ARTICLES

Analysis of one million base pairs of Neanderthal DNA

Richard E. Green[1], Johannes Krause[1], Susan E. Ptak[1], Adrian W. Briggs[1], Michael T. Ronan[2], Jan F. Simons[2], Lei Du[2], Michael Egholm[2], Jonathan M. Rothberg[2], Maja Paunovic[3]‡ & Svante Pääbo[1]

Neanderthals are the extinct hominid group most closely related to contemporary humans, so their genome offers a unique opportunity to identify genetic changes specific to anatomically fully modern humans. We have identified a 38,000-year-old Neanderthal fossil that is exceptionally free of contamination from modern human DNA. Direct high-throughput sequencing of a DNA extract from this fossil has thus far yielded over one million base pairs of hominoid nuclear DNA sequences. Comparison with the human and chimpanzee genomes reveals that modern human and Neanderthal DNA sequences diverged on average about 500,000 years ago. Existing technology and fossil resources are now sufficient to initiate a Neanderthal genome-sequencing effort.

Neanderthals were first recognized as a distinct group of hominids from fossil remains discovered 150 years ago at Feldhofer in Neander Valley, outside Düsseldorf, Germany. Subsequent Neanderthal finds in Europe and western Asia showed that fossils with Neanderthal traits appear in the fossil record of Europe and western Asia about 400,000 years ago and vanish about 30,000 years ago. Over this period they evolved morphological traits that made them progressively more distinct from the ancestors of modern humans that were evolving in Africa[1,2]. For example, the crania of late Neanderthals have protruding mid-faces, brain cases that bulge outward at the sides, and features of the base of the skull, jaw and inner ears that set them apart from modern humans[3].

The nature of the interaction between Neanderthals and modern humans, who expanded out of Africa around 40,000–50,000 years ago and eventually replaced Neanderthals as well as other archaic hominids across the Old World is still a matter of some debate. Although there is no evidence of contemporaneous cohabitation at any single site, there is evidence of geographical and temporal overlap in their ranges before the disappearance of Neanderthals. Additionally, late in their history, some Neanderthal groups adopted cultural traits such as body decorations, potentially through cultural interactions with incoming modern humans[4].

In 1997, a segment of the hypervariable control region of the maternally inherited mitochondrial DNA (mtDNA) of the Neanderthal type specimen found at Feldhofer was sequenced. Phylogenetic analysis showed that it falls outside the variation of contemporary humans and shares a common ancestor with mtDNAs of present-day humans approximately half a million years ago[5,6]. Subsequently, mtDNA sequences have been retrieved from eleven additional Neanderthal specimens: Feldhofer 2 in Germany[7], Mezmaiskaya in Russia[8], Vindija 75, 77 and 80 in Croatia[9,10], Engis 2 in Belgium, La Chapelle-aux-Saints and Rochers de Villeneuve in France[10], Scladina in Belgium[11], Monte Lessini in Italy[12], and El Sidron 441 in Spain[13]. Although some of these sequences are extremely short, they are all more closely related to one another than to modern human mtDNAs[9,11].

This fact, in conjunction with the absence of any related mtDNA sequences in currently living humans or in a small number of early modern human fossils[5,10] strongly suggests that Neanderthals contributed no mtDNA to present-day humans. On the basis of various population models, it has been estimated that a maximal overall genetic contribution of Neanderthals to the contemporary human gene pool is between 25% and 0.1% (refs 10, 14). Because the latter conclusions are based on mtDNA, a single maternally inherited locus, they are limited in their ability to detect a Neanderthal contribution to the current human gene pool both by the vagaries of genetic drift and by the possibility of a sex bias in reproduction. However, both morphological evidence[4,15] and the variation in the modern human gene pool[16] support the conclusion that if any genetic contribution of Neanderthals to modern human occurred, it was of limited magnitude.

Neanderthals are the hominid group most closely related to currently living humans, so a Neanderthal nuclear genome sequence would be an invaluable resource for annotating the human genome. Roughly 35 million nucleotide differences exist between the genomes of humans and chimpanzees, our closest living relatives[17]. Soon, genome sequences from other primates such as the orang-utan and the macaque will allow such differences to be assigned to the human and chimpanzee lineages. However, temporal resolution of the genetic changes along the human lineage, where remarkable morphological, behavioural and cognitive changes occurred, are limited without a more closely related genome sequence for comparison. In particular, comparison to the Neanderthal would enable the identification of genetic changes that occurred during the last few hundred thousand years, when fully anatomically and behaviourally modern humans appeared.

Identification of a Neanderthal fossil for DNA sequencing

Although it is possible to recover mtDNA[18] and occasionally even nuclear DNA sequences[19–22] from well-preserved remains of organisms that are less than a few hundred thousand years old, determination of ancient hominid sequences is fraught with special difficulties and pitfalls[18]. In addition to degradation and chemical damage to the DNA that can cause any ancient DNA to be irretrievable or misread, contamination of specimens, laboratory reagents and instruments with traces of DNA from modern humans must be avoided. In fact, when sensitive polymerase chain reaction (PCR) is used, human

[1]Max-Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany. [2]454 Life Sciences, 20 Commercial Street, Branford, Connecticut 06405, USA. [3]Institute of Quaternary Paleontology and Geology, Croatian Academy of Sciences and Arts, A. Kovacica 5/II, HR-10 000 Zagreb, Croatia.
‡Deceased.

mtDNA sequences can be retrieved from almost every ancient specimen[23,24]. This problem is especially severe when Neanderthal remains are studied because Neanderthal and human are so closely related that one expects to find few or no differences between Neanderthals and modern humans within many regions[25], making it impossible to rely on the sequence information itself to distinguish endogenous from contaminating DNA sequences. A necessary first step for sequencing nuclear DNA from Neanderthals is therefore to identify a Neanderthal specimen that is free or almost free of modern human DNA.

