The EMBO Journal (2009) 28, 2494–2502 | © 2009 European Molecular Biology Organization | Some Rights Reserved 0261-4189/09
www.embojournal.org

THE EMBO JOURNAL

EMBO open

EMBO Member's Review

The Neandertal genome and ancient DNA authenticity

Richard E Green, Adrian W Briggs, Johannes Krause, Kay Prüfer, Hernán A Burbano, Michael Siebauer, Michael Lachmann and Svante Pääbo*

Max-Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Recent advances in high-throughput DNA sequencing have made genome-scale analyses of genomes of extinct organisms possible. With these new opportunities come new difficulties in assessing the authenticity of the DNA sequences retrieved. We discuss how these difficulties can be addressed, particularly with regard to analyses of the Neandertal genome. We argue that only direct assays of DNA sequence positions in which Neandertals differ from all contemporary humans can serve as a reliable means to estimate human contamination. Indirect measures, such as the extent of DNA fragmentation, nucleotide misincorporations, or comparison of derived allele frequencies in different fragment size classes, are unreliable. Fortunately, interim approaches based on mtDNA differences between Neandertals and current humans, detection of male contamination through Y chromosomal sequences, and repeated sequencing from the same fossil to detect autosomal contamination allow initial large-scale sequencing of Neandertal genomes. This will result in the discovery of fixed differences in the nuclear genome between Neandertals and current humans that can serve as future direct assays for contamination. For analyses of other fossil hominins, which may become possible in the future, we suggest a similar 'boot-strap' approach in which interim approaches are applied until sufficient data for more definitive direct assays are acquired.

The EMBO Journal (2009) **28,** 2494–2502. doi:10.1038/emboj.2009.222; Published online 6 August 2009
Subject Categories: genome stability & dynamics
Keywords: ancient DNA; DNA contamination; evolution; genome; Neandertal

*Corresponding author. MPI for Evolutionary Anthropology, MPI, Deutscher Platz 6, D-04103 Leipzig, Germany. Tel.: +49 341 3550 501; Fax: +49 341 3550 550; E-mail: paabo@eva.mpg.de

Received: 29 June 2009; accepted: 10 July 2009; published online: 6 August 2009

Ancient DNA and authenticity

The presence of DNA in ancient remains was initially shown by staining of DNA in histological samples of Egyptian mummies (Pääbo, 1984) and by extraction and cloning in bacterial plasmids of DNA from the extinct quagga and mummies (Higuchi *et al*, 1984; Pääbo, 1985). However, it was only the advent of the polymerase chain reaction (PCR) (Mullis and Faloona, 1987) that made it possible to reproduce results. Through PCR it became possible to ensure that the sequence determined did not contain errors (Pääbo and Wilson, 1988) and that the DNA sequences determined were indeed derived from the organism under study. Early results included the determination of DNA sequences from extinct mammals such as marsupial wolves (Thomas *et al*, 1989), moas (Cooper *et al*, 1992), and mammoths (Hagelberg *et al*, 1994; Höss *et al*, 1994; Krause *et al*, 2006) allowing resolution of the phylogenetic relationships between these extinct organisms and extant species.

Along with these successes came the realization that the ability of the PCR to amplify few or even single template molecules meant that when ancient specimens contain little or no endogenous DNA, the DNA amplified could be partially or wholly derived from exogenous DNA contaminating a specimen and be mistaken for endogenous DNA (Pääbo *et al*, 1989). For example, reports of dinosaur DNA sequences (Woodward *et al*, 1994) proved to be derived from human DNA contaminating the fossil or performed experiments (Zischler *et al*, 1995). These and other similar experiences served as cautionary tales for the growing field. As a remedy, 'criteria of authenticity' for the study of ancient DNA were suggested (Pääbo *et al*, 1989) and the community converged on a set of laboratory practices to prevent contamination that have developed over time (Cooper and Poinar, 2000; Hofreiter *et al*, 2001b; Pääbo *et al*, 2004). They include, for example, strict spatial separation of ancient DNA extraction from other experiments and UV irradiation and bleach treatment of extraction areas to minimize the extent of contamination. Often, the ancient DNA extraction facilities fulfill clean room requirements in that they operate under positive pressure using filtered air, require personnel to wear sterile clothing and face shields, and to work in laminar flow hoods. The criteria for authenticity include independent replication of results within a laboratory and in many cases, such as when particularly noteworthy or surprising results are obtained, in an independent second laboratory. Such practices have generally served the field well.

In the case of hominins, that is modern humans and their close relatives such as Cro-Magnons, Neandertals, and *Homo floresiensis* (Brown *et al*, 2004), the issue of authenticity is particularly acute because they can be expected to be identical to current humans for much or almost all of their genome. For example, although morphologically distinct, Neandertals were so closely related to people living today that most Neandertal DNA fragments retrieved from a fossil are expected to carry no sequence differences to the corresponding human sequences (Pääbo, 1999). As human DNA is a very common contaminant in fossils and laboratory experiments, this makes it particularly challenging to ascertain the authenticity of Neandertal DNA sequences. Indeed, for Cro-Magnons and other modern humans, the problems are so severe that over the past 15 years we have been pessimistic over the prospects of ever reliably determining such DNA sequences (Pääbo *et al*, 2004). For Neandertals, the situation is more tractable because their mitochondrial (mt) genome proved to be different from that of any modern human studied to date (Krings *et al*, 1997), making it possible to determine Neandertal mtDNA sequences confidently.

Recently, high-throughput sequencing techniques have become available that allow large numbers of DNA sequences to be determined (Margulies *et al*, 2005; Bentley *et al*, 2008). They rely on the construction of sequencing libraries by the ligation of DNA adapters to the ends of DNA molecules in a sample. These adapters then serve as priming sites both for amplification and for sequencing, which occur either on beads or on a solid surface in which each bead or cluster on a surface represents an amplified copy of a single original template molecule.

