# Sequencing and analysis of Neanderthal genomic DNA

James P. Noonan[1,2], Graham Coop[3], Sridhar Kudaravalli[3], Doug Smith[1], Johannes Krause[4], Joe Alessi[1], Feng Chen[1], Darren Platt[1], Svante Pääbo[4], Jonathan K. Pritchard[3] and Edward M. Rubin[1,2]

1. US DOE Joint Genome Institute, Walnut Creek, CA
2. Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA
3. Department of Human Genetics, University of Chicago, Chicago, IL
4. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

One sentence summary: Recovery and analysis of multiple Neanderthal autosomal sequences using a metagenomic approach reveals that modern humans and Neanderthals split ~400,000 years ago, without significant evidence of subsequent admixture.

Our knowledge of Neanderthals is based on a limited number of remains and artifacts from which we must make inferences about their biology, behavior and relationship to ourselves. Here we describe the characterization of these extinct hominids from a new perspective based on the development, high-throughput sequencing and analysis of a Neanderthal metagenomic library. Several lines of evidence indicate the 66,643 bp of hominid sequence so far identified in the library is of Neanderthal origin, the strongest being the identification of sequence differences in humans at sites where Neanderthal and chimpanzee genomic sequences are identical. These findings enabled us to calculate the human-Neanderthal divergence time based on multiple, randomly distributed autosomal loci. Our analyses suggest that on average the Neanderthal genomic sequence we obtained and the reference human genome sequence share a most recent common ancestor ~770,000 years ago, and that the human and Neanderthal ancestral populations split ~400,000 years ago, prior to the emergence of anatomically modern humans. This study contributes to our understanding of the evolutionary relationship of *Homo sapiens* and *Homo neanderthalensis* and signifies the beginning of Neanderthal genomics.

Neanderthals are the closest hominid relatives of modern humans, and possibly as late as thirty thousand years ago, humans and Neanderthals coexisted in Europe and western Asia (1). Since that time, our species has spread across the Earth, far surpassing any previous hominid or primate species in numbers, technological development and environmental impact, while Neanderthals have vanished (2). Molecular studies of Neanderthals have been exclusively constrained to comparison of human and PCR-amplified Neanderthal mitochondrial sequences, which suggest that the most recent common ancestor of humans and Neanderthals existed ~500,000 years ago, well before the emergence of modern humans (3-5). Further analyses of mitochondrial data, including comparison of mitochondrial sequences obtained from several Neanderthals and early modern humans, suggest little or no admixture between Neanderthal and modern human populations in Europe (3, 4, 6, 7). However, a major limitation of these studies is that mitochondrial sequences only reflect maternal inheritance of a single locus. Accordingly, in the absence of Neanderthal autosomal and Y-chromosome sequences, the assessment of human-Neanderthal admixture remains incomplete. Mitochondrial data also provide no access to the molecular differences between humans and Neanderthals that would help to reveal biological features unique to each. These insights await the recovery of Neanderthal genomic sequence.

The introduction of high-throughput sequencing technologies and recent advances in metagenomic analysis of complex DNA mixtures now provide a strategy to recover genomic sequence from ancient remains (8-11). In contrast to previous efforts to obtain ancient sequences by direct analysis of extracts (3-6, 12), metagenomic libraries allow the immortalization of DNA isolated from precious ancient samples, obviating the need for

repeated destructive extractions (10). In addition, once an ancient DNA fragment is cloned into a metagenomic library, it can be distinguished from contamination that might be introduced during subsequent PCR amplification or sequencing by the vector sequences linked to each library-derived insert (Fig. 1A).

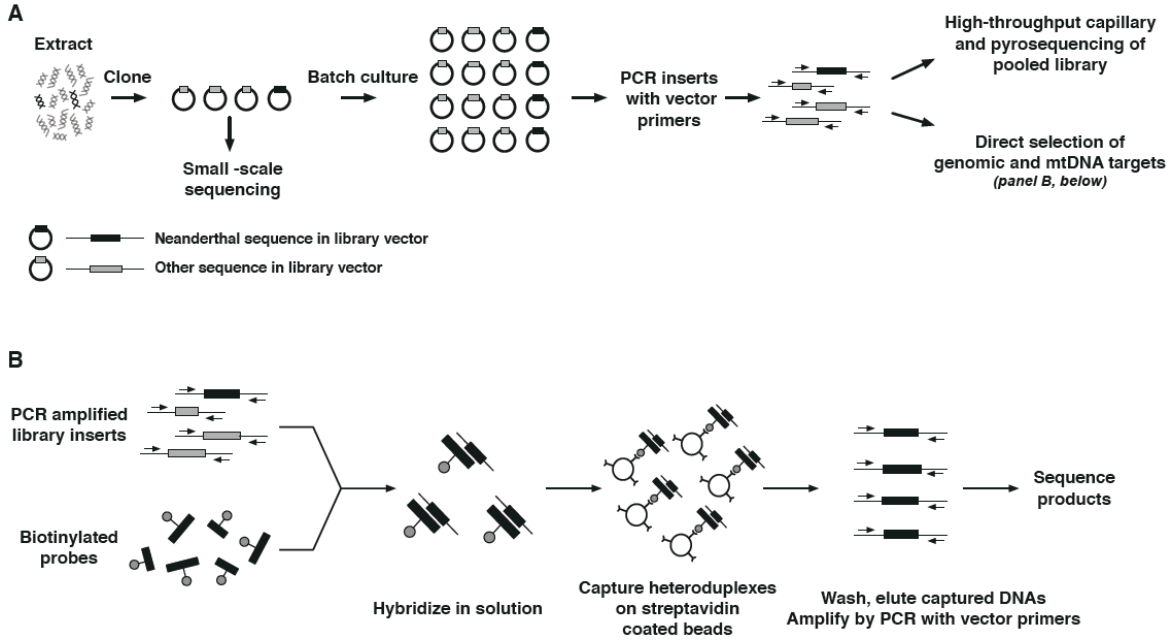


**Figure 1. (A)** Generation of ancient metagenomic library DNAs for direct selection and pyrosequencing. **(B)** Recovery of Neanderthal genomic sequences from library NE1 by direct genomic selection.

In this study we apply an amplification independent direct cloning method to construct a Neanderthal metagenomic library using DNA extracted from a 38,000-year old specimen from Vindija, Croatia (5, 13). We have recovered 66,643 bp of Neanderthal genome sequence from this library by high-throughput sequencing and have isolated specific Neanderthal sequences by direct genomic selection. Several lines of evidence indicated that the hominid sequences in this library were largely Neanderthal, rather than modern human contamination. Mitochondrial PCR analysis of the extract used to build the

library, using an amplicon of similar size as the average hominid sequence identified in the library (Fig. 2), revealed that only 2% of the products were from contaminating modern human DNA, while the remaining 98% were Neanderthal (14). Signatures of damage in the hominid sequences that are characteristic of ancient DNA also suggested that they were ancient. Finally and most importantly, comparison of hominid sequences from the library to orthologous human and chimpanzee genomic sequences identified human-specific substitutions at sites where the hominid sequence was identical to chimpanzee, enabling us to make estimates of the human-Neanderthal divergence time (3, 4).

| | **Individual Clones** | **Batch Culture** | |
|---|---|---|---|
| Sequencing chemistry | Sanger | Sanger | Pyrosequencing |
| Reads | 9984 | 19,200 | 1,474,910 |
| Average insert | 111 bp | 111 bp | n.a. |
| Average BLAST hit | 52 bp | 52 bp | 48 bp |
| Unique loci | 139 | 32 | 1169 |
| Total unique hominid sequence | 7282 bp | 2660 bp | 56,701 bp |

**Table 1.** Amount of unique Neanderthal sequence obtained from library NE1 by Sanger sequencing of individual clones, as well as Sanger sequencing and pyrosequencing of clones in batch culture.

We initially assessed the Neanderthal genomic sequence content of library NE1 by Sanger sequencing of individual clones, which allowed individual library inserts to be completely sequenced and thus provided a direct measure of hominid insert size. We sequenced 9984 clones and obtained 139 (1.4%) with significant BLAST hit similarity to the human genome ($E \leq 1e\text{-}3$), yielding 7,282 bp of hominid sequence (15). The average library insert size was 111 bp and the average hit size to the human genome was 52 bp

(Fig. 2 and Table 1). The small average size of these putatively ancient Neanderthal fragments is similar to results we previously obtained from two Pleistocene cave bear libraries, in which the average library insert size was between 100 and 200 bp while BLAST hits to reference carnivore genome sequences were on average 69 bp (Fig. 2; 10). The small BLAST hit sizes and insert sizes in both cave bear and Neanderthal metagenomic libraries are consistent with the degradation of ancient genomic DNA into small fragments over tens of thousands of years.

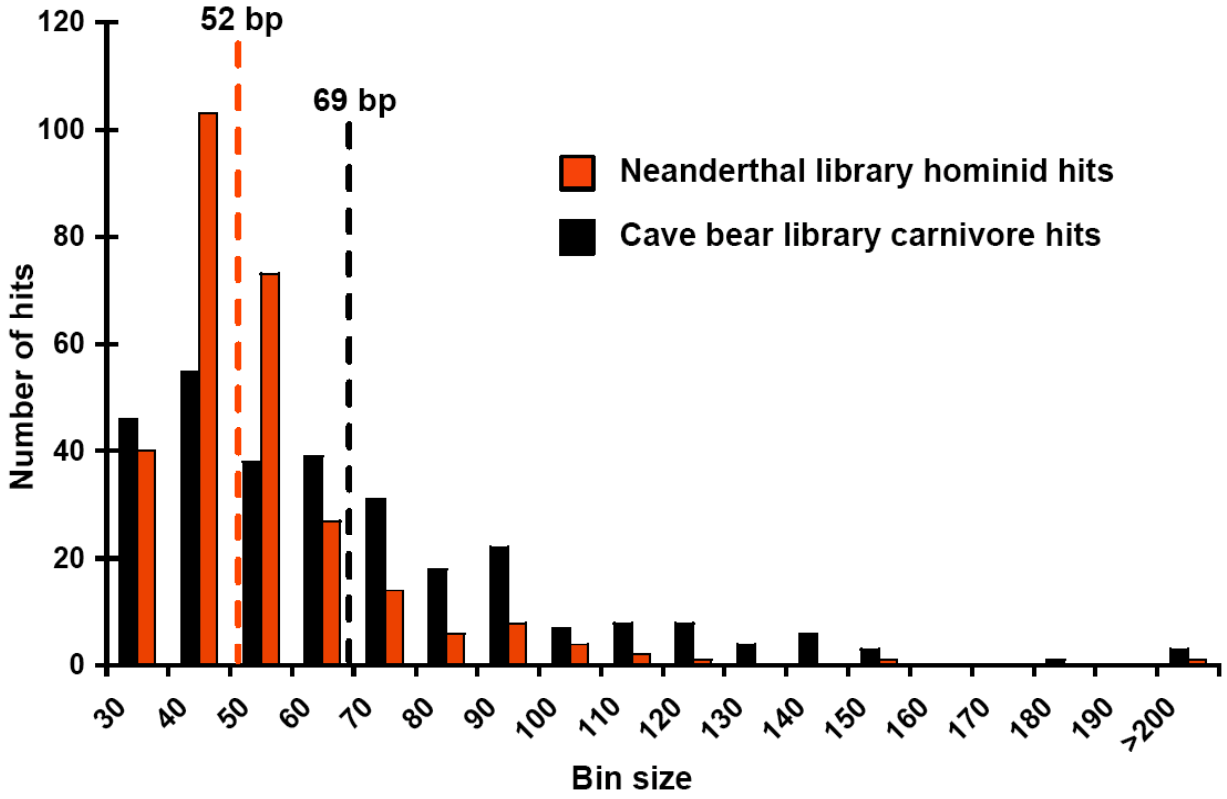


**Figure 2.** Size distribution, plotted in 10-bp bins, of Neanderthal and cave bear sequences obtained from metagenomic libraries by Sanger sequencing of individual clones. The average hit size in each case indicated by a dotted line.

Sanger sequencing of individual clones from library NE1 suggested that it contained sufficient amounts of Neanderthal sequence to conduct a random sequence survey of the Neanderthal genome. Such a survey would provide many unlinked autosomal loci that could be used to make estimates of human-Neanderthal divergence time and an assessment of the degree of admixture between Neanderthal males and early modern human females. However, the low sequence yield of library NE1 required sequencing a very large number of clones to obtain enough Neanderthal genome sequence for these analyses. We therefore carried out deep sequencing of pooled inserts from library NE1 using massively parallel pyrosequencing. To obtain pooled inserts, we amplified transformed NE1 library DNA in liquid batch culture and recovered library inserts from purified plasmid DNA by PCR (Fig. 1A). We generated 1.47 million pyrosequencing reads, compared each to the human genome sequence by MEGABLAST and obtained 7880 hits. Assembly of these reads and reanalysis of the resulting scaffolds by BLASTN produced 1169 unique Neanderthal loci, yielding 56,701 bp of Neanderthal genomic sequence (13).

The pyrosequencing approach, while generating significant amounts of sequence, does so at an increased error rate compared to Sanger sequencing (11). Identification of human-specific substitutions and estimates of human-Neanderthal divergence times from genomic data could be affected by such errors. To assess the quality of Neanderthal pyrosequencing data, we generated consensus sequences from pyrosequencing reads overlapping the same Neanderthal genomic locus and filtered out low quality positions in the resulting contigs ($Q < 15$). To determine if these contigs contained additional errors not detectable by quality score filtering, we also analyzed by Sanger sequencing 19,200 clones

from the same batch culture used to generate the pyrosequencing data. This sequencing yielded an additional 32 unique Neanderthal loci and 130 loci (6.2 kb) also represented in the pyrosequencing data (Table 1). Sanger sequencing and pyrosequencing results for these 130 Neanderthal loci agreed at 99.89% of ungapped positions. In addition, Sanger sequencing and pyrosequencing yielded Neanderthal sequences that were nearly equally divergent from the human reference sequence (pyrosequencing = 0.47% divergence, Sanger sequencing = 0.49%). These results indicate that the frequency of single-base errors is likely no greater in Neanderthal genomic sequence obtained from assembled, quality filtered pyrosequencing data compared to that obtained from Sanger sequencing.

The low complexity of library NE1 made these analyses possible, as it resulted in a limited number of clones in the library that were amplified by batch culture and PCR and then sequenced in depth. We estimated that the coverage obtained in library NE1 (~0.00002x) is significantly lower than that previously obtained in cave bear metagenomic libraries prepared from samples of similar age to the Neanderthal sample used here (10). The low coverage in library NE1 is more likely due to the quality of this particular library rather than a general feature of ancient DNA. Nevertheless, we were able to obtain substantial amounts of authentic Neanderthal genomic sequence from the library by deep sequencing.