We tested more than 70 Neanderthal bone and tooth samples from different sites in Europe and western Asia for bio-molecular preservation by removing samples of a few milligrams for amino acid analysis. The vast majority of these samples had low overall contents of amino acids and/or high levels of amino acid racemization, a stereoisomeric structural change that affects amino acids in fossils, indicating that they are unlikely to contain retrievable endogenous DNA[26]. However, some of the samples are better preserved in that they contain high levels of amino acids (more than 20,000 p.p.m.), low levels of racemization of amino acids such as aspartate that racemize rapidly, as well as amino acid compositions that suggest that the majority of the preserved protein stems from collagen.

From 100–200 mg of bone from six of these specimens we extracted DNA and analysed the relative abundance of Neanderthal-like mtDNA sequences and modern human-like mtDNA sequences by performing PCR with primer pairs that amplify both human and Neanderthal mtDNA with equal efficiency. The amplification products span segments of the hypervariable region of the mtDNA in which all Neanderthals sequenced to date differ from all contemporary humans. From subsequent cloning into a plasmid vector and sequencing of more than a hundred clones from each product, we determined the ratio of Neanderthal-like to modern human-like mtDNA in each extract. We used two different primer pairs that amplify fragments of 63 base pairs and 119 base pair to gauge the contamination levels for different lengths of DNA molecules.

Figure 1 shows that the level of contamination differs drastically among the samples. Whereas only around 1% of the mtDNA present in three samples from France, Russia and Uzbekistan was Neanderthal-like, one sample from Croatia and one from Spain contained around 5% and 75% Neanderthal-like mtDNA, respectively. One bone (Vi-80) from Vindija Cave, Croatia, stood out in that ~99% of the 63-base-pair mtDNA segments and ~94% of the 119-base pair segments are of Neanderthal origin. Assuming that the ratio of Neanderthal to contaminating modern human DNA is the same for mtDNA as it is for nuclear DNA, the Vi-80 bone therefore yields DNA fragments that are predominantly of Neanderthal origin and provided that the contamination rate was not increased during the downstream sequencing process, the extent of contamination in the final analyses is below ~6%.

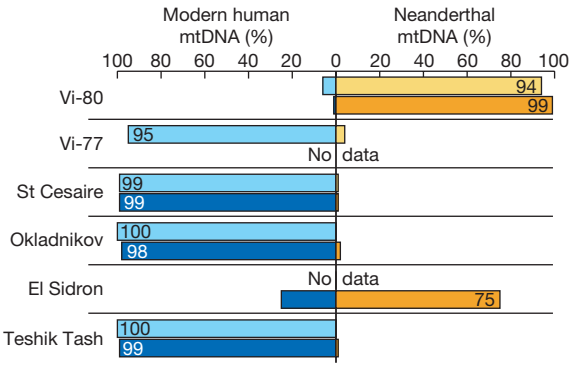

Figure 1 | Ratio of Neanderthal to modern human mtDNA in six hominid fossils. For each fossil, primer pairs that amplify a long (119 base pairs; upper lighter bars) and short (63 base pairs; lower darker bars) product were used to amplify segments of the mtDNA hypervariable region. The products were sequenced and determined to be either of Neanderthal (yellow) or modern human (blue) type.

The Vi-80 bone was discovered by M. Malez and co-workers in layer G3 of Vindija Cave in 1980. It has been dated by carbon-14 accelerator mass spectrometry to 38,310 ± 2,130 years before present and its entire mtDNA hypervariable region I has been sequenced[10]. Out of 14 Neanderthal remains from layer G3 that we have analysed, this bone is one of six samples that show good bio-molecular preservation, while the other eight bones show intermediate to bad states of preservation that do not suggest the presence of amplifiable DNA. Preservation conditions in Vindija Cave thus vary drastically from bone to bone, a situation that may be due to different extents of water percolation in different parts of the cave.

Direct large-scale DNA sequencing from the Vindija Neanderthal

Because the Vi-80 Neanderthal bone extract is largely free of contaminating modern human mtDNA, we chose this extract to perform large-scale parallel 454 sequencing[27]. In this technology, single-stranded libraries, flanked by common adapters, are created from the DNA sample and individual library molecules are amplified through bead-based emulsion PCR, resulting in beads carrying millions of clonal copies of the DNA fragments from the samples. These are subsequently sequenced by pyrosequencing on the GS20 454 sequencing system.

For several reasons, the 454 sequencing platform is extremely well suited for analyses of bulk DNA extracted from ancient remains[28]. First, it circumvents bacterial cloning, in which the vast majority of initial template molecules are lost during transformation and establishment of clones. Second, because each molecule is amplified in isolation from other molecules it also precludes template competition, which frequently occurs when large numbers of different DNA fragments are amplified together. Third, its current read length of 100–200 nucleotides covers the average length of the DNA preserved in most fossils[29]. Fourth, it generates hundreds of thousands of reads per run, which is crucial because the majority of the DNA recovered from fossils is generally not derived from the fossil species, but rather from organisms that have colonized the organism after its death[20,30]. Fifth, because each sequenced product stems from just one original single-stranded template molecule of known orientation, the DNA strand from which the sequence is derived is known[28]. This provides an advantage over traditional PCR from double-stranded templates, in which the template strand is not known, because the frequency of different nucleotide misincorporations can be deduced. For example, using 454 sequencing, the rate at which cytosine is converted to uracil and read as thymine can be distinguished from the rate at which guanine is converted to xanthine and read as adenine, whereas this is impossible using traditional PCR or bacterial cloning. This is important since nucleotide conversions and misincorporations in ancient DNA are caused by damage that affects different bases differently[28,31] and this pattern of false substitutions can be used to estimate the relative probability that a particular substitution (that is, the observation of a nucleotide difference between DNA sequences) represents the authentic DNA sequence of the organism versus an artefact from DNA degradation.

