Supplementary Information

Genetic History of an Archaic Hominin group from Denisova Cave in Siberia

Reich D*, Green RE*, Kircher M*, Krause J*, Patterson N*, Durand EY*, Viola B*, Briggs AW, Stenzel U, Johnson PLF, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin J-J, Kelso J, Slatkin M, Pääbo S

* Contributed equally

Table of Contents **1**

Supplementary Information 1

DNA extraction, library preparation and sequencing of the Denisova phalanx.

Martin Kircher*, Udo Stenzel, Qiaomei Fu and Johannes Krause

* To whom correspondence should be addressed (Martin.Kircher@eva.mpg.de)

DNA extraction and library preparation

A total of 40 mg of bone was removed from beneath the surface of the Denisova phalanx by a sterile dentistry drill in our clean room facility, where procedures that minimize contamination from present-day human DNA are rigorously implemented[1]. In this facility, DNA was extracted as described in ref. 2 and was treated with the enzyme uracil-DNA-glycosylase (UDG)[3], which removes uracil residues from DNA to leave abasic sites[4], as well as the enzyme endonuclease VIII (Endo VIII), which cuts the DNA at the 5´ and 3´ sides of abasic sites. Subsequent incubation with *T4* polynucleotide kinase and *T4* DNA polymerase was used to generate blunt and 5´-phosphorylated ends amenable to adaptor ligation. Since the great majority of uracil residues occur close to the ends of ancient DNA molecules, this procedure leads only to a moderate reduction in the average lengths of the molecules in the library but a several-fold reduction in nucleotide misincorporation due to the removal of uracil residues from the library[3].

Two independent libraries were created using this approach (SL3003 and SL3004) with a modified Illumina multiplex protocol[5]. A 7nt-index (5'-GTCGACT-3') not available outside of the clean room, as well as outer adapter sequences required for sequencing, were then added by a PCR reaction that was set up inside the clean room but performed outside the clean room.

Illumina Sequencing and primary data processing

DNA sequencing was performed on the Illumina Genome Analyzer IIx platform. The libraries SL3003 and SL3004 were sequenced using 2×101 + 7 cycles on two flow cells (12 lanes SL3003, 4 lanes SL3004) according to the manufacturer's instructions for multiplex sequencing on the Genome Analyzer IIx (FC-104-400x v4 sequencing chemistry and PE-203-4001 v4 cluster generation kit). The protocol was followed except an indexed control PhiX 174 library (index 5'- TTGCCGC-3') was spiked into each lane, yielding 2-3% control reads in each lane.

The sequencing data were analyzed starting from QSEQ sequence files and CIF intensity files from the Illumina Genome Analyzer RTA 1.6 software. The raw reads were aligned to the corresponding PhiX 174 reference sequence to obtain a training data set for the base caller Ibis[6], which was then used to call bases and quality scores. Raw sequences called by Ibis 1.1.1 for the two paired end reads were subjected to an index read filtering step where the index read was required to match the index with at most one error[5]. The two reads in each cluster were then merged (including removal of adapter sequences and dimers) by requiring at least an 11nt overlap between the two reads. In the overlapping sequence, quality scores were combined and the base with the highest base quality score called. Only sequences merged in this way were used for further analysis. The small proportion of molecules longer than 191nt was thus discarded.

Merged reads were aligned with BWA[7] to the human genome (NCBI 36/*hg18*) and chimpanzee genome (CGSC 2.1/*pantro2*) using default parameters. These alignments were converted to

SAM/BAM format[8] with BWA's *samse* command and subsequently analyzed for PCR duplicates. Both libraries were sequenced with low redundancy of individual molecules. The few PCR duplicates obtained (identified based on their outer genomic coordinates) were consensus-called (incorporating sequence quality scores) to further increase sequence accuracy.

For the two libraries, this resulted in a total of 111,466,516 unique sequences that were mapped to the human genome (SL3003: 75,514,616; SL3004: 35,951,900), altogether resulting in 6.6 Gb of sequence (SL3003: 4.1Gb; SL3004 2.5Gb). After we restricted to the 82,227,320 sequences with a mapping quality of at least 30 (SL3003: 55,582,157; SL3004: 26,645,163) , this resulted in a total of 5.2 Gb (~1.9×) of filtered sequence data (SL3003 3.2Gb, SL3004 2.0Gb). The number of sequences unambiguously mapped to the chimpanzee genome with a mapping quality of ≥ 30 is 72,304,848, which is 87.9% of that reported for the human genome.

Access to the raw sequence data from the Denisova individual

The alignments of reads to *hg18* and *pantro2* are available in BAM format from http://genome.ucsc.edu/Denisova.

References for SI 1

1. Green, R.E. et al., The Neandertal genome and ancient DNA authenticity. *EMBO J* **28**, 2494 (2009)
2. Rohland, N. and Hofreiter, M., Comparison and optimization of ancient DNA extraction. *Biotechniques* **42**, 343 (2007).
3. Briggs, A.W., et al., Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res **38**, e87 (2010).
4. Lindahl, T., Ljungquist, S., Siegert, W., Nyberg, B. and Sperens, B., DNA N-glycosidases: properties of uracil-DNA glycosidase from Escherichia coli. *J Biol Chem* **252**, 3286 (1977).
5. Meyer, M. and Kircher, M., Ilumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010**, pdb.prot5448 (2010).
6. Kircher, M., Stenzel, U. and Kelso J., Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* **10**, R83 (2009).
7. Li, H. and Durbin, R., Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754 (2009).
8. Li, H., et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (2009).

Supplementary Information 2

Genetic divergence of various hominins from the human reference genome.

Richard E. Green*

* To whom correspondence should be addressed (ed@soe.ucsc.edu)

Strategy for estimating genetic divergences

To estimate the genetic divergence between hominins and the human genome reference sequence, we use a methodology similar to that described previously[1]. Briefly, we generate three-way alignments between the reference chimpanzee *C*, reference human *H*, and the individual under examination *Q*. These alignments are examined for sites that differ in one of the three individuals. Because the reference chimpanzee and human sequences are high-quality genome sequences, we assume that they have a negligible error rate and that the differences that are specific to human or chimpanzee reflect true evolutionary divergences. Differences that are specific to *Q*, however, are often due to sequencing error because these data derive from low-coverage sequencing data and, for some individuals, ancient DNA that is heavily affected by base deamination[2]. We thus ignore these sites for the purpose of computing genetic divergence.

We have made several improvements to the methodology used to generate the three-way alignments, with the goal of increasing the efficiency of handling data and allowing a more comprehensive view of genome divergence. The improved methodology is as follows. First, we use the Enredo-Pecan-Ortheus (EPO) 6-way primate whole genome alignments[3] to generate an inferred human/chimpanzee common ancestor sequence, HCCA. This sequence contains the inferred human/chimpanzee base at sites where this is available and the human genome base where this information is missing (for example, due to missing data in chimpanzee). Further, we construct a base-by-base annotation that describes the evidence underlying the common ancestor inference. This annotation summarizes the number of human, chimpanzee, and outgroup sequences aligning at each position. For the genetic divergence calculations described in this note, we are primarily concerned with positions that are annotated by a single human and chimpanzee sequence and at least one outgroup sequence. These are regions of confident one-to-one human and chimpanzee orthology where the common ancestor inference can be most reliably inferred. The number of bases in HCCA annotated in each class is shown in Table S2.1.

Table S2.1: Summary of the Human/Chimpanzee Common Ancestor (HCCA) alignment

Human segments	Chimpanzee segments	Outgroup segments	Number of bases	Notes
0	0	0	1,504,832,580	Primarily unaligned repetitive sequence
1	**1**	**1**	**134,736,301**	**used for divergence calculation**
1	**1**	**2**	**1,083,328,619**	**used for divergence calculation**
1	**1**	**≥3**	**3,027,770**	**used for divergence calculation**
1	1	0	43,301,312	No outgroup
1	0	0	20,878,082	Human-specific sequence
1	0	1	19,425,712	Chimpanzee deletion
1	0	2	12,773,330	Chimpanzee deletion

Note: The largest categories of bases are shown. The classes of bases used for divergence calculations are highlighted.

As a check on the robustness of our inferences, and in particular to assess whether our inferences are biased by the presence of a substantial error rate in a sample *Q* that we are comparing to *hg19*, we also analyzed a whole genome alignment of Craig Venter (*HuRef*) against *hg19*[4]. The counts for the alignment of *HuRef* to the *hg19* reference genome are shown in Figure S2.1. Interestingly, although the *HuRef* error rate is far lower than for our low-coverage sequencing data described below, its error rate still appears to be higher than that of *hg19*, as reflected in more inferred substitutions on *HuRef* than *hg19*. The excess of inferred substitutions in *HuRef* is especially high on chromosome X, which we hypothesize is due to the lower coverage of this chromosome due to the fact that the sequenced individual is male.

These alignment data provide a means to assess overall sequencing error in each dataset. Assuming that an equal number of true evolutionary divergences have occurred on the *H* and *Q* lineages, we can count the excess number of substitutions inferred to have occurred on the *Q* branch, considering these to be sequence error. We can then divide by the total number of aligned bases to provide an estimate of sequencing error. We use this idea in the analyses below.

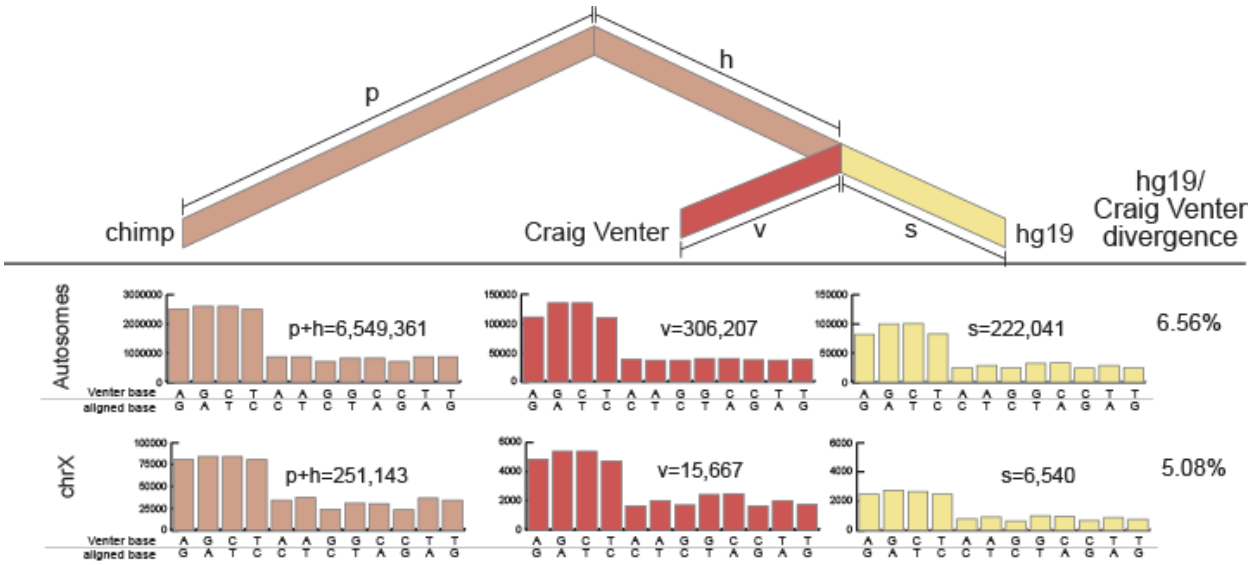


Figure S2.1: Genome divergence between *HuRef* (Craig Venter) and *hg19*. The complete genome sequence of *HuRef* and *hg19* were aligned. This alignment was then annotated using the HCCA human/chimpanzee common ancestor alignment to supply the chimpanzee base. At each position of one-to-one human/chimpanzee orthology, the corresponding *hg19*, *pantro2*, and *HuRef* allele were examined. The number of lineage-specific substitutions is shown. The number of transversions specific to each lineage is also shown. Note the higher number of substitutions localizing to the *HuRef* branch compared with *hg19*, indicating higher error rates in *HuRef*. The effect is pronounced on chromosome X, where the *HuRef* coverage is poorer as the individual is a male.

Genetic divergence estimates of diverse hominins to the human reference sequence *hg19*

To estimate the genetic divergences of diverse samples *Q* to the human reference sequence, we align each read for an individual, *Q*, to the inferred common ancestor sequence (HCCA). Because all parts of the human genome are represented within HCCA (either by the common ancestor base or the *hg19* base), we regard unique placement within this genome sequence as evidence of unique and correct placement within human and chimpanzee for regions of one-to-one human and chimpanzee orthology. We take the reference human and chimpanzee aligned

base from the EPO alignment to HCCA. In this way, the multiple sequence alignment between *Q, C,* and *H* is induced by the single, pairwise alignment between *Q* and HCCA. Conveniently, this circumvents the biases inherent in progressive multiple sequence alignment. We use a combination of alignment tools to map reads: BWA[5] for the present-day human samples and Denisova and ANFO[1], a specialized fast-mapper for ancient DNA, for the Vindija Neandertal samples. We require a map-quality of ≥30 to HCCA for the two ancient samples and ≥60 for the paired-end HGDP samples for further consideration. Finally, we only consider bases at positions of one-to-one human chimpanzee orthology that are covered by at least one outgroup sequence. This filtering is designed to ensure that all analyzed alignments of *Q* are unambiguously placed in the genomes of the human and chimpanzee.

For each individual, *Q*, we also examine the distribution of sequence coverage across the genome, which differs for each sample primarily due to the total amount of data collected. To avoid analyzing sites of possible copy number variation or mapping difficulties, we set a cut-off such that sites above the 95th percentile of coverage are excluded. We also set a base quality cutoff for each sample that excludes the lowest 5th percentile of all base observations for each sample. This requires finer filtering resolution than is possible for the PHRED base quality score distribution. Therefore, at the cutoff quality score for each base, a base is accepted with a probability that allows 95% of all bases of that type to be included. The coverage and base quality cutoffs are shown in Table S2.2.

For each site in HCCA for which there is a single human and chimpanzee genomic base, we take the first, randomly chosen base from *Q* that passes the filtering criteria. This strategy avoids the complication of heterozygous sites, which could otherwise bias divergence estimation since mappers have increased sensitivity in detecting reads carrying the reference allele. The substitution spectra for Vindija and Denisova data obtained in this way are shown in Figure S2.2.

Table S2.2: Base filtering cutoffs used for divergence estimates

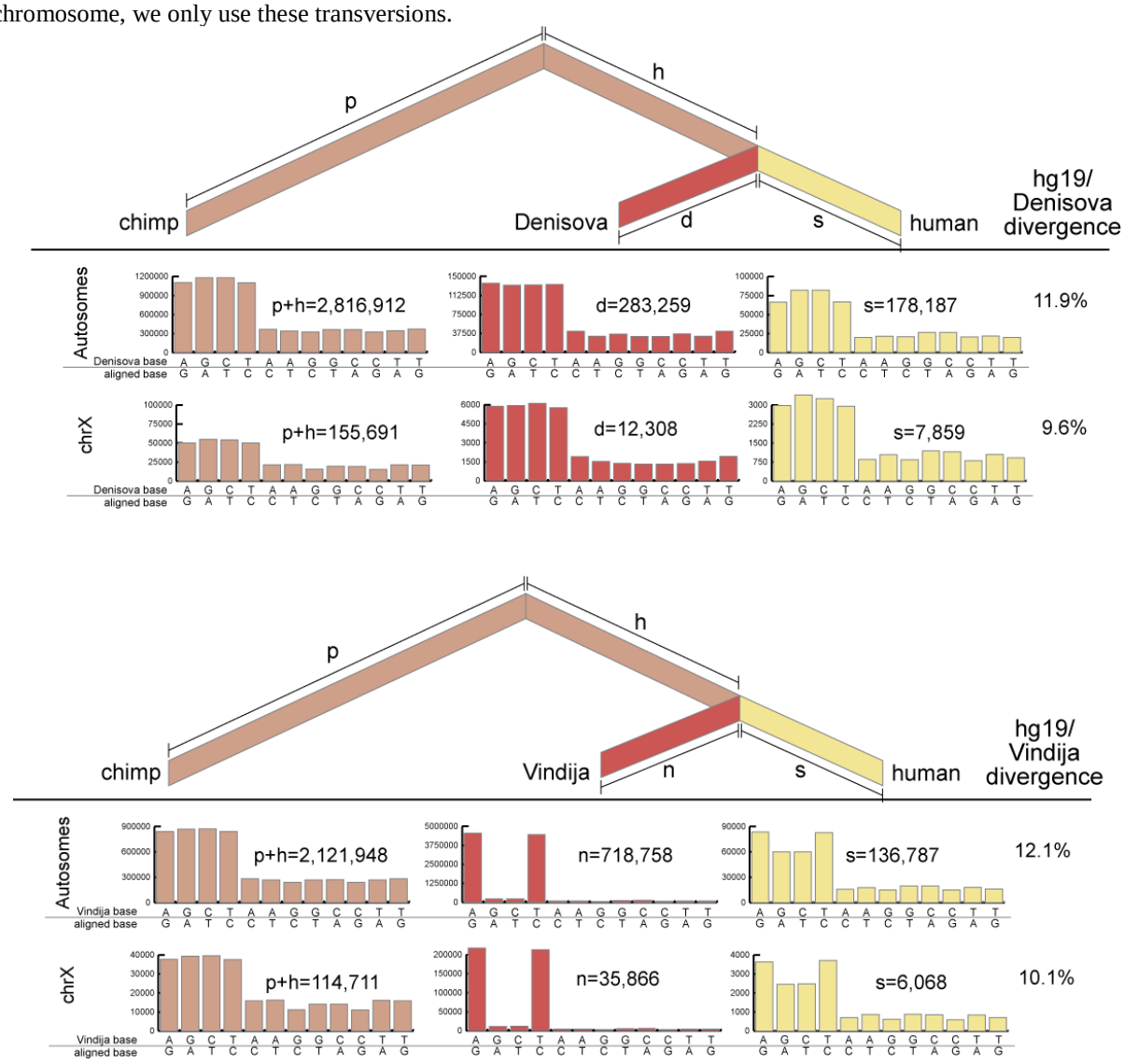
Sample (Abbreviation)	A (Pr)	C (Pr)	G (Pr)	T (Pr)	Maximal coverage
Denisova	40 (1.000)	40 (1.000)	40 (1.000)	40 (1.000)	6
Vindija	27 (0.428)	26 (0.049)	27 (0.308)	27 (0.579)	4
HGDP00778 (Han)	16 (0.489)	14 (0.239)	17 (0.003)	15 (0.11)	8
HGDP00542 (Papuan1)	13 (0.051)	10 (0.119)	15 (0.434)	13 (0.880)	8
HGDP00927 (Yoruba)	17 (0.692)	14 (0.440)	18 (0.562)	16 (0.985)	9
HGDP01029 (San)	17 (0.830)	15 (0.914)	18 (0.649)	16 (0.877)	12
HGDP00521 (French)	17 (0.317)	16 (0.985)	18 (0.024)	17 (0.515)	10
HGDP00456 (Mbuti)	17 (0.041)	14 (0.504)	17 (0.704)	16 (0.379)	8
HGDP00998(Native American)	18 (0.210)	14 (0.126)	17 (0.147)	17 (0.589)	4
HGDP00665 (Sardinian)	19 (0.789)	15 (0.302)	18 (0.474)	17 (0.200)	6
HGDP00491 (Bougainville)	18 (0.810)	14 (0.288)	17 (0.445)	16 (0.291)	6
HGDP00711 (Cambodian)	18 (0.717)	14 (0.303)	17 (0.331)	16 (0.398)	6
HGDP01224 (Mongolian)	18 (0.371)	15 (0.789)	17 (0.051)	16 (0.090)	6
HGDP00551 (Papuan2)	17 (0.188)	14 (0.661)	17 (0.932)	16 (0.885)	6

Note: For each aligned genomic position, the first randomly selected base to satisfy these filtering criteria was chosen for our divergence analysis. For each base, the quality score cutoff and a probability of acceptance at that cutoff is shown in parentheses. No site was considered that was covered by more reads than is listed in the column "Maximal coverage".

Genetic divergence to the African and European portions of the human reference sequence

The divergence estimates computed here are based on all segments of the reference human genome *hg19* that pass the unambiguous orthology filters as described above. However, *hg19* is in fact a mosaic of Bacterial Artificial Chromosomes of different ancestries[1]. A concern is thus that the genetic divergence estimates may be different depending on the underlying ancestry of the human genomic region in question. Thus, we recomputed the divergence estimates using only the regions of *hg19* confidently inferred to be of European and African ancestry, as described previously[1]. Since our previous annotation of which sections of the human genome reference sequence were of European or African ancestry was carried out for *hg18*, the genome annotation of these segments was transferred to *hg19* using liftOver (http://genome.ucsc.edu).

Figure S2.2: Genome divergence between *hg19* and two archaic hominins, Denisova and Vindija. Sequences from the Denisova individual (above) and Vindija Neandertals (below) were aligned to the HCCA sequence and lineage-specific substitutions of each type are shown. We observe an extreme excess of C to T and G to A errors in the Vindija sample compared with the Denisova sample, which is expected since the Denisova sample was UDG treated during library preparation in order to remove uracil residues. The number of transversions specific to each lineage are indicated above each histogram. To calculate genetic divergences for both the autosomes and the X chromosome, we only use these transversions.



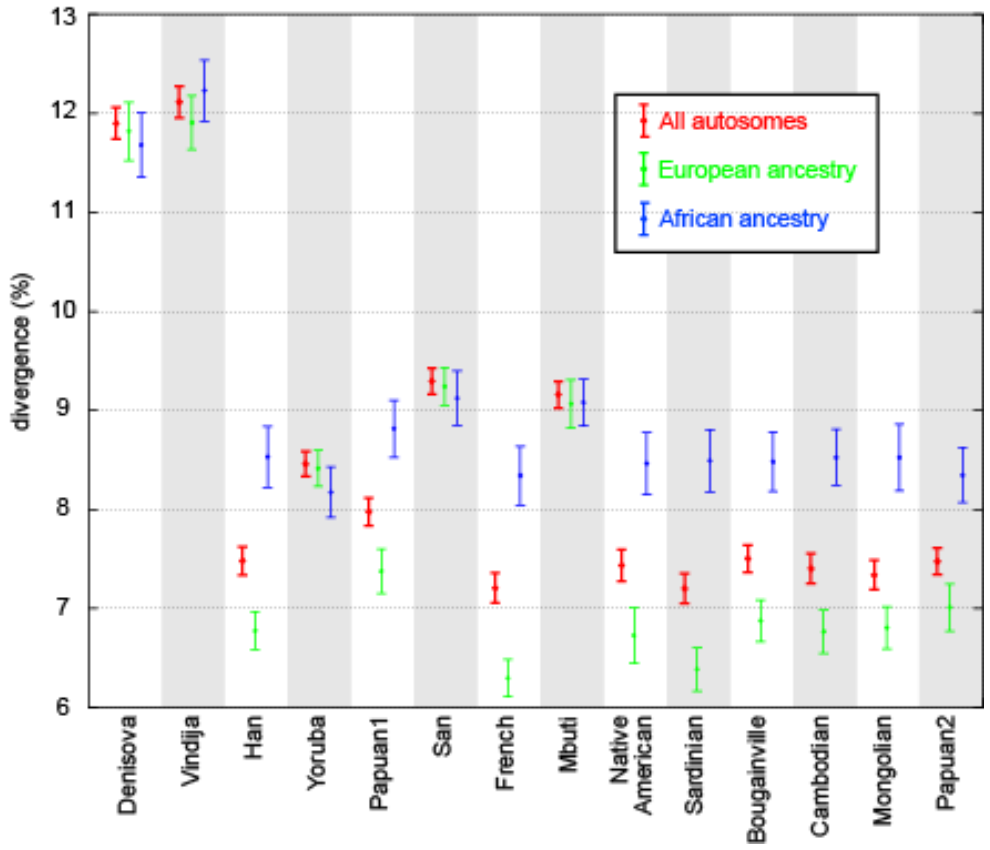


Figure S2.3: Divergence to the reference human genome of Denisova, Vindija, and diverse present-day humans as a fraction of the total divergence of the human lineage since the human-chimpanzee common ancestor. For each sample, we calculate divergence using only transversion substitutions. The error bars correspond to 95% confidence intervals from a Block Jackknife over 10Mb segments.

Figure S2.3 and Table S2.3 use this approach to compare the divergence of *hg19* to diverse hominins. The analysis confirms that both Denisova and the Vindija Neandertal are more deeply diverged to the *hg19* than any present-day humans. We also present these results restricted to the segments of *hg19* that are confidently assigned to be of African and European ancestry[1].

Table S2.3: Genetic divergence of diverse hominins to the segments of the human reference sequence (hg19) as a fraction of human-chimpanzee genetic divergence

	All autosomes		European segments		African Segments	
Sample (Abbreviation)	**Divergence**	**Std. Err.**	**Divergence**	**Std. Err.**	**Divergence**	**Std. Err.**
Denisova	11.90%	0.08%	11.82%	0.15%	11.68%	0.17%
Vindija	12.11%	0.08%	11.91%	0.14%	12.23%	0.16%
HGDP00778 (Han)	7.48%	0.07%	6.78%	0.10%	8.53%	0.16%
HGDP00927 (Yoruba)	8.46%	0.06%	8.42%	0.09%	8.18%	0.13%
HGDP00542 (Papuan1)	7.98%	0.07%	7.38%	0.11%	8.81%	0.15%
HGDP01029 (San)	9.29%	0.07%	9.24%	0.10%	9.13%	0.14%
HGDP00521 (French)	7.21%	0.08%	6.30%	0.10%	8.34%	0.15%
HGDP00456 (Mbuti)	9.16%	0.07%	9.07%	0.12%	9.08%	0.12%
HGDP00998(Native American)	7.44%	0.08%	6.73%	0.14%	8.46%	0.16%
HGDP00665 (Sardinian)	7.20%	0.08%	6.39%	0.11%	8.49%	0.16%
HGDP00491 (Bougainville)	7.50%	0.07%	6.88%	0.11%	8.48%	0.15%
HGDP00711 (Cambodian)	7.41%	0.08%	6.77%	0.11%	8.53%	0.15%
HGDP01224 (Mongolian)	7.34%	0.08%	6.80%	0.11%	8.53%	0.17%
HGDP00551 (Papuan2)	7.48%	0.07%	7.01%	0.12%	8.35%	0.14%

Note: Analyses are restricted to transversions. This table presents the values plotted in Figure S2.3 in numerical form.

Variation in genetic divergence across loci

We also computed the variation in genetic divergence between diverse hominins and the human reference sequence *hg19* over smaller intervals. Figure S2.4 presents divergence in 100kb bins, calculated across bins containing at least 50 informative transversion sites. We observe a similar profile for Vindija and Denisova, each of which show more deeply diverging segments than any of the panel of current humans sequenced here.