Currently, the most common application of these techniques in ancient DNA research is direct shot-gun sequencing of random DNA fragments extracted from a fossil (Green *et al*, 2006; Poinar *et al*, 2006; Stiller *et al*, 2006; Miller *et al*, 2008). The advent of the high-throughput approach to ancient DNA analyses makes it important to revisit the criteria of authenticity. For example, reproduction of results in the same or different laboratories as a prerequisite for publication is not practicable for large-scale DNA sequencing of random molecules because of constraints on time, costs, and sample materials. Nonetheless, there are means to control and assess the extent of contamination. Here, we discuss the measures taken to control contamination in the initial sequencing efforts aimed at showing the feasibility of sequencing the Neandertal genome. We describe technical improvements incorporated into the ongoing genome sequencing effort, and use published and unpublished Neandertal DNA sequence data to describe means of assessing contamination that we are convinced will allow the reliable determination of a useful Neandertal genome sequence.

MtDNA as an inroad to the Neandertal genome

The mitochondrial genome is only 1/200 000 the size of the nuclear genome, occurs in many copies per cell, is maternally inherited without recombination, and has been extensively used in PCR-based studies of ancient DNA, including Neandertals. These studies have shown that Neandertal mtDNAs fall outside the variation found among extant humans (Krings *et al*, 1997, 2000; Ovchinnikov *et al*, 2000; Serre *et al*, 2004; Orlando *et al*, 2006). Thus, there are substitutions seen in all or most Neandertal mtDNAs but not in current humans, and others seen in all or most current human mtDNAs but not in Neandertals. By PCR amplification of mtDNA regions containing such diagnostic positions and subsequent cloning and sequencing of multiple independent clones from the PCR products, such substitutions can be used to estimate the relative amounts of Neandertal versus human mtDNA in a Neandertal fossil extract (Figure 1). The level of contamination observed at diagnostic positions is thus as-

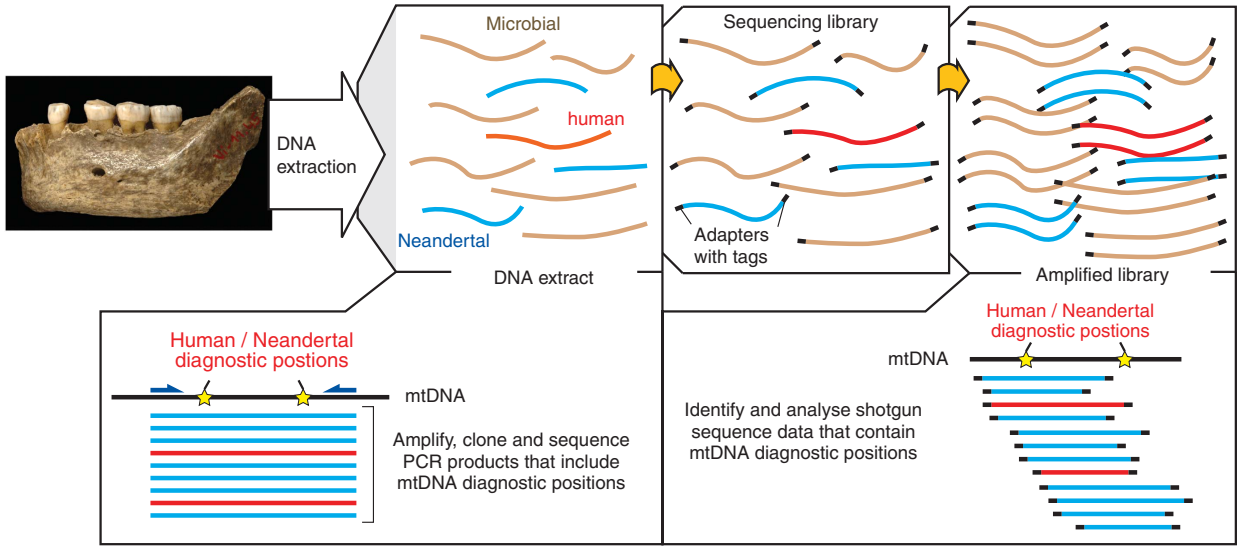

Figure 1 Estimates of human mtDNA contamination in Neandertal extracts. DNA extracts of Neandertal bones contain a large excess of microbial DNA (brown), at most a few percent of Neandertal DNA (blue) and generally variable amounts of contaminating DNA from current humans (red). Traditionally, contamination has been assayed through PCR directly from DNA extract from fossil bone (left lower panel). Accumulation of large numbers of reads from high-throughput sequencing allows a direct estimate of mtDNA contamination in the sequencing library (right lower panel). Once human/Neandertal diagnostic nuclear genome positions are learned, this strategy can be extended to nuclear DNA sequences.

sumed to represent the contamination level across the mitochondrial genome, allowing authenticity to be determined even for DNA fragments from regions in which no fixed differences between Neandertals and current humans occur. Thus, for the initial work addressing the feasibility of large-scale Neandertal sequencing, extracts of Neandertal bones were prepared under clean room conditions and analysed for contamination by amplification of mtDNA regions containing diagnostic positions. The extent of mtDNA contamination was estimated to be below 1% for two extracts from a ~38 000-year-old Neandertal bone from Vindija Cave, Croatia (Serre *et al*, 2004). These were then sent to other laboratories to be directly sequenced by high-throughput technologies (Green *et al*, 2006) or cloned in plasmid vectors and subjected to sequencing (Noonan *et al*, 2006).

However, analysis of contamination in DNA extracts cannot show contamination in subsequent laboratory steps, for example, in library construction and sequencing. A particular concern is that libraries from Neandertal DNA extracts contain at most a few percent of Neandertal DNA whereas the rest stems from microbes that have colonized bones after the death of the Neandertal. Therefore, contamination of a Neandertal library with even tiny quantities of a library that contains 100% human DNA will greatly affect the results. For the ongoing genome sequencing effort, novel sequencing adapters were therefore designed and used for library construction (Briggs *et al*, 2007) that contain a unique four-nucleotide tag (TGAC) at their 3′-ends. Thus, each sequence determined from a Neandertal library should start with these four bases. Indeed, subsequent work in our laboratory has shown that carry-over of low amounts of library molecules from one sequencing run to the next can occur, making the use of such tags particularly important. A further advantage of constructing sequencing libraries using project-specific tags is that this provides a 'snapshot' of the molecules that are present in the extract at the time of library construction. Thus, estimates of contamination and other parameters can be determined once and applied to the library as it is being used for further sequencing or other experiments.

