almost all vertebrates. Thus, analyses of the nuclear genome have greater potential for providing information in several important respects. First, if we want to investigate adaptive changes that resulted in certain phenotypes (e.g., long hair in woolly mammoth compared to the living elephants), the nuclear genome, rather than the mitochondrial one, is the place to look at. Second, the complete population history, including the male lineage, can only be revealed if Y-chromosomal and, ideally, autosomal sequences are studied in addition to mtDNA sequences. Third, nuclear sequences may inform us about the mode and timing of past speciation events, as recently shown for humans and chimpanzees [37]. Thus, nuclear sequences show great promise for ancient DNA analyses.

Unfortunately, it is not yet clear whether palaeogenomics can actually realize its potential, as several problems limit its applicability. First, non-replicated shotgun sequences contain a large number of errors. Thus, studying potential adaptations will always require confirmation of sequence positions via either PCR or high-coverage genomic sequencing, although the latter approach will most likely remain too expensive in the near future. Therefore, for studying potential adaptations, a candidate gene approach, as recently taken for the study of mammoth hair colour [42], may be more appropriate. The large error rate also limits the potential for using palaeogenomic sequences in the identification of SNPs for population genetic studies, as recently suggested [38], since many SNPs will actually be false positives. Yet, as no other clearly better method exists for this purpose, this may indeed be a reasonable application, while keeping in mind that SNP typing in ancient DNA may suffer from problems of contamination, as mentioned above. Second, due to the random sampling of the genome by shotgun sequencing, palaeogenomics is not in itself suitable for population studies. This is unfortunate, because only ancient DNA provides direct evidence of how the genetic composition of populations changes over time, making it one of the most interesting aspects of ancient DNA research (e.g., [2,16,25,43,45]). Third, the high costs of producing complete genomic sequences

prevent extending this approach to large numbers of samples. This problem is exacerbated with ancient DNA, as most samples will have only a small proportion of endogenous sequences between 1% and 6% [10,31,32], rather than the 50% recovered for the analyzed mammoth bone [38]. Thus, the amount of sequence data necessary to obtain similar coverage is 20–100 times higher for ancient samples than for modern DNA. Even as the costs for genomic sequencing rapidly decrease, the costs to sequence a complete palaeogenome will remain substantial in the near future.

Yet given the rapidly growing number of modern genome sequences (http://www.ncbi.nlm.nih.gov/Genomes/), low-coverage palaeogenomes may become increasingly important for studying extant species with extinct close relatives, such as modern humans and Neanderthals. In this way, the total number of substitutions that need to be studied in order to uncover the genetic basis for human specific traits can be reduced substantially. For example, the genomes of humans and chimpanzees differ at about 40 million positions, whereas this number is about tenfold lower when comparing humans and Neanderthals. The extinct aurochs, the ancestor of domestic cattle, represents another example where the genome sequence of an extinct species may be of high value (see also the contributions by Bollongino and Geigl in this issue). Finding positions that have been selected during cattle domestication would be greatly facilitated if, in addition to only the cattle genome, the genome of the extinct aurochs and a third bovid genome, from a species basal to aurochs and modern cattle, would be available.

In conclusion, nuclear palaeogenomics will probably remain a rather small research field. Just as all of biology is not genomics, not all of ancient DNA will become palaeogenomics. However, within this limitation, it clearly has the potential to yield some intriguing results. If the Neanderthal genome project succeeds – even only in the form of a 'probabilistic genome' – we will learn much about our own species and probably also about Neanderthals. Even if the resulting genomes are full of errors, low-coverage genomic sequences of extinct species – such as the mammoth or aurochs – would provide valuable resources for further research, e.g. by identifying potential SNPs that can then be verified using PCR or by providing better insight into the biology of extant species.

I will close on a cautionary note. Even if we obtained a high-quality genome of an extinct organism, this will not mean that we can bring this species back to life, as often claimed. The mammoth will never roam the earth again. The best we can do is trying to understand its biology, evolution and maybe the reasons for its extinction. Perhaps this information will provide us with better tools necessary to prevent the extinction of the millions of species that currently live on our planet.

Acknowledgements

I thank Adrian Briggs, Eva-Maria Geigl, Christine Green, Ed Green, Johannes Krause, Tomislav Maricic, Svante Pääbo, Holger Römpler, and Mathias Stiller for helpful comments, Knut Finstermeier for help with the figure design, Johannes Krause, Joshua Pollack, and Ingo Ebersberger for providing unpublished results on the revised mammoth mtDNA genome sequence and the Max Planck Society for financial support. This paper is dedicated to Reinhold Rau († 2006), a passionate naturalist and the greatest expert ever on the extinct quagga, who was also involved in the first ancient DNA study.

References

[1] J.J. Austin, A.J. Ross, A.B. Smith, R.A. Fortey, R.H. Thomas, Problems of reproducibility – does geologically ancient DNA survive in amber-preserved insects? Proc. R. Soc. Lond. B Biol. Sci. 264 (1997) 467–474.