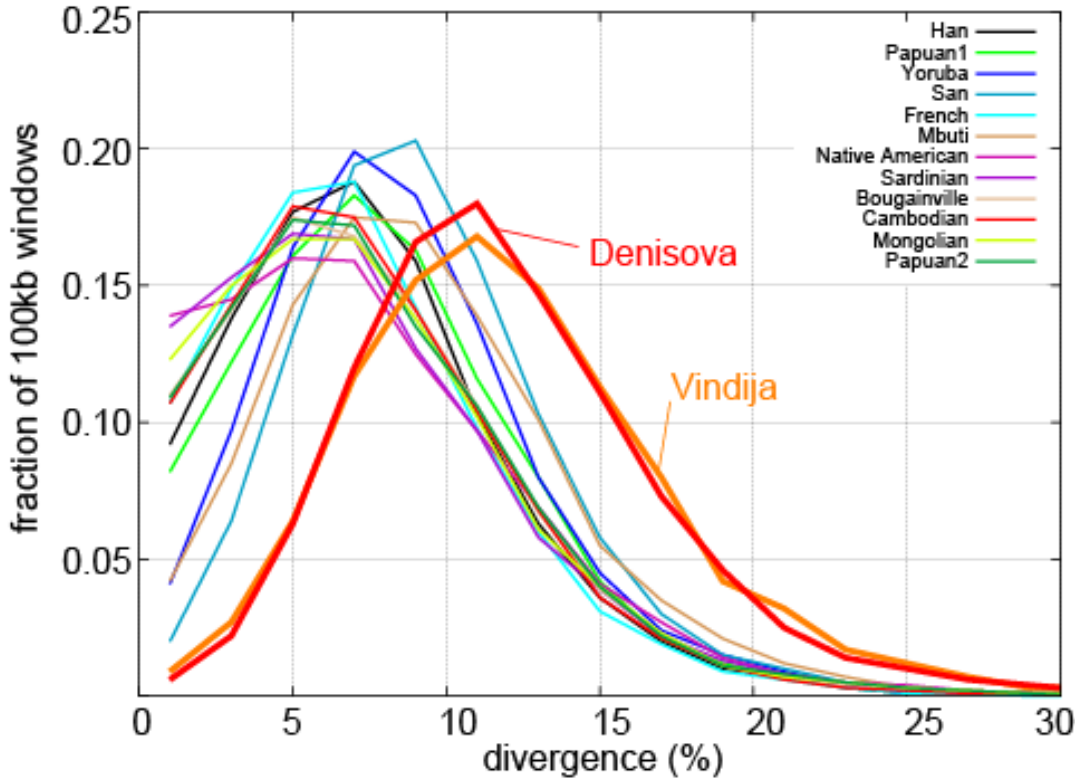


Figure S2.4: Variation in genetic divergence over 100kb windows for Denisova, Vindija, and diverse present-day humans as a fraction of human-chimpanzee divergence. For each sample, we calculate divergence in 100kb windows. Only windows containing at least 50 informative sites are considered.

Sequencing error rate estimates

The overall rate of sequencing error in each sample can also be estimated. Assuming that an equal number of true substitutions have occurred on lineages *H* and *Q* since they diverged, and that there is no error in the human reference sequence *H*, any excess in *Q* can be attributed to sequencing error (here, we are ignoring the slightly shorter branch lengths in the archaic hominins, which is expected to result in a very small underestimate of the error rate for Denisova and Vindija). It is important to recognize that because our allele-picking strategy for each dataset takes a single, random base that passes the quality criteria, these error estimates describe the error within the data that pass those quality criteria. These thresholds are different for each sample (Table S2.2), in order to maintain overall base composition which would otherwise be skewed by the variable ability to call each base with equal reliability[6]. Table S2.4 shows the inferred error rate for each sample both for all sites and restricted to transversions (transversion-based analysis, with its characteristic lower error rate, is used for most of our inferences).

Table S2.4: Sequencing error rates inferred for each sample

Dataset	***H***	***Q***	***H(tv)***	***Q(tv)***	**Total aligned nucleotides**	**Error at all sites**	**Error at transversions**
Craig Venter *(HuRef)*	585,634	798,078	222,041	306,207	1,689,774,029	0.000126	0.000050
Craig Venter chrX	16,853	35,982	6,540	15,667	62,396,720	0.000307	0.000470
Denisova	475,708	819,616	178,187	283,259	833,094,104	0.000413	0.000127
Vindija	422,934	10,201,038	136,787	718,758	632,205,127	0.015467	0.000940
Han	350,340	2,774,594	133,910	1,622,646	955,199,554	0.002538	0.001570
Papuan1	361,034	5,145,515	223,316	2,080,997	934,216,140	0.005121	0.003159
Yoruba	397,655	2,787,004	151,857	1,588,838	953,400,021	0.002506	0.001518
San	462,329	2,958,369	176,332	1,696,237	1,005,575,044	0.002482	0.001523
French	322,767	1,849,922	123,552	1,033,175	908,610,018	0.001681	0.001008
Mbuti	226,057	1,510,980	85,551	916,002	523,153,868	0.002456	0.001598
Native American	151,561	1,157,148	57,371	696,435	428,915,752	0.002344	0.001500
Sardinian	165,978	909,486	62,833	528,401	486,971,839	0.001527	0.000962
Bougainville	227,517	1,335,137	86,260	773,135	635,195,155	0.001744	0.001089
Cambodian	228,729	1,368,464	87,061	808,663	649,955,726	0.001754	0.001118
Mongolian	182,062	1,000,951	68,912	594,788	522,749,460	0.001567	0.001013
Papuan2	221,106	1,370,149	83,504	846,126	615,597,218	0.001867	0.001247

Note: These alignment data are from the autosomes at positions of one-to-one human and chimpanzee orthology whose common ancestor sequence is supported by at least one outgroup sequence. The Craig Venter data derive from a whole-genome alignment of the Craig Venter genome against *hg19* and not Illumina sequence reads. Because *HuRef* is a male, the chromosome X coverage is one-half the autosomal average. Thus, the inferred error rate is expected to be higher, as we observe.

References for SI 2

1. Green, R. E. et al., A draft sequence of the Neandertal genome. *Science* **328,** 710 (2010).
2. Hofreiter, M., Jaenicke, V., Serre, D., Haeseler Av, A. and Pääbo, S., DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res* **29**, 4793 (2001).
3. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. and Birney, E., Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814 (2008).
4. Levy, S. et al., The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
5. Li, H. and Durbin, R., Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**, 1754 (2009).
6. Kircher, M., Stenzel, U., Kelso, J., Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.* **10**, R83 (2009).

Supplementary Information 3

Estimates of present-day human contamination.

Philip Johnson*, Svante Pääbo and Richard E. Green

* To whom correspondence should be addressed (plfjohnson@emory.edu)

Mitochondrial DNA contamination estimates

All DNA sequences from each of the two libraries (SL3003, SL3004) were assembled using *mia*, as described previously[1]. The sequences built into each assembly were filtered to remove low-quality bases (Q < 20) and putative nuclear insertions of mtDNA sequences were removed using BWA[2] to identify DNA sequences whose best match was against sequences other than the mtDNA of *hg18*. Potential PCR duplicate sequences were identified based on having identical start and end coordinates and strand orientation, and such examples were collapsed into a single sequence. The assembly was then used to assess the level of human mtDNA contamination. For each fragment that covered one of 276 diagnostic positions in which the Denisova phalanx mtDNA sequence[3] differs from at least 99% of a world-wide panel of 311 contemporary human mtDNAs, we examined the diagnostic position to infer if the sequence matched the Denisova individual or modern humans. The counts for each are shown in Table S3.1.

Table S3.1: mtDNA contamination rate estimates for each library

Library	# fragments that match Denisova	# fragments that match present-day humans	Estimated % contamination from present-day humans
SL3003	7,421	12	0.16
SL3004	5,036	6	0.12

Note: For each library, the number of fragments that match to the human and Denisova consensus sequences are shown.

Sex determination and Y chromosome contamination estimates

To determine the sex of the Denisova individual, we focused on 157 DNA segments that we previously identified as unique to the human Y chromosome. These amount to a total of 111,132 base pairs and were identified as any DNA sequence of at least 500 base pairs where all 30-mer subsequences differ by at least 3 mismatches from any non-Y chromosome sequence, and that are not within a repetitive element using the rmsk327 table from the UCSC annotation of *hg18*[4].

We counted all sequences that mapped within these regions. If the sequence data derive from a random sample of DNA from a male, the frequency of Y chromosome sequences should be:

$$(111{,}132 / 2{,}800{,}000{,}000) \times 0.5 = 0.000019845$$

Thus, approximately 2 in 100,000 hominin sequences should fall within these regions if the Denisova individual is a male. GC rich sequences tend to be overrepresented in ancient DNA[1]. Since these regions unique to the Y chromosome have a GC content of 51.3% GC, higher than the human genome average of 40.9%, we expect to over-estimate the presence of male-derived sequences. Thus, we have a conservative estimate of contamination if the bone is from a female.

For each Denisova library, we count the total number of sequences that mapped to *hg18* to arrive at an expected number under the assumption that the individual is male. We then test whether the

observed number is consistent with this expectation. The data are shown in Table S3.2. For both libraries, we can reject the hypothesis that the Denisova individual is a male.

Given that the Denisova individual is female, we assume that the Y chromosomal sequences that we observe are due to male contamination. If male contamination is equally likely to fall anywhere in the genome, the rate of accumulation of male contaminating sequences within these regions can be used to estimate contamination:

$$y = c \times \mathrm{Y} \times n$$

where y is the number of hits in the Y-unique regions, c is the percentage of male contamination, Y is the fraction of the genome in the Y-unique regions, and n is the number of reads. The male human contamination estimates are shown in Table S3.2.

Table S3.2: Male contamination rate estimates for each library

Library	Total *hg18* mapped sequences >= 30 nt	chrY hits expected from a male	chrY hits observed	Estimate of % male contamination (95% confidence interval)
SL3003	73,005,587	1,449	0	0.00 (0.00-0.25)
SL3004	35,049,154	696	3	0.43 (0.09-1.26)

Note: For each library, the number of sequences mapping within the Y-chromosome unique regions is shown, and we also show the expected number of matches for a male individual. The ratio gives our estimate of the male contamination rate.

Autosomal contamination estimates

We begin with alignments to the human-chimpanzee common ancestor sequence (HCCA; SI 2), which eliminates bias toward aligning fragments (and thus alleles) that match the reference genome. Given the high coverage, we restrict to sites that are inferred to be fixed differences from chimpanzee on the basis of the 5 modern humans from CEPH-HGDP that we previously sequenced[4]. We further filter the Denisova data on map quality >30 and base quality >30.

Data and intuition

Our data can be summarized by the counts of the number of sites matching each possible pattern (i.e. d_1 derived + a_1 ancestral out of n_1 alleles in SL3003 and d_2 derived + a_2 ancestral out of n_2 alleles in SL3004) (Table S3.3). These counts allow us to form a simple estimator of (heterozygosity + sequencing error +contamination) by taking the percentage of derived alleles found in library #1 at sites for which library #2 has only ancestral alleles. As the number of ancestral alleles in library #2 increases, the chance of a site being heterozygous decreases, but sequencing error remains. We develop a likelihood-based estimator to use all the data and maximize power to separate contamination from heterozygosity and sequencing error.

Likelihood estimator

Our estimator follows a similar procedure to that used in the Neandertal paper (ref. 4; SOM 7), with the exception that, since the two Denisova libraries (SL3003 and SL3004) derive from different physical extracts, we allow for the possiblity of different contamination (c_1, c_2) and error (ε_1, ε_2) rates for each each library. The two libraries still share the nuisance evolutionary parameters (p_{ad}, p_{dd}). We refer to a human-like allele as "derived" and a chimpanzee-like allele as "ancestral"; however, this notation is technically incorrect since these substitutions have not been polarized by an outgroup.

Let $\Omega = \{c_1, c_2, p_{ad}, p_{dd}, \varepsilon_1, \varepsilon_2, f\}$ denote the set of all parameters, where:

- $c_j \rightarrow$ contamination rate. A given read from library j will be from a (contaminating) human with probability c_j and from the Denisova individual with probability $1-c_j$.
- $p_{ad} \rightarrow$ probability of the Denisova individual being heterozygous, given than humans and chimpanzees differ at this site.
- $p_{dd} \rightarrow$ probability of the Denisova individual being homozygous for the human allele, given that humans and chimpanzees differ at this site.
- $\varepsilon_j \rightarrow$ probability of an error in library j. We observe the human allele when the truth is chimpanzee (or vice versa) with probability ε_j.
- $f \rightarrow$ probability of a contaminating allele being human-like. When examining only sites of fixed differences between humans and chimpanzees, $f = 1$.

We write the probability of the observed numbers of derived alleles (n_d) as the product of the probabilities of the L individual sites, conditional on the number of reads (n) sampled from each library (second subscript $\in \{1,2\}$) at each site:

$$\text{lik}(\Omega) = \Pr(n_{1,1,d}, \ldots, n_{L,2,d} \mid n_{1,1}, \ldots, n_{L,2}, \Omega) = \prod_i \Pr(n_{i,\{1,2\},d} \mid n_{i,\{1,2\}}, \Omega) \quad \text{(S4.1)}$$

Dropping the subscript i for ease of notation, we condition on the true derived allele frequency, t, and assume that contamination and sequencing error occur independently:

$$\Pr(n_{\{1,2\},d} \mid n_{\{1,2\}}, \Omega) = \sum_{t=0}^{2} \Pr(t \mid p_{ad}, p_{dd}) \Pr(n_{1,d} \mid t, n_1, c_1, \varepsilon_2, f) \Pr(n_{2,d} \mid t, n_2, c_2, \varepsilon_2, f) \quad \text{(S4.2)}$$

The first term inside the sum (the probability of the truth) is a simple function of the parameters:

$$\Pr(t \mid p_{ad}, p_{dd}) = \begin{cases} 1 - p_{ad} - p_{dd} & t = 0 \\ p_{ad} & t = 1 \\ p_{dd} & t = 2 \end{cases} \quad \text{(S4.3)}$$

The second term inside the sum, the probability of the observed number of derived alleles in each library $j \in \{1,2\}$, follows a binomial distribution:

$$\Pr(n_{j,d} \mid t, n_j, c_j, \varepsilon_j) = \binom{n_j}{n_{j,d}} q_t^{n_{j,d}} (1-q_t)^{n_j - n_{j,d}} \quad \text{(S4.4)}$$

$$\begin{aligned} q_2 &= c_j f_j (1-\varepsilon_j) + c_j (1-f_j)\varepsilon + (1-c_j)(1-\varepsilon_j) \\ q_1 &= c_j f_j (1-\varepsilon_j) + c_j (1-f_j)\varepsilon + (1-c_j)(1-\varepsilon_j)/2 + (1-c_j)\varepsilon_j/2 \\ q_0 &= c_j f_j (1-\varepsilon_j) + c_j (1-f_j)\varepsilon + (1-c_j)\varepsilon_j \end{aligned} \quad \text{(S4.5)}$$

The overall likelihood of the data given the parameters can be calculated from (S4.1), by substituting (S4.2), (S4.3), (S4.4) and (S4.5) in turn.

Finally we estimate our parameters of interest (c_1 and c_2) by maximizing the likelihood of the data over all parameters $\{c_1, c_2, p_{ad}, p_{dd}, \varepsilon_1, \varepsilon_2, f\}$. We reduce the number of dimensions by recalling that $f = 1$ for fixed sites and estimating the error parameters independently using sites at which three different bases are observed. Confidence intervals for c_1 and c_2 can be generated using a likelihood ratio test of the global maximum likelihood to the profile likelihood ($\ell(c_1, c_2) = \max_{p_{ad}, p_{dd}}[\text{lik}(\Omega)]$) and comparing to a χ^2 distribution with 2 degrees of freedom.

Table S3.3: Autosomal estimates of heterozygosity + sequencing error + contamination

	Derived/Ancestral allele counts in test library			
Test library/ Reference library	**2 = Reference library coverage**	**3 = Reference library coverage**	**4 = Reference library coverage**	**5 = Reference library coverage**
SL3003/SL3004	480/21441 (2.2%)	54/4881 (1.1%)	11/660 (1.6%)	0/47 (0%)
SL3004/SL3003	393/20331 (1.9%)	116/9636 (1.2%)	33/3135 (1.0%)	5/561 (0.9%)

Note: The cells in the table show the total number of derived vs. ancestral alleles (counts and percentage) in one "test" library for sites at which the other "reference" library has no derived alleles and the number of ancestral alleles in the column header. The sites analyzed in this table always have the derived allele in 5 present-day humans (San, Yoruba, Han, Papuan and French)[4].

Results: error rate

To estimate the error rate for the purpose of this analysis, we use triallelic sites. First we restrict to sites where the human-chimpanzee alleles form a transition and divide the number of reads containing a third allele (implying a transversion error) by the total number of reads at these sites. We then repeat the procedure for sites where the human-chimp alleles form a transversion to yield a transition error rate. The estimated error rates are shown in Table S3.4.

Table S3.4: Transition and transversion error rate estimates in the two Denisova libraries

	ε_1 **(SL3003)**	ε_2 **(SL3004)**
transversion error rate	0.00038	0.00038
transition error rate	0.0018	0.0017

Despite UDG treatment to filter out ancient DNA damage, error rates are slightly higher at transitions than transversion. The error rates in the libraries are similar, if not indistinguishable.

Results: contamination

Given the error rates, we can estimate contamination rates by examining sites covered by a total read coverage between from 1 and 6. This encompasses 95% of the sites covered by Denisova data. This filter is designed to avoid mapping or genome assembly artifacts that are often coincident with high coverage sites (ref. 5). The triangle in Figure S3.1 indicates the maximum likelihood estimate, and the dashed blue lines indicate approximate 95% confidence intervals.

If we subdivide by read coverage (Figure S3.2), the inferred contamination rate changes slightly, which may be indicative of variation in ancient DNA preservation rates across the genome. Assuming that contamination remains at a constant level, then lower amounts of recovered and sequenced ancient DNA would yield both lower coverage and a higher contamination rate. Contamination estimates for sites with coverages of 1 through 6 are presented in Table S3.5.

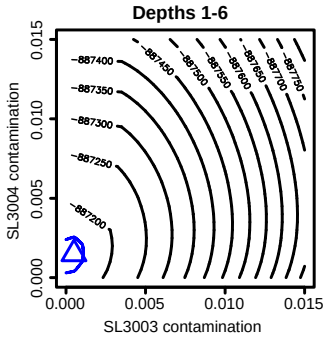


Figure S3.1: Maximum-likelihood estimates of human contamination in the two Denisova sequence libraries.

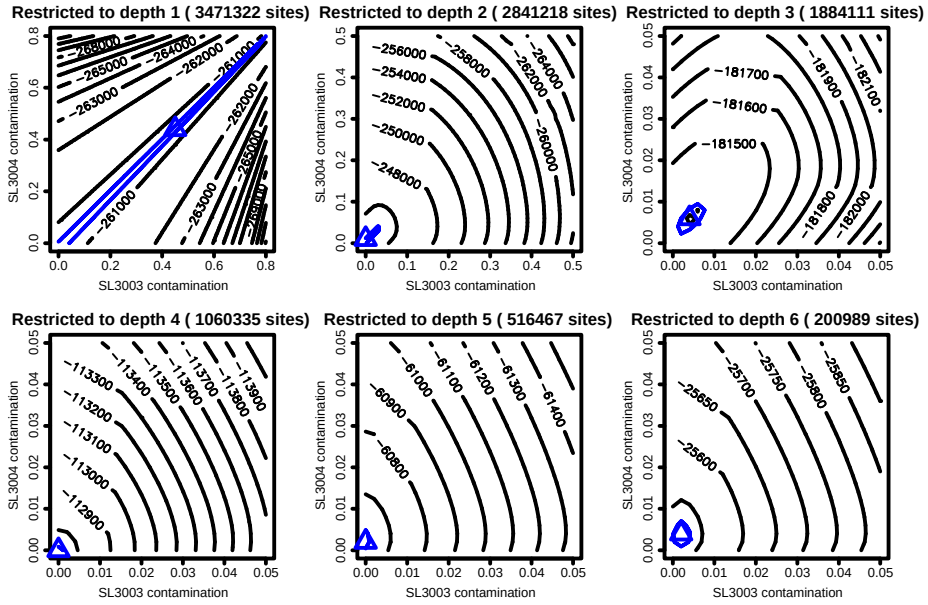


Figure S3.2: Maximum-likelihood estimates of human contamination in the two Denisova sequence libraries, subdivided by sequence depth at the sites under consideration.

Table S3.5 Nuclear DNA estimates of human contamination in the two Denisova libraries

	Maximum likelihood estimate	95% confidence interval
SL3003	0.0003	$(0, 9\times10^{-4})$
SL3004	0.0010	$(5\times10^{-4}, 2\times10^{-3})$

References for SI 3

1. Green, R.E. et al., A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**, 416 (2008).
2. Li, H. and Durbin, R., Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754 (2009).
3. Krause, J. et al., The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464**, 894 (2010).
4. Green, R.E. et al., A draft sequence of the Neandertal genome. *Science* **328**, 710 (2010).
5. Bentley, D.R. et al., Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53 (2008).

Supplementary Information 4

A catalog of ancestral features in the Denisova genome.

Martin Kircher*, Udo Stenzel and Janet Kelso

* To whom correspondence should be addressed (Martin.Kircher@eva.mpg.de)

Identification of changes on the human lineage

We identified positions that have changed on the hominin lineage since separation from apes and more distantly related primates using whole genome alignments for human (*hg18*), chimpanzee (*pantro2*), orangutan (*ponabe2*) and rhesus macaque (*rhemac2*) as described in ref. 1. Briefly, multi-species whole genome alignments, based on either *hg18* or *pantro2*, were screened for differences between the human and chimpanzee sequence, and the lineage on which the change occurred was assigned based on two out-groups (the orangutan and rhesus macaque). We extracted 15,216,383 single nucleotide differences (SNDs) and 1,364,433 insertion or deletion differences (indels) from the human-based alignment, and 15,523,445 SNDs and 1,507,910 indels from the chimpanzee-based alignment. We retained only (i) positions identified in both human-based and chimpanzee-based alignments where (ii) no gaps are present within a 5nt-window of the event, and (iii) where there is sequence available for both out-groups and where these sequences are consistent. In the case of indels we required that (v) indel length does not vary between species and that (vi) the indel sequence is not marked as a repeat. This generates a set of 10,535,445 SNDs and 479,863 indels inferred to have occurred on the human lineage.

Identification of positions with Denisova sequence coverage

To reduce the effects of sequencing error, we used the alignments of the Denisova phalanx reads to the human and chimpanzee reference genomes to construct human-based and chimpanzee-based consensus sequences from multiple reads of the same Denisova molecule (SI 1), and joined overlapping fragments to construct "minicontigs". In this process, overlapping alignments were merged along the common reference to create a single multi-sequence alignment. For each column of the alignment, the number of gaps was counted, and if half the reads or more showed a gap, a gap (resulting in a deletion or no insertion, as appropriate) was called. If fewer than half the reads showed a gap, the most likely diallele per column was calculated follows:

Define the likelihood of diallele XY as:

$$\boldsymbol{L}_{XY} = \prod_i \frac{\boldsymbol{p}(\boldsymbol{b}_i, \boldsymbol{X}) + \boldsymbol{p}(\boldsymbol{b}_i, \boldsymbol{Y})}{2} \tag{S4.1}$$

$$\boldsymbol{p}(\boldsymbol{b}_i, \boldsymbol{X}) = \frac{10^{\frac{-\boldsymbol{q}_i}{10}}}{3} \boldsymbol{if}\ \boldsymbol{b}_i \neq \boldsymbol{X} \tag{S4.2}$$

$$= 1 - 10^{\frac{-\boldsymbol{q}_i}{10}}\ \boldsymbol{if}\ \boldsymbol{b}_i = \boldsymbol{X}$$

Here, i ranges over the overlapping reads, b_i is the base of read i in the current column, and q_i is its quality score on the Phred scale.

We then define the probability of each diallele as:

$$\boldsymbol{p}_{XY} = \frac{\boldsymbol{L}_{XY}\,\boldsymbol{p}_{prior}(\boldsymbol{XY})}{\sum_{UV}\boldsymbol{L}_{UV}\,\boldsymbol{p}_{prior}(\boldsymbol{UV})} \tag{S4.3}$$

where

$$\begin{aligned} p_{prior}(XY) &= \frac{1}{1000}\ if\ X \neq Y \\ &= \frac{999}{1000}\ if\ X = Y \end{aligned} \tag{S4.4}$$

Then we call the most probable diallele, express it as a IUPAC ambiguity code, and calculate its quality score as:

$$\boldsymbol{Q} = 10\log_{10}(1-\boldsymbol{p}_{XY}) \tag{S4.5}$$

We used the resulting minicontigs to extract the Denisova sequence homologous to the human-lineage-specific changes from both the human and the chimpanzee minicontig alignments.

We further filtered the data as follows:
(i) The Denisova sequence obtained for a specific site from the human-based and chimpanzee-based alignments was required to be identical and to have a PHRED quality score >30.
(ii) All positions that fall within 5 nucleotides of the ends of minicontigs were excluded to minimize alignment errors and substitutions due to nucleotide misincorporations.
(iii) Positions that fall within 5 nucleotides of insertions or deletions (*i.e.* gaps) in the minicontig alignments were excluded.

Using this filtered dataset, we have Denisova sequence coverage for 4,267,431 of the 10,535,445 substitutions and 105,372 of the 479,863 indels inferred to have occurred on the human lineage.

Electronic access to the catalog
The full catalog of sites where the human reference sequence *hg18* carries the derived allele relative to apes and other primates, annotated by the allelic state in Denisova and Neandertal, is available for download from http://bioinf.eva.mpg.de/download/DenisovaGenome/.

Annotation
We annotated all SNDs and indels using the Ensembl v54 annotation for *hg18* and Ensembl v55 for *pantro2* (in cases where no human annotation was available). A set of 16,762 CCDS genes (Consensus Coding Sequence project of EBI, NCBI, WTSI, and UCSC), each representing the longest annotated coding sequence for the respective gene, was used for downstream analyses.

Amino acid substitutions
We identified 35,523 SNDs in the coding regions of the human CCDS set. There are 21,354 synonymous substitutions and 14,169 non-synonymous substitutions. Non-synonymous amino acid substitutions that rose to nearly 100% frequency (are fixed) in present-day humans since the separation from Neandertals might be of special interest as they may represent targets of recent selection in humans. We therefore excluded all non-synonymous substitutions where current

humans are known to vary (dbSNP v131), and identified 129 fixed, non-synonymous amino substitutions from a total of 2,176 positions in 119 genes where the Denisova carries the ancestral (chimpanzee) allele (Table S4.1)

Table S4.1: Changes in the coding sequences of CCDS genes (n=129) for which the Denisova individual is ancestral and present-day humans are all fixed for the derived state

The table is sorted by Grantham scores (GS), which classifies amino acid changes as radical (>150), moderately radical (101-150), moderately conservative (51-100), or conservative (1-50)[2]. Genes with multiple substitution changes are highlighted. Genomic coordinates are zero-based.