Unfortunately, because of logistical constraints associated with the locations of the emerging technologies, it was necessary during the initial exploratory work to take scrupulously prepared DNA extracts from our clean room and use them to construct libraries in laboratories elsewhere. Subsequently, once a Neandertal genome sequencing effort based on direct sequencing of DNA extracts was initiated, the construction of tagged libraries for high-throughput sequencing was established in the clean room environment in our laboratory (Briggs *et al*, 2007). Indeed, work by Wall and Kim (2007) confirmed the need to perform library construction under clean room conditions using tagged library adapters. They showed that the DNA sequences from the two extracts that were prepared in our clean room and shown to carry similar and low mtDNA levels of contamination before being sent to other laboratories, showed evidence of contamination by current human DNA in at least one (Green *et al*, 2006) of the two data sets.

Once a tagged sequencing library is constructed, the next step is to estimate the level of contamination in the library. One way is to perform sequencing and identify DNA fragments indicative of contamination. The first, most obvious means of directly assessing contamination is by examining mtDNA sequences. However, until recently only about 600 bp of the Neandertal mtDNA were known, yielding only a small number of sites that could distinguish Neandertal mtDNA from current human mtDNA. Consequently, few DNA fragments among random sequence reads were informative with respect to contamination estimates. The later determination of the 16 565 nt that make up the complete mtDNA from this Neandertal individual (Green *et al*, 2008) dramatically increased the number of informative positions to 133. Using these positions, it is possible to identify enough informative fragments in data sets of reasonable size to estimate levels of mtDNA contamination. Table I shows estimates of contamination using these 133 sites from several Neandertal libraries as well as estimates directly from the extracts from which they were prepared. The corresponding extract and library estimates are generally in agreement. However, there is one notable exception: in the library constructed without the tagged adapters outside our clean room facility in 2006 (Green *et al*, 2006), 8 of 75 fragments carrying one or more of the 133 informative positions indicate current human contamination, giving an estimate of 11% contamination (CI 4.7–20%). This confirms that contamination was introduced into this dataset as suggested (Wall and Kim, 2007), presumably during library construction outside the clean room.

Nuclear DNA contamination estimates

A limitation inherent to extrapolation of mtDNA contamination estimates to the nuclear genome is that the ratio of

Table I MtDNA contamination and mtDNA to nuclear DNA ratios in some DNA extracts and sequencing libraries used to study the Neandertal genome

Extract	N	H	Extract cont.	Library	N	H	Library cont.	Nuclear-mtDNA ratio
A	111	1	0.8% (0.0–4.9%)	A.1	67	8	10.7% (4.7–19.9%)	375
				A.2	4	0	0% (0.0–60.2%)	222
B	103	0	0.0% (0.0–3.5%)	BC.1	22	0	0.0% (0.0–15.4%)	186
C	112	0	0.0% (0.0–3.2%)	BC.2	1822	7	0.4% (0.2–0.8%)	157
D	152	8	5.0% (2.2–9.6%)	DEF.1	30	1	3.2% (0.1–16.7%)	419
E	100	1	1.0% (0.0–5.4%)					
F	174	8	4.4% (1.9–8.5%)					

Six extracts of Neandertal bone Vindija Vi33.16 (A–F) were prepared and analysed with respect to mtDNA contamination using PCR. N and H refer to Neandertal- and current human-like clones of mtDNA amplification products, respectively. These extracts were used to construct libraries used for sequencing. Library A.1 was constructed outside the clean room facility using standard 454 sequencing adapters and is published in Green *et al* (2006). The other libraries were constructed in the clean room using tagged adapters. Library designations refer to the extracts used to construct them. N and H refer to Neandertal- and current human-like mtDNA fragments, respectively. For each library the mtDNA to nuclear DNA ratios are given. For contamination estimates, 95% confidence intervals are given in parentheses.

mtDNA to nuclear DNA may differ among different tissues as well as between different bones (Schwarz *et al*, 2009). Thus, although analysis of mtDNA sequences yields reliable estimates of the extent of contamination of the mtDNA, the level of nuclear DNA contamination can be under- or over-estimated if the contaminating DNA source contains less or more mtDNA, respectively, than the endogenous DNA. For example, as we have pointed out (Green *et al*, 2006), if a contaminating source of DNA in the proof-of-principle data set contained low amounts of mtDNA relative to the endogenous bone DNA, the level of contamination can be higher than indicated by the mtDNA assay alone. In fact, though 375 nuclear DNA fragments were seen for each mtDNA fragment seen in the proof-of-principle data set, a subsequently produced bar-coded library from the same extract generated in our clean room has yielded 222 nuclear DNA fragments for each mtDNA fragment (Table I). If one takes the entire excess of nuclear fragments to represent contamination, that is, assumes that the contamination was exclusively of mtDNA-free nuclear DNA, this yields an estimate of contamination of 41% in that proof-of-principle data. Although this estimate relies on several tenuous assumptions it serves to show the limitation of extrapolating mtDNA contamination estimates to the nuclear genome.

To achieve reliable nuclear DNA sequences from Neandertals, it is therefore necessary to develop direct nuclear estimates of contamination similar to those for mtDNA. This will become possible once large amounts of DNA sequences from the Neandertal nuclear genome become available because fixed differences between Neandertals and current humans will then be identifiable. On the road to this goal, however, other interim approaches are needed.

Y chromosomal contamination estimates

One such interim approach is available for bones derived from female Neandertals. As females contain no Y chromosome, any Y chromosome sequence from such a bone must be from contaminating DNA derived from a male individual. By comparing the number of such sequences to the total number of sequences that map elsewhere in the genome, it is possible to estimate the levels of male contamination in female bones. However, because the Y chromosome has many regions that are identical or highly similar to those on the X chromosome, it is imperative to avoid misidentifying sequences from the X chromosome as being derived from the Y chromosome. To this end, we have identified regions of Y-unique sequence, totaling 98 kilobases, each of which has $\geqslant$10% sequence difference to all sequences in the human genome outside of the Y chromosome.