We recovered a total of 254,933 unique sequences from the Vi-80 bone (see Supplementary Methods). These were aligned to the human (build 36.1)[32], chimpanzee (build 1)[17] and mouse (build 34.1)[33] complete genome sequences, to environmental sample sequences in the GenBank *env* database (version 3, September 2005), and to the complete set of redundant nucleotide sequences in GenBank *nt* (version 3, September 2005, excluding EST, STS, GSS, environmental and HTGS sequences)[34] using the program BLASTN (NCBI version 2.2.12)[35]. The most similar database sequence for each query was identified and classified by its taxonomic order (Fig. 2) (see Supplementary Methods). No significant nucleotide sequence similarity in the databases was found for 79% of the fossil extract

sequence reads. This is typical of large-scale sequencing both from other ancient bones[20,22,28] and from environmental samples[36,37], although some permafrost-preserved specimens can yield high amounts of endogenous DNA[22]. Sequences with similarity to a database sequence were classified by the taxonomic order of their most significant alignment. Actinomycetales, a bacterial order with many soil-living species, was the most populous order and accounted for 6.8% of the sequences. The second most populous order, to which 15,701 unique sequences or 6.2% of the sequence reads were most similar, was that of primates. All other individual orders were substantially less frequent. Notably, the average percentage identity for the primate sequence alignments was 98.8%, whereas it was 92–98% for the other frequently occurring orders. Thus, the primate reads, unlike many of the prokaryotic reads, are aligned to a very closely related species.

Neanderthal mtDNA sequences

Among the 15,701 sequences of primate origin, we first identified all mtDNA in order to investigate whether their evolutionary relationship to the current human mtDNA pool is similar to what is known from previous analyses of Neanderthal mtDNA. A total of 41 unique DNA sequences from the Vi-80 fossil had their closest hits to different parts of the human mtDNA, and comprised, in total, 2,705 base pairs of unique mtDNA sequence. None of the putative Neanderthal mtDNA sequences map to the two hypervariable regions that have been previously sequenced in Neanderthals. We aligned these mtDNA sequences to the complete mtDNA sequences of 311 modern humans from different populations[38] as well as to the complete mtDNA sequences of three chimpanzees and two bonobos (Supplementary Information). A schematic neighbour-joining tree estimated from this alignment is shown in Fig. 3. In agreement with previous results, the Neanderthal mtDNA falls outside the variation among modern humans. However, the length of the branch leading to the Neanderthal mtDNA is 2.5 times as long as the branch leading to modern human mtDNAs. This is likely to be due to errors in our Neanderthal sequences derived from substitution artefacts from damaged, ancient DNA and from sequencing errors[28].

To analyse the extent to which errors occur in the Neanderthal mtDNA reads, we designed 29 primer pairs (Supplementary Methods) flanking all 39 positions at which the Vi-80 Neanderthal mtDNA sequences differed by substitutions from the consensus bases seen among the 311 human mtDNA sequences. These primer pairs, which are designed to yield amplification products that vary in length between 50 and 98 base pairs (including primers), were used in a multiplex two-step PCR[39] from the same Neanderthal extract that had been used for large-scale 454 sequencing. Twenty five of the PCR products, containing 34 of the positions where the Neanderthal differs from humans, were successfully amplified and cloned, and then six or more clones of each product were sequenced. The consensus sequence seen among these clones revealed the same nucleotides seen by the 454 sequencing at 20 of the 34 positions and no additional differences. Of the 14 positions found to represent errors in the sequence reads, seven were C to T transitions, four were G to A, two were G to T and one was T to C. This pattern of change is typical for ancient DNA, where deamination of cytosine residues[31] and, to a lesser extent, modifications of guanosine residues[28] have been found to account for the majority of nucleotide misincorporations during PCR.

These results also show that the likelihood of observing errors in the sequencing reads is drastically different depending on whether one considers nucleotide positions where a base in the Neanderthal mtDNA sequence differs from both the human and chimpanzee sequences, or positions where the Neanderthal differ from the humans but is identical to the chimpanzee mtDNA sequences. Among the mtDNA sequences analysed, there are 14 positions where the Neanderthal carries a base identical to the chimpanzee, and 13 of those were confirmed by PCR. In contrast, among the remaining 20 positions, where the Neanderthal sequences differed from both humans and chimpanzees, only seven were confirmed. When only PCR-confirmed sequence data are used to estimate the mtDNA tree (Fig. 3), the Neanderthal branch has a length comparable to that of contemporary humans. This suggests that no large source of errors other than what is detected by the PCR analysis affects the sequences.

Using these PCR-confirmed substitutions and a divergence time between humans and chimpanzees of 4.7–8.4 million years[40–42], we estimate the divergence time for the mtDNA fragments determined here to be 461,000–825,000 years. This is in general agreement with previous estimates of Neanderthal–human mtDNA divergence of 317,000–741,000 years[6] based on mtDNA hypervariable region sequences and is compatible with our presumption that the mtDNA sequences determined from the Vi-80 extract are of Neanderthal origin.

Nuclear DNA sequences

We next analysed the sequence reads whose closest matches are to the human or chimpanzee nuclear genomes and that are at least 30 base

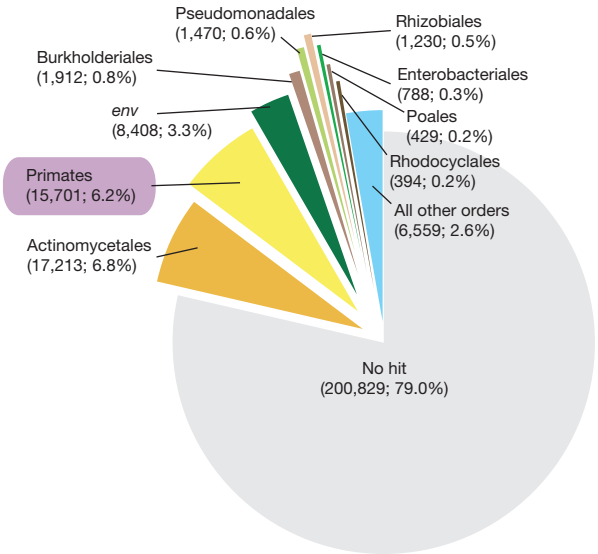

Figure 2 | Taxonomic distribution of DNA sequences from the Vi-80 extract. The taxonomic order of the database sequence giving the best alignment for each unique sequence read was determined. The most populous taxonomic orders are shown.