Human				Chimpanzee				N	Database identifier			Amino acid information				
Base	Chrom.	Strand	Pos	Base	Chrom.	Strand	Pos	Neandertal	Ensembl Transcript (ENST)	Gene ID (CCDS)	SwissProt	Strand	Codon	Pos	AA	GS
C	1	+	160234303	T	1	+	141210333	T	294794	1236	*OLM2B*	-	470	2	W/*	-
C	9	+	124603020	T	9	+	122451444	T	277309	35132	*OR1K1*	+	267	1	R/C	180
T	16	+	55853364	A	16	+	56701776	A	219207	10777	*PLLP*	-	85	2	N/I	149
A	6	+	79634102	G	6	+	79859308	G	369940	34488	*IKBP1*	+	31	1	R/G	125
A	6	+	28033607	T	6	+	28474934	T	244623	4642	*OR2B6*	+	204	2	E/V	121
T	19	+	56195777	A	19	+	56666060	A	391806	42600	*KLK8*	-	27	1	S/C	112
G	5	+	118513136	C	5	+	120549991	C	311085	4125	*DMXL1*	+	1239	2	C/S	112
G	15	+	39585012	T	15	+	38515019	T	263800	10077	*LTK*	-	569	1	R/S	110
T	1	+	89500647	G	1	+	90748162	G	370459	722	*GBP5*	-	497	2	E/A	107
A	17	+	71516433	G	17	+	75622154	G	301607	11737	*EVPL*	-	1483	1	W/R	101
A	13	+	83352655	C	13	+	84340069	C	377084	9464	*SLIK1*	-	330	1	S/A	99
A	1	+	1221067	G	1	+	1209836	G	354980	19	*ACAP3*	-	497	2	L/P	98
A	1	+	89370660	G	1	+	90615166	G	294671	720	*GBP7*	-	559	2	L/P	98
C	10	+	37548307	T	10	+	38070616	T	361713	7193	*AN30A*	+	1165	2	P/L	98
T	16	+	538118	C	16_r	+	5709531	C	219611	10410	*CAN15*	+	427	2	L/P	98
C	17	+	23943903	T	17	-	28754402	T	321765	32594	*SPAG5*	-	162	2	G/E	98
A	19	+	59495378	G	19	+	60019602	G	391745	12887	*LIRA3*	-	103	2	L/P	98
C	22	+	49002257	T	22	+	49544877	T	248846	14087	*GCP6*	-	886	2	G/E	98
A	5	+	86600232	G	5	-	28406904	G	274376	34200	*RASA1*	+	70	2	E/G	98
T	9	+	2719704	C	9	+	2767007	C	382082	6447	*KCNV2*	+	539	2	L/P	98
T	10	+	118311044	A	10	+	117285929	A	369221	7594	*LIPP*	+	414	2	M/K	95
C	17	+	59644188	T	17	+	63501641	T	258991	11658	*TEX2*	-	374	2	G/D	94
T	2	+	241112138	G	2b	+	246936269	G	391987	2536	*ANKY1*	-	467	3	K/N	94
C	4	+	4250211	T	4	+	4284217	T	296358	3372	*OTOP1*	-	417	2	G/D	94
C	1	+	35351279	T	1	+	35619202	T	359858	41302	*ZMYM1*	+	421	2	T/I	89
C	1	+	40499233	T	1	+	40903941	T	372759	449	*FACE1*	+	87	2	T/I	89
G	2	+	40510859	A	2a	+	41372434	A	378715	1806	*NAC1*	-	22	2	T/I	89
A	21	+	29226747	G	21	+	28747960	G	361371	33527	*RN160*	-	1662	2	I/T	89
A	3	+	47444152	G	3	+	48489457	G	265565	2755	*SCAP*	-	140	2	I/T	89
A	9	+	134265412	G	9	+	132451335	G	334270	6948	*TTF1*	-	474	2	I/T	89
C	17	+	3066433	T	17	+	3240106	T	304094	11022	*OR1A1*	+	257	2	T/M	81
C	3	+	99555909	G	3	+	102255729	G	354924	33802	*OR5K4*	+	175	1	H/D	81
G	3	+	198159340	A	3	+	202594456	A	238138	3324	*PIGZ*	-	275	2	T/M	81
C	9	+	126152975	G	9	+	124035499	G	320246	6854	*NEK6*	+	291	1	H/D	81
C	X	+	17678235	T	X	+	17782495	T	380041	35210	*SCML1*	+	202	2	T/M	81
C	X	+	22928705	T	X	+	23118523	T	327968	35214	*DDX53*	+	204	2	T/M	81
C	5	+	75627399	A	5	-	39561669	A	322285	43331	*SV2C*	+	460	2	P/H	77
A	1	+	63831784	C	1	+	64730850	C	371084	625	*PGM1*	+	13	2	Q/P	76
A	14	+	95842916	G	14	+	96604596	G	359933	9944	*ATG2B*	-	1465	1	S/P	74
T	3	+	121952209	C	3	+	125367499	C	283875	3002	*T2EA*	+	41	1	S/P	74
G	18	+	64715493	C	18	+	65628674	C	360242	11996	*C102B*	+	371	2	R/T	71
G	4	+	2919071	C	4	+	3060925	C	314262	33945	*NOP14*	-	493	2	T/R	71
A	1	+	159117069	G	1	+	140167930	G	326245	1211	*ITLN1*	-	206	2	V/A	64
A	17	+	32988030	G	17	-	19834375	G	346661	11321	*SYNG*	-	636	2	V/A	64
A	21	+	41788292	G	21	+	41197009	G	332149	33564	*TMPS2*	-	33	2	V/A	64
T	22	+	39090923	C	22	+	39366823	C	216194	14001	*PUR8*	+	429	2	V/A	64
G	6	+	100475588	A	6	+	101531964	A	281806	5044	*MCHR2*	-	324	2	A/V	64
T	7	+	17341916	C	7	+	17496235	C	242057	5366	*AHR*	+	381	2	V/A	64
A	8	+	10507836	G	8_r	+	5845849	G	382483	43708	*RP1L1*	-	394	2	V/A	64
G	X	+	3249673	A	X	+	3261751	A	217939	14124	*MXRA5*	-	1351	2	A/V	64
G	X	+	50394175	A	X	+	50693925	A	376020	35277	*SHRM4*	-	546	2	A/V	64
C	1	+	46821521	G	1	+	47378043	G	371946	538	*MKNK1*	-	34	2	G/A	60
C	1	+	26564052	T	1	+	26589352	T	329206	279	*ZN683*	-	176	1	A/T	58
A	1	+	43584841	G	1_r	+	8286746	G	372470	483	*TPOR*	+	374	1	T/A	58
T	1	+	118360154	C	1	-	119555017	C	336338	899	*SPG17*	-	1415	1	T/A	58
A	11	+	7463757	G	11	+	7318919	G	329293	7779	*OLFL1*	+	26	1	T/A	58
C	11	+	18295977	T	11	+	18263443	T	352460	7836	*HPS5*	-	2	1	A/T	58

Human				Chimpanzee				N	Database identifier			Amino acid information				
Base	Chrom.	Strand	Pos	Base	Chrom.	Strand	Pos	Neandertal	Ensembl Transcript (ENST)	Gene ID (CCDS)	SwissProt	Strand	Codon	Pos	AA	GS
G	12	+	6754050	A	12	+	7000753	A	203629	8561	*LAG3*	+	181	1	A/T	58
A	14	+	75319511	G	14	+	75532002	G	298832	32124	*TTLL5*	+	958	1	T/A	58
C	15	+	40529603	T	15	+	39548721	T	263805	32208	*ZF106*	-	697	1	A/T	58
G	15	+	78960362	A	15	+	78847888	A	356249	10315	*K1199*	+	150	1	A/T	58
A	16	+	19637267	G	16	+	19834441	G	320394	10580	*IQCK*	+	47	1	T/A	58
T	16	+	65504564	C	16	+	66617266	C	299752	10823	*CAD16*	-	342	1	T/A	58
G	17	+	71264629	A	17	+	75364635	A	200181	11727	*ITB4*	+	1689	1	A/T	58
G	19	+	54363018	A	19	+	54888107	A	252826	33073	*TRPM4*	+	101	1	A/T	58
G	4	+	89627245	A	4	+	91410717	A	264350	3630	*HERC5*	+	619	1	A/T	58
G	10	+	102666423	A	10	+	101264246	A	238961	7500	*F178A*	+	98	1	E/K	56
T	17	+	24983159	C	17	-	27682885	C	269033	11253	*SSH2*	-	1033	1	S/G	56
G	19	+	14671033	A	19	+	15117465	A	292530	12316	*ZN333*	+	83	1	E/K	56
C	4	+	5693149	T	4	+	5786383	T	344408	3382	*LBN*	-	488	1	G/S	56
A	5	+	176731622	G	5	+	179753998	G	398128	43405	*RGS14*	+	549	1	K/E	56
G	8	+	19266005	A	8	+	15599024	A	265807	6009	*SH24A*	+	284	1	E/K	56
T	1	+	94337038	G	1	+	95593362	G	370225	747	*ABCA4*	-	223	1	K/Q	53
G	21	+	42770559	T	21	+	42171255	T	291536	13688	*RSPH1*	-	213	1	Q/K	53
G	7	+	88261633	T	7	+	88409220	T	297203	34678	*CG062*	-	187	1	Q/K	53
T	16	+	82720768	C	16	+	84354684	C	219439	10942	*HSDL1*	-	260	2	N/S	46
G	19	+	40449692	A	19	+	40767034	A	361790	12450	*LSR*	+	424	2	S/N	46
A	19	+	63256998	G	19	+	63932677	G	282326	12969	*ZSCA1*	+	332	2	N/S	46
G	20	+	47001360	A	20	+	46380523	A	371917	13411	*BIG2*	+	124	2	S/N	46
T	22	+	45019740	C	22	+	45430536	C	314567	33670	*CV040*	-	95	2	N/S	46
A	10	+	37548646	G	10	+	38070955	G	361713	7193	*AN30A*	+	1278	2	Q/R	43
C	19	+	60400460	T	19	+	60916698	T	376350	33110	*PTPRH*	-	609	2	R/Q	43
T	4	+	46431919	C	4	-	85955456	C	396533	3472	*CX7B2*	-	16	2	Q/R	43
C	8	+	10506552	T	8_r	+	5844565	T	382483	43708	*RP1L1*	-	822	2	R/Q	43
C	9	+	134267343	T	9	+	132453250	T	334270	6948	*TTF1*	-	229	2	R/Q	43
T	9	+	139259701	G	9	+	137500096	G	344774	35186	*F166A*	-	134	1	T/P	38
G	1	+	55125322	C	1	+	55868939	C	371269	600	*DHC24*	-	20	1	L/V	32
C	11	+	74024946	G	11	+	73005395	G	263681	8233	*DPOD3*	+	393	1	L/V	32
C	19	+	11352605	G	19	+	11682327	G	222139	12260	*EPOR*	-	261	1	V/L	32
C	22	+	41158219	G	22	+	41499623	G	329021	14034	*NFAM1*	-	30	1	V/L	32
A	13	+	49103140	G	13	+	49528651	G	282026	9419	*ARL11*	+	186	2	H/R	29
C	14	+	25987939	T	14	+	25373578	T	267422	32061	*NOVA1*	-	197	1	V/I	29
C	14	+	104588536	G	14	+	105590622	G	392585	9997	*GP132*	-	328	1	E/Q	29
G	15	+	38700151	A	15	+	37608632	A	346991	42023	*CASC5*	+	159	2	R/H	29
G	17	+	71264899	A	17	+	75364905	A	200181	11727	*ITB4*	+	1748	2	R/H	29
G	19	+	14667458	A	19	+	15113915	A	292530	12316	*ZN333*	+	70	2	R/H	29
T	22	+	45183663	C	22	+	45600477	C	262738	14076	*CELR1*	-	1707	1	I/V	29
C	3	+	47137662	T	3	+	48170152	T	330022	2749	*SETD2*	-	653	2	R/H	29
G	3	+	99466160	A	3	+	102166233	A	359776	33800	*OR5H6*	+	115	1	V/I	29
G	5	+	176731601	C	5	+	179753977	C	398128	43405	*RGS14*	+	542	1	E/Q	29
A	7	+	134293530	G	7	+	135457096	G	361675	5835	*CALD1*	+	671	1	I/V	29
G	7	+	146456810	A	7	+	147715341	A	361727	5889	*CNTP2*	+	345	1	V/I	29
T	8	+	19360349	C	8	+	15695152	C	332246	6010	*CGAT1*	-	240	1	I/V	29
A	8	+	22076124	G	8	+	18495549	G	318561	43722	*PSPC*	+	46	1	I/V	29
G	8	+	145211312	C	8	+	144064957	C	355091	43776	*GPAA1*	+	275	1	E/Q	29
C	1	+	156879241	G	1	+	137899626	G	368148	41423	*SPTA1*	-	1531	1	A/P	27
C	6	+	2841353	G	6	+	2911117	G	380698	4478	*SPB9*	-	80	1	A/P	27
C	17	+	24983383	T	17	-	27682661	T	269033	11253	*SSH2*	-	958	2	R/K	26
G	8	+	39683508	A	8	+	36409058	A	265707	6113	*ADA18*	+	649	2	R/K	26
G	X	+	153196801	A	X	+	153627491	A	369915	35448	*TKTL1*	+	317	2	R/K	26
T	1	+	156914833	C	1	+	137934884	C	368148	41423	*SPTA1*	-	265	1	N/D	23
C	11	+	6611344	T	11	+	6490738	T	299441	7771	*PCD16*	-	777	1	D/N	23
A	14	+	57932515	G	14	+	57740382	G	360945	9734	*TO20L*	+	30	1	N/D	23
C	2	+	231682274	T	2b	+	237323413	T	258400	2483	*5HT2B*	-	216	1	D/N	23
A	6	+	160425195	G	6	+	163027524	G	356956	5273	*MPRI*	+	2020	1	N/D	23
A	X	+	150843587	G	X	+	151476159	G	393921	14702	*MAGA4*	+	266	1	N/D	23
A	11	+	18265762	T	11	+	18233187	T	352460	7836	*HPS5*	-	871	2	F/Y	22
C	19	+	3498314	G	19	+	3591210	G	398558	42464	*CS028*	-	326	3	L/F	22
G	3	+	198158891	A	3	+	202594007	A	238138	3324	*PIGZ*	-	425	1	L/F	22
T	1	+	6622636	C	1	+	6699799	C	377577	87	*DJC11*	-	389	1	M/V	21
T	12	+	93975518	C	12	+	96042577	C	393102	9051	*NR2C1*	-	242	1	M/V	21
C	16	+	87474655	T	16	+	89288970	T	268679	10972	*MTG16*	-	482	1	V/M	21
T	12	+	44607998	C	12	-	43858515	C	369367	8748	*SFRIP*	-	584	3	I/M	10
A	20	+	31275866	G	20	+	30217609	G	375454	13216	*SPLC3*	+	108	3	I/M	10
A	20	+	32801189	C	20	+	31822768	C	374796	13241	*NCOA6*	-	823	3	I/M	10
G	4	+	184423846	T	4	+	187919922	T	281445	34109	*WWC2*	+	479	3	M/I	10
C	5	+	54620969	T	5	-	60628686	T	251636	34158	*DHX29*	-	317	3	M/I	10
A	11	+	128345808	T	11	+	128028901	T	392657	31718	*RICS*	-	1140	3	D/E	0
T	4	+	57471854	A	4	-	73606113	A	309042	3509	*REST*	+	98	3	D/E	0

We identify 10 genes affected by two amino acid substitutions that are consistent with being fixed in present-day humans since divergence from the common ancestors of Denisovans:

AN30A (Ankyrin repeat domain-containing protein 30A)
HPS5 (Hermansky-Pudlak syndrome 5 protein)
ITB4 (Integrin beta-4 precursor)
PIGZ (GPI mannosyltransferase 4)
RGS14 (Regulator of G-protein signaling 14)
RP1L1 (Retinitis pigmentosa 1-like 1 protein)
SPTA1 (Spectrin alpha chain, erythrocyte)
SSH2 (Protein phosphatase Slingshot homolog 2)
TTF1 (Transcription termination factor 1)
ZN333 (Zinc finger protein 333)

Interestingly, two of these genes are associated with skin diseases (*HPS5* and *ITB4*), which is similar to the high representation of genes associated with skin diseases in the Neandertal-oriented catalog presented in SOM 11 of Green and colleagues[1].

We also used Grantham scores to categorize the 129 amino acid replacements into classes of chemical similarity[2]. We classified 54 sites as conservative (scores of 0-50), 65 as moderately conservative (scores of 51-100), 8 as moderately radical (scores of 101-150), and 1 as radical (score of >151) (Table S4.1). The only gene with an amino acid substitution that is classified as radical is *OR1K1* (*olfactory receptor, family 1, subfamily K, member 1*), an olfactory receptor with a replacement of arginine by cysteine in one of the extracellular domains.

We believe that each of the rather small number of amino acid substitutions that have become fixed in humans since the divergence from the common ancestor with the Denisova individual are of sufficient interest to warrant further functional investigation.

Stop/Start codon substitutions

We identified one fixed non-synonymous change in a stop codon. In *OLM2B* (*Olfactomedin-like protein 2B precursor*), all present day humans have a loss of a stop-codon at amino acid 470, which is required for the protein to contain the Olfactomedin-like domain (amino acids 493-750). In Denisova, the ancestral stop-codon is present and the protein does not include this domain.

We did not identify fixed, non-synonymous changes in start codons where the Denisova individual carries the ancestral allele. However, at one gene, *Riboflavin kinase* (*RIFK*), Denisova carries an ancestral start-codon (rs2490582) that is lost in about 98% of present-day humans. In addition, there are two genes where some (but not all) present-day humans have gained a start codon relative to Denisova. This includes the melastatin gene (*TRPM1, transient receptor potential cation channel, subfamily M, member 1;* rs4779816 derived allele frequency 88%) and *zinc finger protein 211* (*ZNF211;* rs9749449 derived allele frequency 77%). *TRPM1* encodes an ion channel that maintains normal melanocyte pigmentation; functional variants of this gene that use alternative start positions have been described in human tissues[3] and may be able to compensate for the additional start-codon not being present. *ZNF211* is an as-yet uncharacterized zinc finger protein probably involved in transcriptional regulation.

Insertions and deletions in coding sequence

We identified 69 insertion/deletion events within coding sequences. In 15 cases the Denisova state is ancestral, and for 14 of these, present-day humans are not known to vary in dbSNP 131 (Table S4.2). Twelve of these 14 indels are 3 bases long. Of these, 6 delete exactly one amino acid and the other 6 affect two amino acids while maintaining the reading frame. In *HADHA/ECHB (hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/ enoyl-Coenzyme A hydratase)*, a protein that is responsible for the metabolism of long-chain fatty acids, the first amino acid, which is in the mitochondrial transit peptide region of the protein, is removed. Since the mitochondrial transit peptide is responsible for the transport of the protein from the cytoplasm to the mitochondrion, it is possible that this change affects the cellular localization of this protein. Mutations in this gene are associated with hypoglycemia, hypotonia and lethargy[4]. An entire codon is deleted from *RTTN (rotatin)*, a protein required for the early developmental processes of left-right specification and axial rotation and which may play a role in notochord development[5]. Examples of other three-base deletions are in *AHNK (Desmoyokin)*, a protein involved in neuroblast differentiation, in *EME1 (essential meiotic endonuclease 1 homolog 1)*, involved in DNA replication and repair, *SNG1 (synaptogyrin 1)* involved in short and long-term regulation of neuronal synaptic plasticity, and the spermatogenesis-associated protein *SPT21 (spermatogenesis associated 21)*. Interestingly, several genes in which present-day humans appear to have undergone deletions while Denisova carries the ancestral state are involved in neuronal development and function, spermatogenesis and metabolism.

A particularly striking indel that we detected is a single base deletion in one of the final codons of the membrane protein *ADAM8 (disintegrin and metalloproteinase domain-containing protein 8)*. This indel is predicted to lead to a change of frame in the cytoplasmic portion of the protein, 6 amino acids from the derived C-terminus. Disintegrin and metalloprotease proteins are involved in a variety of biological processes involving cell-cell and cell-matrix interactions, including fertilization, muscle development, and neurogenesis. *ADAM8* has also been linked to inflammation and remodeling of the extracellular matrix (including cancers and respiratory diseases)[6]. A single base pair insertion in *chromosome 17 open reading frame 103 (gene trap locus F3b, GTL3B)*, a protein of unknown function, also results in a change in reading frame.

Table S4.2: 15 indel changes in coding sequences where Denisova has the ancestral state

		Human (*hg18*)			Chimpanzee (*pantro2*)				Database identifier			
Type	Seq. (+)	Chr	Start	End	Chr	Start	End	Denisova state	Ensembl ID (ENST)	Gene ID (CCDS)	SwissProt	Exon
deletion	CTT	1	16599892	16599892	1	16628573	16628576	present	335496	172	*SPT21*	9
deletion	ACT	2	26330629	26330629	2a	26822701	26822704	present	317799	1722	*ECHB*	1
deletion	GAG	6	151715809	151715809	6	154110196	154110199	present	253332	5229	*AKA12*	3
deletion	GAC	8	101275635	101275635	8	99149151	99149154	present	251809	34930	*SPAG1*	9
deletion	C	10	134926669	134926669	10	134553566	134553567	present	368566	31319	*ADAM8*	23
deletion	CTC	11	62060131	62060131	11	60909822	60909825	present	378024	31584	*AHNK*	1
deletion	AGC	17	45807977	45807977	17	49360131	49360134	present	338165	11565	*EME1*	1
deletion	CTC	18	66014830	66014830	18	66961824	66961827	present	255674	42443	*RTTN*	7
deletion	ATC	19	14913983	14913983	19	15369839	15369842	present	248072	12320	*OR7C2*	1
deletion	CAG	19	55573634	55573634	19	56073557	56073560	present	253727	42593	*NR1H2*	4
deletion	ACT	19	58146287	58146287	19	58632140	58632143	present	357666	33096	*Z816A*	3
deletion	CAA	22	38107768	38107768	22	38350855	38350858	present	328933	13989	*SNG1*	4
insertion	AGC	2	79990299	79990302	2a	81651049	81651049	missing	361291	42703	*CTNA2*	6
insertion	GCG	2	95210767	95210770	2a	96095976	96095976	missing	295210	42712	*ZNF2*	4
insertion	G	17	21087327	21087328	17	35020903	35020903	missing	399011	42286	*GTL3B*	3

5´ UTR substitutions and insertion/deletions

We have Denisova sequence data for 5,654 of the 12,045 substitutions in 5´ untranslated regions (UTR's) occurring on the human lineage. Of these, there are 66 positions in 64 genes where the ancestral allele is observed, and present-day humans are consistent with being fixed for the derived allele. Two genes each carry two changes in the 5´ UTR: *ETS2 (human erythroblastosis virus oncogene homolog 29)*, a transcription factor that is involved in stem cell development, apoptosis and tumorigenesis, and *FNBP4 (formin binding protein 4)* a gene with roles in a cell adhesion and GPCR-signaling. Denisova state information was also obtained for 198 of 810 indels in 5´ UTRs. For 24 of these (each in a different gene) the Denisovan individual retains the ancestral state while present-day humans are fixed for the derived allele.

3´ UTR substitutions and insertion/deletions

We have Denisova data for 26,113 of 55,883 substitutions in 3´ UTRs. Among these, there are 283 positions (in 234 genes) where the Denisova individual shows the ancestral state and present-day humans are consistent with being fixed for the derived allele. We also find 37 genes with multiple substitutions, with one gene having 4 substitutions *(PRDM10, PR domain containing 10)*, 10 genes with 3 substitutions, and 26 genes with 2 substitutions. The protein encoded by *PRDM10* is a transcription factor that is implicated in normal somite and craniofacial formation during embryonic development[7], which may be involved in the development of the central nervous system as well as in the pathogenesis of gangliosidosis (GM2, neuronal storage disease)[8]. We also have Denisova data for 1,271 of 5,972 indels in 3´ UTRs, 109 of which show the ancestral state in Denisova while present-day humans are fixed for the derived allele. These indels are located in 108 different genes. Two indels are present in the 3´ UTR of *MMP5 (MAGUK p55 subfamily member 5)*, a protein that may play a role in tight junction biogenesis and in the establishment of cell polarity in epithelial cells.

miRNAs

MicroRNA's (miRNAs) are small non-coding RNAs that regulate gene expression by mRNA cleavage or repression of mRNA translation. These molecules have an important role in mammalian brain and embryonic development. We have Denisova sequence for 143 of the 357 single nucleotide differences seen in 1,685 miRNAs annotated in Ensembl 54 (including 670 miRBase-derived microRNAs), and Denisova agrees with *hg18* at 125 of these sites. Out of the remaining 18 sites, 17 are polymorphic in present-day humans, while one change in miRNA *hsa-mir-564* is fixed in present-day humans for the derived allele. The substitution, however, is unlikely to affect microRNA function as it is located in a small bulge outside of the mature sequence. Denisova sequence is also available for 5 of the 17 insertion/deletion events in miRNAs that occurred on the human lineage. In one case, *hsa-mir-1260*, Denisova carries the ancestral allele while present-day humans are apparently fixed for an insertion of adenosine in the human sequence. This insertion is outside of the mature sequence in an inferred loop structure and is thus not likely to affect function.

Human Accelerated Regions

Human Accelerated Regions (HARs) are regions of the genome conserved throughout vertebrate evolution, which have changed radically since humans and chimpanzees separated from their common ancestor. Earlier results from the Neandertal genome analysis[1] indicated that the acceleration may largely predate the Neandertal-human split. Here we examined the union of 2,613 Human Accelerated Regions (HARs) identified in five different studies[9,10,11,12]. We identified 8,949 single nucleotide changes and 213 indels on the human lineage in these HARs.

Denisova sequence was available for 3,494 changes (3,445 substitutions and 49 indels). Of these, 3,128 are derived in Denisova (89.52%, 95% Wilson 2-sided confidence interval with continuity correction [88.45%, 90.51%]), which is significantly higher than for the complete set (86.64% [86.61%, 86.67%]) of all derived substitutions (3,696,534) and all derived deletions (91,985).

It has been argued that HARs may sometimes not be functionally relevant, but instead may be byproducts of biased gene conversion hotspots changing their genomic locations over evolutionary history[13,14,15,16,17]. To explore this possibility, we restricted our analysis to single nucleotide changes that may be due to biased gene conversion (A/T in chimp to G/C in human). We continue to find that Denisova carries the derive allele more often in HARs than elsewhere in the genome. We find that 1,554 out of 1,719 changes in HARs (90.4% [88.89%,91.73%]) have the derived state in Denisova, which remains significantly higher than for the 1,532,287 out of 1,753,121 (87.40% [87.35%,87.45%] sites genome-wide that have the derived state in Denisova.