Fortunately, the three Neandertal bones used for the bulk of the sequencing of the Neandertal genome derive from females and so this approach can be applied to the shotgun data currently being generated. For one of these bones (Vi33.16), 21 671 548 fragments have currently been identified as being derived from a hominin genome. Of these, two align to these Y-unique regions whereas 380 would be expected if the DNA derived from a male individual. This yields an estimate of 0.5% (CI 0.1–1.9%) for male human contamination. Thus, these bones seem to have levels of nuclear contamination as low as those suggested by the mtDNA assays.

X chromosomal contamination estimates

Assays that allow the detection of female contamination are obviously also desirable. In the case of male Neandertal bones, one such strategy has been suggested. As males are haploid for the X chromosome, heterozygosity should not be observed in overlapping X chromosomal DNA fragments from male DNA samples. In cases in which different alleles are observed, at least one allele must derive from a contaminant. Although this strategy is conceptually attractive, there are several limitations in practice. First, sequencing errors caused by nucleotide misincorporations, common in ancient DNA, or machine error, will appear as a heterozygous position and can be mistaken for contamination. Second, many contaminant molecules will not harbour a sequence difference between Neandertals and humans and thus evade detection. Third, in cases of mismapping of sequences, paralogous regions that have genuine sequence differences may be mistaken for contamination. This approach is therefore unlikely to yield realistic estimates of contamination.

Autosomal contamination estimates

The most direct approach to detect modern human contamination would be to identify autosomal sequence positions in which all Neandertals differ from all or almost all current humans. However, because for so much of their nuclear genome Neandertals share the variation still present in modern humans (Pääbo, 1999), such positions will be rare. Extensive sequence information from several Neandertals will therefore be required before a set of such positions is found.

One interim solution is to use a two-stage approach for any particular Neandertal fossil under study. In the first stage, a targeted capture method such as PEC (Briggs *et al*, 2009) can be used to isolate one or more genomic region of putative Neandertal DNA from sequencing libraries. In these regions, some positions will show a human–chimpanzee difference in which outgroup comparison shows that a substitution occurred on the human lineage after the human–chimpanzee split 5–7 million years ago. At a proportion of these positions the Neandertal will carry the ancestral, ape-like state. These positions can be then genotyped in population samples of humans from around the world to identify the subset of these positions in which all or almost all humans are derived. This final subset represents positions in which this particular Neandertal differs from all or almost all extant humans. Then, in a second stage, these human–Neandertal diagnostic positions can be used to generate estimates of nuclear DNA contamination in sequencing of other, independent libraries from the same Neandertal individual.

A limitation of this approach is that it requires libraries of relatively high genome coverage so that a reasonable number of informative positions can be identified in the first set of experiments, at sufficiently high coverage to be confident that the Neandertal is homozygous, and then independently retrieved in the second set of experiments. It also requires that a large amount of sequencing be performed before estimates can be generated. However, in conjunction with mtDNA and Y chromosomal estimates this approach is currently in our opinion the best way to arrive at realistic estimates of contamination in Neandertal genome sequence

data. Eventually, once a first Neandertal genome sequence is available and information from further Neandertals accumulates, a set of positions in which fixed differences between Neandertals and current humans exist will be identified. These will then serve as future direct estimators of contamination, as is now possible for the Neandertal mitochondrial genome.

Indirect estimates of contamination

In contrast to the approaches above that rely on direct observation of DNA sequences that are indicative of contamination in Neandertal libraries, other approaches have been suggested that use global characteristics of large-scale datasets and compare them to what can be expected for ancient DNA. For example, fragmentation is a universal feature of ancient DNA (Pääbo, 1989). Therefore, *bona fide* ancient DNA sequences will generally be short. Likewise, nucleotide misincorporations resulting from deamination of cytosine residues result in many C to T and G to A substitutions in ancient DNA (Hofreiter *et al*, 2001a; Briggs *et al*, 2007; Brotherton *et al*, 2007). These approaches have the benefit of allowing contamination levels to be estimated from relatively small sets of random sequences. However, many pitfalls that are briefly discussed below make them unreliable as estimators of contamination.

Estimates based on fragment size

Variability in the distribution of fragment lengths in different fossils and even within extracts from a single fossil make the extent of DNA fragmentation a less attractive option for estimating contamination. For example, the fraction of endogenous mtDNA fragments that are above 80 bp in length is ~11% in a Neandertal from El Sidron in Spain, ~27% in a Neandertal from Vindija in Croatia, and ~37% in a Neandertal from Feldhofer in Germany (Figure 2) (Briggs *et al*, 2009). Furthermore, this method is only reliable if the contaminating DNA is not fragmented to an extent similar to that of the endogenous ancient DNA. This is not always the case. For example, in DNA extracts of the El Sidron Neandertal, mtDNA fragments retrieved that are known to be contamination because they carry nucleotide substitutions typical of current humans are as short as the endogenous Neandertal mtDNA; in the Vindija Neandertal some contaminating fragments are longer whereas others are short; and even in the Feldhofer Neandertal, in which the contaminating mtDNA fragments are clearly on average longer than the endogenous ones, the two types of molecules overlap in size (Figure 2) (Briggs *et al*, 2009). Thus, fragment length *per se* is not a reliable estimator of contamination.

Fragment size and divergence to humans

One promising approach is to analyse fragment length in conjunction with other features of the DNA fragments. Wall and Kim (2007) showed that in one of the data sets published in 2006 (Green *et al*, 2006) longer sequences were less diverged from the human reference genome sequence and more often carried the derived allele at positions in which current humans are polymorphic than did shorter sequences. As these are features expected for contaminating human DNA, they suggested that longer sequences were enriched for current human DNA in these data.