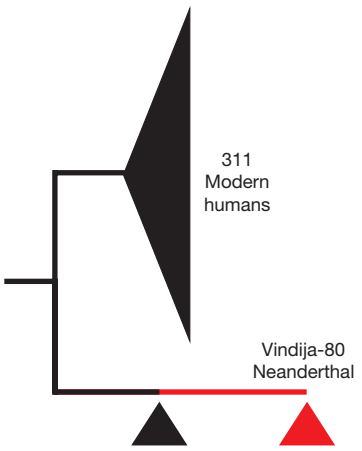

Figure 3 | Schematic tree relating the Vi-80 Neanderthal mtDNA sequences to 311 human mtDNA sequences. The Neanderthal branch length is given with uncorrected sequences (red triangle) and after correction of sequences via independent PCRs (black triangle). Chimpanzee and bonobo sequences (not shown) were used to root the neighbour-joining tree. Several substitution models (Kimura 2-parameter, Tajima-Nei, and Tamura 3-parameter with uniform or gamma-distributed ($\gamma = 0.5–1.1$) rates) yielded bootstrap support values for the human branch from 72–83%.

pairs long. Figure 4 shows where they map to the human karyotype (see Supplementary Methods). Overall, 0.04% of the autosomal genome sequence is covered by the Neanderthal reads—on average 3.61 bases per 10,000 bases. Both X and Y chromosomes are represented, with a lower coverage of 2.18 and 1.62 bases per 10,000, respectively, showing that the Vi-80 bone is derived from a male individual.

The data presented in Fig. 4 show that when the hit density for sequences that have a single best hit in the human genome is plotted along the chromosomes, several suggestive local deviations from the average hit density are seen, which may represent copy-number differences in the Neanderthal relative to the human reference genome. For comparison, we generated 454 sequence data from a DNA sample from a modern human. Interestingly, some of the deviations seen in the Neanderthal are present also in the modern human, whereas others are not. The latter group of sequences may indicate copy-number differences that are unique to the Neanderthal relative to the modern human genome sequence. Thus, when more Neanderthal sequence is generated in the future, it may be possible to determine copy number differences between the Neanderthal, the chimpanzee and the human genomes.

Patterns of nucleotide change on lineages

We generated three-way alignments between all Neanderthal sequences that map uniquely within the human genome and the corresponding human and chimpanzee genome sequences (see Supplementary Methods). An important artefact of local sequence alignments, such as those produced here, is that they necessarily begin and end with regions of exact sequence identity. The size of these regions is a function of the scoring parameters for the alignment. In this case, five bases at both ends of the alignments, amounting to ~14% of all data, needed to be removed (Supplementary Fig. 1) to eliminate biases in estimates of sequence divergence.

Each autosomal nucleotide position in the alignment that did not contain a deletion in the Neanderthal, the human or the chimpanzee sequences and was associated with a chimpanzee genome position with quality score ≥30 was classified according to which species share the same bases (Fig. 5). A total of 736,941 positions contained the same base in all three groups. The next largest category comprises 10,167 positions in which the human and Neanderthal base are identical, but the chimpanzee base is different. These positions are likely to have changed either on the hominid lineage before the divergence between human and Neanderthal sequences or on the chimpanzee lineage. At 3,447 positions, the Neanderthal base differs from both the human and chimpanzee bases, which are identical to each other. As suggested by the analysis of the mtDNA sequences, this category contains positions that have changed on the Neanderthal lineage, as well as a large proportion of errors that derive both from base damage that have accumulated in the ancient DNA and from sequencing errors. At 434 positions, the human base differs from both the Neanderthal and chimpanzee bases, which are identical to each other.