Taken together, these results support the hypothesis that changes in the HARs tend to predate the Denisova-human split slightly more than expected and that differences caused by biased gene conversion tend to be evolutionarily older[1]. Nevertheless, we also identify 104 positions (98 SNPs and 6 indels) where the Denisova individual is ancestral while present-day humans are consistent with being fixed for the derived allele. These are likely to represent very recent changes that have occurred since the Denisova-modern human split, and they merit further study.

Neandertal-Denisova concordance

Of the 10,535,445 SNDs inferred to have occurred on the lineage leading to the human reference genome *hg18*, 4,267,431 (40.51%) positions are covered in the Denisova data while 3,202,190 (30.39%) are covered in Neandertal[1]. The expected overlap from random sampling is 12.31% (40.51% times 30.39%), and thus the actual overlap of 15.61% is higher than expected, which we hypothesize may be due to higher coverage of GC-rich sequences in both data sets. The overlap of indels of 6.05% is also higher than expected from random sampling (3.16%). The Neandertal and the Denisova specimens carry the same assigned state at SNDs in 87.91% of the ancestral positions (Neandertal = Ancestral (A) | Denisova = A) and 97.69% of the derived positions (Neandertal = Derived (D) | Denisova = D). Similarly for indels, p(Neandertal = A | Denisova = A) = 87.64% and p(Neandertal = D | Denisova = D) = 98.60%. Table S4.3 provides details.

Table S4.3: Concordance between Denisova and Neandertal

Single nucleotide changes		
Count	**Denisova**	**Neandertal**
190,836	A	A
32,785	A	D
339,171	A	M
227	A	N
26,245	D	A
1,389,396	D	D
2,279,365	D	M
1,528	D	N
164,555	M	A
1,389,996	M	D
3,204	M	N
534	N	A
818	N	D
1,517	N	M
12	N	N
58	P	A
807	P	D
2,943	P	M
1,189	P	N

Insertion/deletion changes		
Count	**Denisova**	**Neandertal**
2,532	A	A
365	A	D
9,937	A	M
12	A	N
357	D	A
25,642	D	D
65,957	D	M
29	D	N
5,409	M	A
34,218	M	D
382	M	N
4	N	A
23	N	D
458	N	M
56	N	N

Note: A = ancestral, D = derived, M = missing, N = neither chimp nor human state, P = polymorphic in Denisova. Disagreements are highlighted.

Positions where the Neandertal and Denisova data disagree on the ancestral state may be of special interest (32,785 Denisova = A & Neandertal = D; 26,245 Denisova = D & Neandertal = A). These sites show a derived state in the human reference sequence, as well as the derived state in either Denisova or Neandertal but not in both, and may thus reflect standing variation at the time of the separation of the modern human and the Neandertal/Denisova ancestors. Of the 59,030 single nucleotide differences where Neandertal and Denisova disagree, 61 overlap with the coding regions of 63 Ensembl annotated genes (49 of which belong to the CCDS set) and result in a non-synonymous change in the amino acid sequence. Three genes have two such sites:

(1) *RPTN (Repetin),* an matrix protein that is expressed in the epidermis and particularly strongly in eccrine sweat glands, the inner sheaths of hair roots and the filiform papilli of the tongue[18]. *Repetin* was described by Green et al.[1] as one of five genes with two amino acid altering substitutions that have become fixed among humans since the divergence from Neandertals[1]. The same positions are observed in the derived state, however, in the Denisova specimen.

(2) *RGS14 (regulator of G-protein signaling 14),* an integrator of G protein and MAPKinase (Ras/Raf) signaling[19], carries two non-synonymous substitutions that are fixed in present-day humans, ancestral in the Denisova individual, and derived in the Neandertals.

(3) *ZN333 (Zinc finger protein 333)* carries two non-synonymous substitutions that are fixed in present-day humans, ancestral in Denisova, and derived in the Neandertals. *ZN333* is the only known gene containing two KRAB domains, which function in transcriptional repression[20]. In addition to the two coding positions there are several other positions located in the introns, which are also ancestral in the Denisova individual and derived in the Neandertals.

References for SI 4

1. Green, R.E. et al., A draft sequence of the Neandertal genome. *Science* **328**, 710 (2010).
2. Li, W.H., Wu, C.I. and Luo, C.C., A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* **2**, 150 (1985).
3. Oancea, E. et al., TRPM1 forms ion channels associated with melanin content in melanocytes. *Sci Signal* **2**, ra21 (2009).
4. Orii, K.E. et al., Genomic and mutational analysis of the mitochondrial trifunctional protein beta-subunit (HADHB) gene in patients with trifunctional protein deficiency. *Hum Mol Genet* **6**, 1215 (1997).
5. Faisst, A.M., Alvarez-Bolado, G., Treichel, D. and Gruss, P. Rotatin is a novel gene required for axial rotation and left-right specification in mouse embryos. *Mech Dev* **113**, 15 (2002).
6. Koller, G. et al., ADAM8/MS2/CD156, an emerging drug target in the treatment of inflammatory and invasive pathologies. *Curr Pharm Des* **15**, 2272 (2009).
7. Park, J.A. and Kim, K.C., Expression patterns of PRDM10 during mouse embryonic development. *BMB Rep* **43**, 29 (2010).
8. Siegel, D.A., Huang, M.K. and Becker, S.F., Ectopic dendrite initiation: CNS pathogenesis as a model of CNS development. *Int J Dev Neurosci* **20**, 373 (2002).
9. Bird, C.P. et al., Fast-evolving noncoding sequences in the human genome. *Genome Biol* **8**, R118 (2007).
10. Bush, E.C. and Lahn, B.T., A genome-wide screen for noncoding elements important in primate evolution. *BMC Evol Biol* **8**, 17 (2008).
11. Pollard, K.S. et al., Forces shaping the fastest evolving regions in the human genome. *PLoS Genet* **2**, e168 (2006).

12. Prabhakar, S., Noonan, J.P., Paabo, S., and Rubin, E.M., Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**, 786 (2006).
13. Duret, L. and Galtier, N., Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**, 285 (2009).
14. Duret, L. and Galtier, N., Comment on "Human-specific gain of function in a developmental enhancer". *Science* **323**, 714; author reply 714 (2009).
15. Galtier, N. and Duret, L., Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet* **23**, 273 (2007).
16. Noonan, J.P., Regulatory DNAs and the evolution of human development. *Curr Opin Genet Dev* **19**, 557 (2009).
17. Prabhakar, S. et al., Human-specific gain of function in a developmental enhancer. *Science* **321**, 1346 (2008).
18. Huber, M. et al., Isolation and characterization of human Repetin, a member of the Fused gene family of the epidermal differentiation complex. *J Invest Dermatol* **124**, 998 (2005).
19. Shu, F.J., Ramineni, S. and Hepler, J.R., RGS14 is a multifunctional scaffold that integrates G protein and Ras/Raf MAPkinase signalling pathways. *Cell Signal* **22**, 366 (2010).
20. Jing, Z., Liu, Y., Dong, M., Hu, S. and Huang, S., Identification of the DNA binding element of the human ZNF333 protein. *J Biochem Mol Biol* **37**, 663 (2004).

Supplementary Information 5

Segmental duplication analysis of the Denisova genome.

Can Alkan, Tomas Marques-Bonet and Evan E. Eichler*

* To whom correspondence should be addressed (eee@gs.washington.edu)

Methods

We used the whole-genome shotgun sequence detection (WSSD) method to identify regions of >20 kb in length with a significant excess of read depth within 5 kb overlapping windows[1,2]. To apply the WSSD method to the Denisova data, we used the raw reads from the alignments described in SI 1. We discarded any read shorter than 36 bp (n=15,259,082 reads, spanning 112 Mb), and performed WSSD analysis using the remaining 128.5 million "Illuminized" reads from the Denisova genome (computationally fragmented into 36 bp units). This library showed a good correlation with a training set of BAC clones with known copy number in humans (Figure S5.1).

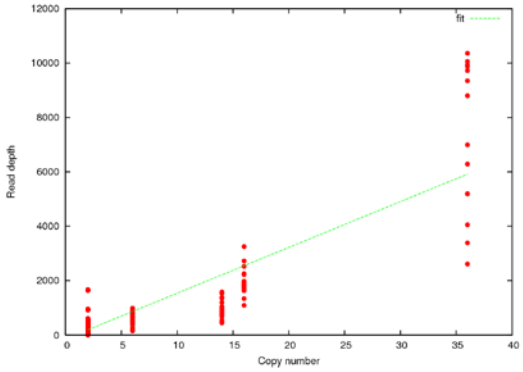


Figure S5.1: Correlation of Denisova read-depth with BAC sequences of known human copy number (R^2=0.74).

To search for segmental duplications, we mapped a total of 46.5 million reads to a repeat masked version of the human genome (NCBI build 35 / *hg17*) using mrFAST with an edit distance of 2. mrFAST is an algorithm that tracks all read map locations allowing read depth to be accurately correlated with copy number in duplicated regions[2].

The key innovation in our WSSD analysis compared with previous reports is to include read depth statistics for a larger number of control regions, allowing us to build better model to correct for GC bias. We defined control regions as intervals where copy number has been fixed at the diploid state (n=2) over the last 25 million years, based on comparison to known segmental duplications in humans, great apes, and Old World monkeys[3], as well as human structural variants from the Database of Genomic

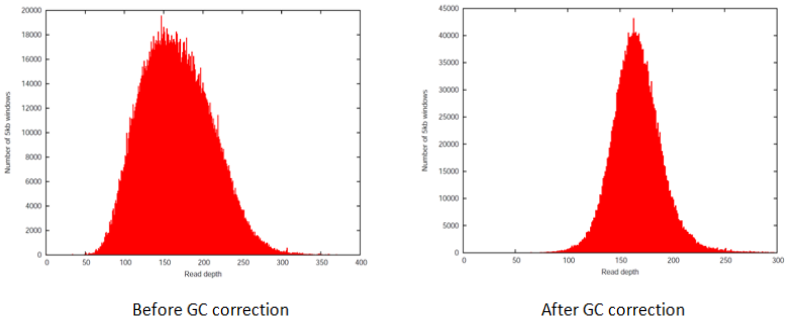


Figure S5.2: Read depth in control regions before and after applying a GC correction. The Denisova data shows a Poisson-like read distribution in 1.56 Gb of control regions with a copy number of 2. Correcting for GC-bias on a large set of diploid control regions tightens the distribution providing a better fit to a Gaussian.

Variants. We applied Loess smoothing to correct GC bias both genome-wide and in control regions, and observed a marked improvement in our ability to quantify read depth (Figure S5.2).

The landscape of segmental duplications comparing archaic and present-day humans

We generated duplication maps for five samples: NA18507 (West African)[4], YH (Han Chinese)[5], Neandertal[6], Denisova and Clint (chimpanzee)[7] (Table S5.1). To identify a duplication in each of these individuals using WSSD, we identified loci where at least 6 out of 7 consecutive overlapping windows of 5 kb each show a read depth that is more than 4 standard deviations greater than the mean. At loci that we identified as segmental duplications by this method, we estimated copy number from the read depth of non-overlapping 1 kb windows. Summary statistics for the different genomes are presented in Table S5.1.

Table S5.1: Mapping statistics of the analyzed samples

Sample	# reads (36 bp each)	Mapped to repeat masked *hg17*	Mean RD*	St. Dev. RD*	Duplicated bp (>10 kb)*	Duplicated bp (>20 kb)*
Denisova	128,513,214	46,570,304	167	24	113,864,467	102,124,360
Neandertal	65,393,768	20,686,477	78	13	107,074,978	98,203,746
Chimpanzee	398,182,534	n.d.	713	196	83,249,454	77,305,367
NA18507	1,776,928,308	556,713,986	2,233	236	109,705,947	100,793,811
NA18507-1.6× [+]	128,515,000	41,953,944	168	24	106,369,935	98,553,892
YH[^]	1,315,249,404	375,234,167	1,489	186	113,417,959	102,743,831
YH-1.6× [+^]	132,187,500	34,993,047	139	20	105,261,458	97,397,936

* GC Corrected, autosomal. [+] Resampled at 1.6×. [^] YH reads are 35 bp. RD: read depth in 5 kb windows.

Overall, we find that the Denisova duplication map is comparable in content and copy number to the other hominins, and is more similar (after copy number correction to account for the human assembly bias) to NA18507. Unsurprisingly, we find that the Denisova genome shares more of its segmental duplications with present-day humans and Neandertals than with chimpanzee. Specifically, it shares about 30% (20.4/67.4) for duplications of >20 kb (Figure S5.3).

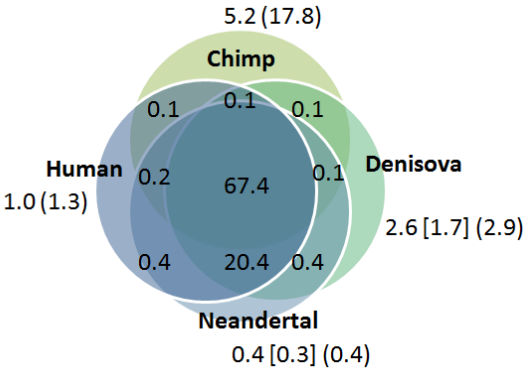


Figure S5.3: Venn diagram of segmental duplications of a present-day human (NA18507), Neandertal, Denisova and chimpanzee. We restrict to duplications of size >20 kb, as sensitivity increases with larger regions[2]. Irrespective of the criteria, the duplication pattern in Denisova shows more similarity to that of present-day humans than the other hominins. In square brackets, duplications with an excess of paralogous variants are reported. In parenthesis, we give copy-number corrected values. All values in Mb.

Denisovans appear to have more population-specific duplications than other hominins

We compared the duplication map for Denisova to that of Neandertals and present-day humans, which we represented by a Yoruba West African NA18507. We observe an increase of duplications seen only in Denisova (1.7 Mb) compared with private duplications in present-day and Neandertals (approximately 2-4 fold more duplications) (Table S5.2). As expected, all three hominins have far fewer private duplications than chimpanzee (5.2 Mb). The proportion of

duplications shared between any two of the three hominin samples but not the third is similar for three pairs of two hominins, and hence in what follows we focus on the more surprising observation of a high degree of Denisova-specific segmental duplications.

Table S5.2: Intersection of duplication maps in four sample for >20 kb duplications

No. samples with duplication	Duplicated	Not duplicated	Length (bp)	No.
One	NA18507	Neandertal, Denisova, Chimpanzee	1,041,510	29
	Neandertal	NA18507, Denisova, Chimpanzee	424,544	17
	Denisova	NA18507, Neandertal, Chimpanzee	2,549,524	74
	Chimpanzee	NA18507, Neandertal, Denisova	5,192,892	81
Two	NA18507, Neandertal	Denisova, Chimpanzee	424,027	12
	NA18507, Denisova	Neandertal, Chimpanzee	548,032	20
	NA18507, Chimpanzee	Neandertal, Denisova	149,659	5
	Neandertal, Denisova	NA18507, Chimpanzee	438,555	17
	Neandertal, Chimpanzee	NA18507, Denisova	0	0
	Denisova, Chimpanzee	NA18507, Neandertal	141,936	4
Three	Neandertal, Denisova, Chimp	NA18507	147,313	6
	NA18507, Denisova, Chimp	Neandertal	183,090	6
	NA18507, Neandertal, Chimp	Denisova	0	0
	NA18507, Neandertal, Denisova	Chimpanzee	20,410,465	399
Four	All four hominins	none	67,363,448	780

We were concerned that the excess of private duplications detected in Denisovans might be an artifact of the different sequencing coverage in the three hominins. To empirically assess the effect of sequencing coverage, we analyzed two present-day humans for which we had high coverage data—NA18507 (Yoruba West African)[4] and YH (Han Chinese)[5]—and compared the duplication maps that we obtained from analysis of all the data to maps from 1.6× coverage (in the range of our coverage of the archaic hominins). Table S5.3 shows that the lower coverage results in missing 4-7% of duplications detected at higher coverage. Moreover, 1-2% of loci are reassigned between the categories of shared or private. Thus, lower coverage reduces sensitivity and specificity for detecting duplications.

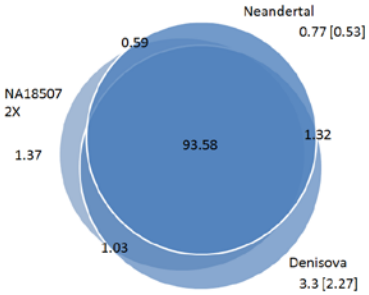


Figure S5.4: Segmental duplications in Neandertal, Denisova, and a present-day human resampled at 1.6× (NA18507). We restrict to >20 kb duplications, and indicate duplications with an excess of paralogous variants in square brackets. All values in Mb.

Table S5.3: Effect of coverage on duplication map length

	High coverage		1.6×	
Duplication map	**ALL**	**> 20 kb**	**ALL**	**> 20 kb**
NA18507 only	4,655,531	1,817,981	6,639,453	3,031,158
YH only	9,153,744	4,411,046	5,989,439	2,450,077
NA18507 & YH	121,159,623	112,158,610	114,330,037	106,673,064

Motivated by the observation that coverage impacts our ability to detect segmental duplications, we compared the archaic hominin duplication maps to that obtained by 1.6× subsampling of Yoruba individual NA18507. Table S5.4 and Figure S5.4 show that even after this reanalysis, we continue to observe that the Denisova genome harbors an excess (~2-3 fold) of individual-specific duplications compared to present-day humans or Neandertals with similar coverage.

To test whether the duplications that are inferred to be private to Denisova have the characteristics expected from true duplications, we examined their sequence divergence to present-day humans. If they are true paralogs, we expect the sequence divergence to be elevated above the genome average. Indeed, ~70% (1.7 Mb / 2.5 Mb) of the Denisova-specific segmental duplications are unusually diverged between Denisovans and present-day humans (Figure S5.5).

Table S5.4: Archaic vs. a 1.6× resampled human

Duplication map	ALL	> 20 kb
Private to NA18507 (1.6×)	3,458,360	3,026,874
Private to Neandertals	2,834,356	2,519,473
Private to Denisovans	8,636,199	8,111,884
NA18507 (1.6×) & Neandertal	1,486,366	1,221,765
NA18507 (1.6×) ^ Denisova	2,473,715	2,105,322
Neandertal and Denisova	3,802,661	3,256,390
All three hominins	98,949,979	98,890,398

Interestingly, 20% of the Denisova-specific duplications map to the *HLA* cluster on chromosome 6, which contains 9 loci of size >20 kb that have unusually high divergence (Figure S5.6).

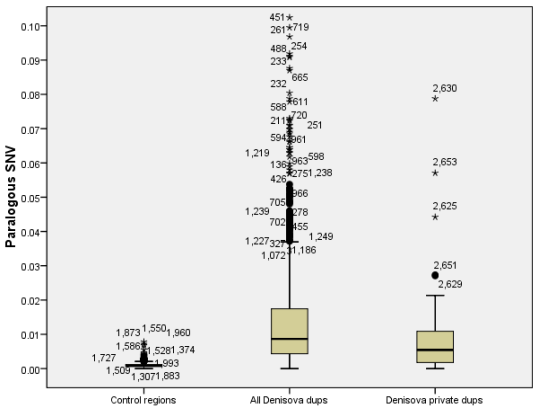


Figure S5.5: Paralogous single nucleotide diversity between Denisova and human. We compare control diploid regions, predicted Denisova duplications, and duplications specific to Denisova.

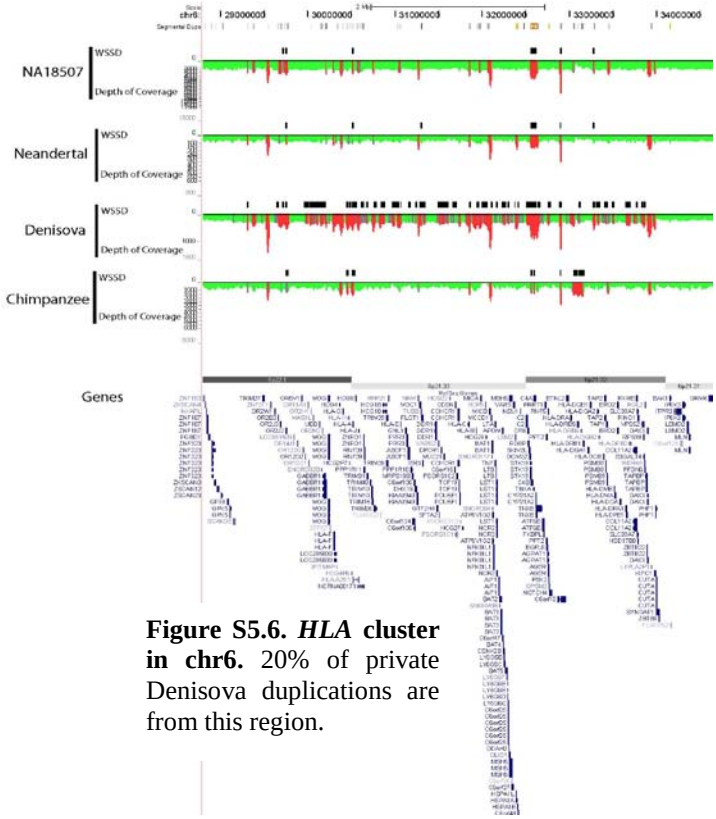


Figure S5.6. *HLA* cluster in chr6. 20% of private Denisova duplications are from this region.

Functional analysis of duplications that are specific to Denisovans

After further excluding known segments duplications that were previously been detected in analysis of 4 human genomes (JDW[8], NA18507[4], YH[5] and JCV[9]), we were left with 31 genic regions of size >20 kb (0.9 Mb of sequence, 1.5 Mb copy number corrected) that showed signatures of increased read-depth and sequence diversity when compared to other hominins (Table S5.5). Many of these duplications overlap genes associated with immune response and environmental interaction, and hence may be worth further exploration.

Table S5.5: Duplicated regions in the Denisova that have not previously been identified.

Gene name	Description	Complete/ Partial	refseq ID (NM#)	Chr	Start	End	Dupli-cation size	Deni-sova copies	Den.-hg*19* diver-gence
B4GALNT3	beta	Partial	173593	12	464000	487646	23646	1.70	0.0037
C18orf22	hypothetical protein LOC79863	Partial	24805	18	75895906	75920076	24170	2.65	0.0052
C2orf62	hypothetical protein LOC375307	Partial	198559	2	219032000	219055111	23111	2.63	0.0049
C6orf15	STG protein	Partial	14070	6	31164000	31187381	23381	3.13	0.0055
CACNA2D4	voltage gated Ca channel α(2) delta4	Partial	172364	12	1865000	1891448	26448	2.10	0.0058
CHST6	carbohydrate (Nacetylglucos. 6O)	Partial	21615	16	74064000	74086000	22000	2.72	0.0037
DEFA4	defensin, alpha 4 preproprotein	Complete	1925	8	6771118	6791445	20327	3.14	0.0066
DEFA5	defensin, alpha 5 preproprotein	Complete	21010	8	6871195	6902000	30805	2.76	0.0065
EIF3F	eukaryotic translation init. fact. 3	Complete	3754	11	7964000	7984316	20316	2.97	0.0061
HCG9	HLA complex group 9	Complete	5844	6	30041740	30069000	27260	3.10	0.0064
HCP5	HLA complex P5	Complete	6674	6	31492343	31559000	66657	2.88	0.0065
HLA	DQA2 MHC, class II, DQ	Complete	20056	6	32814000	32848535	34535	3.31	0.0061
HLA	DPB1MHC, class II, DP	Complete	2121	6	33148000	33205593	57593	3.22	0.0064
HLA	DPA1MHC, class II, DP	Partial	33554	6	33148000	33205593	57593	3.22	0.0064
LOC136242	hypothetical protein LOC136242	Partial	1008270	7	140994000	141017107	23107	3.03	0.0047
OR11A1	olf. receptor, family 11, subfamily A	Complete	13937	6	29465000	29508240	43240	2.84	0.0054
OR12D2	olf. receptor, family 12, subfamily D	Complete	13936	6	29465000	29508240	43240	2.84	0.0054
PNKD	myofibrillogenesis regulator 1 iso. 1	Partial	15488	2	219032000	219055111	23111	2.63	0.0049
TNFRSF10C	TNF receptor superfamily	Partial	3841	8	22983000	23025000	42000	3.13	0.0071
TNFRSF10D	TNF receptor superfamily	Complete	3840	8	23049000	23085506	36506	3.95	0.0126
TRIM26	tripartite motifcontaining 26	Partial	3449	6	30260410	30286136	25726	3.19	0.0033

Note: We list duplications that were not identified in human genomes (JDW, YH, NA18507 and JCV), Neandertal, chimpanzee, or the Database of Genomic Variants. Predicted diploid copy number and divergence are also included.

We were particularly interested in two biomedically relevant loci where the duplication architecture in Denisovans appears to be more similar to chimpanzee than to human.

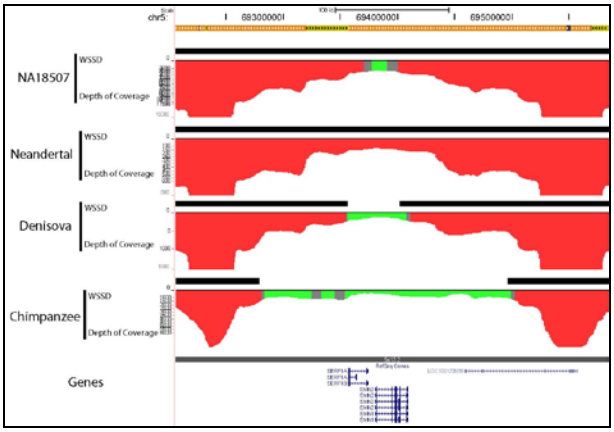


Figure S5.7: *SMA* locus on chromosome 5. Denisova and chimpanzee lack the duplication seen in both present-day humans and Neandertals.

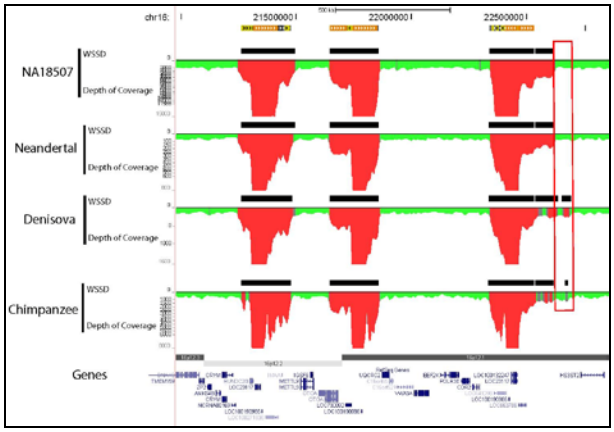


Figure S5.8: 16p12.1 locus. Denisova has an architecture more similar to chimpanzee in this region that is copy number variable in humans (right box).