However, random DNA fragments sequenced from the extract of a Neandertal bone must first be identified as being of hominin rather than bacterial origin before they can be analysed. This step relies on recognizing sequence similarity between Neandertal sequences and either the human or chimpanzee genomes. This is difficult for short fragment, especially if they carry nucleotide misincorporations and sequencing errors, and as a consequence the estimates of the human–Neandertal divergence can be biased upwards for short fragments. To illustrate this, we analysed a data set of 3700 million base pairs determined from a Neandertal fossil on the Roche FLX platform in sequence bins of increasing length (Figure 3). For each length bin, we

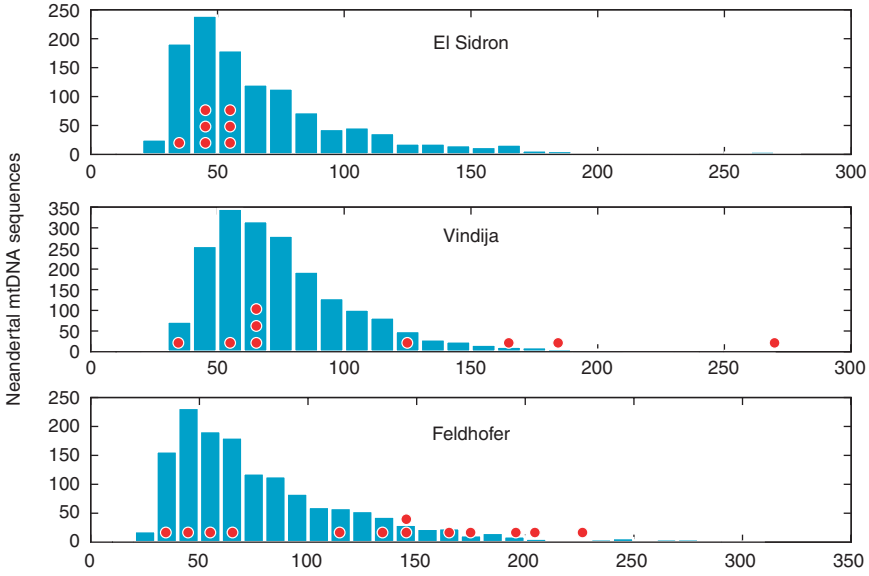

Figure 2 Lengths of Neandertal and human mtDNA fragments. Distributions of mtDNA fragments carrying Neandertal diagnostic positions are shown in blue for three Neandertal fossils. Each red dot represents a single contaminating human mtDNA fragment of the indicated length (data from Briggs *et al*, 2009).

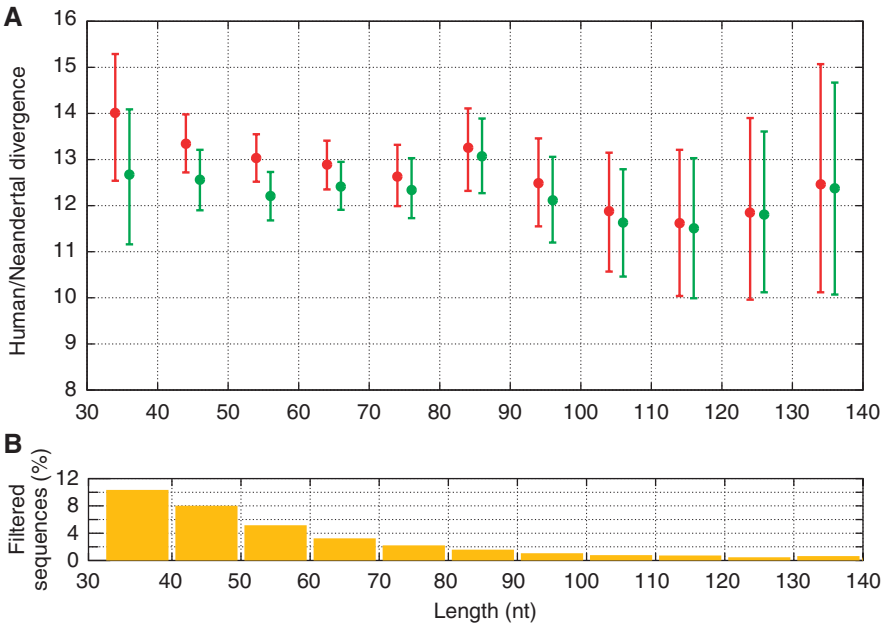

Figure 3 Neandertal/human divergence estimated from sequences of increasing length and score filtering. (**A**) Sequences in each length bin were used to calculate human/Neandertal divergence, given as the percentage of the human lineage back to the human/chimpanzee common ancestor in which the Neandertal sequences diverged. Sequences were filtered for uniqueness in the human and chimpanzee genomes by comparing the best alignment score to the second best score. In red are sequences whose best alignments are at least 1-bit better than the second best, in green with a difference of 5 bits or more. Bars show the 95% confidence interval from 1000 bootstrap replicates of the sequences in each bin. (**B**) Percentage of the sequences in each bin removed when increasing the alignment score filter from 1 to 5 bits. Shorter sequences are more likely to be removed by stricter filtering as they carry less information to place them uniquely in the human and chimpanzee genomes.

analysed the divergence using fragments identified using the same score cut-off for uniqueness in the human and chimpanzee genomes as was used in earlier analyses (Green *et al*, 2006; Noonan *et al*, 2006; Wall and Kim, 2007) as well as a more stringent cut-off. Strikingly, the divergence estimates for the fragments greater than 80 nucleotides in length is only minimally affected by increasing the stringency whereas the discrepancy becomes progressively larger for the shorter fragments. Thus, fragments of length 30–40 nucleotides identified with the lower, less strict cut-off score diverge on average 14.0% back in the past along the lineage to the human–chimpanzee common ancestor whereas the fragments with the higher, stricter cut-off diverge 12.5% back. This indicates that for shorter fragments the higher proportion of sequences that are mapped incorrectly inflate the apparent divergence between the human and Neandertal genomes.