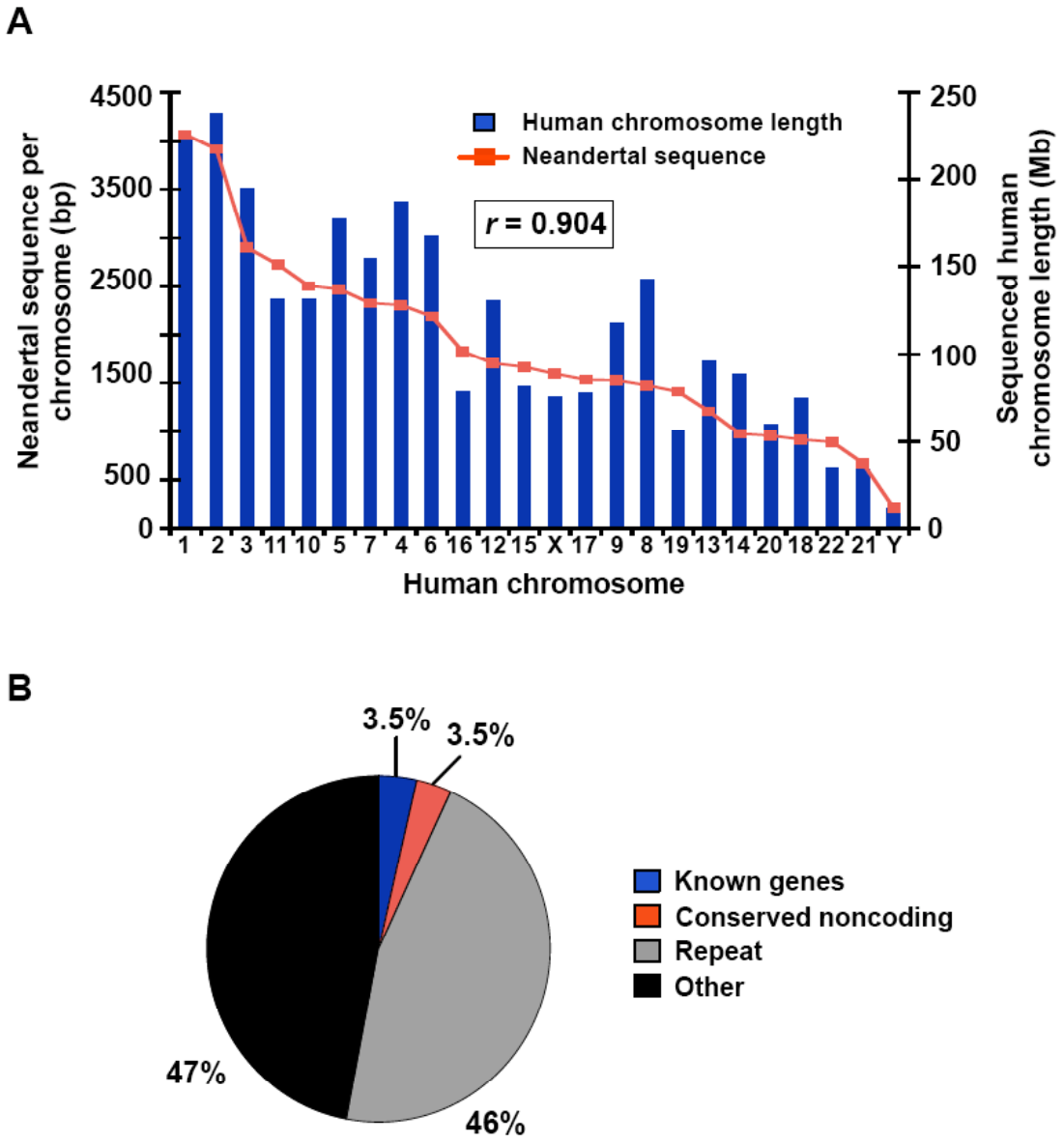


**Figure 3. (A)** Representation of each Neanderthal chromosome in 43.9 kb of NE1 hominid sequences displaying a statistically unambiguous best BLAST hit to the human genome, relative to the total sequenced length of each human chromosome minus gaps. Chromosomes are ranked by the amount of Neanderthal sequence aligned to each. Chromosomes X and Y are shown at half their total length to correct for their haploid state in males relative to the autosomes. The Pearson correlation coefficient (*box*) between the amount of Neanderthal sequence aligned to each human chromosome and the sequenced

length of each chromosome is shown. **(B)** Representation of sequence features in the NE1 hominid sequence shown in **(A)**.

To ascertain if the library NE1 hominid sequence we obtained was a representative sampling of the Neanderthal genome, we identified each NE1 library sequence that had a statistically unambiguous best BLAST hit in the human genome (43,946 bp in 1,039 loci; Table S1) and determined the distribution of these hits across human chromosomes (Fig. 3A). The amount of Neanderthal sequence aligned to each human chromosome was highly correlated with sequenced chromosome length, indicating that the Neanderthal sequences we obtained were randomly drawn from all chromosomes ($r$ = 0.904, Fig. 3A). The hominid hits included Y-chromosome sequences, suggesting that our sample was derived from a Neanderthal male. We annotated each Neanderthal locus according to the annotations (known genes, conserved noncoding sequences and repeats) associated with the aligned human sequence (Table S2). Neanderthal sequences obtained by both Sanger sequencing and pyrosequencing showed a distribution of sequence features consistent with the known distribution of these features in the human genome (Fig. 3B and data not shown). These sequences are therefore likely to represent a random sampling of the Neanderthal genome.

Comparison of authentic Neanderthal sequence with orthologous human and chimpanzee genomic sequences will reveal sites at which Neanderthal is identical to chimpanzee, but at which the human sequence has undergone a mutation since the human-Neanderthal divergence. The number and frequency of human-specific mutations are also critical to dating the human-Neanderthal split. To identify these events, we constructed alignments of orthologous human, Neanderthal and chimpanzee sequences and identified

mutations specific to each lineage by parsimony (16). We identified 34 human-specific substitutions in 37,636 human, Neanderthal and chimpanzee aligned positions, including substitutions on chromosomes X and Y that are not considered in subsequent analyses. We also identified 171 sites with Neanderthal-specific substitutions relative to human and chimpanzee. It has been shown that nucleotides in genuine ancient DNA are occasionally chemically damaged, most frequently due to deamination of cytosine to uracil, resulting in the incorporation of incorrect bases during PCR and sequencing (17). This results in an apparent excess of C to T and G to A mismatches (which are equivalent events) between the ancient sequence and the modern genomic reference sequence. We observe a significant excess of C to T and G to A mismatches (relative to T to C and A to G mismatches) between human and NE1 hominid sequences obtained by both Sanger sequencing and pyrosequencing ($p << 0.0005$, Fisher's exact test; Fig. 4 and Table S3). This accounts for the large number of Neanderthal-specific substitutions we observe and further supports that the hominid sequences are Neanderthal in origin. Importantly, despite the bias toward C to T and G to A events in Neanderthal genomic sequence, the overall frequency of all putative damage-induced events is low (~0.37% of all sites), indicating that the vast majority of human-Neanderthal-chimpanzee aligned positions are not likely to be significantly affected by misincorporation errors.

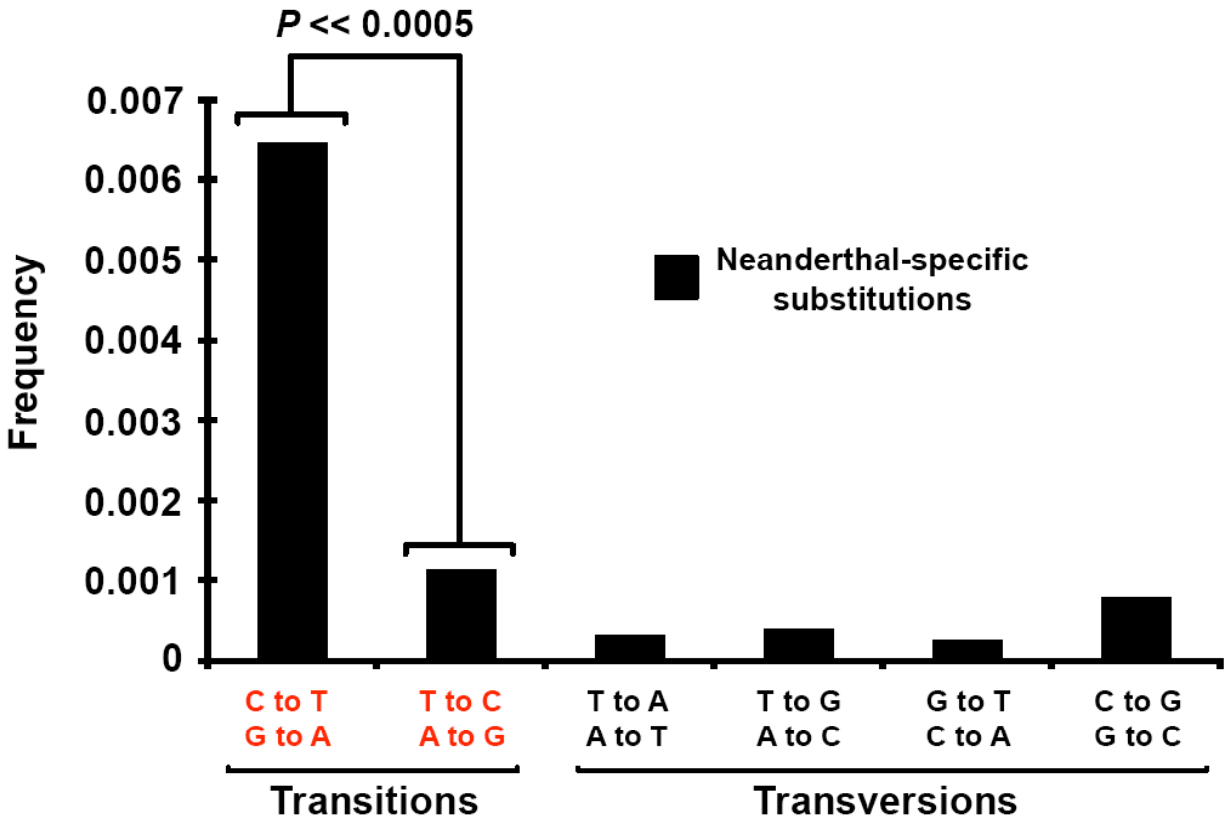


**Figure 4.** Frequency distribution of 171 Neanderthal-specific substitutions observed in 37,636 bp of aligned human, Neanderthal and chimpanzee genomic sequence. There is a highly significant excess of Neanderthal-specific C to T and G to A transitions, compared to T to C and A to G transitions (*red labels*). We cannot determine on which strand the initial substitution event occurred, so complementary substitutions (e.g., C to T and G to A) are considered equivalent events.

We estimated the divergence time of the human and Neanderthal lineages by maximum likelihood (13). We first considered the average coalescence time for the autosomes between the Neanderthal genomic sequence that we obtained and the reference human genome sequence. Based on the observed number of human-specific substitutions, our maximum likelihood estimate of the average time to the most recent common ancestor of these sequences is 770,000 years, with a 95% confidence interval of 490,000 to

1,030,0000 years (Fig. 5A, Fig. 6; 13). This calculation does not make use of Neanderthal-specific changes, since many of those events are due to DNA damage as described above. This estimate makes use of a mutation rate obtained by setting the average coalescence time for human and chimpanzee autosomes to 6.5 million years ago, a value that falls within the range suggested by recent studies (19, 20). Inaccuracies in the human-chimpanzee divergence time would shift all the time estimates and confidence intervals presented here in proportion to the error.

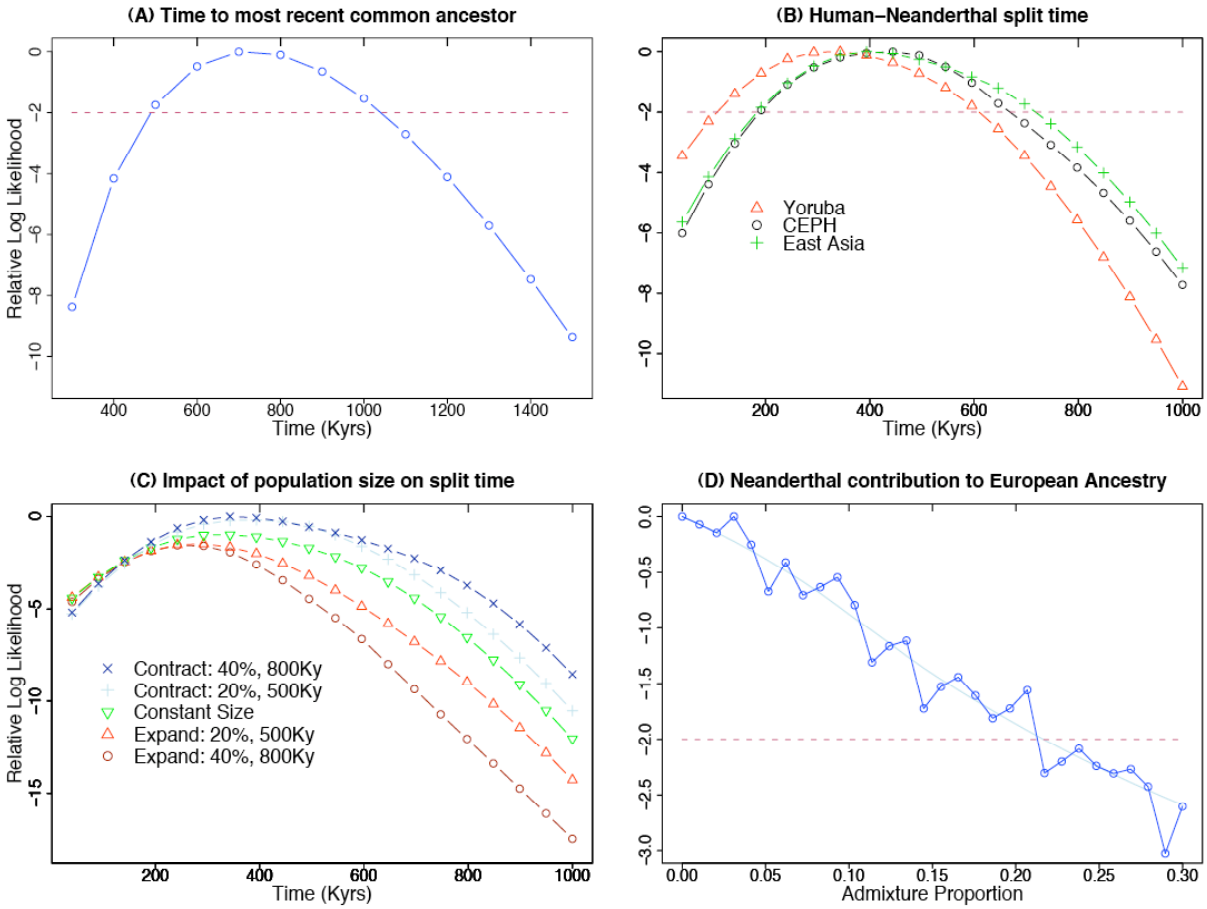


**Figure 5. (A)** Log likelihood curve of the time to the most recent common ancestor of humans and Neanderthal. **(B)** Smoothed relative log likelihood estimates of the split times between different human populations and the Neanderthal population. **(C)** Impact of changes in the ancient population size on split time estimates for 5 models that are consistent with modern polymorphism data. Each curve is the smoothed log likelihood

relative to the maximum over all 5 models. For each model, the text on the plot indicates the degree of expansion/contraction, and the time before present at which the size change occurred. Note that the expansion models are less likely compared to either constant population size or the contraction models. **(D)** The log likelihood estimates of the contribution of the Neanderthal population to the ancestry of Europeans (CEPH). The light blue line is a smoothed version of the estimates. The dashed maroon line on Figures A, B and D represents a 2 log likelihood drop and the region bounded by this line represents the 95% confidence interval around the maximum likelihood estimates.