One of the two regions region, which is located on chromosome 5, contains the *SMN2* gene whose copy number is associated with the severity of spinal muscular atrophy (Figure S5.7). This gene has expanded mainly in the human lineage, with all other non human-primates harboring mostly a single copy (some degree of polymorphism is known in chimpanzee[10]). We

find that Neandertals and present-day humans both share the *SMN2* duplication, while Denisova and chimpanzee both have a single copy (*SMN*). This suggests that the expansion of *SMN2* may have been polymorphic in the ancestral population of modern and archaic hominins.

We also identified a complex copy-number 1.1 Mb variable region on chromosome 16p12.1 whose duplication architecture is similar between chimpanzee and Denisova when compared to present-day human and Neandertals (Figure S5.8). Rearrangement of the region within humans has been associated with cognitive disability and neuropsychiatric disorders as well as rapid evolutionary turnover within the hominin lineage[11]. We did not find evidence for any Neandertal/chimpanzee shared duplication at this locus.

References for SI 5

1. Bailey, J.A., et al., Recent segmental duplications in the human genome. *Science* **297**, 1003 (2002).
2. Alkan, C. et al., Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics* **41**, 1061 (2009).
3. Marques-Bonet, T., et al. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877 (2009).
4. Bentley, D.R. et al., Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53 (2008).
5. Wang, J. et al., The diploid genome sequence of an Asian individual. *Nature* **456**, 60 (2008).
6. Green, R.E. et al., A draft sequence of the Neandertal genome". *Science* **328**, 710 (2010).
7. Mikkelsen, T.S. et al., Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69 (2005).
8. Wheeler, D.A. et al., The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872 (2008).
9. Levy, S. et al., The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254 (2007)
10. Rochette, C.F., Gilbert, N. and Simard, L.R., SMN gene duplication and the emergence of the SMN2 gene occurred in distinct hominids: SMN2 is unique to Homo sapiens. *Hum Genet* **108,** 255 (2001)**.**
11. Girirajan, S. et al., A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet* **42,** 203 (2010).

Supplementary Information 6

Denisovans and Neandertals are sister groups.

Nick Patterson*, Richard E. Green and David Reich

* To whom correspondence should be addressed (nickp@broadinstitute.org)

In this section, we present evidence that Neandertals and Denisovans are sister groups, by which we mean that they are more closely related to each other than either is to present-day humans. To do this, we estimate divergence between all pairs of hominins. We cannot use the procedure of SI 2 for this purpose, since in that method, the genetic divergence for each sample is always calculated relative to the human reference genome, in a way that relies on the assumption that the human reference has a negligible rate of error. Since this procedure is limited to comparisons of genomes to the human reference sequence, SI 2 cannot compare pairs of sequences both of which have substantial rates of error.

Data filtering procedures applied to this and subsequent population genetic analyses

The key population genetic analyses in this study (SI 6, SI 7, SI 8, SI 10 and SI 11) are based on mapping sequencing reads from modern and ancient genomes to the chimpanzee reference sequence (*pantro2*) to avoid biases toward one present-day human group more than another. Here we describe the filters we applied to these data after the mapping to chimpanzee, which are similar to those in our previous reported population genetic analyses of the Neandertal genome[1].

Filtering out reads with potential mapping problems

Each read that we analyze has a mapping quality score (MAPQ) that is generated by either the ANFO or BWA software and that aims to reflect the confidence of its mapping to the chimpanzee genome (SI 1). Based on empirical exploration of the usefulness of these scores, we only use reads with MAPQ values of at least 90 for Neandertal (ANFO mapping), 37 for Denisova (BWA), and 60 for present-day humans (BWA). We also reject reads if alignment to the chimpanzee results in any insertion/deletion difference.

Filtering out nucleotides of low reliability

(a) We do not use nucleotides for which there is no valid nucleotide call for chimpanzee.
(b) For Neandertals, we do not use nucleotides within 5 nucleotides of either end of the reads, because of the elevated rate of ancient DNA degradation errors that we empirically observe[1].
(c) For Denisova, we do not use nucleotides within 1 nucleotide of either end of the read.
(d) For both Neandertals and Denisova, we do not use nucleotides with sequence quality <40.
(e) For present-day humans, we do not use nucleotides with sequence quality $<T_{ij}$, where T_{ij} is a threshold chosen such that half of nucleotides generated from individual i and of allele class j $\{j$ = A, C, G, T$\}$ are less than this value. For nucleotides that have exactly a quality score of T_{ij}, we randomly choose reads to eliminate such that exactly half the reads are dropped.
(f) For the Papuan1 individual from ref. 1, the sequencer had a high error rate at position 34 (41 on the reverse strand). We thus excluded data from position 34 for this individual.

Filtering out CpG dinucleotides and deamination-induced nucleotide misincorporations

(a) We filter out all substitutions of the class that commonly occur at CpG dinucleotides, since recurrent mutation is more likely at these sites, complicating analysis. Specifically, we filter

out sites that are C/T polymorphisms in hominins with the next chimpanzee base being a G, or that are A/G polymorphisms in hominins with the next chimpanzee base being a C.

(b) We filter out transition substitutions (A/G or C/T) from all analyses because of the high rate of ancient DNA degradation at such sites in our Neandertal data (and to a much lesser extent in our Denisova data). For the analysis of the CEPH-Human Genome Diversity Panel genotyping data, however, we do include transitions, because we are analyzing SNPs that are already known to be valid polymorphisms, and at sites that are already known to be real SNP the transition error rate is not expected to have a substantial influence on results.

Filtering out triallelic sites.

Many of our analyses are based on the assumption that at any given site in the genome, there has been at most one mutation in the ancestry of our analyzed samples since humans and chimpanzees diverged. A small proportion of sites, however, have three or more observed alleles, which cannot be explained by a single historical mutation (instead, the data must reflect at least two mutations or sequencing errors). To process such sites, we choose the "ancestral" allele as the one matching chimpanzee, the "derived" allele as the most common allelic class that does not match chimpanzee (counting all reads in all samples independently), and then discard reads that do not match either the "ancestral" or "derived" type. We do not use data from the site at all if either (i) there is a tie in the number of reads supporting two candidates for the derived allele, or (ii) at least 5 reads across samples do not match the "ancestral" or "derived" type.

Estimates of genetic divergence between hominins as a fraction of human-chimpanzee

We consider all positions in the genome where we have at least one high quality sequencing read representing each of 7 hominins (French, Han, Papuan, San, Yoruba, Neandertal, Denisova) as well as a valid base in the chimpanzee genome. For each site, we then choose a read at random to represent each individual. (We treat the Vindija Neandertal data as a single individual, even though in fact it is from a pool of three closely related individuals.) We then count the total number of transversion substitutions between all possible pairs of samples (Table S6.1).

Table S6.1: Genetic divergence for all pairs of samples uncorrected for sequencing error

	Han	Papuan	San	Yoruba	Neandertal	Denisova	Chimp
French	22633	22948	22373	22805	25372	22138	101714
Han		23795	23596	24026	26542	23332	102939
Papuan			23801	24271	26562	23160	102894
San				22042	23832	20445	100000
Yoruba					25136	21748	101328
Neandertal						17963	100077
Denisova							96501

Note: All numbers are normalized such that San-Chimpanzee divergence is 100,000.

Much of this apparent divergence is caused by sequencing error. Suppose that the probability of a sequencing error for hominin i is $e(Q_i)$. Then, the probability of an error contributing to the observed divergence $R(Q_i,Q_j)$ with a second sample j is approximately $e(Q_i)+e(Q_j)$, assuming independence and small error rates per nucleotide. We thus seek correction factors $C(Q_i)$ giving divergences: $D(Q_i,Q_j) = R(Q_i,Q_j)-C(Q_i)-C(Q_j)$. To implement this idea, we assume that there are no sequencing errors in chimpanzee, that all hominins have the same true divergence A to chimpanzee due to a constant rate molecular clock, and that San-Yoruba divergence is 9.13%

that of human-chimpanzee (based on the comparisons of San to the African parts of the human genome reference sequence in SI2, Table S2.3). Thus we seek $C(Q_i)$ such that:

$$D(Q_i,Q_j) = R(Q_i,Q_j) - C(Q_j) - C(Q_j) \quad \text{(S6.1)}$$
$$C(Chimpanzee) = 0 \quad \text{(S6.2)}$$
$$D(Q_i,Chimpanzee) = A \quad \text{(S6.3)}$$
$$0.0913 \times A = D(San,Yoruba) \quad \text{(S6.4)}$$

Algebraic manipulation leads to several expressions that we can use to compute $D(Q_i,Q_j)$ for any pair of samples correcting for sequencing error. From Equations S8.1, S8.2 and S8.3:

$$\begin{aligned} \boldsymbol{D}(\boldsymbol{Q}_i,\boldsymbol{Chimp}) &= \boldsymbol{R}(\boldsymbol{Q}_i,\boldsymbol{Chimp}) - \boldsymbol{C}(\boldsymbol{Q}_i) - \boldsymbol{C}(\boldsymbol{Chimpanzee}) \\ &\Rightarrow \boldsymbol{A} = \boldsymbol{R}(\boldsymbol{Q}_i,\boldsymbol{Chimp}) - \boldsymbol{C}(\boldsymbol{Q}_i) \\ &\Rightarrow \boldsymbol{C}(\boldsymbol{Q}_i) = \boldsymbol{R}(\boldsymbol{Q}_i,\boldsymbol{Chimp}) - \boldsymbol{A} \end{aligned} \quad \text{(S6.5)}$$

From Equations S8.1, S8.4 and S8.5:

$$\begin{aligned} 0.0913 \times \boldsymbol{A} &= \boldsymbol{D}(\boldsymbol{San},\boldsymbol{Yoruba}) = \boldsymbol{R}(\boldsymbol{San},\boldsymbol{Yoruba}) - \boldsymbol{C}(\boldsymbol{San}) - \boldsymbol{C}(\boldsymbol{Yoruba}) \\ &= \boldsymbol{R}(\boldsymbol{San},\boldsymbol{Yoruba}) - \boldsymbol{R}(\boldsymbol{San},\boldsymbol{Chimp}) - \boldsymbol{R}(\boldsymbol{Yoruba},\boldsymbol{Chimp}) + 2\boldsymbol{A} \\ &\Rightarrow \boldsymbol{A} = [\boldsymbol{R}(\boldsymbol{San},\boldsymbol{Chimp}) + \boldsymbol{R}(\boldsymbol{Yoruba},\boldsymbol{Chimp}) - \boldsymbol{R}(\boldsymbol{San},\boldsymbol{Yoruba})]/1.9087 \end{aligned} \quad \text{(S6.6)}$$

With our estimate of A from Equation S6.6, we can use Equation S6.5 to solve for the $C(Q_i)$ sequencing error rates for each individual, and finally use Equation S6.1 to estimate the sequence divergence $D(Q_i,Q_j)$ between all pairs of samples.

Calendar dates for the divergence of genomes

For studying population history, we are interested in the average date when two genomes diverged. For comparisons involving archaic samples, this is not the same as the amount of time during which mutations have had time to accumulate (since mutations stopped occurring when the individuals died). Conveniently, however, our procedure above produces unbiased estimates of the calendar date even when archaic samples are analyzed, since it overestimates the genetic divergence on the branch specific to the archaic samples to exactly the extent that is necessary to make $D(Q_i,Q_j)$ an appropriate calendar date estimate. To understand this, we note that the $C(Q_i)$ terms that we are estimating can be viewed as adding a pseudo-distance to the leaf edges of the tree, which for present-day samples is just the sequencing error rate, but for archaic samples, is the sequencing error rate minus the amount by which the archaic branch has been shortened due to mutations having had less time to accumulate, thus compensating exactly for the shortening of the branches for the archaic samples. We note that this argument is valid for an arbitrary number of ancient hominins, and thus allows an estimate of the mean calendar time to the most recent common ancestor that does not require any knowledge of the age of the bones.

Table S6.2 presents the estimates of divergence for all pairs of genome obtained using this procedure, both as a fraction of the human-chimpanzee divergence date and scaled in years assuming 6.5 million years for human-chimpanzee divergence[2]. Standard errors for these estimates are obtained using a Block Jackknife[3,4], and are in general very small (around 0.0005-0.0007 per base pair, or about a hundred times smaller than our absolute estimates of divergence

time). We caution that these standard errors underestimate the true uncertainty. First, since we calibrate all estimates to San-Yoruba and human-chimpanzee genetic divergence, our standard errors do not reflect our statistical uncertainty about these quantities. Second, the systematic errors may be larger than the standard errors. For example, the Denisova and the Neandertal DNA samples were processed and aligned differently. Differences in alignment could affect the estimates of divergences, and we do not understand the extent of this potential bias.

Table S6.2: Estimated genetic divergence dates for each pair of hominin samples

*As a fraction of the human-chimpanzee divergence**

	Han	Papuan	San	Yoruba	Neandertal	Denisova
French	.0622	.0660	.0907	.0812	.1218	.1255
Han		.0620	.0907	.0811	.1212	.1251
Papuan			.0933	.0842	.1219	.1238
San				.0913	.1237	.1257
Yoruba					.1234	.1254
Neandertal						.0984

1000's of years assuming 6,500 for human-chimp

	Han	Papuan	San	Yoruba	Neandertal	Denisova
F	404	429	590	528	794	818
H		403	589	527	790	815
P			607	547	794	807
S				593	806	819
Y					804	817
N						644

Note: This table presents absolute dates of genetic divergence between a pair of samples dated relative to the present. For present-day samples, this is the same as the genetic divergence of the two samples, whereas for ancient samples which were interred tens of thousands of years ago, the numbers are somewhat larger than the actual separation time between samples.

* Standard errors from a Block Jackknife are in the range 0.0005-0.0007, corresponding to 3-5 thousand years.

The pairwise divergence results in Table S6.2 make it clear that Neandertal and Denisova are sister groups—more closely genetically related to each other on average than either is to modern humans—with estimated divergence from a common ancestor that has a mean calendar date of 644,000 years before present when calibrated by assuming human-chimpanzee genetic divergence of 6.5 million years. This is less than the divergence of both Neandertals and Denisovans to present-day Africans (average of 812,000 years, which is in reasonable agreement with the 825,000 estimate from ref. 1 given that we used a different analysis to obtain that estimate). Table S6.2 also shows that the San are an "outgroup" with about equal divergence to the other present-day humans (Han, French, Papuan and Yoruba) at an average of 595,000 years.

Estimated genetic divergences between Neandertals (Vindija individuals & Mezmaiskaya 1)

Most of our analyses of Neandertal genetic material have been concentrated on three bones from Vindija Cave, Croatia. The only other bone for which we have collected a substantial amount of data (56 Mb) is Mezmaiskaya 1 from Mezmaiskaya Cave, in the Northern Caucasus[1]. To understand how this bone relates to other archaic hominins in terms of its genetic divergence, we carried out the same divergence calculation to that above. To maximize the number of nucleotides available for this analysis, we no longer restrict to sites where we have data from the non-African present-day humans. Instead, we now consider all sites where we have least one sequencing read from each of San, Yoruba, Denisova, Vindija, and Mezmaiskaya 1. The results are presented in Table S6.3, and they agree well with Table S6.2 after taking into account the much larger standard errors due to the limited data set size.

Table S6.3 Estimated genetic divergence dates including Mezmaiskaya 1

1000's of years assuming 6,500 for human-chimp

	Yoruba	Denisova	Vindija	Mez. 1
San	593	839	814	810
Yoruba		832	828	827
Denisova			689	678
Vindija				140

±1 standard error

	Yoruba	Denisova	Vindija	Mez. 1
S	n/a	22	22	21
Y		20	21	20
D			27	27
V				33

Note: Standard errors are about five times larger than in Table S6.2, reflecting our limited data from Mezmaiskaya 1. No error is given for San-Yoruba since this quantity is used for calibration.

The estimated average divergence date between the Vindija and Mezmaiskaya 1 genomes is remarkably low, especially in light of the separation between the fossils both in geographical space and time. The divergence date corresponds to a best estimate of 140,000 ± 33,000 years assuming 6.5 million years for human-chimpanzee divergence. (Taking into account the fact that the Vindija and Mezmaiskaya bones were interred around forty to seventy thousands years ago, the actual average time to the common genomic ancestor at the time of interment was <100,000 years). By contrast, the divergence between pairs of present-day humans in Table S6.2 is 3-4 times greater. Thus, Neandertals across a wide geographic range harbored little heterozygosity compared with modern humans, in line with our previous analyses of mitochondrial DNA[5].

The low divergence between Mezmaiskaya 1 and Vindija also sheds light on the interpretation of the Mezmaiskaya 1 fossil itself. Some researchers have argued that Mezmaiskaya 1 fossil might not be a true Neandertal[6]. However, since Vindija and Mezmaiskaya 1 have an average genetic divergence 140,000 years ago, which is well after the full suite of Neandertal traits appear in the fossil record around 230,000 years ago[7], Mezmaiskaya 1 is likely to be a true Neandertal.

Table S6.4 Estimated genetic divergence dates including pairs of Vindija bones

1000's of years assuming 6,500 for human-chimp

	Yoruba	Vi33.16	Vi33.25	Vi33.26
San	593	860	866	860
Yoruba		851	858	848
Vi33.16			85	80
Vi33.25				93

±1 standard error

	Yoruba	Vi33.16	Vi33.25	Vi33.26
San	n/a	12	12	12
Yoruba		13	13	13
Vi33.16			20	21
Vi33.25				20

Note: No standard error is given for San-Yoruba since this quantity is used for calibration.

We finally computed the divergence between all pairs of bones from Vindija Cave (Table S6.4). The genetic divergence estimates averaging across all pairs of Vindija bones is estimated to be 86,000 years ago. Since the Vindija bones are all about 40,000 years old[1], this suggests that they may have been very closely related, with a best estimate of 46,000 years for the average time since the most recent common genetic ancestor. We caution that this estimate has high uncertainty especially as there is likely to be systematic error. We are restricting to sites that are

unusually highly covered in Vindija—with data from at least two individuals even though the average coverage is 1.3-fold—so we may be enriching for sites affected by alignment error.
Figure 1 presents a neighbor joining tree summarizing these estimates. It is important to recognize that it is merely an approximate representation of genetic distances and does not represent a "true" phylogeny. Indeed a main point of this paper and ref. 1 is that the relationship of present-day humans to Neandertals and Denisovans cannot be faithfully represented as a tree.

Comparisons of genetic divergence estimates from SI 2 and SI 6

To assess the robustness of the pairwise inferences of genetic divergence, we also compared them to the more direct estimates that we obtained in SI 2 by comparison to West African and European parts of the human reference sequence *hg19* under the assumption that the human reference genome has no errors. Table S6.5 shows that the divergence estimates of the two methods are generally concordant for pairs of present-day humans. The most substantial differences are for the comparison of Denisova and present-day humans, where this analysis gives an estimate of 12.5% that is substantially higher than the 11.7% from SI 2. One difference between the two methods is that in this Supplementary Note we are restricting to the subset of the genome where we have data from 7 hominins and chimpanzee—thus restricting to higher coverage segments—whereas in SI 2 we only require data from *hg19* and *pantro2*. Nevertheless, the two methods are in sufficient agreement that our main finding—that Denisovans and Neandertals are more closely related than either to present-day humans—is likely to be robust.

Table S6.5: Estimated genetic divergence dates comparing two methods

Sample 1	**Sample 2**	**Based on comparison to *hg19* (SI 2)**	**Based on pairwise analysis (SI 6)**
European	West African	8.3%	8.1%
Han	West African	8.5%	8.1%
Han	European	6.8%	6.2%
Papuan	West African	8.4%	8.4%
Papuan	European	7.0%	6.6%
San	European	9.2%	9.1%
Neandertal	West African	12.2%	12.3%
Neandertal	European	11.9%	12.2%
Denisova	West African	11.7%	12.5%
Denisova	European	11.8%	12.5%

Note: All estimates are given as a fraction of the genetic divergence between humans and chimpanzee.

References for SI 6

1. Green, R. E. et al., A draft sequence of the Neandertal genome. *Science* **328**, 710 (2010).
2. Goodman, M., The genomic record of Humankind's evolutionary roots. *Am J Hum Genet* **64**, 31 (1999).
3. Busing, F.M.T.A., Meijer, E. and van der Leeden, R., Delete-m jackknife for unequal m. *Statistics and Computing* **9**, 3 (1999).
4. Kunsch, H.K., The jackknife and the bootstrap for general stationary observations. *Ann Statist* **17**, 1217 (1989).
5. Briggs, A. W. et al., Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**, 318 (2009).
6. Hawks, J. and Wolpoff, M.H., Brief communication: paleoanthropology and the population genetics of ancient genes. *Am J Phys Anthropol* **114**, 269 (2001).
7. Hublin, J.-J., The origin of Neandertals. *Proc Natl Acad Sci USA* **106**, 16022 (2009).

Supplementary Note 7

Denisovans have a distinct history from Neandertals.

Nick Patterson, Richard E. Green and David Reich*

* To whom correspondence should be addressed (reich@genetics.med.harvard.edu)

This section analyzes the extent to which Denisovans and Neandertals have distinct population histories. We first learn about the population relationships among the present-day and ancient hominins that we sequenced, taking advantage of the power that comes from analyzing four sequences together instead of pairwise comparisons[1]. A key result is that Neandertals are more closely related than Denisovans to the archaic population that contributed genetic material to present-day non-Africans[2]. We also document the extensive loss of genetic diversity that occurred in the common ancestry of all the Neandertals for which we have data since they separated from the ancestors of Denisovans.

Population relationships among archaic and present-day hominins

In SI 6 we computed the genetic divergences between many pairs of hominins to show that the Vindija and Mezmaiskaya 1 Neandertals are much more closely related to each other on average than either is to Denisovans or present-day humans. However, we were not able to discern the relationships to the other Neandertals for which we have less data (Feldhofer 1 and El Sidron 1253, with only a couple of megabases of nuclear data each[2]). We also were not able to learn the relationship of Neandertals and Denisovans to the archaic population that contributed genetic material to the ancestors of all non-Africans[2].

We explored how Denisova, Vindija, Mezmaiskaya 1, Feldhofer 1, and El Sidron 1253 relate to each other by taking advantage of the parsimony-based technique of "quartet puzzling"[1], which has more power to discern relationships than pairwise analysis. The key idea is that when data from at least four individuals are compared, more than one phylogenetic relationship is possible. For any alignment of 3 hominin reads and chimpanzee in the order H_1-H_2-H_3-chimpanzee, there are three possible cluster patterns at sites where two of the hominin reads carry the derived (non-chimpanzee) allele B and one carries the ancestral allele A. The relative rates of these three classes provide information about the population relationships. We use the terminology BBAA, BABA and ABBA to denote which pair of samples carries the derived allele. If the samples are related according to the tree $(H_1,H_2)H_3$, we expect that BBAA sites will occur most often, and that BABA and ABBA sites will occur less often (but at a non-zero rate due to incomplete lineage sorting and/or migration[3]).

To implement the quartet puzzling idea, we examine all nucleotides for which we have at least one read passing sequence quality filters for each of the hominins we were analyzing. We then compute the expected number of sites with 2 copies of the derived and 1 copy of the ancestral allele, assuming that we randomly draw a single read to represent each individual. Since two copies of each allele are observed across the three hominins and the chimpanzee, the effect of sequencing error is expected to be negligible (SI 10). To quantify the relative rates of BBAA, BABA and ABBA sites, we denote the observed number of sites of each class as n_{BBAA}, n_{BABA} and n_{ABBA}, and define an "*E*-statistic" whose standard error we estimate by a Block Jackknife[4,5]:

$$E(H_1, H_2, H_3, chimpanzee) = \frac{n_{BBAA} - 0.5(n_{BABA} + n_{ABBA})}{n_{BBAA} + 0.5(n_{BABA} + n_{ABBA})} \quad \text{(S7.1)}$$

The "*E*-statistic" quantifies the excess of BBAA over the average of the other classes of sites, and must be at >0 to support a history in which the populations from which samples H_1 and H_2 are drawn from "sister groups" relative to the population from which sample H_3 is drawn. (By sister groups, we mean which pair of the three hominin samples is most closely related of the three possible pairs.) We interpret *E*-statistics that are more than *Z*=3 standard deviations greater than 0 as statistically significant evidence that for H_1 and H_2 are sister groups.

Table S7.1 presents results for selected sets of 3 hominins. The *E*-statistic analyses show that (i) all 3 Vindija Neandertals are equally closely related within the limits of our resolution, (ii) Denisova and all Neandertals are sister groups relative to present-day Africans (a pool of all reads from San, Yoruba and Mbuti), and (iii) Vindija, Feldhofer 1, Mezmaiskaya 1 and El Sidron 1253 form a clade relative to Denisova. Thus, all the Neandertal samples for which we have collected DNA sequence data fall into a "Neandertal" group relative to Denisova. This further strengthens the finding in SI 6, where we used divergence data to show that Mezmaiskaya 1 and Vindija are more closely related to each other genetically than either is to Denisova. With the more powerful quartet-puzzling approach presented in this note, we are now able to generalize the result to additional Neandertals for which we have less data.

Table S7.1: Quartet puzzling to discern the relationships of archaic and modern hominins

Samples (H_1, H_2, H_3)	n_{BBAA}	n_{BABA}	n_{ABBA}	E	Std. Err	Z-score for sister groups
Vi33.26,Vi33.16,Vi33.25	697	614	672	0.04	0.02	1.7
Denisova, All Vindija, African *	53412	24954	26517	0.35	0.01	**62.2**
Denisova, Mezmaiskaya 1, African	878	426	410	0.35	0.02	**18.0**
Denisova, El Sidron 1253, African	42	36	18	0.21	0.09	2.3
Denisova, Feldhofer 1, African	34	18	20	0.28	0.12	2.4
All Vindija, Mezmaiskaya 1, Denisova	924	80	81	0.84	0.01	**62.0**
All Vindija, El Sidron 1253, Denisova	49	9	4	0.76	0.06	**12.6**
All Vindija, Feldhofer 1, Denisova	44	5	8	0.73	0.07	**10.7**
All Vindija, Mezmaiskaya 1, African	1385	103	100	0.86	0.01	**81.2**
All Vindija, El Sidron 1253, African	73	6	3	0.89	0.03	**27.3**
All Vindija, Feldhofer 1, African	65	3	6	0.86	0.04	**22.1**

Notes: Number of counts of each class is rounded to the nearest integer. Values that are significant at more than Z = 3 standard deviations are highlighted. There is strong evidence for "H_1" and "H_2" forming sister groups in all rows except for the three Vindija samples, which are consistent with deriving from a single population so that no clear phylogeny is evident.

* The counts for n_{BBAA}, n_{BABA}, and n_{ABBA} reported in this row are somewhat larger than the numbers reported in the text (although the ratios are the same). This reflects the fact that here we use a pool of all Africans (instead of just Yoruba), and compute the expected probability of each substitution class (instead of randomly sampling a read).