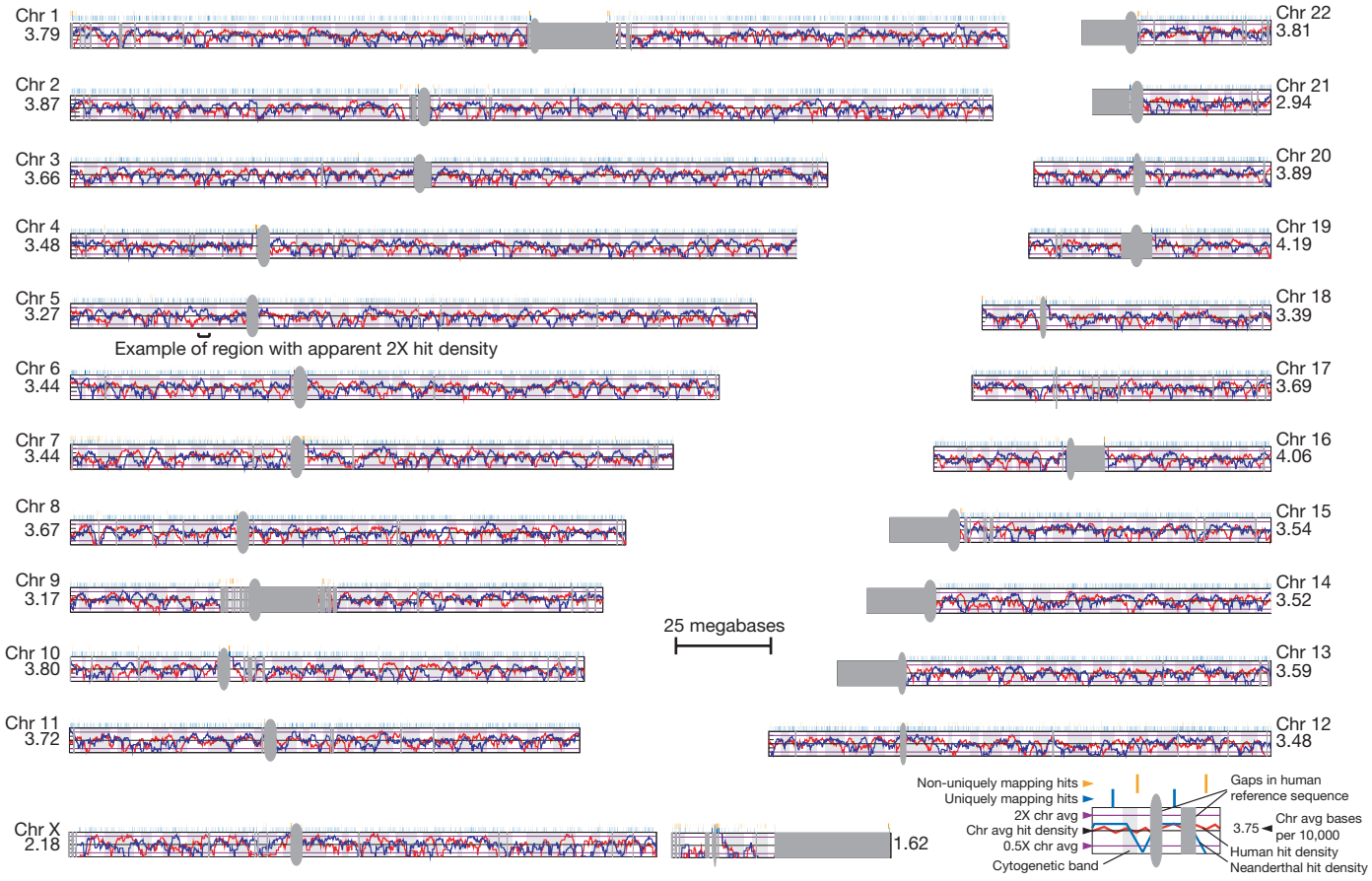

Figure 4 | Location on the human karyotype of Neanderthal DNA sequences. All sequences longer than 30 nucleotides whose best alignments were to the human genome are shown. The blue lines above each chromosome mark the position of all alignments that are unique in terms of bit-score within the human genome. Orange lines are alignments that have more than one alignment of equal bit-score. To the left of each chromosome, the average number of Neanderthal bases per 10,000 is given. Lines (Neanderthal, blue; human, red) within each chromosome show the hit density, on a log-base 2 scale, within sliding windows of 3 megabases along each chromosome. The centre black lines indicate the average hit-density for the chromosomes. The purple lines above and below indicate hit densities of 2X and 1/2X the chromosome average, respectively. On chromosome 5, an example of a region of increased sequence density is highlighted. Sequence gaps in the human reference sequence are indicated by dark grey regions. Chromosomal banding pattern is indicated by light grey regions.

These positions are likely to have changed on the human lineage after the divergence from Neanderthal. Finally, a total of 51 positions contain different bases in all three groups.

Because the 454 sequencing technology allows the base in a base pair from which a sequence is derived to be determined, the relative frequencies of each of the 12 possible categories of base changes can be estimated for each evolutionary lineage. As seen in Fig. 5, the patterns of the chimpanzee-specific and human-specific changes are similar to each other in that the eight transversional changes are of approximately equal frequency and about fourfold less frequent than each of the four transitional changes, yielding a transition to transversion ratio of 2.04, typical of closely related mammalian genomes[43]. For the Neanderthal-specific changes the pattern is very different in that mismatches are dominated by C to T and G to A differences. Thus, the pattern of change seen among the Neanderthal-specific alignment mismatches is typical of the nucleotide substitution pattern observed in PCR of ancient DNA.

Consistent with this, modern human sequences determined by 454 sequencing show no excess amount of C to T or G to A differences (Supplementary Fig. 2), indicating that lesions in the ancient DNA rather than sequencing errors account for the majority of the errors in the Neanderthal sequences. Assuming that the evolutionary rate of DNA change was the same on the Neanderthal and human lineages, the majority of observed differences specific to the Neanderthal lineage are artefacts. All Neanderthal-specific changes were therefore disregarded in the subsequent analyses and the Neanderthal sequences were used solely to assign changes to the human or chimpanzee lineage where the human and chimpanzee genome sequences differ and the Neanderthal sequence carries either the human or the chimpanzee base.

Genomic divergence between Neanderthals and humans

Assuming that the rates of DNA sequence change along the chimpanzee lineage and the human lineage were similar, it can be estimated that 8.2% of the DNA sequence changes that have occurred on the human lineage since the divergence from the chimpanzee lineage occurred after the divergence of the Neanderthal lineage. However, although the Neanderthal-specific changes that are heavily influenced by errors are not used for this analysis, some errors in the single-pass sequencing reads from the Neanderthal extract will create positions where the Neanderthal is identical either to human or chimpanzee sequences, and thus affect the estimates of sequence change on the human and chimpanzee lineages. When the effects of such errors in the Neanderthal sequences are quantified and removed (see Supplementary Methods), ~7.9% of the sequence changes along the human lineage are estimated to have occurred after divergence from the Neanderthal. If the human–chimpanzee divergence time is set to 6,500,000 years (refs 40, 41, 44), this implies an average human–Neanderthal DNA sequence divergence time of ~516,000 years. A 95% confidence interval generated by bootstrap re-sampling of the alignment data gives a range of 465,000 to 569,000 years. Obviously, these divergence estimates are dependent on the human–chimpanzee divergence time, which is a much larger source of uncertainty.

We analysed the DNA sequences generated from a contemporary human using the same sequencing protocol as was used for the Neanderthal. Although ancient DNA is degraded and damaged, this comparison controls for many of the aspects of the analysis including sequencing and alignment methodology. In this case, ~7.1% of the divergence along the human lineage is assigned to the time subsequent to the divergence of the two human sequences. The average divergence time between alleles within humans is thus ~459,000 years with a 95% confidence interval between 419,000 and 498,000 years. As expected, this estimate of the average human diversity is less than the divergence seen between the human and the Neanderthal sequences, but constitutes a large fraction of it because much of the human sequence diversity is expected to predate the human–Neanderthal split[25]. Neanderthal genetic differences to humans must therefore be interpreted within the context of human diversity.

Ancestral population size

Humans differ from apes in that their effective population size is of the order of 10,000 while those of chimpanzees, gorillas and orangutans are two to four times larger[45–47]. Furthermore, the population size of the ancestor of humans and chimpanzees was found to be similar to those of apes, rather than to humans[42,48]. The Neanderthal sequence data now allow us to ask if the effective size of the population ancestral to humans and Neanderthals was large, as is the case for apes and the human–chimpanzee ancestor, or small, as for present-day humans.