Our estimate of the average common ancestor time reflects the average time at which the Neanderthal and human reference sequence began to diverge in the common ancestral population, not the actual time of divergence of the ancestral populations that gave rise to Neanderthals and modern humans. To estimate the actual split time of the ancestral human and Neanderthal populations, we constructed a model that incorporated data from the human and Neanderthal reference sequences, as well as genotypes from 210 individuals with genome-wide SNP data collected by the International HapMap Consortium (Table 2; 21). We included the HapMap data as they indicate what proportion of sites in the Neanderthal sequence falls outside of modern human variation. If the ancestral human and Neanderthal populations diverged long ago, before the rise of most modern human genetic diversity captured by the HapMap data, Neanderthal sequence would almost never carry the derived allele, relative to the orthologous chimpanzee sequence, for a human SNP (Table 2).

We constructed a simulation-based composite likelihood framework to estimate the time at which the Neanderthal population split from the ancestral human population, while accounting for the SNP ascertainment process used by HapMap and using appropriate demographic histories for each HapMap population, including bottlenecks or growth (13,

18). Our initial model assumed that there was no admixture between Neanderthals and humans subsequent to the original population split. The composite likelihood framework treats each site independently, which is an excellent approximation in this case since the Neanderthal sequence reads are very short and just 1 out of 906 aligned fragments contains more than one human-specific allele or SNP.

| | | Human Reference | |
|---|---|---|---|
| Neanderthal | *With SNPs* | Ancestral | Derived |
| | Ancestral | 24 | 8 |
| | Derived | 3 | 0 |
| | *Without SNPs* | Ancestral | Derived |
| | Ancestral | 35802 | 20 |
| | Derived | 162 | 476 |

**Table 2.** Summary of all autosomal sites sequenced in Neanderthal and uniquely aligned to the human and chimpanzee reference sequences. The designations "ancestral" and "derived" indicate whether each site is respectively a match or mismatch with chimpanzee. Sites are partitioned into those that overlap a Phase II HapMap SNP (*with SNPs*), and those that do not (*without SNPs*).

Under this model, the maximum likelihood estimates for the divergence of ancestral human and Neanderthal populations are 440,000 years (95% confidence interval of 170,000-620,000 years) based on the European data, 390,000 years (170,000-670,000 years) for East Asians and 370,000 years (120,000-570,000 years) for Yorubans (Fig. 5B and Fig. 6). These values predate the first appearance of anatomically modern humans in

Africa ~160,000 years ago. Since these divergence times are prior to the migration of modern humans out of Africa, the three divergence estimates should all be estimates of the same actual split time. Substantial contamination with modern human DNA would artificially lower these estimates, but 2% contamination, the rate suggested by mitochondrial PCR analysis of the primary extract used to construct the library, would have essentially no impact (13).

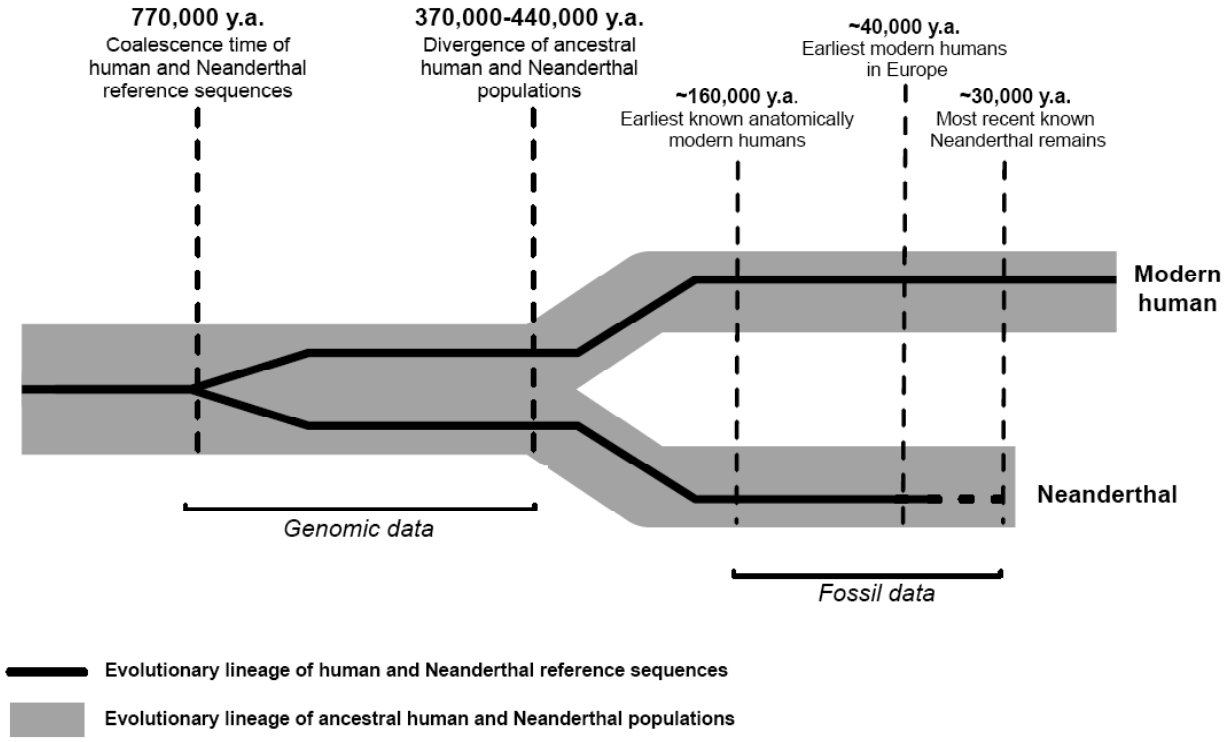


**Figure 6.** Divergence estimates for human and Neanderthal genomic sequences and ancestral human and Neanderthal populations, shown relative to dates of critical events in modern human and Neanderthal evolution. The branch lengths are schematic and not to scale.

Our data include three sites at which Neanderthal carries the derived allele for a polymorphic HapMap SNP. These sites are unlikely to represent modern contamination

because for two of the SNPs the derived allele is found only in Yorubans; also one of the SNPs lies on a fragment that contains a C to T transition in Neanderthals that is characteristic of chemical damage to DNA. These observations indicate that the Neanderthal sequence may often coalesce within the human ancestral tree. Based on simulations of our best-fit model for Yoruba, we estimate that Neanderthal is a true outgroup for approximately 17% (assuming a split time of 340,000 years, the Yoruban estimate) to 26% (assuming a split time of 440,000 years, the CEPH estimate) of the autosomal genome of modern humans, though more data will be required to achieve a precise estimate.

Our estimates of the Neanderthal divergence time might depend heavily on the assumption that the ancestral effective population size of humans was 10,000 individuals. To address this we explored a set of models in which the ancestral human population expanded or contracted at least 200,000 years ago (13). We found that much of the parameter space—though not the original model—could be excluded on the basis of modern human polymorphism data from (18). We repeated our likelihood analysis of the Neanderthal data using models incorporating ancient expansion or contraction that are consistent with modern data, and found that these did not substantially change our population split time estimates (Fig. 5C).

Since Neanderthals coexisted with modern humans in Europe, there has long been interest in whether Neanderthals might have contributed to the European gene pool. If Neanderthal admixture did indeed occur, then this would be evident in our data as an abundance of low frequency derived alleles in Europeans where the derived allele matches Neanderthal. No site in the dataset is of this type, and our maximum likelihood estimate

for the Neanderthal contribution to modern genetic diversity is zero, with a 95% confidence interval of 0-20% (Fig. 5D). This result is consistent with previous comparative studies of human and Neanderthal mitochondrial DNA, which also failed to find substantial positive evidence of a Neanderthal genetic contribution to modern humans, but relied on one locus and could only rule out admixture between Neanderthal females and modern human males (3-6).

Although we have recovered significant amounts of Neandertal genome sequence using a metagenomic approach, many gigabases of sequence would be required to achieve 1x coverage of a single Neanderthal genome by this method. Moreover, our results indicate that ~99.5% of the Neanderthal sequence that would be obtained would be identical to modern human. The human-Neandertal sequence differences that would yield great insight into human biology and evolution are thus rare events in an overwhelming background of uninformative sequence. We therefore explored the potential of metagenomic libraries to serve as substrates to recover specific Neanderthal sequences of interest by targeted methods. To this end, we developed a direct genomic selection approach to recover known and unknown sequences from metagenomic ancient DNA libraries (22). We first attempted to recover specific sequences from a Pleistocene cave bear metagenomic library we previously constructed. We designed PCR probes corresponding to 96 sequences highly conserved among mammals, amplified from the human genome, and hybridized these to PCR-amplified cave bear library inserts produced as described above (Figure 1, A and B). Recovered library DNAs were amplified by PCR and sequenced. We successfully recovered 5 targets consisting of a known enhancer of *Sox9* and conserved sequences near *Tbx3*, *Shh*, *Msx2* and *Gdf6* (Table S4). In principle

these sequences could be derived from contaminating DNA rather than the cave bear library. Critically, the captured cave bear sequences were flanked by library vector sequence, directly demonstrating that these sequences were derived from a cloned library insert and not from contaminating DNA introduced during direct selection (Fig. 1B and Fig. S1).

Based on these results we attempted to recover specific Neanderthal sequences from library NE1. We focused on recovering sequences we had previously identified by shotgun sequencing due to the low complexity of library NE1, and were able to recover 29 of 35 sequences we targeted (Table S4). The authenticity of these sequences was confirmed by the presence of library vector sequences in the reads. Our success in recovering both previously unknown cave bear and known Neanderthal genomic sequences using direct genomic selection indicates that this is a feasible strategy for purifying specific cloned Neanderthal sequences out of a high background of Neanderthal and contaminating microbial DNA. This raises the possibility that, should multiple Neanderthal metagenomic libraries be constructed from independent samples, direct selection could be used to recover Neanderthal sequences from several individuals to obtain and confirm important human-specific and Neanderthal-specific substitutions.

The current state of our knowledge concerning Neanderthals and their relationship to modern humans is largely inference and speculation based on archaeological data and a limited number of hominid remains. Using a metagenomic library-based approach, we were able to obtain sufficient amounts of Neanderthal genomic sequence to date the human-Neanderthal split to ~400,000 years ago. Our analysis, which employed for the first time multiple autosomal loci from Neanderthal, suggested that if admixture between

Neanderthal and early modern human populations occurred, the likely Neanderthal contribution to the modern human gene pool was small. Future Neanderthal genomic studies will provide insight into the profound phenotypic divergence of humans both from the great apes and from our extinct hominid relatives, and may allow us to explore aspects of Neanderthal biology not evident from artifacts and fossils.

**References**


1. P. Mellars, *Nature* **432**, 461, 2004.

2. R.G. Klein and B. Edgar, *The Dawn of Human Culture*, John Wiley and Sons, New York (2002).

3. M. Krings *et al., Cell* **90**, 19 (1997).

4. M. Krings *et al., Proc. Natl. Acad. Sci. USA* **96**, 5581 (1999).

5. S. Pääbo *et al., Annu. Rev. Genet.* **38**, 645 (2004).

6. D. Serre *et al., PLoS Biol.* **2**, e57 (2004).

7. M. Currat and L. Excoffier, *PLoS Biol.* **2**, e421 (2004).

8. S.G. Tringe *et al., Science* **308**, 554 (2005).

9. S.G. Tringe and E.M. Rubin, *Nat. Rev. Genet.* **6**, 805 (2005).

10. J.P. Noonan *et al., Science* **309**, 554 (2005).

11. M. Margulies *et al., Nature* **437**, 376 (2005).

12. H.N. Poinar *et al*., *Science* **311**, 392 (2006).

13. Materials and methods are available as supporting material on *Science* online.

14. S. Pääbo, personal communication.

15. S.F. Altschul *et al., Nucleic Acids Res.* **25**, 3389 (1997).

16. Chimpanzee Sequencing and Analysis Consortium, *Nature* **437**, 69 (2005).

17. M. Hofreiter *et al., Nucleic Acids Res.* **29,** 4793 (2001).

18. B.F. Voight *et al., Proc. Natl. Acad. Sci. USA* **102**, 18508 (2005).

19. S. Kumar, A. Filipski, V. Swarna, A. Walker, S.B. Hedges, *Proc. Natl. Acad. Sci. USA* **102**, 18842 (2005).

20. N. Patterson, D. Richter, S. Gnerre, E. Lander, D. Reich, *Nature* in press; published online 17 May 2006 (10.1038/nature04789).

21. D. Altshuler *et al., Nature* **437**, 1299 (2005).

22. S. Bashiardes *et al., Nat. Methods* **2**, 63 (2005).

23. Neanderthal sequences reported in this study have been deposited in GenBank under accession numbers xxxx-xxxx. We thank members of the Rubin, Pääbo and Pritchard laboratories for insightful discussions and support. J.P.N. was supported by NIH NRSA fellowship 1-F32-GM074367. G.C. and S.K. were supported by R01 HG002772-1 (NIH) to J.K.P. This work was supported by HL066681, Berkeley PGA, NIH Programs for Genomic Application, funded by the National Heart, Lung and Blood Institute, and by the University of California, Lawrence Berkeley National Laboratory and the US Department of Energy Joint Genome Institute under contract number DE-AC02-05CH11231.