[2] I. Barnes, P. Matheus, B. Shapiro, D. Jensen, A. Cooper, Dynamics of Pleistocene population extinctions in Beringian brown bears, Science 295 (2002) 2267–2270.

[3] M. Bunce, T.H. Worthy, T. Ford, W. Hoppitt, E. Willerslev, et al., Extreme reversed sexual size dimorphism in the extinct New Zealand moa Dinornis, Nature 425 (2003) 172–175.

[4] J. Cherfas, Ancient DNA: still busy after death, Science 253 (1991) 1354–1356.

[5] A. Cooper, C. Lalueza-Fox, S. Anderson, A. Rambaut, J. Austin, R. Ward, Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution, Nature 409 (2001) 704–707.

[6] A. Cooper, C. Mourer-Chauviré, G.K. Chambers, A. von Haeseler, A.C. Wilson, S. Pääbo, Independent origins of New Zealand moas and kiwis, Proc. Natl Acad. Sci. USA 89 (1992) 8741–8744.

[7] G. Del Pozzo, J. Guardiola, Mummy DNA fragment identified, Nature 339 (1989) 431–432.

[8] E.M. Geigl, On the circumstance surrounding the preservation and analysis of very old DNA, Archaeometry 44 (2002) 337–342.

[9] M.T. Gilbert, J. Binladen, W. Miller, C. Wiuf, E. Willerslev, et al., Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis, Nucleic Acids Res. 35 (2007) 1–10.

[10] R.E. Green, J. Krause, S.E. Ptak, A.W. Briggs, M.T. Ronan, et al., Analysis of one million base pairs of Neanderthal DNA, Nature 444 (2006) 330–336.

[11] A. Greenwood, C. Capelli, G. Possnert, S. Pääbo, Nuclear DNA Sequences from Late Pleistocene Megafauna, Mol. Biol. Evol. 16 (1999) 1466–1473.

[12] O. Haddrath, A.J. Baker, Complete mitochondrial DNA genome sequences of extinct birds: ratite phylogenetics and the vicariance biogeography hypothesis, Proc. R. Soc. Lond. B: Biol. Sci. 268 (2001) 939–945.