Fragment size and derived alleles

At positions in the genome in which current humans exhibit a single nucleotide polymorphism, the Neandertal will either carry the ancestral allele seen in apes, or the derived allele caused by a mutation in the past. If longer fragments carry derived alleles more frequently than short fragments, this may suggest that the longer fragments are relatively more contaminated with modern human DNA fragments. However, this may also be because of the fact that shorter sequences are more frequently mapped incorrectly to genome sequences. To explore this, we first aligned Neandertal sequences to the human genome and determined the observed fraction of derived alleles that they carry. We then cut them in half *in silico* and again aligned them to the human genome and determined the fraction of derived alleles. Although Neandertal sequences of length 60–78 nt carry the human derived alleles in 17.1% of cases, when the same sequences are artificially reduced to a length of 30–39 nt, they carry the derived alleles in 15.1% of cases (Figure 4B), similar to what is seen in sequences that are already this short (Figure 4A). That this is indeed an artefact caused by how sequences of different lengths are aligned to the human genome is supported by the observation that the segments of the human reference genome to which the sequences are aligned reduce their fraction of derived alleles from 25.7 to 22.7% when the Neandertal sequences are shorter (Figure 4B).

The cause of this phenomenon is that if a Neandertal DNA sequence and reference human genome carry non-matching alleles there is a reduced ability to recognize the Neandertal sequence as being of hominin origin. This effect is stronger for shorter sequences. As derived alleles are of lower frequency than ancestral alleles both in humans and Neandertals, and even lower in Neandertals than in humans, sequences with mismatches represent cases in which the Neandertal is derived relatively more often than sequences in which no mismatches are seen. For example, if Neandertals carry derived alleles at 17% of polymorphic sites and humans at 35% of such sites, then the Neandertal will carry derived alleles at ~28% polymorphic positions in which there are mismatches between the human and Neandertal genomes at only 10% of positions in which the two genomes match each other. Thus, short DNA fragments will appear to carry derived alleles in Neandertals more rarely than long fragments because short fragments are more often lost in the analysis. This will result in an overestimate of contamination rates as calculated by Wall and Kim (2007).

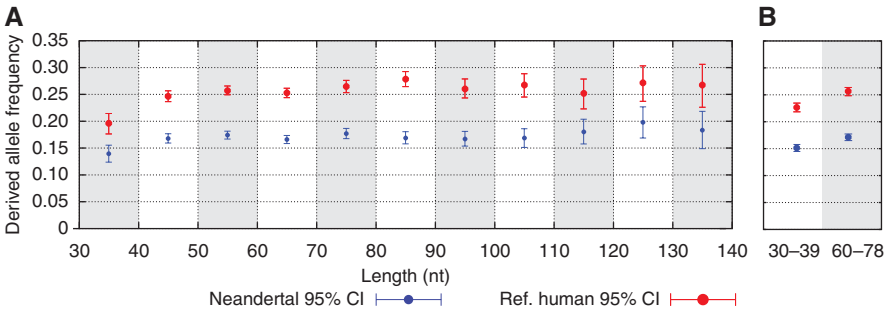

Figure 4 Fraction of human polymorphic positions carrying derived alleles in Neandertal and human DNA sequences. (**A**) Neandertal sequences of increasing length that overlap human polymorphic positions were assessed for having the derived or ancestral (chimpanzee-like) allele. Blue points are for Neandertal data, red points for the corresponding sequences in the human reference genome (hg18). (**B**) Sequences of length 60–78 nucleotides were split in half and re-analysed ('30–39'). Derived alleles are preferentially lost when fragments size is reduced.

Nucleotide misincorporations

C to T and G to A nucleotide misincorporations are a feature often seen in ancient DNA (Hofreiter *et al*, 2001a) and such misincorporations are more frequent at the 5′- and 3′-ends of molecules, respectively (Briggs *et al*, 2007; Brotherton *et al*, 2007). This could in principle provide a means of establishing that DNA is ancient. Unfortunately, this approach is limited in three ways. First, most *bona fide* ancient DNA sequences contain no nucleotide misincorporations. Second, contaminating sequences do occasionally contain nucleotide misincorporations (Malmström *et al*, 2005; Green *et al*, 2008). Third, although the extent of deamination-induced nucleotide misincorporations seems to be correlated positively with the extent of fragmentation of the DNA, both of these features vary substantially between Neandertal specimens (Briggs *et al*, 2009). Indeed, in many cases, contaminating DNA may be both degraded (such as in skin fragments in dust particles) and deaminated (for example, when it has been deposited in or on a fossil or in chemical reagents for a long time). Therefore, similar to fragment size, nucleotide misincorporations are at best a quantitative rather then a qualitative difference between endogenous and contaminating DNA that can not easily be used to estimate contamination, at least on a fragment-by-fragment basis (but see Conclusions and prospects, below).

Conclusions and prospects

When ancient DNA is studied by high-throughput sequencing rather than PCR, the laboratory procedures that have been developed for ancient DNA extraction and contamination prevention over the past 20 years are still of utmost importance. These procedures need to be adhered to up to the point of the construction of libraries using adapters carrying unique tags. Only after such tagged adapters have been added is it safe for libraries to leave clean room facilities for other manipulations and sequencing. Significantly, all potential sources of contamination, starting with the bone itself, through DNA extraction and adaptor ligation can then be considered together in a single 'snapshot' of the contamination of a library. Thus, all later assays of contamination of the same library can be assigned to the same contamination estimate and thereby add to its precision.

Methods that allow specific sequences of interest to be retrieved from such tagged libraries (Hodges *et al*, 2007; Briggs *et al*, 2009; Gnirke *et al*, 2009) make it possible to quickly analyse many sequences of interest from such libraries. Criteria of authenticity that are currently successfully applied to PCR-based studies of ancient DNA, such as reproduction of results from an independent extraction from the same bone, will then be useful just as they have been hitherto in PCR-based studies. In contrast, these criteria are not easily applicable to high-throughput shot-gun sequencing of entire ancient genomes. This is a particular problem for the Neandertal genome but applies also to other ancient genomes, such as mammoths (Miller *et al*, 2008), because all mammals including humans share conserved DNA sequence elements that may confuse results.