Supporting Online Material

Materials and Methods

Figs. S1 to S7

Tables S1 to S11

# Supporting online material

# Materials and methods

## Cloning and sequencing of Neanderthal genomic DNA

*Construction of Neanderthal metagenomic library NE1*

Extracts for library construction were prepared and cloned as described in (6, 10). Briefly, extracted ancient DNAs were end-repaired using the Epicentre End-It kit (Epicentre, Madison, WI) and end-repaired samples were blunt-end ligated into pMCL200 without any size-selection. 1 μL of resuspended ligation reaction was electroporated into DH10B Electromax™ cells (Invitrogen, Carlsbad, CA). Transformed cells were grown at 37 degrees for 1 hour. Cells (250-300 μl) were spread on LB agar plates containing 20 μg/ml of chloramphenicol and 50 mg/ml of x-gal. Plating 250 μL of transformation culture yielded 1200 colonies, indicating that the overall complexity of the library was very low (approximately 4800 colonies per 1 mL transformation from 1 μL ligation). Individual white recombinant colonies were selected and picked into 384-well microtiter plates containing LB/glycerol (7.5%) media containing 20 μg/ml of chloramphenicol using the Q-Bot™ multitasking robot (Genetix, Dorset, U.K.). For details on production sequencing protocols see http://www.jgi.doe.gov/sequencing/protocols/.

*Generating pooled metagenomic library DNAs from batch culture*

We transformed 1 μL of each metagenomic library ligation reaction and expanded the resulting 1 mL SOC cultures into 200 mL batch cultures in 2xYT medium. 100-200 μL of transformation culture were used to inoculate each 200 mL batch culture. Cultures were grown for 12-16 hours overnight and plasmid DNAs recovered using the Qiagen Plasmid Maxi Kit (Qiagen, Valencia, CA). To compare representation before and after expansion of the library in batch culture, we plated aliquots of the transformation culture and the 200 mL 2xYT culture and sequenced 384 clones from each. For Sanger sequencing of library DNA grown in batch culture, purified plasmid pools were transformed into DH10B cells, cells were cultured and plated, and 19,200 colonies picked and sequenced as described above and in (10).

*Preparation of library inserts for pyrosequencing and direct selection*

Inserts were recovered by PCR amplification across the insert using Platinum Taq High-Fidelity polymerase (Invitrogen) and primers flanking the multiple cloning site. These primers are located relative to the insert such that flanking vector sequence is included in all PCR products. PCR products were purified by agarose gel electrophoresis and recovered using the QiaQuick Gel Extraction Kit (Qiagen). The total number of amplified molecules in the product pool was estimated from the total molecular weight of the purified products and the average product size (approx. 200 bp for library NE1). The total expected number of amplified products corresponding to a specific segment of the cave

bear or Neanderthal genome (i.e., the total increase in coverage due to amplification) was then estimated based on the fraction of clones that contained a cave bear or Neanderthal insert. This value was used to calculate the amount of probe to use in direct selection experiments as described below.

*Pyrosequencing*

PCR amplified library inserts were processed for pyrosequencing on the 454 Life Sciences sequencing platform at the JGI according to the manufacturer's instructions. Pyrosequencing was carried out as described (11).

**Sequence assembly and analysis**

*Initial determination of the hominid sequence content of library NE1 by Sanger sequencing*

We initially sequenced 9984 clones from plating of unamplified transformation cultures using standard Sanger sequencing methods as described above. We identified putative Neanderthal genomic sequences in these clones by BLASTN comparison to the human genome reference sequence (hg17, NCBI build 35) using an expect score threshold of 0.001. Low quality positions in Sanger reads (Phred Q < 20) were replaced with N. We did not identify any duplicate Neanderthal sequences in these clones.

*Identification of Neanderthal sequences in batch culture pyrosequencing data*

We searched for putative Neanderthal genomic sequences in 1,474,910 pyrosequencing reads by MEGABLAST comparison to the human genome using a wordsize of 12 and an expect score threshold of 0.001. We identified 7880 reads with significant MEGABLAST homology to human. However, we observed significant "stacking" in these reads, in that multiple identical reads aligned to the same human locus (approximately 5 reads on average per locus; (Fig. S2A). Stacking affected all reads in the library whether or not they contained putative Neanderthal sequence. This phenomenon is due to the low complexity of library NE1 described above and in the main text. Based on colony yields obtained from plating of transformation cultures, library NE1 contains a relatively small number of clones compared to previous metagenomic libraries we constructed from cave bear DNA (10). This low complexity pool was amplified by batch culture, producing many identical copies of each clone, and the PCR used to recover the inserts generated additional copies of each clone. The stacking we observe is not an inherent artifact of the batch culture process. We amplified the cave bear library used in the direct selection experiments described below and in the main text, and have sequenced inserts from this library by pyrosequencing using the exact protocol employed for library NE1. Based on colony yields and the percentage of targeted sequences recovered by direct selection, this cave bear library is of much higher complexity than library NE1 (5 of 96 targets recovered, suggesting ~0.05X coverage of the cave bear genome). Accordingly, we did not observe significant stacking of pyrosequencing reads from this library (Fig. S2B).

*Assembly of pyrosequencing reads*

We assembled the 7880 reads containing putative Neanderthal genomic sequence using Forge, a whole genome shotgun assembly program that has been tuned for pyrosequencing data at the JGI (D. Platt, unpublished). The sample was treated as a metagenomic collection to allow for the high, non-Poisson depth variation. This relaxed the penalty for excessively deep contigs. The k-mer step size was set to 1 to ensure overlap detection at low sequence coverage for short read lengths. The default word size of 17 was used and all other parameters were left at default settings. After 5' and 3' vector removal, 425,540 bases of raw sequence with an average trimmed read length of 54 bases was assembled into 1425 contigs with average depth across a contig ranging from 1 to 45. The resulting assembly was scrutinized and all positions of internal discrepancy between reads were visually inspected. Likewise, all discrepancies between the consensus sequences and the aligned human sequence were inspected. There was evidence that at least one contig resulted from a collapse of two repeats, but these contigs were discarded. The resulting scaffolds were realigned to the human reference by MEGABLAST. Only those scaffolds in which every constituent read aligned to the same genomic location were retained for further analysis. For final analysis and identification of lineage-specific substitutions, scaffolds were realigned to human using BLASTN (e = 1e-3), as BLASTN yielded more conservative alignments (e.g., fewer gaps) than MEGABLAST. Low quality positions in scaffolds (Q < 15) were replaced with N.

*Comparison of Sanger sequencing and pyrosequencing data*

We analyzed by Sanger sequencing 19,200 clones from the same batch culture used to generate the pyrosequencing data and identified 130 Neanderthal sequences represented in both the Sanger sequence data and the pyrosequencing data by MEGABLAST comparison of the two datasets. After quality screening, we identified 7 mismatches out of 6,196 aligned, ungapped positions (0.11%). At four of these positions, the reference human genome sequence agreed with the base call in the pyrosequencing data. We also compared Sanger sequence data and pyrosequencing data for all 130 clones directly to the human reference genome sequence by MEGABLAST. As reported in the main text, the mismatch frequency in both cases was nearly identical.

*Identification of lineage-specific substitutions*

A Neanderthal sequence was considered to have an "unambiguous best BLAST hit" if the bitscore of the best BLASTN hit in the human genome for that sequence was higher than the bitscores of all other hits. These "best" BLASTN alignments of modern human and Neanderthal sequences were mapped onto the human-chimpanzee aligned positions in multiz 8-way alignments obtained from the UCSC Genome Browser (genome.ucsc.edu). Lineage-specific substitutions were identified by parsimony by identifying positions where at least two of the three aligned positions were identical. We ignored all gaps as gap ascertainment in pyrosequencing data is confounded by homopolymer length estimation errors (11). Sites aligned to low quality positions in chimpanzee (Q < 30) were excluded. We ignored all positions in the first four and last four positions of each human-Neanderthal

BLASTN alignment, as under the default blastn parameters used here these positions are required to be 100% identical and including these positions in the alignment would thus overestimate human-Neanderthal sequence similarity.

We calculated the excess of Neanderthal-specific C to T and G to A transitions over T to C and A to G transitions by Fisher's exact test using the values shown in Table S3. The count of each Neanderthal-specific transition (*k*) was considered to be a subset of the total number of observed lineage-specific substitution events (*n)* (including no substitution in any lineage) in human, Neanderthal or chimpanzee arising from the same presumed ancestral state. For example, a Neanderthal-specific C to T substitution, of the form CTC (where the base order is human-Neanderthal-chimpanzee), arises from a presumed ancestral state of C, which can be maintained in the form CCC or give rise to lineage-specific substitutions of the form CxC, xCC and CCx, where x is any base other than C. Counts were summed for complementary substitutions (e.g., C to T and G to A, as shown in Table S3). The frequencies shown in Figure 4 were simply calculated as *k* / *n*.

**Direct selection**

We carried out direct genomic selection to recover specific targets from cave bear and Neanderthal metagenomic libraries using a protocol adapted from Bashiardes *et al*. (22). Individual probes were prepared by PCR amplification of the sequence of interest from modern human genomic DNA. PCRs were carried out using Platinum Taq polymerase (Invitrogen) and a mixture of dNTPs and biotin-16-dUTP (Roche) added at the following final concentrations: dATP, dGTP, dCTP = 200 μM; dTTP = 130 μM; biotin-16-dUTP = 70 μM.

For direct selection from cave bear metagenomic libraries, probe and target DNAs were denatured and hybridized in 125 mM dibasic sodium phosphate at a molar ratio of 450,000:1 probe:target for 144 hours at 55 degrees. Heteroduplexes were recovered with streptavidin-coated magnetic beads (Invitrogen). The beads were collected with a magnet and washed several times in 0.1X SSC at 55 degrees. After washing, captured metagenomic library insert DNA was eluted by boiling in 0.1X TE and PCR amplified using Platinum Taq High-Fidelity polymerase and the same vector primers used to amplify the inserts out of the library pool. The resulting mixture of PCR products was rehybridized to the biotinylinated probes used in the first round of selection. Selection was multiplexed: biotinylated probes were pooled and hybridized to target DNAs at once, and selected DNAs were sequenced by pyrosequencing. Direct selection of targets known to be in library NE1 was performed in the same manner, except hybridization was carried out for 72 hours at 58 degrees and the molar ratio of probe to target was 2000:1. Captured cave bear and Neanderthal sequences were identified by BLASTN comparison to probe sequences. Due to the low genomic sequence coverage in these libraries, generally only a single clone was recovered for each target, though this clone was often represented by multiple pyrosequencing reads.

## Estimating human-Neanderthal divergence time

### *Overview*

We estimate the average time to the most recent common ancestor (TMRCA) for the autosomal segments of the human reference sequence and the Neanderthal sequence to be 770,000 years. This number represents an average as the underlying TMRCA will vary across the autosomal portions of the genome. The time of last gene flow between humans and Neanderthals is the time at which the two populations split assuming no subsequent admixture. We estimated the population split time between Neanderthal and three human populations: the HapMap Europeans (CEU), the Yorubans (YRI), and the combined Han Chinese and Japanese (ASN). To gain information about the split time we considered SNP data from the Phase II HapMap in the regions that are orthologous to those sequenced in Neanderthal. All of the sites successfully sequenced in Neanderthal are broken into sites that are present and polymorphic in the Phase II HapMap and those that are not. These two groups are further broken down on the basis of whether the type of the human reference sequence is a mismatch with the chimp sequence, and whether the Neanderthal sequence is a mismatch with the chimp sequence. An example of the 2 x 2 tables is given in Table S5. The data for the three HapMap populations are given in Tables S6-S11.

The maximum likelihood estimates (and 95% confidence intervals) for the split time are as follows: 440Kyrs (160-620Kyrs) for CEU, 390Kyrs (160-670Kyrs) for ASN and 340Kyrs (110-570Kyrs) for YRI (Fig. 5B, main text). As these times predate the move out of Africa they represent estimates of the same time split time. The log likelihood ratio, the log likelihood difference between the model with no split (i.e. the Neanderthal ancestral lineage is allowed to coalesce with human lineages any time further in the past than 45Kyrs ago) to that with the maximum likelihood split time can be used to assess our certainty of a split having occurred. For CEU the log likelihood ratio is 6.0, for ASN it is 6.1, and for the YRI it is 3.8. Therefore, all three populations provide strong support for the split model. Note that all of these estimates were obtained assuming an average TMRCA of human and chimpanzee 6.5 million years ago, we also estimated the timing of the split and most recent common ancestor using alternative human-chimpanzee TMRCA dates (see below).

This split time between human and Neanderthal populations means that the Neanderthal sequence will often (in many regions of the genome) be the outgroup to the human populations. To evaluate how often this is the case we performed coalescent simulations with the Neanderthal lineage being separated from the human population (YRI in this case) for 340 or 440Kyrs (the maximum likelihood estimates for the YRI and CEU populations). We found that the Neanderthal was an outgroup to the human tree 17% or 26% of the time respectively. For each simulation we simulate 5000 YRI haplotypes. Note that we do not include the possibility of human-Neanderthal admixture in this simulation, admixture would reduce the probability that the Neanderthal was the outgroup to the human populations.

The log likelihood curves for YRI (Fig. 5B) are less steep at low values of the split time

than those of CEU and ASN. This is because there are two extra HapMap Phase II SNP polymorphic in YRI for which the Neanderthal has the derived allele. This makes the YRI data consistent with a slightly lower value of the split time in YRI. However, the derived allele at this SNP was probably lost by genetic drift in the bottleneck in the CEU and ASN populations, thus the estimates of the split time in the three populations are very consistent with each other. In the region containing the shared SNP (rs2617656) the Neanderthal sequence also has a mutation consistent with DNA degradation, thus the derived allele at this SNP in Neanderthal is unlikely to represent contamination.

Changes in ancient population size can have important consequences for our estimate of the population split time. Patterns of diversity in Modern day human populations contain information about ancient growth/contraction and so we explored simple models of ancient population change using the Hausa resequencing data (who are closely related to the Yoruba) of Voight et al. (2005) (Fig. S3). We concentrated on the Hausa individuals of Voight et al. (2005) (from Cameroon) as the recent bottlenecks present in the histories of Europeans and Asians mean that their patterns of genetic diversity are somewhat less informative about ancient growth or contraction than the African individuals. We used the model of recent growth fitted by Voight et al. (2006). In our model the population size changes some number of generations in the past instantaneously by a fraction $r$ ($r = 1$ is no change). The log likelihood surface for when the change occurs and the population size change parameter $r$ is shown in Fig. S4. The surface demonstrates that we can rule out extreme changes in population size in the past million years. The range of plausible ancient population sizes expands the further in the past that the change happened as the data is less informative about population sizes far in the past.