An intriguing observation in Table S7.1 is that there are fewer BABA than ABBA sites in the {Denisova, Vindija, African, Chimpanzee} alignment, suggesting that present-day Africans share more derived alleles with Neandertals than with Denisovans (if this asymmetry was measured by a *D*-statistic, it would be highly statistically significant as discussed in SI 10). However, we do not interpret this as providing convincing evidence for another ancient gene

flow event, since as we emphasize in the text, it is important for samples H_1 and H_2 to be experimentally and computationally processed in the same way for a *D*-statistic measuring the symmetry of their relationship to a sample H_3 to be valid (the Neandertal and Denisova data sets were generated very differently). In SI 10, we examine this potential pitfall in detail, and find that the observed excess of ABBA over BABA sites is not stable when stratified by the number of reads covering the analyzed sites (Table S10.7). Importantly, however, the excess of BBAA sites over BABA or ABBA sites, leading to the positive *E*-statistic and the conclusion that Neandertals and Denisovans are sister groups, is stable regardless of read coverage.

Figure S7.1 presents the tree that emerges from these studies, illustrating how all the Neandertals analyzed to date are consistent with forming a group, with Denisovans more distantly related.

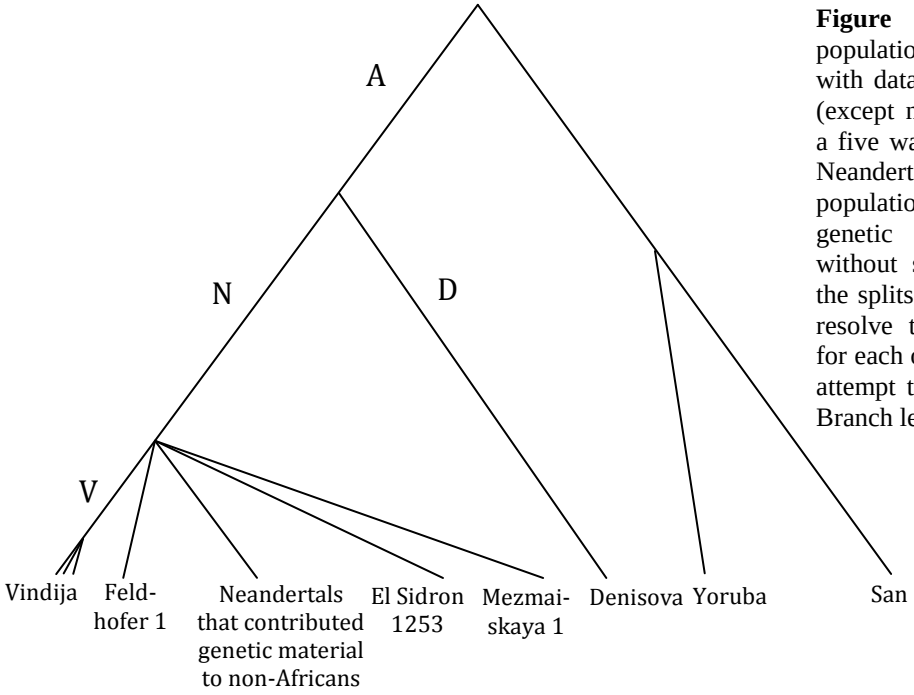


Figure S7.1: A schematic population tree that is consistent with data from diverse hominins (except non-Africans). We show a five way separation among the Neandertals bones and the population that contributed genetic material to Eurasians, without specifying the order of the splits, as we cannot currently resolve them. Labels are given for each of the lineages where we attempt to estimate genetic drift. Branch lengths are not to scale.

The archaic hominins who contributed genetic material to non-Africans were more closely related to Neandertals than to Denisovans

We previously demonstrated that Vindija Neandertals are more closely related to non-Africans than Africans[2]. A possible explanation is a history of gene flow from an archaic population into the ancestors of modern non-Africans around the time of the dispersal of modern humans out of Africa >45,000 years ago. If this explained our data, then we would predict that present-day non-Africans would not have the same relationship to Neandertals and Denisovans (they would be more closely related to the population that contributed gene flow). A second hypothesis is that substructure in the ancestral population of Neandertals and humans dating to several hundred thousand years ago explains the data[2]. If ancient substructure explains the data, then in the simplest scenario the fact that Denisovans and Neandertals are sister groups would predict that Denisovans would share derived alleles with non-Africans to the same extent as Neandertals.

The empirical data are consistent with the predictions of the gene flow hypothesis, while weakening the evidence for ancient substructure. Specifically, Table 1 in the main text shows that Neandertals are substantially closer to non-Africans (average D(*Eurasian, African,*

Neandertal, chimpanzee) = 5.0%) than is the case for Denisovans (average D(*Eurasian, African, Denisova, chimpanzee*) = 1.8%). This is unexpected if ancient substructure explains our observations and Denisovans and Neandertals descend from exactly the same ancestral population since their divergence from the ancestors of present-day Africans .

As a second line of evidence showing that the archaic population that contributed genetic material to all non-Africans was more closely related to Neandertals than to Denisovans, we re-analyzed the 13 regions that we previously identified as candidates for gene flow from Neandertals into the ancestors of present-day non-Africans, based on their harboring deeply diverged haplotypes that are present only in non-Africans[2] (Table S7.2). At each region, we identified tag SNPs that classified a sequencing read as matching the haplotype that is present only in out-of-Africa populations (OOA), or matching the cosmopolitan haplotype present in both non-Africans and Africans (COS). Neandertals match the OOA haplotype in 11 of the 13 regions, far exceeding the expectation in the absence of gene flow[2]. Denisovans match the OOA haplotype in 6 of 13 regions, which is a lower rate than is seen in Neandertals, although still higher than expected. An online table that presents the full list of 190 tag SNPs over the 13 regions, annotated by their allelic status in Neandertals and Denisovans, can be downloaded from http://bioinf.eva.mpg.de/download/DenisovaGenome/. We conclude that Denisovans are less closely related than Neandertals to the archaic group that contributed genes to non-Africans.

Table S7.2: Admixed haplotypes in non-Africans match Neandertals more than Denisovans

Chrom.	Start	End	VAM	VDM	VAN	VDN	DAM	DDM	DAN	DDN	Vindija	Denisova
1	168,110,000	168,220,000	6	10	1	0	3	10	2	0	OOA	OOA
1	223,760,000	223,910,000	1	4	0	0	2	3	0	0	OOA	OOA
4	171,180,000	171,280,000	119	2	0	0	0	0	2	1	OOA	COS
5	28,950,000	29,070,000	19	18	5	0	12	11	5	5	OOA	OOA
6	66,160,000	66,260,000	7	6	0	0	0	0	3	6	OOA	COS
9	32,940,000	33,040,000	8	13	0	0	0	0	12	7	OOA	COS
10	4,820,000	4,920,000	9	5	0	0	0	0	5	9	OOA	COS
10	38,000,000	38,160,000	5	9	1	0	4	8	3	0	OOA	OOA
10	69,630,000	69,740,000	3	2	0	0	3	1	0	0	OOA	OOA
15	45,250,000	45,350,000	5	7	1	0	5	6	3	0	OOA	OOA
17	35,500,000	35,600,000	0	3	1	0	0	2	3	0	OOA	COS
20	20,030,000	20,140,000	0	0	8	5	0	0	8	4	COS	COS
22	30,690,000	30,820,000	0	2	3	1	0	1	5	2	COS	COS

Notes: For each haplotype, we count the number of sites at which the Vindija and Denisova sequence matches the derived or ancestral alleles that tag the out-of-Africa (OOA) or cosmopolitan (COS) haplotypes. We use the following notation: VAM=Vindija ancestral allele matches OOA; VDM=Vindija derive allele matches OOA; VAN=Vindija ancestral allele does not match OOA; VDN=Vindija derived allele does not match OOA (similarly DAM, DDM, DAN, and DDN for the Denisova data). The last two columns provide a qualitative assessment of the haplotype inferred for Denisova and Vindija. Vindija matches 11 of the 13 whereas Denisova matches only 6 of the 13. We note that this table updates Table 5 ref. 2, which was based on a non-final version of the Neandertal data. The qualitative results agree with the previously published table, except that we now also have tag SNPs for a 13th region. This further strengthens the already strong signal of Neandertals being more closely related to non-Africans than to Africans.

The intensity of the population bottleneck in the shared history of Neandertals

In SI 6, we showed that the date of genetic divergence of Mezmaiskaya 1 and the Vindija Neandertals, averaged across the genome, is around 140,000 years ago. This is 3-4 times less

than the average divergence among pairs of present-day humans. Based on this, we argued that the Mezmaiskaya 1 and Vindija Neandertals likely descend from a common ancestral population that experienced an extreme bottleneck. From the quartet puzzling analysis, we have further shown that the population bottleneck occurred not just in the history of Vindija and Mezmaiskaya 1 Neandertals, but also Neandertals including ones from Spain and Germany.

To quantify the intensity of the bottleneck, we estimate the "genetic drift" along each of the labeled lineages in Figure S7.1, where for the purposes of the discussion below we define genetic drift as the probability that two alleles coalesce along a lineage. In the case of a population that has been of constant size N over the interval between population splits Δt, this is $1\text{-}e^{-\Delta t/2N}$ (SI 11). However, our definition of genetic drift is more general, as we do not need to assume that population sizes have been constant in time.

To estimate the amount of genetic drift that occurred in the history of Vindija Neandertals since they diverged from Denisovans—that is, the probability that two alleles from the Vindija population share a common ancestor more recently than the population divergence from Denisovans—we examine alignments of sequencing data from five individuals: any two Vindija bones (denoted V_1 and V_2), Denisova (D), a present-day African (H, a pool of Yoruba, San and Mbuti), and chimpanzee (C). From these five individuals, we examine all three possible subsets of 2 archaic hominins, 1 African and 1 chimpanzee, and restrict to polymorphic sites where 2 copies of the derived allele and 1 copy of the ancestral allele are observed. We use the following notation to denote the sum of the counts in the two rare classes (BABA and ABBA):

$\boldsymbol{n}_{(V_1V_2)H} = \boldsymbol{n}_{V_1H} + \boldsymbol{n}_{V_2H}$ = the sum of BABA and ABBA in a V_1-V_2-H-C alignment

$\boldsymbol{n}_{(V_1D)H} = \boldsymbol{n}_{V_1H} + \boldsymbol{n}_{DH}$ = the sum of BABA and ABBA in a V_1-D-H-C alignment

$\boldsymbol{n}_{(V_2D)H} = \boldsymbol{n}_{V_2H} + \boldsymbol{n}_{DH}$ = the sum of BABA and ABBA in a V_2-D-H-C alignment

In the absence of genotyping error, mapping error, or recurrent mutation, BABA and ABBA sites reflect incomplete lineage sorting. The rate at which such sites occur is expected to decrease in proportion to the amount of genetic drift that occurred in the history of the two archaic samples since they diverged from the ancestors of modern Africans. We denote:

K = Probability of observing a BABA or ABBA substitution at a nucleotide, conditional on two archaic samples coalescing prior to the root of the tree in Figure S7.1.

p_V = Probability of two Vindija lineages coalescing prior to the split from Mezmaiskaya 1 in Figure S7.1 (that is, not coalescing on the lineage labeled V).

p_N = Probability of a Vindija and Mezmaiskaya 1 lineage coalescing prior to the split from Denisova in Figure S7.1 (that is, not coalescing on the lineage labeled N).

p_A = Probability of a Vindija and Denisova lineage coalescing prior to the split from present-day Africans in Figure S7.1 (that is, not coalescing on the lineage labeled A).

With these definitions, we can see that the expected values are:

$$\boldsymbol{E}[\boldsymbol{n}_{(V_1V_2)H}] = \boldsymbol{p}_V \boldsymbol{p}_N \boldsymbol{p}_A \boldsymbol{K} \quad \text{(S7.3)}$$

$$\boldsymbol{E}[0.5(\boldsymbol{n}_{(V_1D)H} + \boldsymbol{n}_{(V_2D)H})] = \boldsymbol{p}_A \boldsymbol{K} \quad \text{(S7.4)}$$

We define the ratio of Equations S7.3 and S7.4 as $\hat{G}_{VN}$. For large data sets such as those in this study, the expectation of the ratio is approximately the same as the ratio of expectations. Thus:

$$\hat{G}_{VN} = \frac{\eta_{(V_1V_2)H}}{0.5\left(\eta_{(V_1D)H} + \eta_{(V_2D)H}\right)} \qquad E[\hat{G}_{VN}] \approx p_V p_N \tag{S7.5}$$

We can carry out the same type of analysis for sequence data for any set of 5 individuals, and focus on three such alignments in the analyses below (denoting Mezmaiskaya 1 as "M"):

V_1-V_2-D-H-C (yielding the quantities $\eta_{(V_1V_2)H}$, $\eta_{(V_1D)H}$ and $\eta_{(V_2D)H}$ as presented above)
V_1-V_2-M-H-C (yielding the quantities $\eta_{(V_1V_2)H}$, $\eta_{(V_1M)H}$ and $\eta_{(V_2M)H}$)
V- M-D-H-C (with Vindija data pooled, yielding the quantities $\eta_{(VM)H}$, $\eta_{(VD)H}$ and $\eta_{(MD)H}$)

We can use these quantities to obtain two additional estimators:

$$\hat{G}_V = \frac{\eta_{(V_1V_2)H}}{0.5\left(\eta_{(V_1M)H} + \eta_{(V_2M)H}\right)} \qquad E[\hat{G}_V] \approx p_V \tag{S7.6}$$

$$\hat{G}_N = \frac{\eta_{(VM)H}}{0.5\left(\eta_{(VD)H} + \eta_{(MD)H}\right)} \qquad E[\hat{G}_N] \approx p_N \tag{S7.7}$$

Table S7.3 reports the estimates of coalescence probabilities along various lineages that emerge from empirical computation of the $\hat{G}$ values. We estimate that the probability that Vindija and Mezmaiskaya 1 alleles coalesce more recently than their split from Denisova is 64.7 ± 2.8%, the probability that two Vindija alleles coalesce more recently than their split from Mezmaiskaya 1 is 21.1 ± 6.2%, and the probability that two Vindija alleles coalesce more recently than their split from Denisova is 76.0 ± 1.2%. Thus, the genetic drift on the Neandertal lineage since the split from Denisovans is estimated to be far more than the 15-20% probability of coalescence in Eurasians since their separation from West Africans[6]. In other words, even if we restrict to the history that occurred prior to the divergence of Vindija and Mezmaiskaya Neandertals, we must conclude that Neandertals experienced a stronger bottleneck in their common history than the "out of Africa" bottleneck that has affected all present-day non-Africans.

Table S7.3: Estimates of coalescence probability on archaic lineages

Statistic	Quantity being estimated	Basic estimate of coalescence probability	Corrected estimate of coalescence probability
$1-\hat{G}_V$	$1-p_V$ (prob. of 2 Vindija lineages coalescing more recently than split from Mez. 1)	21.1 ± 6.2%	n/a
$1-\hat{G}_N$	$1-p_N$ (prob. of Vindija & Mez. 1 coalescing more recently than Denisova split)	64.7 ± 2.8%	n/a
$1-\hat{G}_{VN}$	$1-p_Vp_N$ (prob. of 2 Vindija lineages coalescing more recently than Denisova split)	76.0 ± 1.2%	82.3 ± 1.6%

Note: $\hat{G}_V$ and $\hat{G}_{VN}$ are averaged over all possible pairs of Vindija bones, and we compute a standard error with a Block Jackknife.

The estimates of genetic drift (coalescence probability) in the first column of Table S7.3 are conservative minima, as the computations were carried out assuming that sites of the rare substitution classes (BABA and ABBA) are genuinely reflecting incomplete lineage sorting. In fact, sites like this can also be generated by sequencing error, mapping error, and recurrent mutation. Assuming as a first approximation that these processes contribute an error term *e* equally to the numerator and denominator of the $\hat{G}$ statistics, they are expected to bring the ratios closer to 1 than is appropriate. Thus, we expect the numbers in the first column to underestimate the coalescence probability (that is, the genetic drift) specific to each lineage.

To correct for false-positive incomplete lineage sorting events, we carry out a new analysis restricting to sites in the genome where we have coverage from 2 reads of one of the Vindija samples (V_{1a}, V_{1b}), 1 read from a second Vindija sample (V_2), one read from Denisova (D), one read from a present-day African (H, represented by a pool of our San, Yoruba and Mbuti data), and chimpanzee (C). We then analyze the rates of rare classes in 4 read alignments:

$\boldsymbol{n}_{(V_1D)H} = \boldsymbol{n}_{V_1H} + \boldsymbol{n}_{DH}$ is the sum of BABA and ABBA in a V_1-D-H-C alignment
$\boldsymbol{n}_{(V_2D)H} = \boldsymbol{n}_{V_2H} + \boldsymbol{n}_{DH}$ is the sum of BABA and ABBA in a V_2-D-H-C alignment
$\boldsymbol{n}_{(V_1V_2)H} = \boldsymbol{n}_{V_1H} + \boldsymbol{n}_{V_2H}$ is the sum of BABA and ABBA in a V_1-V_2-H-C alignment
$\boldsymbol{n}_{(V_{1a}V_{1b})H} = \boldsymbol{n}_{V_{1a}H} + \boldsymbol{n}_{V_{1b}H}$ is the sum of BABA and ABBA in a V_{1a}-V_{1b}-H-C alignment

In the absence of sequencing error, $\boldsymbol{n}_{(V_{1a}V_{1b})H}$ should be half of $\boldsymbol{n}_{(V_1V_2)H}$. This is because in half of cases, by chance, we expect reads to perfectly match because they are from the same haplotype.

We now define three new statistics, $\hat{\boldsymbol{Q}}$, $\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{G}}_{VN}{}^{corrected}$:

$$\hat{\boldsymbol{Q}} = \frac{\boldsymbol{n}_{(V_{1a}V_{1b})H}}{\boldsymbol{n}_{(V_1V_2)H}} \qquad \hat{\boldsymbol{R}} = \frac{\boldsymbol{n}_{(V_1V_2)H}}{0.5\left(\boldsymbol{n}_{(V_1D)H} + \boldsymbol{n}_{(V_2D)H}\right)} \qquad \hat{\boldsymbol{G}}_{VN}{}^{corrected} = \hat{\boldsymbol{R}} \Big/ \left[1 + \left(\frac{\hat{\boldsymbol{Q}} - 0.5}{1 - \hat{\boldsymbol{Q}}}\right)\left(1 - \hat{\boldsymbol{R}}\right)\right] \qquad \text{(S7.8)}$$

The first two of these statistics $\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{R}}$ have the following expectations under the assumption that error *e* contributes to the same extent to the numerator and denominator:

$$\boldsymbol{E}[\hat{\boldsymbol{Q}}] \approx \frac{0.5\boldsymbol{p}_V \boldsymbol{p}_N \boldsymbol{p}_A \boldsymbol{K} + \boldsymbol{e}}{\boldsymbol{p}_V \boldsymbol{p}_N \boldsymbol{p}_A \boldsymbol{K} + \boldsymbol{e}} \qquad \Rightarrow \boldsymbol{e} \approx \boldsymbol{p}_V \boldsymbol{p}_N \boldsymbol{p}_A \boldsymbol{K}\left(\frac{\boldsymbol{E}[\hat{\boldsymbol{Q}}] - 0.5}{1 - \boldsymbol{E}[\hat{\boldsymbol{Q}}]}\right) \qquad \text{(S7.9)}$$

$$\boldsymbol{E}[\hat{\boldsymbol{R}}] \approx \frac{\boldsymbol{p}_V \boldsymbol{p}_N \boldsymbol{p}_A \boldsymbol{K} + \boldsymbol{e}}{\boldsymbol{p}_A \boldsymbol{K} + \boldsymbol{e}} \approx \frac{1 + \left(\frac{\boldsymbol{E}[\hat{\boldsymbol{Q}}] - 0.5}{1 - \boldsymbol{E}[\hat{\boldsymbol{Q}}]}\right)}{\frac{1}{\boldsymbol{p}_V \boldsymbol{p}_N} + \left(\frac{\boldsymbol{E}[\hat{\boldsymbol{Q}}] - 0.5}{1 - \boldsymbol{E}[\hat{\boldsymbol{Q}}]}\right)} \qquad \text{(S7.10)}$$

It is now easy to see that $\boldsymbol{E}\left[\hat{\boldsymbol{G}}_{VN}{}^{corrected}\right] = \boldsymbol{p}_V \boldsymbol{p}_N$. We estimate this quantity by averaging over all 6 possible pairs of Vindija bones (for sites where each bone covered twice, there are two choices for the bone covered once). We compute standard errors using a Block Jackknife[4,5]. We obtain $\hat{\boldsymbol{Q}}$ = 83.6 ± 1.4%, $\hat{\boldsymbol{R}}$ = 41.0 ± 2.3%, and $\left(1 - \boldsymbol{p}_V \boldsymbol{p}_N\right) \approx \left(1 - \hat{\boldsymbol{G}}_{VN}{}^{corrected}\right)$ = 82.3 ± 1.9%.

Thus, the probability of two Vindija lineages coalescing more recently than their split from Denisova is estimated to be around 82.3 ± 1.9%, after correcting for error. This high coalescence probability genetic drift cannot be entirely explained by a bottleneck specific to Vindija cave, as 64.7 ± 2.8% is a conservative minimum for the probability that Vindija and Mezmaiskaya 1 alleles coalesce more recently than their split from Denisova (Table S7.3). Taken together, these results imply an extreme bottleneck in the common history of Neandertals.

Preliminary evidence that Denisovans experienced their own population bottleneck

To learn about the coalescence probability of two alleles in the Denisovan individual since the historical divergence of Denisovan and Neandertal ancestors, we restricted analysis to sites in the genome where we had data from two distinct reads from two samples; e.g. 2 from a single Vindija bone and 2 from Denisova. The relative rates of heterozygous genotypes in the two individuals is informative about the amount of genetic drift in the lineage of the two samples since they divergence, and we show the results of this analysis in Table S7.4.

Table S7.4 shows that the highest rate of shared heterozygotes is between San and Yoruba at 12.2 ± 0.3%, a small reduction compared with the expectation of 16.7% for two individuals from the same population (for the same population, there is a 25% chance that the two sampled reads will represent distinct haplotypes in both individuals, which we multiply by the 66.7% probability of double heterozygotes). The high rate of shared heterozygotes reflects the relatively small amount of genetic drift since San and Yoruba diverged. In contrast, between San and Vindija the proportion of shared heterozygotes is 3.0 ± 0.3%, between San and Denisova 2.0 ± 0.2%, and between Vindija and Denisova 4.1 ± 0.3% Thus, there are few shared polymorphisms between the archaic and modern hominins, or between the two archaic hominins, suggesting a substantial probability of coalescence and high drift on many of the archaic hominin lineages.

We next examined sites where 3 copies of the derived allele (and 1 copy of the ancestral) were observed, suggesting a polymorphism in the ancestral population where the derived allele is observed in one of the two analyzed individuals but not the other (Table S7.4). This computation allows us to estimate which population may have experienced more drift based on the one with a higher proportion of fixed sites, although a caveat is that strong inbreeding in the recent ancestry of any one sample could cause it to appear to have experienced more genetic drift than is true for the population as a whole. Comparing San and Yoruba, 49.0 ± 0.4% of the heterozygous sites are in Yoruba, suggesting similar amounts of genetic drift since these two populations diverged. However, there has been much more genetic drift in the history of Neandertal and Denisovans than in present-day humans since the two groups diverged. For example, in a pairwise comparison of San to Vindija, only 24.1 ± 1.0% of sites that we find to be heterozygous in one of the samples are heterozygous in Vindija, and in a pairwise comparison of San to Denisova, an even lower proportion of 20.9 ± 0.8% sites are estimated to be heterozygous in Denisova.

These data suggest that Denisovans may have experienced a comparable amount of genetic drift since divergence from modern humans as the Vindija Neandertals. A caveat to the analysis presented in Table S7.4, however, is that we only have access to data from a single Denisovan, and inbreeding in this individual's recent ancestors could cause us to overestimate Denisovan-specific genetic drift. Thus, we view the finding of high genetic drift specific to Denisovans as an intriguing result, which deserves future follow-up analysis once substantial amounts of nuclear genomic data from more Denisovan individuals becomes available.

Table S7.4: Relative probability of coalescence in two samples since they diverged

Sample A for which we have two reads	**Sample B for which we have two reads**	**Statistic 1: % of sites with 2 derived and 2 ancestral reads where A, B are both heterozygous**	**Statistic 2: % of sites with 3 derived and 1 ancestral reads in which B is heterozygous**	**Comments on results**
San	Yoruba	12.2 ± 0.3%	49.0 ± 0.4%	The rate of doubly heterozygous sites is close to the theoretical maximum of 16.7% if San and Yoruba were in the same population, suggesting large population sizes since their divergence (Statistic 1). The fact that Statistics 2 is close to 50% suggests about the same coalescent probability in San and Yoruba since their divergence.
Papuan1	Yoruba	9.7 ± 0.4%	62.2 ± 0.5%	The "out of Africa" bottleneck is reflected in a reduced rate of doubly heterozygous sites compared to San-Yoruba (Statistic 1). Most coalescence is in non-Africans (fewer heterozygous sites in Papuan1 than in Yoruba) (Statistic 2)
San	Vindija	3.0 ± 0.3%	24.1 ± 1.0%	The strong bottleneck in Vindija since divergence from present-day humans is reflected in the greatly reduced rate of doubly heterozygous sites (Statistic 1) and the fact that Vindija has many fewer heterozygous sites than San (Statistic 2).
San	Denisova	2.0 ± 0.2%	20.9 ± 0.8%	Two alleles from Denisova have a much higher probability of coalescence than two alleles from San since their common divergence (Statistic 2).
Vindija	Denisova	4.1 ± 0.4%	54.1 ± 0.9%	The high rate of coalescence in Vindija and Denisova since they diverged is seen in the low rate of doubly heterozygous sites (Statistic 1), and the fact that they retain a similar proportion of ancestral polymorphism (Statistic 2).

Note: For analyses involving Vindija, we average results over three different individuals and compute standard errors using a Block Jackknife.

References for SI 7

1. Strimmer, K. and von Haeseler, A., Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol Biol Evol*, **13,** 964 (1996).
2. Green, R. E. et al., A draft sequence of the Neandertal genome. *Science* **328**, 710 (2010).
3. Pamilo, P. and Nei, M., Relationships between gene trees and species trees. *Mol Biol Evol* **5**, 568 (1988).
4. Busing, F.M.T.A., Meijer, E. and van der Leeden, R., Delete-m jackknife for unequal m. *Statistics and Computing* **9**, 3 (1999).
5. Kunsch, H.K., The jackknife and the bootstrap for general stationary observations. *Ann Statist* **17**, 1217 (1989).
6. Keinan, A., Mullikin, J.C., Patterson, N. and Reich, D., Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* **39,** 1251 (2007).

Supplementary Information 8

Denisovans share more derived alleles with Melanesians than with other groups.