We applied a method[42] that co-estimates the ancestral effective population size and the split time between Neanderthal and human populations (Fig. 6a; see Supplementary Methods). As seen in Fig. 6b, we recover a line describing combinations of population sizes and split times compatible with the data and lack power to be more

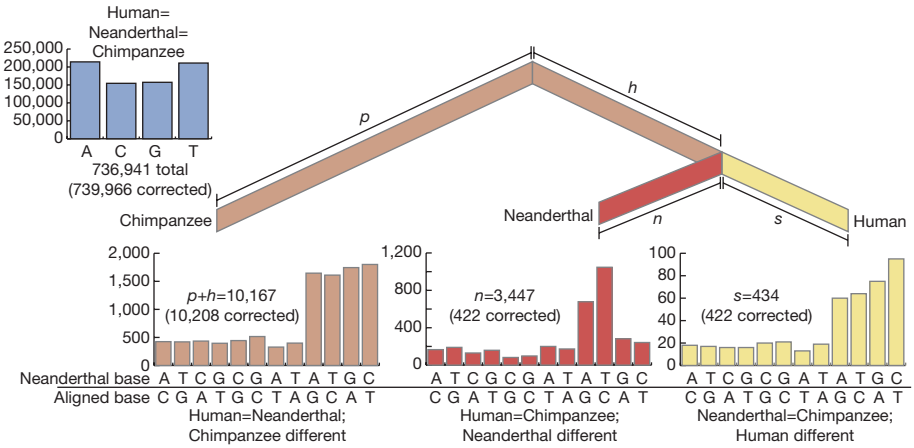

Figure 5 | Schematic tree illustrating the number of nucleotide changes inferred to have occurred on hominoid lineages. In blue is the distribution of all aligned positions that did not change on any lineage. In brown are the changes that occurred either on the chimpanzee lineage (p) or on the hominid lineage (h) before the human and Neanderthal lineages diverged. In red are the changes that are unique to the Neanderthal lineage (n), including all changes due to base-damage and base-calling errors. In yellow are changes unique to the human lineage. The distributions of types of changes in each category are also given. The numbers of changes in each category, corrected for base-calling errors in the Neanderthal sequence (see Supplementary Methods), are shown within parentheses.

precise (see Supplementary Methods and Results). Using this line we can estimate the ancestral population size, given estimates about the population split time from independent sources. If we use a split time of 400,000 years inferred from the fossil record (J. J. Hublin, personal communication), then our point estimate of the ancestral population size is ~3,000. Given uncertainty in both the sequence divergence time and the population split time, our estimate of the ancestral population size varies from 0 to 12,000.

These results suggest that the population ancestral to present-day humans and Neanderthals was similar to present-day humans in having a small effective size and thus that the effective population size on the hominid lineage had already decreased before the split between humans and Neanderthals. Therefore, the small effective population size seen in present-day human samples may not be unique to modern humans, but was present also in the common ancestor of Neanderthals and modern humans. We speculate that a small effective size, perhaps associated with numerous expansions from small groups, was typical not only of modern humans but of many groups of the genus *Homo*. In fact, the origin of *Homo erectus* may have been associated with genetic or cultural adaptations that resulted in drastic population expansions as indicated by their appearance outside Africa around two million years ago.

Neanderthal sequences and human polymorphisms

Another question that can be addressed with these data is how often the Neanderthal has the ancestral allele (that is, the same allele seen in the chimpanzee) versus the derived (or novel) allele at sites where humans carry a single nucleotide polymorphism (SNP). The latter case identifies SNPs that were present in the common ancestor of Neanderthals and present-day humans. Using the SNPs that overlap with our data from two large genome-wide data sets (HapMap[49], 786 SNPs and Perlegen[50], 318 SNPs), we find that the Neanderthal sample has the derived allele in ~30% of all SNPs. This number is presumably an overestimate since the SNPs analysed were ascertained to be of high frequency in present-day humans and hence are more likely to be old. Nevertheless, this high level of derived alleles in the Neanderthal is incompatible with the simple population split model estimated in the previous section, given split times inferred from the fossil record. This may suggest gene flow between modern humans and Neanderthals. Given that the Neanderthal X chromosome shows a higher level of divergence than the autosomes (R.E.G., unpublished observation), gene flow may have occurred predominantly from modern human males into Neanderthals. More extensive sequencing of the Neanderthal genome is necessary to address this possibility.

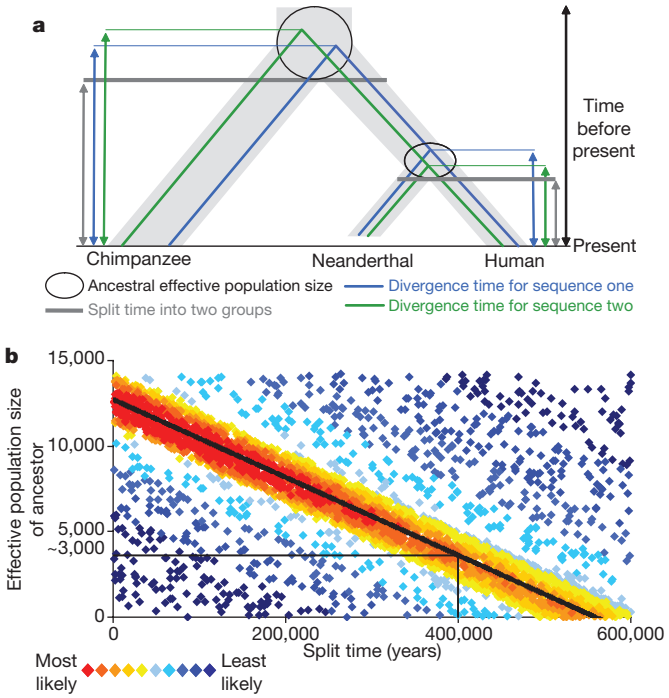

Figure 6 | Estimate of the effective population size of the ancestor of humans and Neanderthals. a, Schematic illustration of the model used to estimate ancestral effective population size. By split time, we mean the time, in the past, after which there was no more interbreeding between two groups. By divergence, we mean the time, in the past, at which two genetic regions separated and began to accumulate substitutions independently. Effective population size is the number of individuals needed under ideal conditions to produce the amount of observed genetic diversity within a population. **b**, The likelihood estimates of population split times and ancestral population sizes. The likelihoods are grouped by colour. The red–yellow points are statistically equivalent based on the likelihood ratio test approximation. The black line is the line of best fit to red–yellow points (see Supplementary Methods). This graph is scaled assuming a human–chimpanzee average sequence divergence time of 6,500,000 years.