To explore further how ancient changes in population size would affect our results we performed the split time inference for the YRI data under a model with an 20% reduction or a 20% increase in population size ($r = 0.8$ and $r = 1.2$ respectively) 500Kyrs ago and an 40% reduction or a 40% increase in population size ($r = 0.6$ and $r = 1.4$ respectively) 800Kyrs ago. The relative log likelihood curves for these four models are shown in Fig. 5C. The ancient expansion model is somewhat less likely than the constant or contracting ancient population models, suggesting that a constant or slightly reduced ancient population size is the best fit to the data. As can be seen in Fig. 5C a smaller ancient population size leads to a higher estimate of the split time. A longer split time is needed when the population size is small to be compatible with the observed level of divergence between human and Neanderthal.

To investigate the signal of admixture of data we estimated the proportion of ancestry due to the Neanderthals in the CEU data. The estimate of the ancestry proportion is zero, but there is little certainty about this proportion. The signal of admixture in the CEU data would be low frequency derived alleles also present the Neanderthal sequence. There is little evidence of this signal in our data. The only SNP position polymorphic in CEU at which the Neanderthal has the derived allele is at reasonable frequency; also the allele is present in YRI and ASN making it an unlikely candidate for a recent Neanderthal contribution to CEU ancestry. The shared derived allele at this SNP more likely represents an ancestral polymorphism that has not been lost or fixed in humans. We performed simulations to assess the power to detect non-zero ancestry proportions and found that we

have low power to detect admixture with data sets like ours (results not shown). Our power is considerably decreased by the fact that we have imperfect knowledge of the human diversity in our regions. Low frequency alleles are underrepresented in HapMap Phase II and so we cannot be certain that low frequency alleles introduced by Neanderthal admixture are not present in our regions. If instead of having ascertained SNPs we had full resequencing data in humans for our regions then our power would be improved.

***Statistical methods***

*Estimating the average time of the most recent common ancestor (TMRCA) of the human reference sequence and Neanderthal*

We ignore Neanderthal-specific base changes as the majority of these represent degradation of the ancient DNA, and so to estimate average TMRCA we restrict ourselves to human-specifc changes. The divergence between chimpanzee and human is 1.30% across our autosomal regions. Assuming an average time to the most recent common ancestor of 6.5 million years (and a generation time of 25 years) this gives a mutation rate of $2.5 \times 10^{-8}$ per base per generation. The divergence we observe between the human reference sequence and the Neanderthal-human reference sequence ancestor is 0.077%, about 12% of the distance to the human-chimpanzee ancestor. We estimate the average number of generations to the most recent common ancestor of the human reference sequence and the Neanderthal by

$$\frac{N}{\mu L} \qquad (1)$$

where $N$ is the number of human specific mutations on the human reference sequence compared to the Neanderthal sequence ($N = 28$) and $L$ is the total number of bases sequenced in Neanderthal ($L = 36494$). This results in an estimated 30,690 generations since the common ancestor of human reference sequence and Neanderthal, or (assuming 25 years per generation) 770,000 years. Note, that the generation time is actually unimportant in this calculation as the mutation rate could instead have been estimated in years, but the per generation mutation rate is needed in the next section. To obtain the likelihood surface for the average TMRCA ($\mathbf{E}(T)$) we assume that N is drawn from a Poisson with mean $\mu L \mathbf{E}(T)$. There is still disagreement on the average TMRCA of human and chimpanzee (S1, 19, 20 and references therein). If the date of the average TMRCA is somewhat different this in turn leads to slightly different estimates of the time of the average most recent human-Neanderthal ancestor. For example, if the average time to the most recent common ancestor of human and chimpanzee were instead 7 million years then the average time to the most recent common ancestor of human and Neanderthal would be 840,000 years, while if it were 6 million years the time to the Neanderthal-human common ancestor would be 710,000 years.

*Estimating the population split time between human and Neanderthal populations*

We estimated the time at which the Neanderthal and human population split (known hereafter as the split time). This is different (and more recent in time) than the time of the most recent common ancestor of the Neanderthal and present day human individuals. For the moment we ignore the possibility of more recent admixture and concentrate on inferring the time at which the human and Neanderthal populations split. The approach we take uses the information from the differences specific to the human reference sequence compared to Neanderthal. In addition, the information from sites that are known to be polymorphic in human populations was used in this analysis. We took a simulation-based approach to calculate the likelihood of the data for different values of the split time. The models used in estimating the split time are shown in Fig. S6. We used a mutation rate of 2.5 10e-8 based on a human-chimpanzee time to the most recent common ancestor of 6.5 million years. We show the effects of varying the mutation rate below. The simulations were performed using **ms** (S2) and all analysis and figures were produced in **R**.

The analysis is complicated by the fact that the Phase II SNPs in a region do not represent all of the polymorphisms in humans in that region but only a subset that was ascertained during SNP discovery efforts. For example of the 13 non-SNP sites in Europeans, that have the derived allele in the human reference sequence and the ancestral allele in the Neanderthal sequence ($W_{AD}$; Table S5 and Table S6), some of these might represent fixed differences between human and Neanderthal but some fraction of these are polymorphic within humans but have not been ascertained as SNPs in Phase II. In addition, sites that are not present in Phase II HapMap at which the Neanderthal, but not the human reference sequence, has a mismatch with chimpanzee have to be ignored (treated as missing data) as a high proportion of these apparent mutations will be due to the degradation of the Neanderthal DNA.

To analyze the data we took a simulation-based approach to evaluate the likelihood of the data under various values of the split times. We calculated the likelihood surface for the time of no gene flow between the Neanderthal population and each HapMap population separately (CEU European, YRI Yoruban, ASN the combined Japanese and Han samples). We adopted a composite approach to the inference and treated each site independently. As with any approach that ignores dependences in the data this will lead to a overly peaked likelihood surface, resulting in too narrow a confidence region. Given the many short genomic regions sequenced in Neanderthal this assumption is likely to be an excellent approximation.

To find the likelihood of our data for a population split time T, we need to find the probability of the 2 x 2 table of sites that are not Phase II SNPs, $P(N, H_{ref}, \text{not Ascert}|T)$ and for Phase II SNPs the probability of their frequencies and the allele of the Neanderthal and human reference sequence for these SNPs, $P(N, H_{ref}, freq, \text{Ascert}|T)$. To calculate these probabilities we performed a large number of coalescent simulations for a range of values of the split time. For each split time we counted the proportion of simulations that gave rise to a particular configuration of the human reference sequence, the Neanderthal sequence and if the simulated site was determined to be a SNP the frequency of the SNP.

We performed the following simulation M times for each value of the split time $T$.

- With probability $p_C(T)$ (the probability of a fixed difference between chimpanzee and the common ancestor of the human population and Neanderthal, see below) we added one to the count in the $W_{DD}$ bin. If we chose to do this we are finished for this iteration.

  If not:

- Using the demographic model for the HapMap population of interest, we simulated a coalescent genealogy of a single base pair (using the per site mutation rate) for Neanderthal sequence, the human reference sequence (a particular sequence), and the HapMap population.

- If the site did not segregate we added one to the $W_{AA}$ bin. If the site segregated we recorded whether the site had the derived allele in the reference sequence, and whether the Neanderthal has the derived allele. Using the simulated frequency in the HapMap sample, $x$, we chose to call the site a SNP with probability $P(Ascert|freq)$ (see below).

- If the site was not chosen to be an ascertained SNP we added one to the count in the relevant part of the 2 x 2 non-SNP table. If the site was chosen to be a SNP we added this to the count in histogram bin for frequency broken down by the human reference allele and the Neanderthal allele.

We divided the counts in all of the count tables and histograms by M. These normalized count tables and histograms are our estimated probabilities $P(N, H_{ref}$, not Ascert$|T)$ and $P(N, H_{ref}, freq$, Ascert$|T)$. The likelihood of the data for $T$ can then be calculated by treating the observed data as multinomial draws from the calculated probabilities. The above procedure is time consuming, since few of the simulated sites segregate. To make the algorithm more efficient we used the technique of importance sampling. Rather than simulating sites unconditional on segregating we simulated sites conditional on segregating and then weighted the contribution of each simulation by the probability that a mutation occurred on the genealogy generated for that simulation (S3). The various terms can then be calculated from these re-weighted simulations. This procedure results in a likelihood identical to that given by the procedure given above. We performed 5 million coalescent simulations for each value of $T$ for each HapMap population; this was a sufficient number of simulations to ensure that multiple runs of the procedure resulted in similar answers.

Our dating of the split time is dependent on the average human-chimpanzee TMRCA, which determines the mutation rate used. We redid our analysis for ASN using a mutation rate of $2.3 \times 10^{-8}$ (a human-chimpanzee TMRCA of 7 million years) and $2.7 \times 10^{-8}$ (6 million years) to show the effects of changing the mutation rate and the two log likelihood curves are shown in Fig. S5. The change in mutation rate leads to only a small shift in the maximum likelihood estimate. However the uncertainty in the mutation rate should be

borne in mind when considering the uncertainty in our estimates of the split time. In addition, the model of demography was fitted by Voight et al. (2005) using a mutation rate calculated on the basis of an average human-chimpanzee TMRCA of 6 million years. Thus the demographic model would be slightly changed if an average human-chimpanzee TMRCA of 7 million years was assumed consistently through all the analyses. We assumed a generation time of 25 years throughout the analysis. However, our results are reasonably robust to this assumption, as our split time estimates are really estimates in coalescent time, i.e. the time in years scaled by the effective population size and generation time which are approximately linearly related. For example, decreasing the generation time would decrease the mutation rate per generation which in turn would lead to a higher estimate of the effective population size $N$, so that the product of $N$ and generation time (and hence the scale our split time estimate) remains roughly constant.

*Probability of a mismatch between chimpanzee and the common ancestor of all of the human population and the Neanderthal*

$P(N, H_{ref}$ , not Ascert$|T)$ for $W_{AA}$ and $W_{DD}$ include the probability of no mutation or a mutation between chimpanzee and the common ancestor of human and Neanderthal. We use the observed divergence between human and chimpanzee reduced to account for the time to the most recent common ancestor all of the human population and the Neanderthal. This means that we do not have to model the ancestral population size of human and chimpanzee and so reduce the complexity of our model. The probability of observing a fixed difference between chimpanzee and the common ancestor all of the human population is taken to be

$$p_C(T) = (1 - 0.987) - \mathbf{E}(1 - \exp(\theta T_{MRCA}/2)) \qquad (2)$$

where $T_{MRCA}$ is the time to the most recent common ancestor of all of the human population and the Neanderthal. The expectation is calculated as the mean over simulations of the Neanderthal and the HapMap population for a split time $T$.

*Ancient variation in human population size*

Voight et al. (2005) did not investigate possible models for ancient population history of humans, but changes in ancient population size can have important consequences for our estimate of the population split time. To investigate the possibility of ancient changes in human population size we evaluated the likelihood of the Hausa "locus-pair" resequencing data from Voight et al. 2005 (18) under simple models of ancient growth or contraction (kindly provided by A. DiRienzo). Our simple model is shown in Fig. S3. The population (before the recent expansion) changes size by a fraction $r$ at a time $W$ generations in the past. We used the Hausa folded frequency spectrum (the number of copies of the minor allele at every segregating site) as the data to estimate the model. The locus pair regions were specifically chosen to be noncoding and so provide a good data set to evaluate such models. To simplify the analysis we treat the segregating sites independently, this will somewhat inflate our certainty in our parameters, but given that the 50 sequenced regions are short and widely distributed across the genome this should not be a serious problem.

Even this simple model has a number of parameters. Changing one of the parameters (*W, r, N*) will lead to different estimates of the other parameters as they are highly correlated. For example, a smaller ancient population would lead to fewer segregating sites unless the modern population size (*N*) is larger. To explore over a range of plausible parameters while keeping the problem computationally tractable we chose to vary *r* and *W* while adjusting *N* to keep the observed number of segregating sites constant. In doing this we keep the time of the modern day expansion the same in coalescent units, and so the exact timing of the expansion in generations is changing. This is a sensible procedure as much of the information about the recent expansion comes from the frequency spectrum (which depends upon the timing of demography in coalescent time units). To estimate the *N* needed to get the observed number of segregating sites for a given value of *r* and *W* we found the value of *N* where the expected number of segregating sites per locus pair matches that observed in the Hausa data.

We then evaluated the probability of the folded frequency spectrum of the Hausa for the parameters *r* and *W* and *N*. To obtain the probability of the observed folded frequency spectrum, we treated the minor allele frequency at each site as an independent draw from the estimated folded frequency spectrum. We simulated 30 chromosomes at a single site (with a mutation rate of $2.63 \times 10^{-8}$ per generation (18) under the chosen demographic model and recorded the number of times that an allele at a site segregates at a given minor frequency. As the information about the segregating sites has already been incorporated into the analysis we evaluated the probability of the frequency spectrum conditional on the site segregating and we ignored monomorphic sites. Once again to make the algorithm computationally efficient we simulated segregating sites and reweighted the simulation by the probability that a mutation occurs on the genealogy (S3).

*Simulations for robustness to contamination*

We simulated 600 datasets with a Neanderthal sequence, a human reference sequence, and 120 haplotypes (the CEU sample size). Each dataset consisted of fully linked regions (i.e. no recombination within a region, free recombination between the regions) with lengths matched to those sequenced in Neanderthals. The ascertainment policy was applied to each segregating site in turn to decide whether it was to be kept as a SNP in our simulated data set (i.e. we retained the segregating site with frequency, *freq* with probability $P(Ascert|freq)$). The simulations had a split time of 470,000 years and the CEU demography, and the same mutation rate used above.

To test the robustness of the inference method to low levels of human contamination of the Neanderthal sequence data we took the 600 simulated data sets and with a 2% probability we replaced the Neanderthal sequence in a region by that of a randomly chosen HapMap sequence. We then estimated the maximum likelihood estimate of the split time for each of the simulated contamination data sets.