For sequencing ancient genomes we suggest a two-phase approach, much as was done for the Neandertal mitochondrial genome, in which initial work identified differences to current human mtDNAs and such differences were later applied to directly estimate contamination. For the first phase of genome sequencing, several direct contamination estimates, where each in itself is less than comprehensive, will be applied in concert. For the Neandertal genome, this includes the determination of mtDNA contamination, the detection of male contamination in bones of females, and capture methods that allow positions diagnostic of contamination in one particular individual to be identified and subsequently used in other libraries from the same individual. Eventually, once a Neandertal genome sequence is determined to high coverage, capture approaches can be applied to other Neandertals to identify enough positions that are fixed among Neandertals and differ from current humans. At that point such positions can be used to estimate contamination in Neandertal libraries even before they are subjected to other analyses. However, even then, some possible technical concerns need to be addressed. For example, because the sequences retrieved from ancient bones tend to be rich in the nucleotides G and C (Green *et al*, 2008), it needs to be determined to what extent such preservation biases are equally representative of endogenous and contaminating DNA, and thus whether a 'correction factor' might be required when extrapolating contamination estimates derived from high-coverage diagnostic positions to the entire genome.

In contrast to the direct estimates that we describe and advocate above, indirect measures based on the extent of fragmentation or modification of the DNA are at best supportive in nature. Particularly, comparisons of features between longer and shorter DNA fragments suffer from the fact that shorter fragments are more difficult to identify and correctly align to genome sequences of extant species.

One interesting question is whether it will be possible to estimate contamination in analyses of early hominins other than Neandertals, such as other archaic human forms or early modern humans. Conceivably, this may be possible by 'bootstrapping' oneself from the mtDNA to the nuclear DNA much as is done for the Neandertal genome. If extracts from a specimen can be identified for which deep high-throughput sequencing of mtDNA shows that a single mtDNA genome is present with minimal or absent indication of any additional mtDNA, this shows that the DNA preparation derives from a single individual. This individual is either the ancient individual from which the samples stems or a single recent human contaminating the specimen or extract. In this situation, fragmentation and nucleotide misincorporations may have a helpful role. Although individual ancient DNA fragments cannot be reliably distinguished from modern contaminants based on these features, the knowledge that all sequences in a dataset derive from a single individual will allow the overall fragmentation and misincorporation patterns to be analysed. If it can be shown that these patterns fall in a range typical of ancient, minimally contaminated specimens, and outside the range seen in contaminating DNA from specimens found and curated under conditions similar to the specimen being studied, then the DNA sequences are likely to be ancient. The mitochondrial and nuclear DNA sequences thus determined can then serve as an inroad to targeted studies of other, less well preserved specimens of the same hominin group. We are thus hopeful that it may become possible to sequence not only the Neandertal genome to high coverage, but also to study genomes of other ancient human forms provided that uncontaminated specimens that allow very deep sequencing can be found.

Materials and methods

MtDNA contamination assay

Before library production, DNA extracts were analysed for contamination using primers that amplify Neandertal as well as modern human mitochondrial control region sequences. The products were cloned and over 100 individual clones sequenced for each experiment. Such results are shown in Table I. To assay mtDNA contamination in the shotgun sequencing reads we score all positions in which the recently determined Neandertal mtDNA sequence differs from more than 99% of 311 humans (Green *et al*, 2008). Positions in which the fragment sequenced carries T or A residues and the human or Neandertal mtDNAs C or G residues, respectively, are excluded from the analyses because these may be caused by deaminated C residues in the extracted DNA (Hofreiter *et al*, 2001a). In addition, any positions in which the human and Neandertal mtDNAs differ by an insertion in which two or more of the same base exist in one of the species are excluded because homopolymer length is difficult to score by 454 sequencing.

Orthology and alignments

To identify orthologous sequences in the three genomes, each DNA sequence that had a best match to a single region of the human genome by megablast (Zhang *et al*, 2000) (bit-score difference of 1 or 5) was similarly compared in the chimpanzee genome. If it had a unique best match also in the chimpanzee by the same criteria, then these two alignment positions were used if they are reciprocally orthologous in the human–chimpanzee whole genome alignments (UCSC hg18vsPanTro2 and panTro2vsHg18). This removed about 25% of the sequences initially identified as Neandertal.

To align the DNA sequences from the three genomes, we refrained from progressive alignment approaches (for example, *clustalw*) because divergence estimates were found to be very sensitive to alignment order. Instead, we implemented a three-dimensional dynamic programming alignment (Lipman *et al*, 1989; Durbin *et al*, 1998) that simultaneously maximizes the similarity between Neandertal, human, and chimpanzee sequences. Finally, 5.2% of the remaining sequences that represent possible chimaeras or contain large numbers of insertions and deletions were removed.

Neandertal DNA divergence

To estimate the average divergence between Neandertal and human DNA sequences, we use alignment positions in which the human and chimpanzee genomes differ. Assuming equal evolutionary rates on the human and chimpanzee lineages, we then determine the fraction of substitutions on the human lineage in which the Neandertal shares the human state versus the chimpanzee state (Green *et al*, 2006), restricting the analysis to positions in which the human base differs from the chimpanzee base by a transversion because of the increased rate of C to T and G to A transitions seen in ancient DNA.