Nick Patterson*, Heng Li, Swapan Mallick and David Reich*

* To whom correspondence should be addressed (nickp@broadinstitute.org) or David Reich (reich@genetics.med.harvard.edu)

In this section, we provide two independent lines of evidence that Denisovans are more closely related to Melanesians (Papuans and Bougainville islanders) than to other geographically dispersed present-day humans. We show that a likely explanation for this pattern is a second archaic gene flow event into modern humans: above and beyond the signal of Neandertal-related gene flow into the ancestors of all non-Africans that we reported previously[1]. We show that to explain this pattern, the gene flow must have been from a population more closely related to Denisovans than to Neandertals. We finally estimate the percentage of Denisovan-related ancestry in Melanesians that would be needed to explain these patterns.

Genotyping data shows that Melanesians have a different relationship to archaic lineages than do other non-Africans

We compared the sequencing data from Denisovans and Neandertals to 938 unrelated individuals from 53 populations from the CEPH-Human Genome Diversity Panel (CEPH-HGDP), using previously published data from an Illumina 650Y SNP array that had been run on these samples[2].

We mapped all 642,690 SNPs that passed data quality filters to the chimpanzee reference genome, *pantro2*. After overlapping these SNPs with our Denisova and Neandertal (Vindija) sequencing data[1], and applying the same set of data quality filters as in SI 6, we had 255,077 SNPs available for analysis, at all of which we had an allele call for Neandertal, Denisova and chimpanzee. At each of these sites, we represented Neandertal and Denisova with a single randomly chosen read. The results below report data from both transitions and transversion substitutions to maximize the number of sites available for analysis, since the number of SNPs was a limit to the power of our analysis. Reassuringly, when we repeated our key analyses restricting to transversion substitutions, we obtained qualitatively similar results (but noisier).

We carried out a Principal Component Analysis (PCA) on chimpanzee, Neandertal and Denisova, without using data from present-day humans at all. The top two eigenvectors from this PCA determine a plane. Using the SNP weights from the PCA, we can then project the CEPH-HGDP samples onto the plane, a now-standard technique that is described in ref. 3. This allows us to explore the relationship of diverse present-day humans relative to archaic hominins and chimpanzee, and to test if the genetic differences among present-day human populations are correlated to the differences among these non-modern humans.

Figure S8.1 presents results for selected present-day human populations: sub-Saharan Africans (San, Yoruba and Mbuti), two Melanesian populations (Papuans and Bougainville islanders), and two other non-African populations (French and Han). There are two main patterns. The first is that the African samples separate from the rest, as expected based on our previous finding that Neandertals are more closely related to non-Africans[1]. The second is a surprise: we find that the two Melanesian populations separate from the other non-Africans.

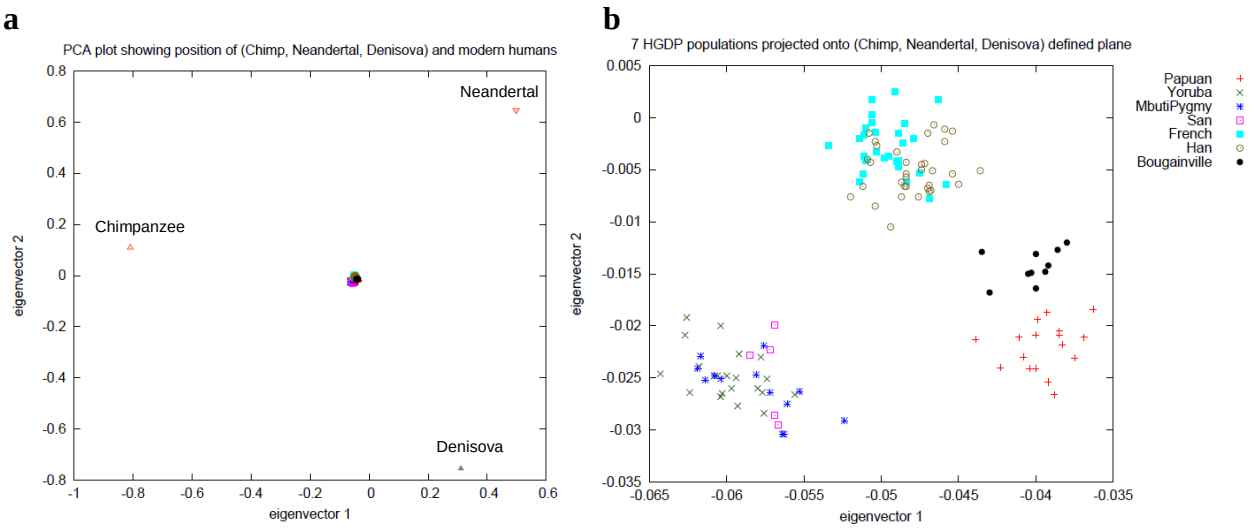


Figure S8.1: (a) Principal Components Analysis (PCA) of 255,077 SNPs known to be polymorphic in present-day humans where we also have data from Neandertal, Denisova and chimpanzee. Present-day humans all appear at the center of the plot when projected onto the top two eigenvectors. (b) Magnification of the central portion of the plot shows that present-day humans separate into three clusters in relation to archaic hominins and chimpanzees: "Africans", "Melanesians" and other "Non-Africans".

To assess the generality of this result in a more diverse set of populations, in Figure S8.2 (a reproduction of Figure 2 in the main text) we plot the mean of eigenvectors 1 and 2 for each of the 53 CEPH-HGDP populations, and color the populations by geography: "African" (n=7) (San, Mbuti, Biaka, Bantu Kenya, Bantu South Africa, Yoruba and Mandenka), "Papuan" (n=1), "Bougainville" (n=1), and other "Non-Africans" (n=44; all other populations). We continue to observe three clusters. We interpret the fact that Bougainville Melanesians appear intermediate between the Papuans and the other Non-Africans as reflecting a history of mixture between Melanesians and East Eurasians, consistent with previous studies of Melanesian populations[2,4].

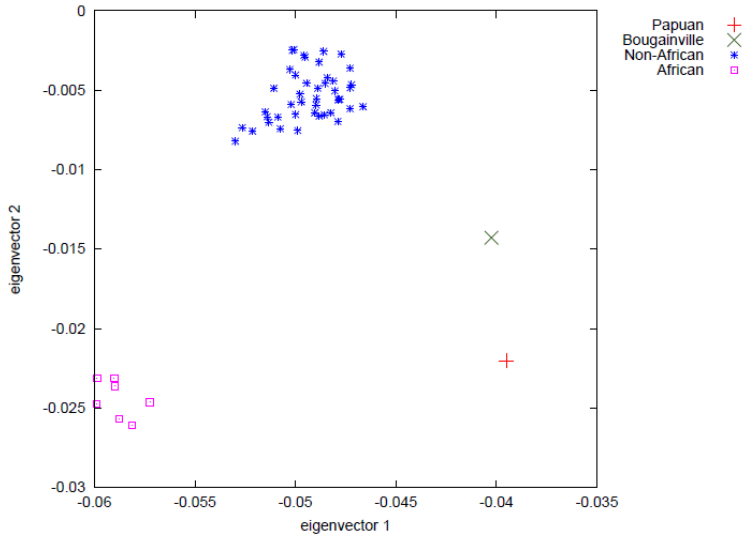


Figure S8.2: Projection of all 53 populations from the CEPH-HGDP panel onto eigenvectors 1 and 2, defined based on analysis of chimpanzee, Neandertal and Denisova. We represent each population by the means of all samples, which reduces noise compared with Figure S8.1 where individual samples are plotted. The 7 "African" populations are San, Mbuti, Biaka, Bantu Kenya, Bantu South Africa, Yoruba and Mandenka, and the 44 "Non-African" populations are all remaining groups except for Papuan and Bougainville. This figure reproduces Figure 2 in the main text.

The PCA results provide qualitative evidence for Melanesians having a different pattern of average genetic relationship to archaic populations than do other non-Africans. To use the

genotyping data to develop a formal statistical test for whether there is a different relationship to archaic humans comparing Melanesians and other non-Africans, we carried out a *4 Population Test* that assesses whether 4 populations are consistent with being related by an unrooted phylogenetic tree ((1, 2), (3, 4))[5]. Denote the minor allele frequency of SNP *i* in population *j* as $\hat{p}_j^{\,i}$. Then, the allele frequencies in all four populations are $(\hat{p}_1^{\,i}, \hat{p}_2^{\,i}, \hat{p}_3^{\,i}, \hat{p}_4^{\,i})$. If the proposed phylogenetic tree is correct, the frequency difference $(\hat{p}_1^{\,i} - \hat{p}_2^{\,i})$ should be uncorrelated to the frequency difference $(\hat{p}_3^{\,i} - \hat{p}_4^{\,i})$, as the differences reflect random drift on the lineages relating populations (1, 2), and (3, 4), which have non-overlapping histories.

To implement the *4 Population Test*, we compute a statistic that is a generalization of the *D*-statistic of ref. 1. Specifically, for the experiment in which we randomly draw a single allele to represent each population, we compute the expected number of "BABA" sites where populations 1=3 match for the alternative "B" allele and 2=4 match for the reference "A" allele, and "ABBA" sites where 2=3 and 1=4. If populations (1, 2) form a clade relative to (3, 4), then the difference between the two classes is expected to be consistent with 0. If the $\hat{D}(1,2,3,4)$ statistic measuring this is more than 3 standard deviations from 0 (using a standard error from a Block Jackknife[6,7]) we reject the hypothesis that the 4 populations are related by a simple tree:

$$\hat{D}(1,2,3,4) = \frac{n_{BABA} - n_{ABBA}}{n_{BABA} + n_{ABBA}} = \frac{\sum_{i=1}^{n}\left(\hat{p}_1^{\,i}(1-\hat{p}_2^{\,i})\hat{p}_3^{\,i}(1-\hat{p}_4^{\,i}) - (1-\hat{p}_1^{\,i})\hat{p}_2^{\,i}\hat{p}_3^{\,i}(1-\hat{p}_4^{\,i})\right)}{\sum_{i=1}^{n}\left(\hat{p}_1^{\,i}(1-\hat{p}_2^{\,i})\hat{p}_3^{\,i}(1-\hat{p}_4^{\,i}) + (1-\hat{p}_1^{\,i})\hat{p}_2^{\,i}\hat{p}_3^{\,i}(1-\hat{p}_4^{\,i})\right)} \quad \text{(S8.1)}$$

Table S8.1 presents results for selected sets of populations 1, 2, 3 and 4 (we choose populations for which we also have sequence data, allowing us to compare results from genotyping data (left side) to sequencing data (right side). The first block is provided as a negative control illustrating how sets of populations can pass the *4 Population Test*. The second block shows that Neandertals are more closely related to non-Africans than to Africans as we reported earlier[1]. The third block shows that Denisovans are closer to Papuans than to Chinese.

Table S8.1: 4 Population Test results on CEPH-HGDP genotyping data

Modern humans		**More distant relatives**		**Genotyping data**		**Sequencing data**		
H_1	**H_2**	**H_3**	**H_4**	**D-stat**	**Z-score**	**D-stat**	**Z-score**	**Conclusions from these tests**
San	Yoruba	Neandertal	Chimpanzee	0.5%	1.6	-0.3%	-0.6	Examples of sets of populations that pass the *4 Population Test*
French	Han	Neandertal	Denisova	0.4%	0.9	-0.9%	0.9	
French	Yoruba	Neandertal	Chimpanzee	2.9%	8.3	4.6%	6.9	Neandertals are closer to Non-Africans than to Africans
Papuan1	Yoruba	Neandertal	Chimpanzee	3.8%	8.4	4.0%	4.9	
Han	Papuan1	Denisova	Chimpanzee	-3.4%	-7.7	-5.2%	-5.8	Denisovans are closer to Melanesians than to other non-Africans.
Han	Papuan1	Neandertal	Denisova	3.3%	6.1	7.1%	6.4	

We considered whether these findings could be an artifact of "ascertainment bias", whereby the fact that the SNPs genotyped on the Illumina 650Y array were selected in a complicated way for medical genetics purposes, could confound inferences about population relationships[8]. While ascertainment bias is a serious concern in certain types of population genetics analyses—in

particular studies that infer human population expansions and contractions—we do not believe that it can explain our findings of Denisovans being more closely related to Melanesians than to other present-day humans. Firstly, Melanesian samples were not used (as far as we are aware) as part of SNP ascertainment. Secondly, ancient DNA was certainly not used as part of the array design and thus the last row of Table S8.1, which shows that the allele frequency differences between Han and Papuans are highly significantly correlated to those between Neandertals and Denisovans (Z = 6.1), is difficult to explain as an artifact.

As a further line of evidence for the robustness of these findings to the potential confounder of ascertainment bias, we also compared the *4 Population Test* results on the left side of Table S8.1, with *D*-statistic analyses of sequencing data collected from these same populations, which cannot be affected by ascertainment bias. The signs of the statistics are expected to be consistent, and indeed this is what we observe. (The absolute values of the *D*-statistics differ between the genotyping and sequencing analyses, which is not unexpected given that SNP arrays on average sample more common polymorphisms where empirically the signals of gene flow are weaker.)

Sequencing data confirms that Denisovans are closer to Melanesians than to other humans

To further understand the relationship of diverse present-day humans to Denisovans and Neandertals, we analyzed low-pass sequencing data from 12 present-day humans: 5 that we previously sequenced to 4-6× coverage[1] and 7 that we newly sequenced to 1-2x coverage (SI 9). Analysis of these data in conjunction with sequences from the Denisova phalanx and 3 Vindija bones provides further evidence that Denisovans are more closely related to Melanesians than to any of the other present-day human populations for which we have data.

We computed *D*-statistics for all 66 possible pairs of the twelve present-day human samples (H_1, H_2), testing them for consistency with being a clade in a tree where the proposed alternative clade is (Neandertal, Chimpanzee) or (Denisova, Chimpanzee) (Table S8.2). We classified the samples H_1 and H_2 by geographic region: 3 "African" (San, Yoruba, Mbuti), 3 "Melanesian" (Papuan1, Papuan2, Bougainville), and 6 "Eurasian" (French, Han, Karitiana, Sardinian, Cambodian, Mongolian). The designation "Eurasian" is used as a shorthand, as we use it to also indicate Karitiana who are Native Americans, but who are thought to descend from a Eurasian population prior to 15,000 years ago. The *D*-statistics reveal three broad patterns.

(i) Within regions there are no differences in how present-day humans relate to ancient bones. Within Africa, within Melanesia, and within Eurasia, the *D*-statistics are not significant.

(ii) Neandertals are more closely related to the archaic population that contributed genes to all Eurasians than are Denisovans. This is seen in the fourth block of Table S8.2. For example the highly significantly skewed statistic *D*(Han, San, Neandertal, Chimpanzee) = 5.5 ± 0.6%, becomes attenuated when we replace Neandertal with Denisova: *D*(Han, San, Denisova, Chimp) = 1.8 ± 0.5%. The reduction is highly significant (Z = 6.6 standard deviations).

(iii) Denisovans are more closely related to Melanesians than to other non-Africans. This is seen in the fifth block of statistics in Table S8.2, corresponding to Eurasian-Melanesian comparisons. Here, we find that the archaic ancestry in Melanesians is more correlated to Denisovans than to Neandertals: average *D*(Eurasian, Melanesian, Neandertal, Denisova) = 4.2%. Thus, Melanesians and Eurasians harbor different mixtures of archaic ancestry.

Table S8.2: Statistics comparing present-day humans to ancient bones for all pairs of samples

			H_3=Neandertal				H_3=Denisova			
	H_1	**H_2**	**n_{BABA}**	**n_{ABBA}**	**D-stat**	**Z**	**n_{BABA}**	**n_{ABBA}**	**D-stat**	**Z**
Eurasian	French	Han	17,214	17,602	-1.1%	-1.4	27,250	27,265	0.0%	0.0
/ Eurasian	French	Karitiana	3,482	3,435	0.7%	0.5	5,207	5,062	1.4%	1.3
	French	Sardinian	4,887	4,857	0.3%	0.3	7,398	7,333	0.4%	0.5
	French	Cambodian	8,267	8,383	-0.7%	-0.7	12,641	12,813	-0.7%	-0.9
	French	Mongolian	6,015	6,023	-0.1%	-0.1	9,367	9,252	0.6%	0.7
	Han	Karitiana	3,441	3,169	4.1%	**3.0**	5,117	4,857	2.6%	2.4
	Han	Sardinian	5,027	4,799	2.3%	2.0	7,522	7,411	0.7%	0.8
	Han	Cambodian	7,334	7,060	1.9%	1.9	10,982	10,961	0.1%	0.1
	Han	Mongolian	5,227	5,188	0.4%	0.3	7,981	8,059	-0.5%	-0.5
	Karitiana	Sardinian	1,116	1,085	1.4%	0.7	1,559	1,627	-2.1%	-1.2
	Karitiana	Cambodian	1,683	1,707	-0.7%	-0.4	2,371	2,460	-1.8%	-1.2
	Karitiana	Mongolian	1,128	1,195	-2.9%	-1.3	1,765	1,742	0.7%	0.4
	Sardinian	Cambodian	2,592	2,670	-1.5%	-1.0	3,935	3,925	0.1%	0.1
	Sardinian	Mongolian	1,966	2,027	-1.5%	-0.9	3,036	3,057	-0.3%	-0.3
	Cambodian	Mongolian	2,811	2,804	0.1%	0.1	4,442	4,342	1.1%	1.0
Melanesian	Papuan1	Papuan2	5,000	5,182	-1.8%	-1.6	8,034	8,424	-2.4%	-2.5
/ Melanesian	Papuan1	Bougainville	5,887	6,225	-2.8%	-2.7	9,347	9,430	-0.4%	-0.5
	Papuan2	Bougainville	3,351	3,284	1.0%	0.8	5,319	5,140	1.7%	1.5
African	San	Yoruba	23,690	23,855	-0.3%	-0.6	39,042	39,019	0.0%	0.1
/ African	San	Mbuti	7,910	7,611	1.9%	2.4	12,665	12,404	1.0%	1.4
	Yoruba	Mbuti	7,360	7,071	2.0%	2.2	11,511	11,646	-0.6%	-0.8
Eurasian	French	San	25,242	22,982	4.7%	**7.6**	39,838	38,495	1.7%	**3.4**
/ African	French	Yoruba	21,794	19,890	4.6%	**6.9**	34,262	33,078	1.8%	**3.6**
	French	Mbuti	8,068	7,113	6.3%	**7.0**	12,296	11,762	2.2%	**3.0**
	Han	San	25,081	22,470	5.5%	**8.5**	38,815	37,439	1.8%	**3.4**
	Han	Yoruba	21,741	19,412	5.7%	**7.9**	33,182	32,184	1.5%	2.8
	Han	Mbuti	7,851	6,746	7.6%	**8.4**	11,537	10,954	2.6%	**3.5**
	Karitiana	San	5,149	4,775	3.8%	**3.5**	7,722	7,683	0.3%	0.3
	Karitiana	Yoruba	4,383	4,199	2.1%	1.8	6,566	6,639	-0.6%	-0.6
	Karitiana	Mbuti	1,577	1,473	3.4%	1.8	2,368	2,360	0.2%	0.1
	Sardinian	San	6,892	6,337	4.2%	**4.3**	10,625	10,491	0.6%	0.8
	Sardinian	Yoruba	6,037	5,522	4.5%	**4.2**	9,362	9,064	1.6%	2.0
	Sardinian	Mbuti	2,562	2,400	3.3%	2.2	4,028	3,784	3.1%	2.6
	Cambodian	San	11,362	10,379	4.5%	**5.6**	17,647	16,922	2.1%	**3.0**
	Cambodian	Yoruba	10,048	9,150	4.7%	**5.3**	15,468	14,806	2.2%	**3.1**
	Cambodian	Mbuti	4,235	3,641	7.5%	**6.5**	6,329	5,850	3.9%	**4.0**
	Mongolian	San	8,312	7,545	4.8%	**5.4**	12,812	12,497	1.2%	1.7
	Mongolian	Yoruba	7,232	6,531	5.1%	**5.5**	11,138	10,804	1.5%	1.9
	Mongolian	Mbuti	3,077	2,765	5.3%	**3.9**	4,514	4,505	0.1%	0.1
Eurasian	French	Papuan1	15,523	15,548	-0.1%	-0.1	23,509	25,470	-4.0%	**-5.7**
/ Melanesian	French	Papuan2	7,638	8,066	-2.7%	-2.6	11,651	13,380	-6.9%	**-8.2**
	French	Bougainville	8,020	8,491	-2.9%	-2.9	12,261	13,554	-5.0%	**-6.5**
	Han	Papuan1	15,059	14,677	1.3%	1.5	22,262	24,198	-4.2%	**-5.8**
	Han	Papuan2	7,169	7,082	0.6%	0.6	10,461	11,987	-6.8%	**-7.7**
	Han	Bougainville	7,353	7,435	-0.6%	-0.5	10,889	12,022	-4.9%	**-5.8**
	Karitiana	Papuan1	3,242	3,352	-1.7%	-1.2	4,595	5,185	-6.0%	**-5.2**
	Karitiana	Papuan2	1,522	1,658	-4.3%	-2.2	2,201	2,641	-9.1%	**-5.8**
	Karitiana	Bougainville	1,577	1,717	-4.3%	-2.4	2,229	2,671	-9.0%	**-5.9**
	Sardinian	Papuan1	4,335	4,439	-1.2%	-0.9	6,485	7,044	-4.1%	**-4.2**
	Sardinian	Papuan2	2,447	2,647	-3.9%	-2.6	3,714	4,150	-5.5%	**-4.5**
	Sardinian	Bougainville	2,531	2,762	-4.4%	**-3.0**	3,877	4,336	-5.6%	**-4.9**
	Cambodian	Papuan1	6,968	6,895	0.5%	0.5	10,269	11,103	-3.9%	**-4.4**
	Cambodian	Papuan2	3,713	3,891	-2.3%	-1.8	5,457	6,272	-6.9%	**-6.5**
	Cambodian	Bougainville	3,847	3,994	-1.9%	-1.6	5,751	6,333	-4.8%	**-4.7**
	Mongolian	Papuan1	5,050	5,060	-0.1%	-0.1	7,498	8,269	-4.9%	**-5.0**
	Mongolian	Papuan2	2,783	2,852	-1.2%	-0.8	4,192	4,758	-6.3%	**-5.3**
	Mongolian	Bougainville	2,813	3,066	-4.3%	-2.9	4,234	4,847	-6.8%	**-6.0**
Melanesian	Papuan1	San	21,985	20,366	3.8%	**5.1**	35,923	32,841	4.5%	**7.2**
/ African	Papuan1	Yoruba	19,107	17,646	4.0%	**4.9**	30,995	28,186	4.7%	**7.4**
	Papuan1	Mbuti	6,826	6,133	5.3%	**5.4**	10,836	9,752	5.3%	**6.2**
	Papuan2	San	10,641	9,351	6.5%	**6.9**	17,304	15,266	6.3%	**8.4**
	Papuan2	Yoruba	9,393	8,272	6.3%	**6.4**	15,380	13,545	6.3%	**8.5**
	Papuan2	Mbuti	3,832	3,324	7.1%	**5.4**	6,124	5,233	7.8%	**7.2**
	Bougainville	San	11,296	10,020	6.0%	**6.8**	17,770	16,058	5.1%	**7.1**
	Bougainville	Yoruba	9,936	8,805	6.0%	**6.4**	15,784	14,050	5.8%	**8.1**
	Bougainville	Mbuti	4,216	3,596	7.9%	**6.8**	6,498	5,633	7.1%	**6.7**

Gene flow from Denisovan relatives into the ancestors of Melanesians is the most parsimonious explanation for these observations

To gain insight into the history that could be responsible for these observations, we first considered an "ancient substructure" model. If ancient substructure explained our data, it would imply that the ancestors of present-day humans were highly differentiated hundreds of thousands of years ago when they separated from the ancestors of Denisovans and Neandertals, after which point there was little further interaction. If the ancestors of present-day humans did not fully homogenize genetically since that time, some present-day populations might retain a greater degree of relatedness to archaic hominins than others, explaining the different relationship of Neandertals to Africans and non-Africans that we first documented in ref. 1. An even more complicated ancient substructure scenario could potentially explain the specific patterns in Melanesians. An ancient substucture model consistent with the data is presented in SI 11.

We cannot formally rule out ancient substructure with the analyses presented here. However, we believe that ancient substructure is a less parsimonious explanation for our observations than at least two episodes of archaic gene flow: the first into the ancestors of all present-day non-Africans, and the second into the ancestors of Melanesians. An important observation in this context is that Denisovans and Neandertals are sister groups (SI 6). If they were perfect sister groups, we would expect them to have exactly the same relationships to present-day human populations as each other. However, they have qualitatively very different relationships (Table S8.2), which makes our results difficult to explain by ancient structure, although more complicated scenarios of ancient structure could be consistent with the results as shown in SI 11.

Under the hypothesis that gene flow explains these observations, we can infer that its direction must have been, at least in part, from Denisovan relatives into the ancestors of Melanesians. To understand why this is the case, we focus on the statistic *D*(Eurasian, Melanesian, African, Chimpanzee), which does not use archaic hominin data at all, and which we compute at maximal precision by pooling data from the 6 Eurasian samples, the 3 Melanesian samples, and the 3 Africans to represent these geographic groupings. (We checked that this pooling strategy does not bias our results relative to comparisons of pairs of individuals.) We observe a 3.4 ± 0.3% higher rate of matching of Africans to Eurasians than to Melanesians (Z = 10.8). A more distant relationship of Africans to Melanesians is expected from gene flow into Melanesian ancestors, but not from the reverse direction of gene flow.

Gene flow from modern humans into the ancestors of Denisovans is not only unsupported by the *D*-statistics, but is also historically implausible. The Denisova phalanx is more than 30,000 years old, and in our opinion is likely to be more that 50,000 years old (SI 12). The more ancient age estimate is older, and the more recent age estimate is only slightly younger than the age of the oldest confirmed modern human remains outside of Africa and the Levant. It is difficult to envision a plausible scenario in which the Denisovan population could have ancestry from a modern human group that experienced mixture in an area near where Melanesians live now, and then migrated to Siberia in just a few thousand years.

If gene flow explains our observations, another implication is that the patterns observed in Melanesians must reflect at least two archaic gene flow events, and cannot simply be a stronger manifestation of the gene flow that affected the ancestors of all non-Africans[1]. To see this, we considered the hypothesis that after the initial mixture event that affected the ancestors of all

non-Africans (e.g. in the Levant), the proportion of archaic ancestry in the ancestors of non-Africans was higher than today, and that this pulse of mixture was then diluted (except in Melanesians) by subsequent migrations out of Africa. However, this scenario would predict that the archaic ancestry in Melanesians would have exactly the same source as in other non-Africans, and this cannot explain the patterns we observe, since we find that the archaic affinities in Melanesians are different in nature than in other non-Africans. In particular, when we compute all 18=6×3 possible statistics of the form *D*(Eurasian, Melanesian, Neandertal, Denisova), we find that 7 of 18 are skewed from zero at a significance of Z>3. This contrasts with what would be expected if the archaic ancestry in Eurasians and Melanesians came from the same source, which would predict no difference in relatedness to Neandertals and Denisovans. Thus, the archaic material in Melanesians reflects a different mixture of ancestries than in Eurasians.