*Estimating the contribution of Neanderthal admixture to modern human ancestry*

To assess the contribution of Neanderthal admixture we implemented a scheme similar to that described above but allowing a proportion of CEU ancestral lineages to be descended from the Neanderthal population. To model this we use the scheme shown in Fig. S7. The Neanderthal contribution ($p$) to modern human ancestry is a one-time event 1600 generations ago. We fixed the split time ($T$) between the human and Neanderthal populations to 470,000 years. We then used the simulation procedure described above to assess the likelihood of the CEU data, the Neanderthal sequence and the human reference sequence given various values of $p$. A more comprehensive method would evaluate the likelihood over values of $T$ and $p$. However, our estimate of the time is somewhat robust to low levels of admixture and our certainty of the admixture proportion is low so we chose not to do this. The Neanderthal population is assumed to have had a constant effective size of 10,000. Our results will be somewhat affected by this assumption, but given our very limited information about admixture this is not of concern.

*Demographic histories of the HapMap populations*

Our models of the HapMap population histories and mutation rate were chosen from those found to be likely by Voight et al. (2005) (Fig. S6; 18). An effective population size of $N$ = 10,000, a mutation rate of $2.5 \times 10^{-8}$ per site per generation (giving a per-site population size scaled mutation rate $\theta = 4N\mu = \sim 0.001$) and a generation time of 25 years were used. For the CEU population a bottleneck occurred 800 generations (0.04 coalescent units) in the past reducing the population size to 10% of its current size for 800 generations before returning to its original size. For the ASN population the bottleneck is the same but the population size is reduced to 5% of its current size during the bottleneck. For YRI population we used a model of recent expansion occurring 940 generations in the past with the population growing exponentially to a population size of 22,000. Note that the Voight et al. 2005 (18) African demographic model was estimated for the Hausa population rather than the Yoruba, but the two populations are genetically very close thus the model should be appropriate.

For the analysis of all three populations the ancestral effective population size of human and Neanderthal was assumed to be constant at 10,000 individuals after any of the demographic events described above. For estimating the split time the demography of the Neanderthal population is unimportant as there is only a single sequence from this population. The ancient nature of the Neanderthal sequence is irrelevant in our analysis, because the mutations that have arisen on the Neanderthal lineage are treated as missing data. Therefore, the divergence time of the Neanderthal sequence does not need to be reduced to account for the ancient nature of the sequence. As all of the split times we evaluate our date for are $T \geq 45{,}000$ years the Neanderthal lineage cannot coalesce with any human lineage before this time so the ancient nature of the sample is fully accounted for in our simulations.

*Two example command lines for **ms***

CEU command line with no admixture, the split between human and Neanderthal occurs 2*NT* generations

ms 122 ndraws -s 1 -I 2 1 121 -en 0.02 2 0.1 -en 0.04 2 1 -ej T/2 1 2

CEU command line with a proportion *p* admixture the split between human and Neanderthal
occurs 2*NT* generations

ms 122 ndraws -s 1 -I 2 1 121 -en 0.02 2 0.1 -en 0.04 2 1 -es 0.04 2 (1-p) -ej 0.04001 1 3 -ej T/2 3 2

Note that **ms** uses a coalescent time scale of 4*N* rather than 2*N*, hence the scaling of numbers by a factor of a half.

*Ascertainment*

To avoid assumptions about the form of the ascertainment panel, which may be highly variable across Phase II SNPs, we use the following method (also used by Voight *et al.* 2006 (S4). We write the probability of ascertainment given the frequency $P(Ascert|freq)$ as:

$$P(Ascert \mid freq) = \frac{P(freq \mid Ascert)P(Ascert)}{P(freq)} \qquad (3)$$

We simulated the frequency spectrum of the derived allele for a sample size of the chosen HapMap population under the demographic model for that HapMap population; this provides $P(freq)$. Note that the admixture event is included in these simulations when we are assessing the likelihood of admixture. We find the frequency spectrum for the derived allele present in HapMap phase II; this gives us $P(freq|Ascert)$. To find the probability of ascertainment, $P(Ascert)$ we need to find what proportion of polymorphic sites are in Phase II HapMap. To estimate the probability that a site segregates, we simulate a single site under our population demographic model for the HapMap population (note no Neanderthal sequence is simulated here), and our per-site estimate of $\theta$ and count the proportion of simulations where a site segregates. This provides the expected probability that a site segregates. To obtain $P(Ascert)$ we divide the fraction of sites that have a polymorphic Phase II SNP in the regions sequenced by the expected probability that a site segregates.

**Supplemental References**

S1. H. Innan and H. Watanabe, *Mol. Biol. Evol.* **23**, 1040 (2006).
S2. R.R. Hudson, *Bioinformatics* **18**, 337 (2002).
S3. L. Markovtsova, P. Marjoram, S. Tavare, *Mol. Biol. Evol.* **18**, 1132 (2001).
S4. B.F. Voight, S. Kudaravalli, X. Wen, J.K. Pritchard, *PLoS Biol.* **4**, e72 (2006).

## Supplemental Figures

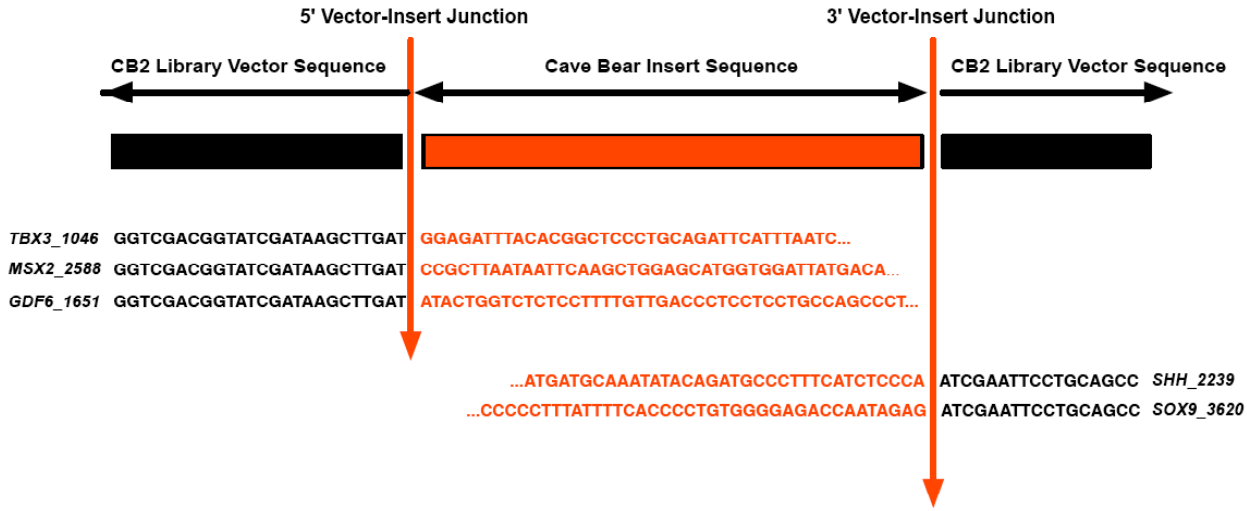


**Figure S1.** Recovery of cave bear conserved sequences by direct genomic selection. Examination of pyrosequencing reads reveals cave bear library vector sequence (black) followed by captured cave bear genomic sequence (red), directly demonstrating that these sequences are derived from a cloned library insert and not from exogenous contamination.

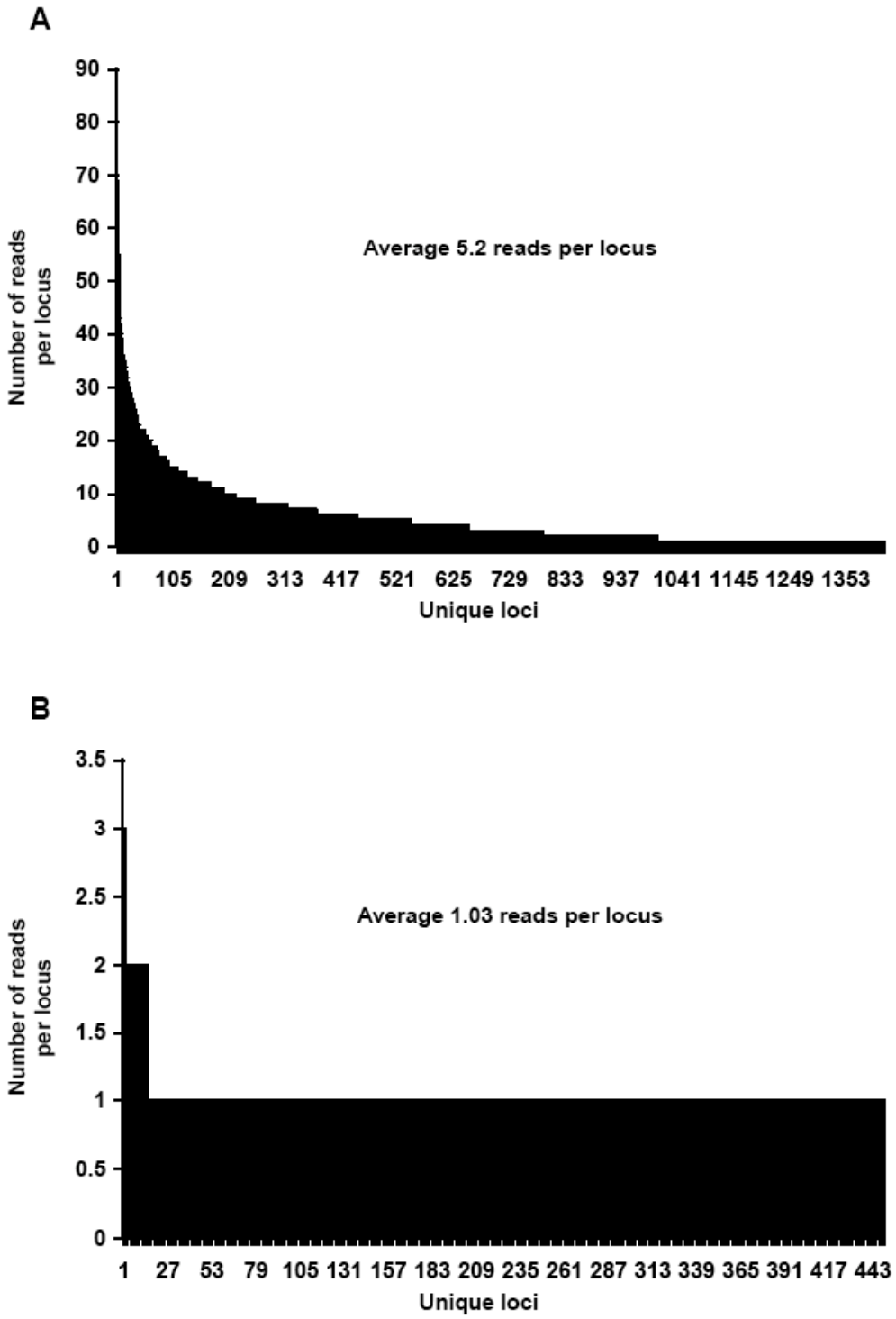


**Figure S2.** Degree of stacking in pyrosequencing reads obtained from ancient DNA metagenomic libraries. **(A)** Distribution of the number of reads per BLAST hit observed in 7880 reads from Neanderthal library NE1 with significant BLAST hit similarity to the human genome. **(B)** Distribution of the number of reads per BLAST hit observed in 467 pyrosequencing reads from the cave bear library (CB2) described in ref. 10. The stacking

observed in library NE1 is significantly greater than that observed in the cave bear library ($P$ = 1.4e-38, one-tailed $t$-test).

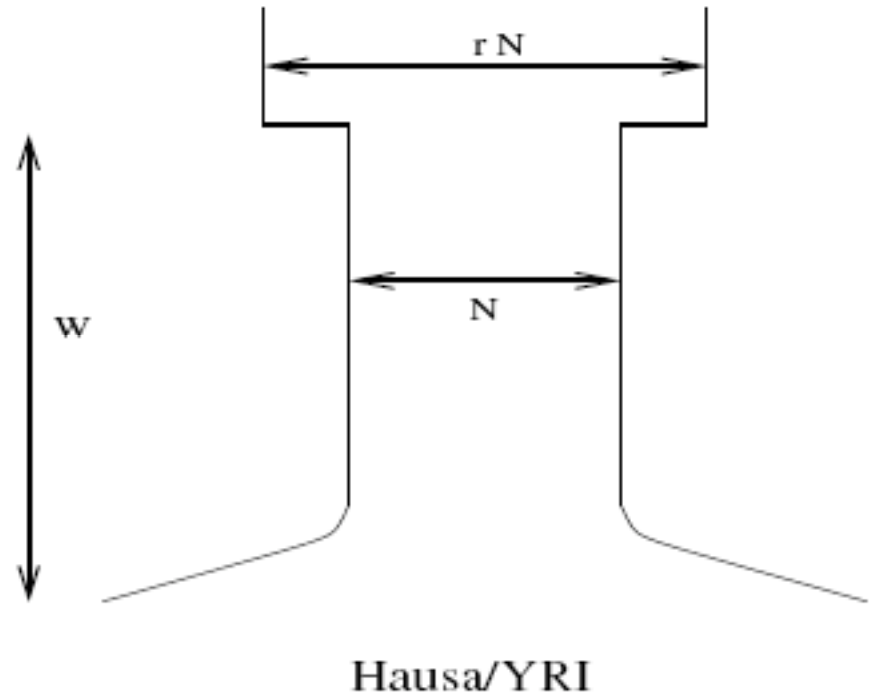


**Figure S3.** The population model used to explore the possibility of changes in the ancestral population size of humans. We compute the likelihood of the Hausa polymorphism data as a function of $W$ and $r$ in Fig. S4 below. The parameters of the Hausa population demography are the same as given for the YRI in Fig. S6.