[13] R. Higuchi, B. Bowman, M. Freiberger, O.A. Ryder, A.C. Wilson, DNA sequences from the quagga, an extinct member of the horse family, Nature 312 (1984) 282–284.
[14] R.G. Higuchi, L.A. Wrischnik, E. Oakes, M. George, B. Tong, A.C. Wilson, Mitochondrial DNA of the extinct quagga: relatedness and extent of postmortem change, J. Mol. Evol. 25 (1987) 283–287.
[15] M. Hofreiter, V. Jaenicke, D. Serre, A. Haeseler Av, S. Pääbo, DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA, Nucleic Acids Res. 29 (2001) 4793–4799.
[16] M. Hofreiter, S. Münzel, N.J. Conard, J. Pollack, M. Slatkin, et al., Sudden replacement of cave bear mitochondrial DNA in the Late Pleistocene, Curr. Biol. 17 (2007) R122–R123.
[17] M. Hofreiter, D. Serre, H.N. Poinar, M. Kuch, S. Pääbo, Ancient DNA, Nat. Rev. Genet. 2 (2001) 353–359.
[18] M. Höss, O. Handt, S. Pääbo, Recreating the Past by PCR, in: K. Mullis, F. Ferre, R. Gibbs (Eds.), The Polymerase Chain Reaction, Birkhauser, Boston, 1994, pp. 257–264.
[19] L. Huynen, C.D. Millar, R.P. Scofield, D.M. Lambert, Nuclear DNA sequences detect species limits in ancient moa, Nature 425 (2003) 175–178.
[20] M. Ingman, H. Kaessmann, S. Pääbo, U. Gyllensten, Mitochondrial genome variation and the origin of modern humans, Nature 408 (2000) 708–713.
[21] V. Jaenicke-Despres, E.S. Buckler, B.D. Smith, M.T.P. Gilbert, A. Cooper, et al., Early allelic selection in maize as revealed by ancient DNA, Science 302 (2003) 1206–1208.
[22] J. Krause, P.H. Dear, J.L. Pollack, M. Slatkin, H. Spriggs, et al., Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae, Nature 439 (2006) 724–727.
[23] J.A. Leonard, N. Rohland, S. Glaberman, R.C. Fleischer, A. Caccone, M. Hofreiter, A rapid loss of stripes: the evolutionary history of the extinct quagga, Biol. Lett. 1 (2005) 291–295.
[24] J.A. Leonard, O. Shanks, M. Hofreiter, E. Kreuz, L. Hodges, et al., Animal DNA in PCR reagents plagues ancient DNA research, J. Archaeol. Sci. 34 (2007) 1361–1366.
[25] J.A. Leonard, R.K. Wayne, A. Cooper, Population genetics of ice age brown bears, Proc. Natl Acad. Sci. USA 97 (2000) 1651–1654.
[26] T. Lindahl, Instability and decay of the primary structure of DNA, Nature 362 (1993) 709–715.
[27] H. Malmström, J. Stora, L. Dalen, G. Holmlund, A. Gotherström, Extensive human DNA contamination in extracts from ancient dog bones and teeth, Mol. Biol. Evol. 22 (2005) 2040–2047.
[28] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, et al., Genome sequencing in microfabricated high-density picolitre reactors, Nature 437 (2005) 376–380.
[29] K.B. Mullis, F.A. Faloona, Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction, Methods Enzymol. 155 (1987) 335–350.
[30] M. Nei, Stochastic errors in DNA evolution and molecular phylogeny, in: H. Gershowitz, DL. Rucknagel, RE. Tashian (Eds.), Evolutionary perspectives and the new genetics, Alan R. Liss, Inc., New York, 1986, p. 133–147.
[31] J.P. Noonan, G. Coop, S. Kudaravalli, D. Smith, J. Krause, et al., Sequencing and analysis of Neanderthal genomic DNA, Science 314 (2006) 1113–1118.
[32] J.P. Noonan, M. Hofreiter, D. Smith, J.R. Priest, N. Rohland, et al., Genomic sequencing of Pleistocene cave bears, Science 309 (2005) 597–599.
[33] S. Pääbo, Molecular cloning of Ancient Egyptian mummy DNA, Nature 314 (1985) 644–645.
[34] S. Pääbo, Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification, Proc. Natl Acad. Sci. USA 86 (1989) 1939–1943.
[35] S. Pääbo, H. Poinar, D. Serre, V. Jaenicke-Despres, J. Hebler, et al., Genetic Analyses From Ancient DNA, Annu. Rev. Genet. 38 (2004) 645–679.
[36] S. Pääbo, A.C. Wilson, Polymerase chain reaction reveals cloning artefacts, Nature 334 (1988) 387–388.
[37] N. Patterson, D.J. Richter, S. Gnerre, E.S. Lander, D. Reich, Genetic evidence for complex speciation of humans and chimpanzees, Nature 441 (2006) 1103–1108.
[38] H.N. Poinar, C. Schwarz, J. Qi, B. Shapiro, R.D.E. MacPhee, et al., Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA, Science 311 (2006) 392–394.
[39] M. Pruvost, R. Schwarz, V.B. Correia, S. Champlot, S. Braguier, et al., Freshly excavated fossil bones are best for amplification of ancient DNA, Proc. Natl Acad. Sci. USA 104 (2007) 739–744.
[40] H. Roempler, P.H. Dear, J. Krause, M. Meyer, N. Rohland, et al., Multiplex amplification of ancient DNA, Nat. Protoc. 1 (2006) 720–728.
[41] E.I. Rogaev, Y.K. Moliaka, B.A. Malyarchuk, F.A. Kondrashov, M.V. Derenko, et al., Complete mitochondrial genome and phylogeny of Pleistocene Mammoth *Mammuthus primigenius*, PLoS Biol 4 (2006) e73.
[42] H. Roempler, N. Rohland, C. Lalueza-Fox, E. Willerslev, T. Kuznetsova, et al., Nuclear gene indicates coat-color polymorphism in mammoths, Science 313 (2006) 62.
[43] T. Schöneberg, M. Hofreiter, A. Schulz, H. Roempler, Learning from the past: Evolution of GPCR functions, Trends Pharmacol. Sci. 28 (3) (2007) 117–121.
[44] D. Serre, A. Langaney, M. Chech, M. Teschler-Nicola, M. Paunovic, et al., No evidence of Neandertal mtDNA contribution to early modern humans, PLoS Biol. 2 (2004) 313–317.
[45] B. Shapiro, A.J. Drummond, A. Rambaut, M.C. Wilson, P.E. Matheus, et al., Rise and fall of the Beringian steppe bison, Science 306 (2004) 1561–1565.
[46] M. Stiller, R.E. Green, M. Ronan, J.F. Simons, L. Du, et al., Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA, Proc. Natl Acad. Sci. USA 103 (2006) 13578–13584.
[47] S.G. Tringe, E.M. Rubin, Metagenomics: DNA sequencing of environmental samples, Nat. Rev. Genet. 6 (2005) 805–814.
[48] P. Wandeler, S. Smith, P.A. Morin, R.A. Pettifor, S.M. Funk, Patterns of nuclear DNA degeneration over time – A case study in historic teeth samples, Mol. Ecol. 12 (2003) 1087–1093.
[49] E. Willerslev, A. Cooper, Ancient DNA, Proc. Biol. Sci. 272 (2005) 3–16.
[50] S.R. Woodward, N.J. Weyand, M. Bunnell, DNA sequence from Cretaceous period bone fragments, Science 266 (1994) 1229–1232.
[51] H. Zischler, M. Höss, O. Handt, A. von Haeseler, A.C. van der Kuyl, J. Goudsmit, Detecting dinosaur DNA, Science 268 (1995) 1192–1193 (discussion 4).