Acknowledgements

We are indebted to Jim Mullikin for suggesting the analysis of Neandertal sequences shortened in size *in silico*; to Christine Verna for the mandible picture in Figure 1; to Udo Stenzel for technical assistance; to the ancient DNA groups in Leipzig and the Neandertal Genome Analysis Consortium for their enthusiasm and input; to Linda Vigilant and Adam Willkins for input on the manuscript; and to the Presidential Innovation Fund of the Max Planck Society for making the Neandertal Genome Project possible.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ *et al* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Pavao R, Brajkovic D, Kucan Z, Gusic I, Schmitz RW, Doronichev VB, Golovanova LV, de la Resilla M, Fortea J, Rosas A, Pääbo S (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325:** 318–321

Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prufer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci USA* **104:** 14616–14621

Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, Austin J, Cooper A (2007) Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res* **35:** 5717–5728

Brown P, Sutikna T, Morwood MJ, Soejono RP, Jatmiko, Saptomo EW, Due RA (2004) A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature* **431:** 1055–1061

Cooper A, Mourer-Chauvire C, Chambers GK, von Haeseler A, Wilson AC, Paabo S (1992) Independent origins of New Zealand moas and kiwis. *Proc Natl Acad Sci USA* **89:** 8741–8744

Cooper A, Poinar HN (2000) Ancient DNA: do it right or not at all. *Science* **289:** 1139

Durbin R, Eddy S, Krogh A, Mitchison G (1998) *Biological Sequence Analysis*. Cambridge: Cambridge University Press

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27:** 182–189

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Pääbo S (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* **444:** 330–336

Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, Prufer K, Siebauer M, Burbano HA, Ronan M, Rothberg JM, Egholm M, Rudan P, Brajkovic D, Kucan Z, Gusic I *et al* (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134:** 416–426

Hagelberg E, Thomas MG, Cook Jr CE, Sher AV, Baryshnikov GF, Lister AM (1994) DNA from ancient mammoth bones. *Nature* **370:** 333–334

Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC (1984) DNA sequences from the quagga, an extinct member of the horse family. *Nature* **312:** 282–284

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR (2007) Genome-wide *in situ* exon capture for selective resequencing. *Nat Genet* **39:** 1522–1527

Hofreiter M, Jaenicke V, Serre D, Haeseler Av A, Pääbo S (2001a) DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res* **29:** 4793–4799

Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S (2001b) Ancient DNA. *Nat Rev Genet* **2:** 353–359

Höss M, Paabo S, Vereshchagin NK (1994) Mammoth DNA sequences. *Nature* **370:** 333

Krause J, Dear PH, Pollack JL, Slatkin M, Spriggs H, Barnes I, Lister AM, Ebersberger I, Pääbo S, Hofreiter M (2006) Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature* **439:** 724–727

Krings M, Capelli C, Tschentscher F, Geisert H, Meyer S, von Haeseler A, Grossschmidt K, Possnert G, Paunovic M, Pääbo S (2000) A view of Neandertal genetic diversity. *Nat Genet* **26:** 144–146

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Pääbo S (1997) Neandertal DNA sequences and the origin of modern humans. *Cell* **90:** 19–30

Lipman DJ, Altschul SF, Kececioglu JD (1989) A tool for multiple sequence alignment. *Proc Natl Acad Sci USA* **86:** 4412–4415

Malmström H, Stora J, Dalen L, Holmlund G, Götherstrom A (2005) Extensive human DNA contamination in extracts from ancient dog bones and teeth. *Mol Biol Evol* **22:** 2040–2047

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC *et al* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380

Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, Tomsho LP, Packard MD, Zhao F, Sher A, Tikhonov A, Raney B, Patterson N, Lindblad-Toh K, Lander ES, Knight JR, Irzyk GP, Fredrikson KM, Harkins TT, Sheridan S *et al* (2008) Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456:** 387–390

Mullis KB, Faloona FA (1987) Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. *Methods Enzymol* **155:** 335–350

Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S, Pritchard JK, Rubin EM (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science* **314:** 1113–1118

Orlando L, Darlu P, Toussaint M, Bonjean D, Otte M, Hänni C (2006) Revisiting Neandertal diversity with a 100 000 year old mtDNA sequence. *Curr Biol* **16:** R400–R402

Ovchinnikov IV, Götherstrom A, Romanova GP, Kharitonov VM, Liden K, Goodwin W (2000) Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* **404:** 490–493

Pääbo S (1989) Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc Natl Acad Sci USA* **86:** 1939–1943

Pääbo S (1984) Über den Nachweis von DNA in altägyptischen Mumien. *Das Altertum* **30:** 213–218

Pääbo S (1985) Molecular cloning of ancient Egyptian mummy DNA. *Nature* **314:** 644–645

Pääbo S (1999) Human evolution. *Trends Cell Biol* **9:** M13–M16

Pääbo S, Higuchi RG, Wilson AC (1989) Ancient DNA and the polymerase chain reaction. The emerging field of molecular archaeology. *J Biol Chem* **264:** 9709–9712

Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M (2004) Genetic analyses from ancient DNA. *Annu Rev Genet* **38:** 645–679

Pääbo S, Wilson AC (1988) Polymerase chain reaction reveals cloning artefacts. *Nature* **334:** 387–388

Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, Rampp M, Miller W, Schuster SC (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* **311:** 392–394

Schwarz C, Debruyne R, Kuch M, McNally E, Schwarcz H, Aubrey AD, Bada J, Poinar H (2009) New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. *Nucleic Acids Res* **37:** 3215–3229

Serre D, Langaney A, Chech M, Teschler-Nicola M, Paunovic M, Mennecier P, Hofreiter M, Possnert G, Pääbo S (2004) No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biol* **2:** E57

Stiller M, Green RE, Ronan M, Simons JF, Du L, He W, Egholm M, Rothberg JM, Keates SG, Ovodov ND, Antipina EE, Baryshnikov GF, Kuzmin YV, Vasilevski AA, Wuenschell GE, Termini J, Hofreiter M, Jaenicke-Despres V, Pääbo S (2006) Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. *Proc Natl Acad Sci USA* **103:** 13578–13584

Thomas RH, Schaffner W, Wilson AC, Pääbo S (1989) DNA phylogeny of the extinct marsupial wolf. *Nature* **340:** 465–467

Wall JD, Kim SK (2007) Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genet* **3:** 1862–1866

Woodward SR, Weyand NJ, Bunnell M (1994) DNA sequence from Cretaceous period bone fragments. *Science* **266:** 1229–1232

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7:** 203–214

Zischler H, Hoss M, Handt O, von Haeseler A, van der Kuyl AC, Goudsmit J, Pääbo S. (1995) Detecting dinosaur DNA. *Science* **268:** 1192–1193