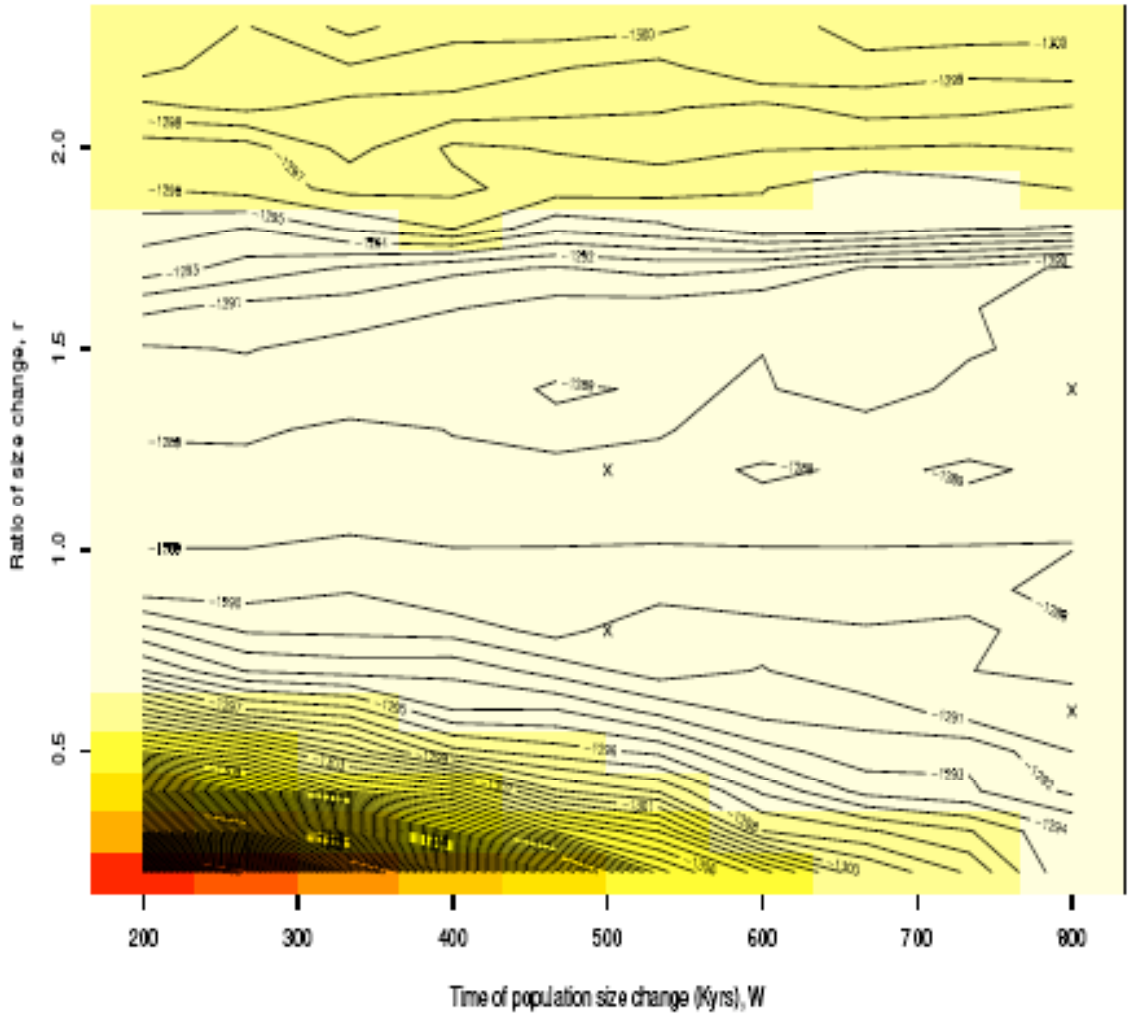


**Figure S4.** Log likelihood surface for the Hausa ancient population change model, varying *W* on the x-axis and *r* on the y-axis (see Fig. S3). Crosses mark the parameter values at which the YRI split time with Neanderthals was evaluated in Fig. 5C.

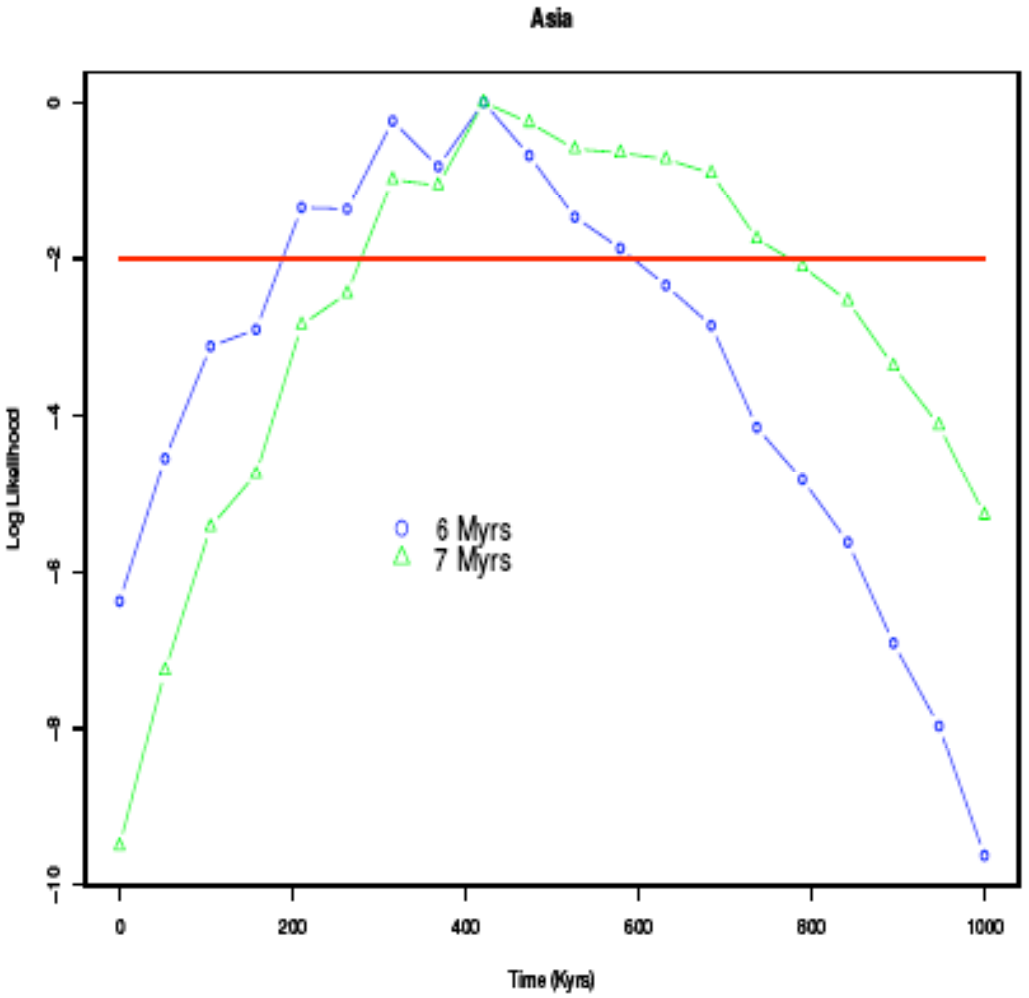


**Figure S5.** Log likelihood surface for the split time between the ASN population and Neanderthal, assuming two different average TMRCAs between human and chimpanzee.

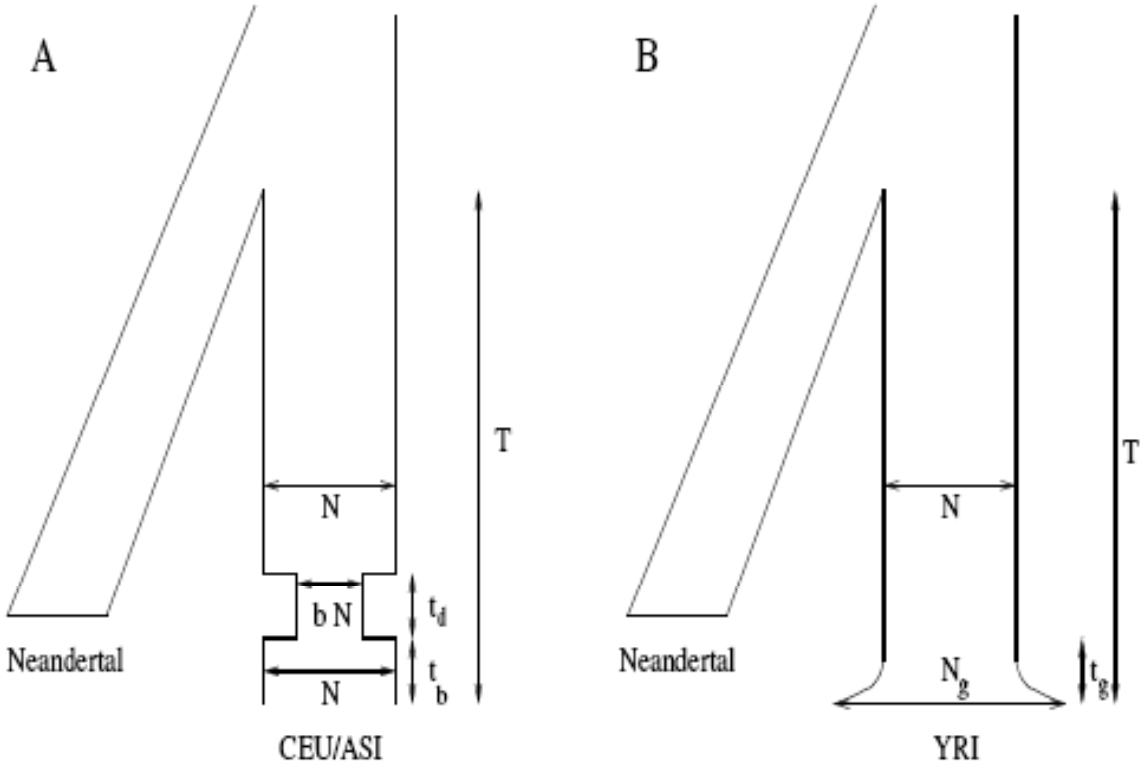


**Figure S6.** The population models used in estimating the split time. A) The model for CEU and ASN data. B) The model for YRI data. The parameters of the modern human population are taken from Voight *et al*. (2005). A) $t_b = t_d = 800$ generations for the CEU $b = 0.1$; for the ASN $t_b = t_d = 800$ generations $b = 0.05$. B) For the YRI $t_g = 940$ generations, $N_g \sim 22{,}000$. For all three populations $N = 10{,}000$ unless otherwise stated.

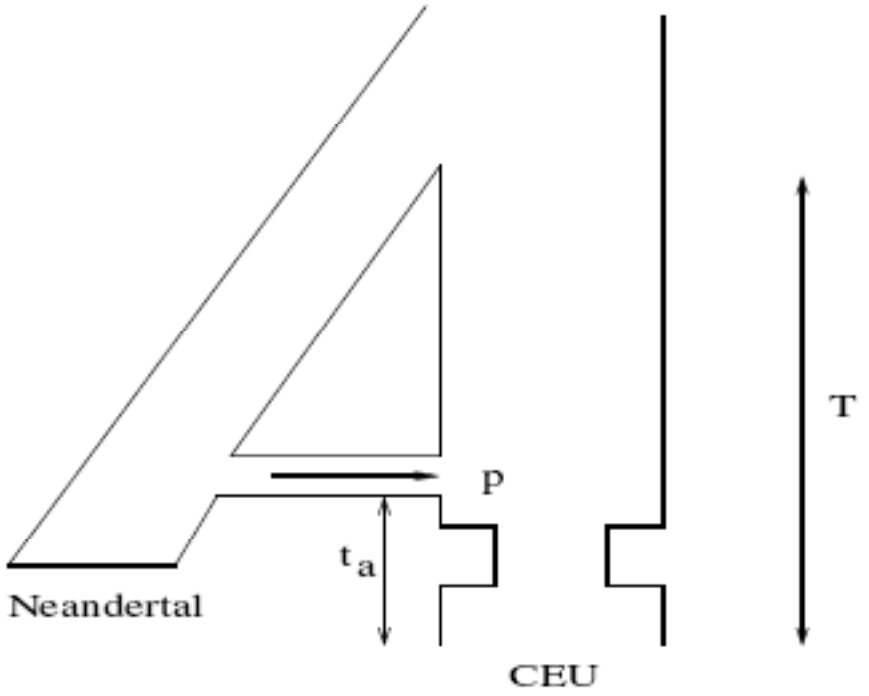


**Figure S7.** The population model used in estimating the admixture proportion. The parameters of the CEU population demography are the same as given in Fig. SxA. The Neanderthals contribute a proportion p in the ancestry of CEU individuals $t_a$ generations in the past, where $t_a$ is chosen to be 1600 generations ago.

## Supplemental Tables

(note that Tables S1 and S4 are provided in a separate Microsoft Excel workbook due to their size).

**Table S1. (provided as a separate Excel spreadsheet).** Results of BLASTN comparison (-e 1e-3) of Neanderthal library and human genome sequences (hg17). This table contains all unambiguous best Neanderthal hits and corresponding segment names to the human genome. BLASTN alignments are truncated as described in the Methods. The Neanderthal sequences are quality screened; low quality positions are replaced with N. The Neanderthal and human sequences are aligned by BLASTN relative to the plus strand in hg17 (NCBI build 35). hg17 coordinates in this file are 1-based. The first line in the file describes the fields.

| hg17_chr | hg17_chr_start | hg17_chr_stop | NE1_segment | Gene_Name |
|---|---|---|---|---|
| chr1 | 35641354 | 35641416 | segment_17 | KIAA0319L |
| chr1 | 152071303 | 152071349 | segment_59 | HCN3 |
| chr1 | 157945127 | 157945194 | segment_61 | USP21 |
| chr1 | 167913116 | 167913159 | segment_63 | FMO2 |
| chr2 | 26610530 | 26610599 | segment_106 | OTOF |
| chr2 | 55756159 | 55756182 | segment_114 | SMEK2 |
| chr2 | 144990135 | 144990204 | segment_145 | ZFHX1B |
| chr2 | 200160313 | 200160353 | segment_164 | FLJ32063 |
| chr2 | 208453662 | 208453711 | segment_168 | C2orf31 |
| chr2 | 238567924 | 238567994 | segment_179 | RAMP1 |
| chr3 | 125535980 | 125536003 | segment_220 | KALRN |
| chr5 | 251984 | 252030 | segment_300 | AK126844 |
| chr5 | 73961510 | 73961574 | segment_324 | ENC1 |
| chr5 | 132187213 | 132187250 | segment_339 | APXL2 |
| chr6 | 24570943 | 24570991 | segment_373 | GPLD1 |
| chr6 | 139350776 | 139350801 | segment_402 | REPS1 |
| chr7 | 3864659 | 3864707 | segment_412 | SDK1 |
| chr7 | 97891433 | 97891456 | segment_440 | NPTX2 |
| chr8 | 10214805 | 10214852 | segment_468 | MSRA |
| chr8 | 11739708 | 11739756 | segment_469 | CTSB |
| chr8 | 23057990 | 23058046 | segment_473 | TNFRSF10D |
| chr8 | 101275572 | 101275620 | segment_487 | SPAG1 |
| chr8 | 131097874 | 131097924 | segment_496 | FAM49B |
| chrX | 64744085 | 64744123 | segment_1014 | MSN |
| chr11 | 874075 | 874143 | segment_596 | CHID1 |
| chr11 | 2997036 | 2997064 | segment_598 | CARS |
| chr11 | 63893635 | 63893664 | segment_625 | RPS6KA4 |
| chr11 | 111604571 | 111604594 | segment_651 | PTS |
| chr11 | 125668006 | 125668055 | segment_660 | TIRAP |
| chr12 | 52972641 | 52972669 | segment_680 | NFE2 |
| chr13 | 24642959 | 24642994 | segment_708 | FLJ25477 |
| chr14 | 99195518 | 99195553 | segment_751 | KIAA1822 |
| chr14 | 104272405 | 104272431 | segment_757 | ADSSL1 |
| chr15 | 62236040 | 62236067 | segment_778 | PPIB |
| chr16 | 66434565 | 66434588 | segment_824 | THAP11 |
| chr17 | 7167748 | 7167798 | segment_844 | KIAA1787 |
| chr17 | 38215049 | 38215107 | segment_859 | CNTD1 |
| chr19 | 18803837 | 18803870 | segment_906 | RENT1 |
| chr19 | 19169410 | 19169446 | segment_907 | RFXANK |

| | | | | |
|---|---|---|---|---|
| chr19 | 19765999 | 19766051 | segment_908 | ZNF506 |
| chr19 | 54348949 | 54349000 | segment_920 | HRC |
| chr19 | 60635581 | 60635607 | segment_923 | LOC284296 |
| chr19 | 61753280 | 61753341 | segment_926 | ZFP28 |
| chr19 | 63111979 | 63112049 | segment_927 | ZNF417 |
| chr20 | 19650326 | 19650349 | segment_938 | SLC24A3 |
| chr21 | 16024303 | 16024344 | segment_955 | AF170562 |
| chr22 | 22447881 | 22447942 | segment_974 | MMP11 |
| chr22 | 49333499 | 49333522 | segment_987 | MAPK8IP2 |