Rationale and prospects for a Neanderthal genome sequence

We demonstrate here that DNA sequences can be generated from the Neanderthal nuclear genome by massive parallel sequencing on the 454 sequencing platform. It is thus feasible to determine large amounts of sequences from this extinct hominid. As a corollary, it is possible to envision the determination of a Neanderthal genome sequence. For several reasons, we believe that this would represent a valuable genomic resource.

First, a Neanderthal genome sequence would allow all nucleotide sequence differences as well as many copy-number differences between the human and chimpanzee genomes to be temporally resolved with respect to whether they occurred before the separation of humans from Neanderthals, or whether they occurred after or at the time of separation. The latter class of changes is of interest, because some of them will be associated with the emergence of modern humans. A Neanderthal genome sequence would therefore allow the research community to determine whether DNA sequence differences between humans and chimpanzees that are found to be functionally important represent recent changes on the human lineage. No data other than a Neanderthal genome sequence can provide this information.

Second, the fact that Neanderthals carry the derived allele for a substantial fraction of human SNPs suggests a method of identifying genomic regions that have experienced a selective sweep subsequent to the separation of human and Neanderthal populations. Such selective sweeps in the human genome will make the variation in these regions younger than the separation of humans and Neanderthals. As we show above, in regions not affected by sweeps a substantial proportion of polymorphic sites in humans will carry derived alleles in the Neanderthal genome sequence, whereas no sites will do so in regions affected by sweeps. This represents an approach to identifying selective sweeps in humans that is not possible from other data.

Third, once large amounts of Neanderthal genome sequence is generated, it will become possible to estimate the misincorporation probabilities for each class of nucleotide differences between the Neanderthal and chimpanzee genomes with high accuracy by analysing regions covered by many reads such as mtDNA, repeated genome regions of high sequence identity, as well as single-copy regions covered by multiple reads. Once this is done, the confidence that any particular nucleotide position where the Neanderthal differs from human as well as chimpanzee is correct can be reliably estimated. In combination with future knowledge about the function of genes and biological systems, comprehensive information from the Neanderthal genome will then allow aspects of Neanderthal biology to be deciphered that are unavailable by any other means.

Are fossil and technical resources today sufficient to imagine the determination of a Neanderthal genome sequence? The results presented here are derived from approximately one fifteenth of an extract

prepared from ~100 mg of bone. To achieve one-fold coverage of the Neanderthal genome (3 gigabases) without any further improvement in technology, about twenty grams of bone and 6,000 runs on the current version of the 454 sequencing platform would be necessary. Although this is at present a daunting task, technical improvements in the procedures described here that would make the retrieval of DNA sequences of the order of ten times more efficient can easily be envisioned (our unpublished results). In view of that prospect, we have recently initiated a project that aims at achieving an initial draft version of the Neanderthal genome within two years.

Received 14 July; accepted 11 October 2006.