**Table S2.** Fragments of Neanderthal genes recovered from library NE1. Neanderthal sequences are identified by segment as shown in Table S1.

| Events | Counts |
|---|---|
| *C to T & G to A* | |
| CTC + GAG | 117 |
| CCC + CxC + xCC + CCx + GGG + GxG + xGG + GGx | 18123 |
| *T to C & A to G* | |
| TCT + AGA | 22 |
| TTT + TxT + xTT + TTx + AAA + AxA + xAA + AAx | 19513 |

Order of bases (e.g., "CTC") = Human-Neanderthal-Chimpanzee
x = any mismatched base

**Table S3.** 2x2 table used to calculate the excess of Neanderthal-specific C to T and G to A transitions by Fisher's exact test in 37,636 human, Neanderthal and chimpanzee aligned positions.

**Table S4. (provided as a separate Excel spreadsheet)** Recovery of cave bear **(A)** and Neanderthal **(B)** sequences by direct genomic selection. Positive reads were identified by BLASTN comparison of all pyrosequencing reads with the probe sequences. Only reads with identifiable library vector sequences were considered in the analysis. 10,710 total pyrosequencing reads with vector sequence were recovered for cave bear and 16,552 for Neanderthal. Neanderthal sequences are identified by segment as shown in Table S1.

| | | Human Reference | |
|---|---|---|---|
| Neanderthal | *With SNPs* | Ancestral | Derived |
| | Ancestral | $S_{AA}$ | $S_{AD}$ |
| | Derived | $S_{DA}$ | $S_{DD}$ |
| | *Without SNPs* | Ancestral | Derived |
| | Ancestral | $W_{AA}$ | $W_{AD}$ |
| | Derived | $W_{DA}$ | $W_{DD}$ |

**Table S5.** 2x2 layout of counts of all autosomal sites sequenced in Neanderthal and uniquely aligned to the human and chimpanzee reference sequences. The designations "ancestral" and "derived" indicate whether each site is respectively a match or mismatch with chimpanzee. Sites are partitioned into those that overlap a Phase II HapMap SNP (*with SNPs*), and those that do not (*without SNPs*).

| | | Human Reference | |
|---|---|---|---|
| Neanderthal | *With SNPs* | Ancestral | Derived |
| | Ancestral | 24 | 7 |
| | Derived | 1 | 0 |
| | *Without SNPs* | Ancestral | Derived |
| | Ancestral | 35802 | 21 |
| | Derived | 163 | 476 |

**Table S6.** Overlap of all autosomal sites sequenced in Neanderthal and uniquely aligned to human and chimpanzee, partitioned into those that overlap a Phase II HapMap SNP in the CEPH population versus those that do not.

| | | Human Reference | |
|---|---|---|---|
| Neanderthal | *With SNPs* | Ancestral | Derived |
| | Ancestral | 24 | 8 |
| | Derived | 3 | 0 |
| | *Without SNPs* | Ancestral | Derived |
| | Ancestral | 35802 | 20 |
| | Derived | 161 | 476 |

**Table S7.** Overlap of all autosomal sites sequenced in Neanderthal and uniquely aligned to human and chimpanzee, partitioned into those that overlap a Phase II HapMap SNP in the Yoruba population versus those that do not.

| | | Human Reference | |
|---|---|---|---|
| Neanderthal | *With SNPs* | Ancestral | Derived |
| | Ancestral | 23 | 7 |
| | Derived | 1 | 0 |
| | *Without SNPs* | Ancestral | Derived |
| | Ancestral | 35803 | 21 |
| | Derived | 163 | 476 |

**Table S8.** Overlap of all autosomal sites sequenced in Neanderthal and uniquely aligned to human and chimpanzee, partitioned into those that overlap a Phase II HapMap SNP in the East Asian population versus those that do not.

| rs number | Chromosome | Position | Derived allele frequency | SNP category |
|---|---|---|---|---|
| rs10864790 | 1 | 227359798 | 0.440678 | $S_{AA}$ |
| rs1226901 | 2 | 168333331 | 0.122807 | $S_{AA}$ |
| rs4616587 | 3 | 24323469 | 0.208333 | $S_{AA}$ |
| rs9869276 | 3 | 58376116 | 0.0727273 | $S_{AA}$ |
| rs17692869 | 5 | 4707629 | 0.457627 | $S_{AA}$ |
| rs2247650 | 5 | 96255506 | 0.491667 | $S_{AA}$ |
| rs10949285 | 6 | 9858898 | 0.228814 | $S_{AA}$ |
| rs9361666 | 6 | 81605980 | 0.175 | $S_{AA}$ |
| rs13215357 | 6 | 83534333 | 0.4 | $S_{AA}$ |
| rs17066362 | 6 | 106368462 | 0.0916667 | $S_{AA}$ |
| rs10945578 | 6 | 159306643 | 0.440678 | $S_{AA}$ |
| rs992770 | 7 | 90855324 | 0.22807 | $S_{AA}$ |
| rs17168150 | 7 | 134116318 | 0.108333 | $S_{AA}$ |
| rs709822 | 8 | 11739722 | 0.266667 | $S_{AA}$ |
| rs767604 | 8 | 82444633 | 0.125 | $S_{AA}$ |
| rs1562430 | 8 | 128457034 | 0.358333 | $S_{AA}$ |
| rs7826096 | 8 | 139482898 | 0.194915 | $S_{AA}$ |
| rs3781107 | 10 | 26630940 | 0.133333 | $S_{AA}$ |
| rs12417975 | 11 | 56269471 | 0.241667 | $S_{AA}$ |
| rs276975 | 16 | 84785466 | 0.075 | $S_{AA}$ |
| rs12935334 | 16 | 84824742 | 0.546296 | $S_{AA}$ |
| rs10405239 | 19 | 61113005 | 0.108333 | $S_{AA}$ |
| rs12482487 | 21 | 43250163 | 0.0169492 | $S_{AA}$ |
| rs2284015 | 22 | 35421073 | 0.258333 | $S_{AA}$ |
| rs11884956 | 2 | 1545433 | 0.732759 | $S_{AD}$ |
| rs711967 | 3 | 136680788 | 0.395833 | $S_{AD}$ |
| rs2648829 | 8 | 129217093 | 0.816667 | $S_{AD}$ |
| rs2297786 | 10 | 104669968 | 0.62931 | $S_{AD}$ |
| rs10219574 | 12 | 114657286 | 0.991667 | $S_{AD}$ |
| rs2809070 | 13 | 82837752 | 0.172414 | $S_{AD}$ |
| rs2795550 | 16 | 6357702 | 0.383929 | $S_{AD}$ |
| rs6465839 | 7 | 101194913 | 0.25 | $S_{DA}$ |

**Table S9.** Detailed information for the CEPH SNPs shown in Table S6.

| rs number | Chromosome | Position | Derived allele frequency | SNP category |
|---|---|---|---|---|
| rs10864790 | 1 | 227359798 | 0.35 | $S_{AA}$ |
| rs1226901 | 2 | 168333331 | 0.00892857 | $S_{AA}$ |
| rs4616587 | 3 | 24323469 | 0.266667 | $S_{AA}$ |
| rs17692869 | 5 | 4707629 | 0.0545455 | $S_{AA}$ |
| rs2247650 | 5 | 96255506 | 0.594828 | $S_{AA}$ |
| rs10949285 | 6 | 9858898 | 0.0423729 | $S_{AA}$ |
| rs9361666 | 6 | 81605980 | 0.0333333 | $S_{AA}$ |
| rs13215357 | 6 | 83534333 | 0.0932203 | $S_{AA}$ |
| rs17066362 | 6 | 106368462 | 0.206897 | $S_{AA}$ |
| rs10945578 | 6 | 159306643 | 0.305085 | $S_{AA}$ |
| rs992770 | 7 | 90855324 | 0.805556 | $S_{AA}$ |
| rs17166812 | 7 | 94596491 | 0.0666667 | $S_{AA}$ |
| rs17168150 | 7 | 134116318 | 0.0254237 | $S_{AA}$ |
| rs709822 | 8 | 11739722 | 0.266667 | $S_{AA}$ |
| rs767604 | 8 | 82444633 | 0.0833333 | $S_{AA}$ |
| rs1562430 | 8 | 128457034 | 0.458333 | $S_{AA}$ |
| rs7826096 | 8 | 139482898 | 0.266667 | $S_{AA}$ |
| rs1886945 | 10 | 4065839 | 0.0416667 | $S_{AA}$ |
| rs12417975 | 11 | 56269471 | 0.216667 | $S_{AA}$ |
| rs4366498 | 11 | 118900467 | 0.666667 | $S_{AA}$ |
| rs12935334 | 16 | 84824742 | 0.196429 | $S_{AA}$ |
| rs16944715 | 16 | 85122157 | 0.0517241 | $S_{AA}$ |
| rs10405239 | 19 | 61113005 | 0.0333333 | $S_{AA}$ |
| rs2284015 | 22 | 35421073 | 0.308333 | $S_{AA}$ |
| rs711967 | 3 | 136680788 | 0.309091 | $S_{AD}$ |
| rs6536248 | 4 | 158807827 | 0.289474 | $S_{AD}$ |
| rs2648829 | 8 | 129217093 | 0.75 | $S_{AD}$ |
| rs2297786 | 10 | 104669968 | 0.716667 | $S_{AD}$ |
| rs10736787 | 11 | 72251721 | 0.00833333 | $S_{AD}$ |
| rs10219574 | 12 | 114657286 | 0.75 | $S_{AD}$ |
| rs2809070 | 13 | 82837752 | 0.175 | $S_{AD}$ |
| rs2795550 | 16 | 6357702 | 0.0762712 | $S_{AD}$ |
| rs6465839 | 7 | 101194913 | 0.525424 | $S_{DA}$ |
| rs8063656 | 16 | 11802947 | 0.175 | $S_{DA}$ |
| rs2617656 | 19 | 58714724 | 0.0416667 | $S_{DA}$ |

**Table S10.** Detailed information for the Yoruba SNPs shown in Table S7.

| rs number | Chromosome | Position | Derived allele frequency | SNP category |
|---|---|---|---|---|
| rs10864790 | 1 | 227359798 | 0.640449 | $S_{AA}$ |
| rs1226901 | 2 | 168333331 | 0.198864 | $S_{AA}$ |
| rs4616587 | 3 | 24323469 | 0.361111 | $S_{AA}$ |
| rs17692869 | 5 | 4707629 | 0.642045 | $S_{AA}$ |
| rs2247650 | 5 | 96255506 | 0.567073 | $S_{AA}$ |
| rs10949285 | 6 | 9858898 | 0.0337079 | $S_{AA}$ |
| rs9361666 | 6 | 81605980 | 0.1 | $S_{AA}$ |
| rs13215357 | 6 | 83534333 | 0.769663 | $S_{AA}$ |
| rs17066362 | 6 | 106368462 | 0.0898876 | $S_{AA}$ |
| rs10945578 | 6 | 159306643 | 0.0393258 | $S_{AA}$ |
| rs992770 | 7 | 90855324 | 0.430233 | $S_{AA}$ |
| rs17168150 | 7 | 134116318 | 0.235632 | $S_{AA}$ |
| rs709822 | 8 | 11739722 | 0.00561798 | $S_{AA}$ |
| rs767604 | 8 | 82444633 | 0.438202 | $S_{AA}$ |
| rs1562430 | 8 | 128457034 | 0.168539 | $S_{AA}$ |
| rs7826096 | 8 | 139482898 | 0.0222222 | $S_{AA}$ |
| rs1886945 | 10 | 4065839 | 0.2 | $S_{AA}$ |
| rs12417975 | 11 | 56269471 | 0.244444 | $S_{AA}$ |
| rs4366498 | 11 | 118900467 | 0.94382 | $S_{AA}$ |
| rs12935334 | 16 | 84824742 | 0.409091 | $S_{AA}$ |
| rs10405239 | 19 | 61113005 | 0.0833333 | $S_{AA}$ |
| rs12482487 | 21 | 43250163 | 0.0224719 | $S_{AA}$ |
| rs2284015 | 22 | 35421073 | 0.359551 | $S_{AA}$ |
| rs11884956 | 2 | 1545433 | 0.921348 | $S_{AD}$ |
| rs711967 | 3 | 136680788 | 0.955056 | $S_{AD}$ |
| rs2648829 | 8 | 129217093 | 0.308989 | $S_{AD}$ |
| rs2297786 | 10 | 104669968 | 0.539326 | $S_{AD}$ |
| rs10219574 | 12 | 114657286 | 0.722222 | $S_{AD}$ |
| rs2809070 | 13 | 82837752 | 0.634831 | $S_{AD}$ |
| rs2795550 | 16 | 6357702 | 0.476471 | $S_{AD}$ |
| rs6465839 | 7 | 101194913 | 0.0222222 | $S_{DA}$ |

**Table S11.** Detailed information for the East Asian SNPs shown in Table S8.