1. Bischoff, J. L. *et al.* The Sima de los Huesos hominids date to beyond U/Th equilibrium (>350 kyr) and perhaps to 400–500 kyr: New radiometric dates. *J. Archaeol. Sci.* **30,** 275–280 (2003).
2. Hublin, J-J. (ed.) *Climatic Changes, Paleogeography, and the Evolution of the Neandertals* (Plenum Press, New York, 1998).
3. Franciscus, R. G. (ed.) *Neanderthals* (Oxford Univ. Press, Oxford, 2002).
4. Hublin, J. J., Spoor, F., Braun, M., Zonneveld, F. & Condemi, S. A late Neanderthal associated with Upper Palaeolithic artefacts. *Nature* **381,** 224–226 (1996).
5. Krings, M. *et al.* Neandertal DNA sequences and the origin of modern humans. *Cell* **90,** 19–30 (1997).
6. Krings, M., Geisert, H., Schmitz, R. W., Krainitzki, H. & Pääbo, S. DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen. *Proc. Natl Acad. Sci. USA* **96,** 5581–5585 (1999).
7. Schmitz, R. W. *et al.* The Neandertal type site revisited: Interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. *Proc. Natl Acad. Sci. USA* **99,** 13342–13347 (2002).
8. Ovchinnikov, I. V. *et al.* Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* **404,** 490–493 (2000).
9. Krings, M. *et al.* A view of Neandertal genetic diversity. *Nature Genet.* **26,** 144–146 (2000).
10. Serre, D. *et al.* No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biol.* **2,** 313–317 (2004).
11. Orlando, L. *et al.* Revisiting Neandertal diversity with a 100,000 year old mtDNA sequence. *Curr. Biol.* **16,** R400–R402 (2006).
12. Caramelli, D. *et al.* A highly divergent mtDNA sequence in a Neandertal individual from Italy. *Curr. Biol.* **16,** R630–R632 (2006).
13. Lalueza-Fox, C. *et al.* Neandertal evolutionary genetics: mitochondrial DNA data from the Iberian peninsula. *Mol. Biol. Evol.* **22,** 1077–1081 (2005).
14. Currat, M. & Excoffier, L. Modern humans did not admix with Neanderthals during their range expansion into Europe. *PLoS Biol.* **2,** e421 (2004).
15. Stringer, C. Modern human origins: progress and prospects. *Phil. Trans. R. Soc. Lond. B* **357,** 563–579 (2002).
16. Takahata, N., Lee, S. H. & Satta, Y. Testing multiregionality of modern human origins. *Mol. Biol. Evol.* **18,** 172–183 (2001).
17. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
18. Pääbo, S. *et al.* Genetic analyses from ancient DNA. *Annu. Rev. Genet.* **38,** 645–679 (2004).
19. Greenwood, A. D., Capelli, C., Possnert, G. & Paabo, S. Nuclear DNA sequences from late Pleistocene megafauna. *Mol. Biol. Evol.* **16,** 1466–1473 (1999).
20. Noonan, J. P. *et al.* Genomic sequencing of Pleistocene cave bears. *Science* **309,** 597–599 (2005).
21. Rompler, H. *et al.* Nuclear gene indicates coat-color polymorphism in mammoths. *Science* **313,** 62 (2006).
22. Poinar, H. N. *et al.* Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* **311,** 392–394 (2006).
23. Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M. & Pääbo, S. Ancient DNA. *Nature Rev. Genet.* **2,** 353–359 (2001).
24. Malmstrom, H., Stora, J., Dalen, L., Holmlund, G. & Gotherstrom, A. Extensive human DNA contamination in extracts from ancient dog bones and teeth. *Mol. Biol. Evol.* **22,** 2040–2047 (2005).
25. Pääbo, S. Human evolution. *Trends Cell Biol.* **9,** M13–M16 (1999).
26. Poinar, H. N., Höss, M., Bada, J. L. & Pääbo, S. Amino acid racemization and the preservation of ancient DNA. *Science* **272,** 864–866 (1996).
27. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
28. Stiller, M. *et al.* Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. *Proc. Natl Acad. Sci. USA* **103,** 13578–13584 (2006).
29. Pääbo, S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc. Natl Acad. Sci. USA* **86,** 1939–1943 (1989).
30. Höss, M., Dilling, A., Currant, A. & Pääbo, S. Molecular phylogeny of the extinct ground sloth *Mylodon darwinii*. *Proc. Natl Acad. Sci. USA* **93,** 181–185 (1996).
31. Hofreiter, M., Jaenicke, V., Serre, D., Haeseler Av, A. & Pääbo, S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* **29,** 4793–4799 (2001).
32. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
33. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
34. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* **34,** D16–D20 (2006).
35. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
36. Beja, O. *et al.* Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.* **2,** 516–529 (2000).
37. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
38. Ingman, M. & Gyllensten, U. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. *Nucleic Acids Res.* **34,** D749–D751 (2006).
39. Krause, J. *et al.* Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature* **439,** 724–727 (2006).
40. Kumar, S., Filipski, A., Swarna, V., Walker, A. & Blair Hedges, S. Placing confidence limits on the molecular age of the human-chimpanzee divergence. *Proc. Natl Acad. Sci. USA* **102,** 18842–18847 (2005).
41. Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441,** 1103–1108 (2006).
42. Wall, J. D. Estimating ancestral population sizes and divergence times. *Genetics* **163,** 395–404 (2003).
43. Yang, Z. & Yoder, A. D. Estimation of the transition/transversion rate bias and species sampling. *J. Mol. Evol.* **48,** 274–283 (1999).
44. Innan, H. & Watanabe, H. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. *Mol. Biol. Evol.* **23,** 1040–1047 (2006).
45. Kaessmann, H., Wiebe, V., Weiss, G. & Pääbo, S. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature Genet.* **27,** 155–156 (2001).
46. Yu, N., Jensen-Seaman, M. I., Chemnick, L., Ryder, O. & Li, W-H. Nucleotide diversity in gorillas. *Genetics* **166,** 1375–1383 (2004).
47. Fischer, A., Pollack, J., Thalmann, O., Nickel, B. & Pääbo, S. Demographic history and genetic differentiation in apes. *Curr. Biol.* **16,** 1133–1138 (2006).
48. Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164,** 1645–1656 (2003).
49. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
50. Hinds, D. A. *et al.* Whole genome patterns of common DNA variation in diverse human populations. *Science* **307,** 1072–1079 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are indebted to G. Coop, W. Enard, I. Hellmann, A. Fischer, P. Johnson, S. Kudaravalli, M. Lachmann, T. Maricic, J. Pritchard, J. Noonan, D. Reich, E. Rubin, M. Slatkin, L. Vigilant and T. Weaver for discussions. We thank A. P. Derevianko, C. Lalueza-Fox, A. Rosas and B. Vandermeersch for fossil samples. We also thank the Croatian Academy of Sciences and Arts for support and the Innovation Fund of Max Planck Society for financial support. 454 Life Sciences thanks NHGRI for continued support for the development of this platform, as well as all of its employees who developed the sequencing system. R.E.G. is supported by an NSF postdoctoral fellowship in Biological Informatics.

Author Contributions M.P. provided Neanderthal samples and palaeontological information; J.M.R. and S.P. conceived of and initiated the 454 Neanderthal sequencing approach; M.T.R. developed the library preparation method, and generated and processed the sequencing data; J.F.S. planned and coordinated library preparation and sequencing activities; L.D. processed and transferred data between 454 Life Sciences and the MPI; M.E. supervised, planned and coordinated research between MPI and 454 Life Sciences; J.K. and A.W.B. extracted ancient DNA and performed analyses in the ''Identification of a Neanderthal fossil for DNA sequencing'' section; J.K. and R.E.G. performed analyses in the ''Neanderthal mtDNA sequences'' section; R.E.G. performed the analyses in the sections ''Direct large-scale DNA sequencing'' to ''Genomic divergence between Neanderthals and humans''; S.E.P. performed analyses in the sections ''Ancestral population size'' and ''Neanderthal sequences and human polymorphisms''; S.P. conceived of the ideas presented in the section ''Rationale and prospects for a Neanderthal genome sequence'', and initiated, planned and coordinated the study; R.E.G., S.E.P., J.K. and S.P. wrote the paper.

Author Information Neanderthal fossil extract sequences were deposited at EBI with accession numbers CAAN01000001–CAAN01369630. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to R.E.G. (green@eva.mpg.de).