Strategy for estimating a proportion of gene flow

Under the assumption that the patterns we observe are due to gene flow, we estimated the proportion of archaic ancestry in present-day non-Africans (both Eurasians and other non-Melanesians). The model we considered was one in which there was an initial pulse of Neandertal-related material into the ancestors of all non-Africans, followed by a second pulse only into the ancestors of Melanesians (see SI 11 for a parametric exploration of this model).

To estimate the proportion of gene flow, we extended the "*S*-statistic" methodology developed in SOM 18 of ref. 1. Our *S*-statistics are closely related to our *D*-statistics, and like them, examine all nucleotides on the autosomes where we have at least one sample from each of the populations {H_1,H_2,H_3,*Chimpanzee*}. We then restrict to the *n* sites that are biallelic transversion substitutions. Denoting the chimpanzee allele as "A", and using $\hat{\boldsymbol{p}}^{i}{}_1$, $\hat{\boldsymbol{p}}^{i}{}_2$ and $\hat{\boldsymbol{p}}^{i}{}_3$ to denote the empirically observed frequencies of the non-chimpanzee allele "B" in each populations at nucleotide *i*, we define an expression measuring the expected excess of sites with a "BABA" over an "ABBA" pattern, assuming that we randomly draw one allele to represent each sample:

$$\hat{\boldsymbol{S}}(\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{H}_3,\boldsymbol{Chimpanzee}) = \boldsymbol{n}_{\boldsymbol{BABA}} - \boldsymbol{n}_{\boldsymbol{ABBA}} = \sum_{i=1}^{n}\left[\hat{\boldsymbol{p}}^{i}{}_1(1-\hat{\boldsymbol{p}}^{i}{}_2)\hat{\boldsymbol{p}}_3{}^{i} - (1-\hat{\boldsymbol{p}}^{i}{}_1)\hat{\boldsymbol{p}}^{i}{}_2\hat{\boldsymbol{p}}^{i}{}_3\right] = \sum_{i=1}^{n}(\hat{\boldsymbol{p}}^{i}{}_1 - \hat{\boldsymbol{p}}^{i}{}_2)\hat{\boldsymbol{p}}^{i}{}_3 \quad \text{(S8.2)}$$

If samples H_1 and H_2 form a clade relative to sample H_3, then we expect an equal rate of BABA and ABBA sites and this statistic has an expectation of 0. Alternatively, if there has been gene flow into the ancestors of H_1 or H_2, we expect an excess of one class or another, allowing us to quantify the gene flow. We can assess whether the excess is significant by a Block Jackknife[6,7]. We note that $\hat{\boldsymbol{S}}$ is just the numerator of the *D*-statistic defined in ref. 1.

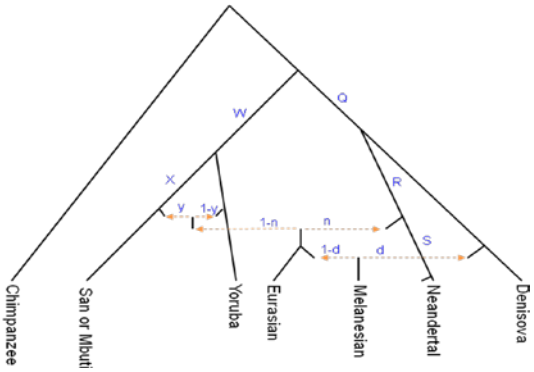


Figure S8.3: An "Admixture Graph" relating populations of interest. Genetic drift, defined as the variance in allele frequency due to sampling across generations, is indicated in upper case letters for selected lineages (X, W, Q, R and S). Mixture proportions are indicated in lower case: "*n*" the proportion of Neandertal-related ancestry in Eurasians, "*d*" the additional Denisova-related gene flow into Melanesians, and "*y*" the proportion of non-African ancestry that is more closely related to San than Yoruba (*n*, *d* and *y* are all likely to be small but possibly non-zero). This figure is not drawn to scale, but this does not affect our inferences, which do not depend on assumptions about the timing of splits.

To compute the expected value of $\hat{S}(H_1, H_2, H_3, Chimpanzee)$, we use "Admixture Graph" theory. As described in ref. 5, an Admixture Graph (Figure S8.3) is a generalization of a phylogenetic tree: a representation of population relationships that uses solid edges to indicate which populations are related by descent from common ancestors, and dotted lines to indicate mixture events. Admixture Graphs provide the information that is needed to compute the expected values of correlations in allele frequency across modern populations, without making assumptions about how population sizes changed along lineages, or assuming timings of events[5].

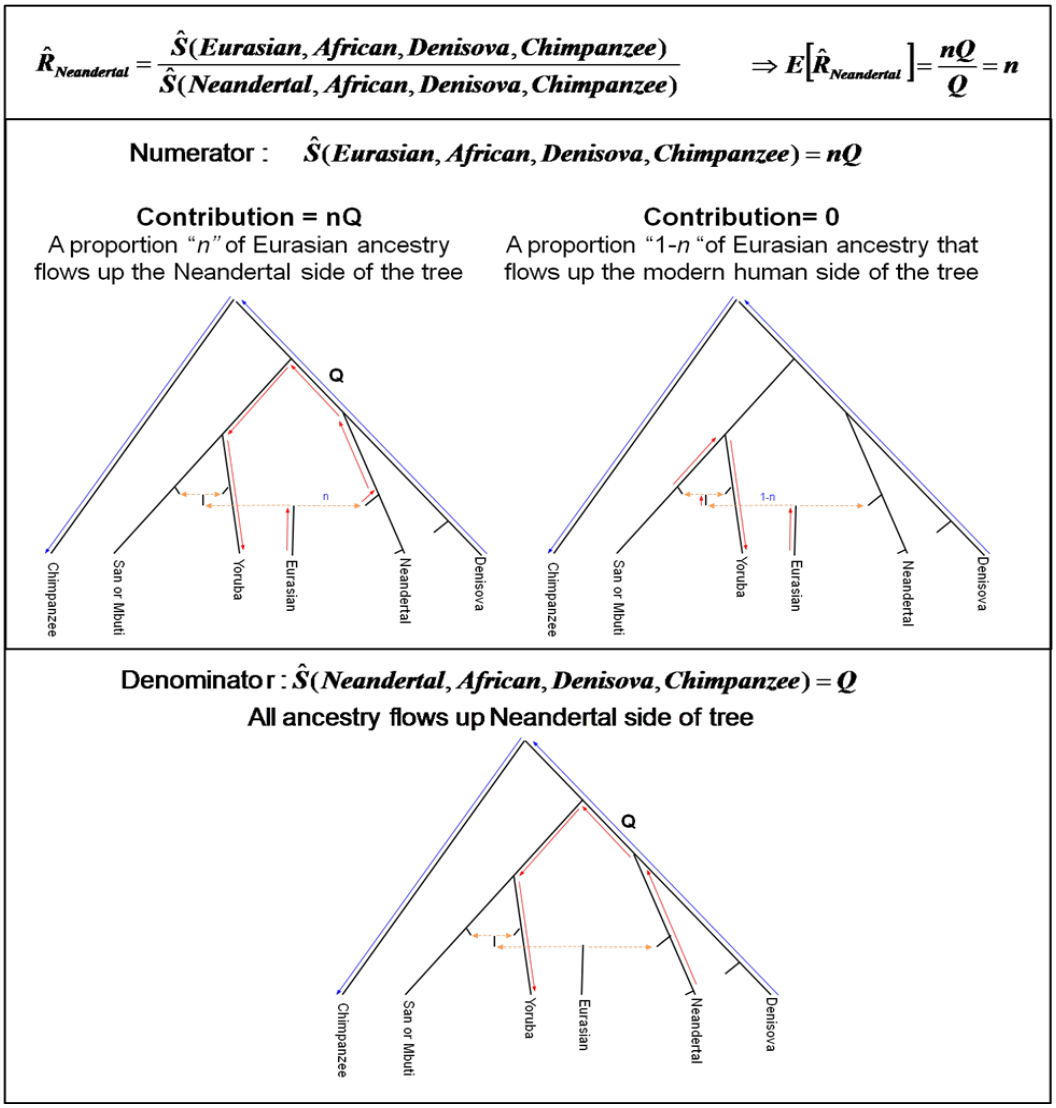


Figure S8.4: The expected values of the $R_{\text{Neandertal}}$ statistic that we use for estimating the proportion n of Neandertal-related mixture in all non-Africans. We trace historical differences through the Admixture Graph of Figure S8.3. Regions where red and blue lines overlap are correlated drifts and contribute to the expectation.

An improved estimate of the proportion of Neandertal-related ancestry in all non-Africans

To estimate the proportion n of archaic ancestry in non-Africans, we compute a ratio ($\hat{R}_{Neandertal}$) of two S-statistics, examining all sites where we have DNA sequence data from at least one sample from each of five groups: Eurasian, African, Denisova, Neandertal and chimpanzee.

$$\hat{R}_{Neandertal} = \frac{\hat{S}(Eurasian, African, Denisova, Chimpanzee)}{\hat{S}(Neandertal, African, Denisova, Chimpanzee)} \quad \text{(S8.3)}$$

Intuitively, we think of the ratio $\hat{R}_{Neandertal}$ as measuring how far of the way a Eurasian population is toward having the allele frequency correlation patterns with Africans, Denisovans, and chimpanzees that is characteristic of a 100% Neandertal. The expectation of both the numerator and denominator can be computed as shown in Figure S8.4 from the overlap between the history relating the first two populations (red arrows) and the second two (blue arrows). For a Eurasian population, a proportion *n* of their ancestry (their Neandertal-derived ancestry) travels along the lineage where the red and blue arrows overlap, thus contributing an expected value of *nQ* to the numerator where *Q* is a number proportional to the variation in allele frequencies due to random sampling of alleles from generation to generation that occurred on that lineage (genetic drift). For a Neandertal population, all of the ancestry travels along the path with genetic drift *Q*, thus contributing an expected value of *Q* to the denominator. The expected ratio is thus:

$$E[\hat{R}_{Neandertal}] \approx eQ/Q = n \quad \text{(S8.4)}$$

We computed $\hat{R}_{Neandertal}$ for all possible combinations of the 6 Eurasian and 3 African populations, and obtained largely consistent results (Table S8.3). To obtain a maximally precise estimate, we also pooled all reads from the 6 Eurasians at each nucleotide to form an "All Eurasia" pool, and all reads from Africans to form an "All Africa" pool, and obtained of $\hat{R}_{Neandertal}$ = 3.0 ± 0.6%. This estimate is substantially larger than the 1.7 ± 0.2% that we inferred in ref. 1 by computing a similar statistic, where we noted that our estimate was a conservative minimum. To understand why the previous estimate was a minimum whereas the present statistic is unbiased, we observe that the statistic computed in ref. 1 (Equation S18.4 of SOM 18) was very similar to Equation S10.3, with the only difference being that a second Vindija individual was used instead of Denisova. By tracing genetic drift paths as shown in Figure S8.4, it is easy to see that the expectation of the ref. 1 statistic is *n(Q+R)/(Q+R+S)*, whereas it is *nQ/Q* for our new statistic, explaining why the old statistic underestimated *n*. The underestimate may have been substantial, since *S* is the genetic drift on the lineage specific to the Vindija Neandertals, which we showed in SI 6 and SI 7 may be quite large.

Table S8.3: Estimates of Neandertal-related mixture proportion in non-Africans

Eurasian	All Africa	Std. Err.	Mbuti	Std. Err.	San	Std. Err.	Yoruba	Std. Err.
All Eurasia pool	3.0%	0.6%	3.6%	0.9%	2.2%	0.8%	2.3%	0.7%
Cambodian	4.4%	1.0%	4.8%	1.4%	3.1%	1.2%	4.4%	1.1%
French	2.6%	0.7%	2.9%	1.2%	1.7%	0.9%	2.2%	0.8%
Han	3.2%	0.9%	4.3%	1.1%	2.4%	1.0%	2.5%	1.0%
Karitiana	0.9%	1.1%	0.8%	1.7%	0.4%	1.3%	0.8%	1.3%
Mongolian	4.0%	1.0%	2.5%	1.5%	3.4%	1.2%	2.9%	1.1%
Sardinian	2.6%	0.9%	2.7%	1.5%	2.1%	1.2%	2.0%	1.0%

Note: Values for the Karitiana are lower than in other non-Africans but are within 2 standard deviations of the mean.

An estimate of the proportion of Denisova-related ancestry in Melanesians

To estimate the proportion *d* of Denisova-related ancestry in Melanesians, we define $\hat{R}_{Denisova}$:

$$\hat{R}_{Denisova} = \frac{\hat{S}(Melanesian, Eurasian, San\ or\ Mbuti, Chimpanzee)}{\hat{S}(Archaic, Yoruba, San\ or\ Mbuti, Chimpanzee)} \quad \text{(S8.5)}$$

We can compute an expected value for the numerator by tracing of genetic drift paths through the Admixture Graph, as shown in Figure S8.5 (this is analogous to Figure S8.4 where we performed the same type of analysis for $\hat{R}_{Neandertal}$). We define:

n = proportion of Neandertal-related ancestry in all non-Africans
d = proportion of Denisova-related ancestry in Melanesians
a = proportion of the Melanesian genome due to archaic gene flow = 1-(1-*d*)(1-*n*)
y = proportion of the modern human ancestors of present-day non-Africans that are more closely related to San than to Yoruba (this is small, but possibly non-zero)
W = genetic drift on the lineage ancestral to San and Yoruba
X = genetic drift before the divergence of Eurasian ancestors from a putative San-related ancestral population of Eurasians, but after San-Yoruba divergence

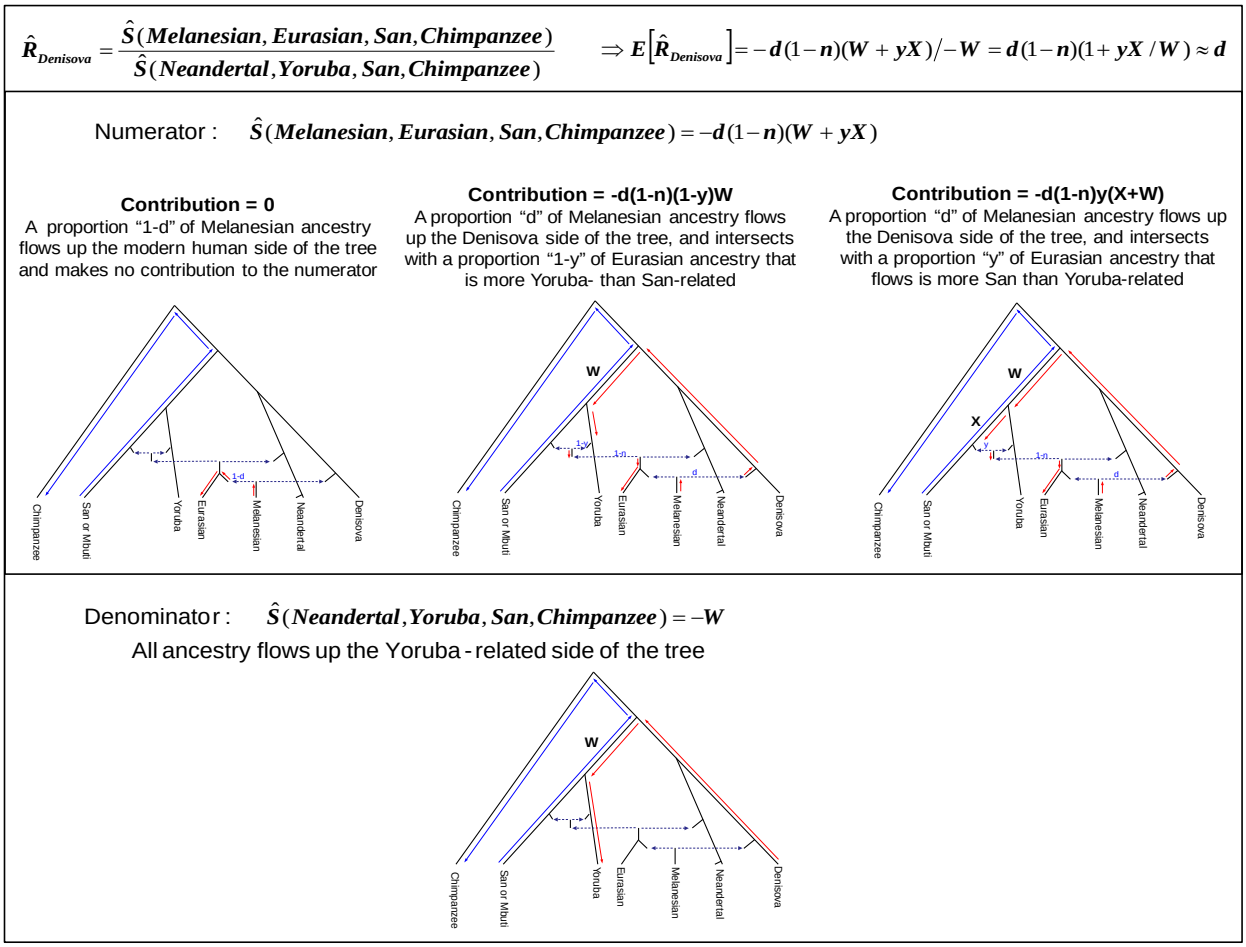


Figure S8.5: The expected values of the R_{Denisova} statistic that is informative for estimating the proportion *d* of Denisova-related ancestry in all non-Africans. To compute the expected value of the statistic, we again trace historical differences through the Admixture Graph of Figure S8.3. Regions where red and blue lines overlap are correlated drifts and contribute to the expectation.

Figure S8.5 shows that $\hat{S}(\textbf{\textit{Melanesian}},\textbf{\textit{Eurasian}},\textbf{\textit{San or Mbuti}},\textbf{\textit{Chimpanzee}})$ is expected to equal -*d(1-n)(W+yX)* and that $\hat{S}(\textbf{\textit{Archaic}},\textbf{\textit{Yoruba}},\textbf{\textit{San or Mbuti}},\textbf{\textit{Chimpanzee}})$ is expected to equal -*W*. The ratio then has an expected value of close to *d*(1-*n*), since we know that *yX/W* is small (*y* <<1 and *X/W* is on the order of 1 or less):

$$\boldsymbol{E}\left[\hat{\boldsymbol{R}}_{\textit{Denisova}}\right]=\frac{-\boldsymbol{d}(1-\boldsymbol{n})(1+\boldsymbol{yX})\boldsymbol{W}}{-\boldsymbol{W}}=\boldsymbol{d}(1-\boldsymbol{n})(1+\boldsymbol{y}\,\boldsymbol{X}/\boldsymbol{W})\approx\boldsymbol{d}(1-\boldsymbol{n}) \tag{S8.6}$$

To obtain an estimator of *d*, we compute the quantity $\hat{\boldsymbol{R}}_{\textit{Denisova}}/\left(1-\hat{\boldsymbol{R}}_{\textit{Neandertal}}\right)$, whose expected value is clearly *d*. To do this, we restrict to nucleotides in the genome covered in six groups: Melanesian, Eurasian, Yoruba, San or Mbuti, Neandertal, Denisova and chimpanzee.

An estimate of the total proportion of archaic ancestry in Melanesians

The total proportion of archaic material *a* in Melanesians, combining the Neandertal-related flow into all Eurasians and the Denisova-related flow specifically into Melanesians, can now be computed as a new statistic:

$$\hat{\boldsymbol{R}}_{\textit{Archaic}}=1-\left(1-\frac{\hat{\boldsymbol{R}}_{\textit{Denisova}}}{1-\hat{\boldsymbol{R}}_{\textit{Neandertal}}}\right)\left(1-\hat{\boldsymbol{R}}_{\textit{Neandertal}}\right)=\hat{\boldsymbol{R}}_{\textit{Neandertal}}+\hat{\boldsymbol{R}}_{\textit{Denisova}} \tag{S8.7}$$

The expected value of this statistic is the total proportion of archaic ancestry in the history of Melanesians. Thus:

$$\boldsymbol{E}[\hat{\boldsymbol{R}}_{\textit{Archaic}}]\approx 1-(1-\boldsymbol{n})(1-\boldsymbol{d})=\boldsymbol{a} \tag{S8.8}$$

Table S8.4 presents joint estimates of the mixture proportions *n*, *d*, and *a* using subsets of the genome where we have data from at least one representative of each of the following six populations: Eurasian, Melanesian, Yoruba, San or Mbuti, Neandertal, Denisova and Chimpanzee. To increase precision, we not only report results for specific combinations of samples, but also for pools of "All Eurasia" (6 samples), "All Melanesia" (3 samples), "All Archaic" (3 Vindija samples and Denisova), and "San or Mbuti" (2 samples).

Pooling all samples, we estimate *n* = E[$\hat{\boldsymbol{R}}_{\textit{Neandertal}}$] = 2.5 ± 0.6%, *d* = E[$\hat{\boldsymbol{R}}_{\textit{Denisova}}/1-\hat{\boldsymbol{R}}_{\textit{Neandertal}}$] = 4.8 ± 0.5%, and *a* = E[$\hat{\boldsymbol{R}}_{\textit{Archaic}}$] = 7.4 ± 0.8%. The estimate of *n* from this computation is not identical to Table S8.3, reflecting the fact that we are analyzing a reduced set of nucleotides to ensure that all analyzed sites also have data available from Melanesians.

An interesting feature of Table S8.4 is that the choice of non-Africans that we analyze results in a somewhat broader range of Denisova-related mixture proportions than would be expected from our standard errors. Varying the Eurasian sample we obtain *d* = 2.3-6.1% (P=0.06 for heterogeneity), and varying the Melanesian sample we obtain *d* = 3.2-6.4% (P=0.0013; Table S8.4). This highlights the possibility that our Block Jackknife standard errors, while statistically valid, may not capture the full uncertainty in our estimates, since there are also systematic errors

due for example to different sample processing, as we have discussed previously[1]. It may also represent different proportions of archaic ancestry in different non-African samples (for example, less in the Bougainville than in the Papuan Melanesians, which is well documented[4]). In any case, our results consistently support two findings: (a) that the Denisova-related mixture specifically into Melanesians *d* is higher than the Neandertal-related mixture proportion *n* in all non-Africans, and (b) the archaic mixture proportion in all Melanesians is in the range a = 6-9%.

Table S8.4: Combined estimates of archaic mixture in present-day humans

Pooling strategy	**Eurasian sample**	**Melanesian sample**	$\hat{R}_{Neandertal}$ **(*n*) percentage of non-African genes contributed by Neandertal relatives**	$\hat{R}_{Denisova}/1-\hat{R}_{Neandertal}$ **(*d*) percentage of Melanesian gene pool contributed by Denisova relatives**	$\hat{R}_{Archaic} = \hat{R}_{Neandertal} + \hat{R}_{Denisova}$ **(*a*) = 1-(1-n)(1-d) = total percentage of Melanesian gene pool of either Neandertal or Denisova origin**
All samples	All Eurasia	All Melanesia	2.5 ± 0.6%	4.8 ± 0.5%	7.4 ± 0.8%
Eurasians separately	Cambodian	All Melanesia	4.3 ± 1.0%	4.5 ± 0.8%	8.7 ± 1.4%
	French	All Melanesia	2.3 ± 0.8%	6.1 ± 0.7%	8.3 ± 1.1%
	Han	All Melanesia	2.9 ± 0.9%	5.2 ± 0.6%	8.2 ± 1.1%
	Karitiana	All Melanesia	1.0 ± 1.1%	2.3 ± 1.1%	3.3 ± 1.7%
	Mongolian	All Melanesia	2.9 ± 1.0%	3.7 ± 0.9%	6.6 ± 1.4%
	Sardinian	All Melanesia	2.3 ± 0.9%	4.8 ± 0.9%	7.0 ± 1.3%
	Test for	*heterogeneity**	*P=0.35*	*P=0.06*	*P=0.14*
Melanesians separately	All Eurasia	Papuan1	1.8 ± 0.7%	6.4 ± 0.6%	8.2 ± 1.0%
	All Eurasia	Papuan2	2.7 ± 0.8%	4.0 ± 0.8%	6.7 ± 1.2%
	All Eurasia	Bougainville	2.5 ± 0.9%	3.2 ± 0.7%	5.7 ± 1.3%
	Test for	*heterogeneity**	*P=0.59*	*P=0.0013*	*P=0.26*

Notes: Africans are represented by Yoruba, San and Mbuti to estimate $\hat{R}_{Neandertal}$, and a San and Mbuti pool to estimate $\hat{R}_{Denisova}$.

* To assess whether the estimates of mixture proportion are heterogeneous, we compute χ^2 tests with 5 degrees of freedom for the Eurasian samples, and 2 degrees of freedom for the Melanesian samples.

References for SI 8

1. Green, R. E. et al., A draft sequence of the Neandertal genome. *Science* **328**, 710 (2010).
2. Li, J.Z. et al., Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100 (2008).
3. Patterson, N. et al., Genetic structure of a unique admixed population: implications for medical research. *Hum Mol Genet* **19**, 411 (2010).
4. Friedlaender, J.S. et al., The genetic structure of Pacific Islanders. *PLoS Genet.* **4**, e19 (2008).
5. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. and Singh, L., Reconstructing Indian population history. Nature **461**, 489 (2009).
6. Busing, F.M.T.A., Meijer, E. and van der Leeden, R., Delete-m jackknife for unequal m. *Statistics and Computing* **9**, 3 (1999).
7. Kunsch, H.K., The jackknife and the bootstrap for general stationary observations. *Ann Statist* **17**, 1217 (1989).
8. Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. and Nielsen, R., Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**, 1496 (2005).

Supplementary Information 9

Low coverage sequencing of seven present-day humans.

Matthias Meyer, Udo Stenzel and Martin Kircher*

* To whom correspondence should be addressed (Martin.Kircher@eva.mpg.de)

Library Preparation

DNA was obtained for each of seven individuals from the CEPH-Human Genome Diversity Panel (HGDP): HGDP00456 (Mbuti**)**, HGDP00998 (Karitiana Native American), HGDP00665 (Sardinia), HGDP00491 (Bougainville Melanesian), HGDP00711 (Cambodian), HGDP01224 (Mongolian) and HGDP00551 (Papuan). DNA was sheared into small fragments (200-400 bp) using a BioRaptur UCD-200 (Diagenode). Shearing was performed four times for seven minutes at a "HIGH" setting with an ON/OFF interval of 30 seconds. Illumina multiplex sequencing libraries were prepared from the sheared samples according to the protocol described by Meyer and Kircher[1]. For each sample (except HGDP00998 which was used without size selection) a narrow band around 300 bp was excised from a 2% agarose gel after adapter ligation to obtain inserts of optimal size for sequencing. DNA was extracted from the gel slices using the QIAquick Gel Extraction kit (Qiagen).

Illumina sequencing of present-day humans

Each of these Illumina multiplex libraries was sequenced using 2×101 + 7 cycles on one flow cell according to the manufacturer's instructions for multiplex sequencing on the Genome Analyzer IIx platform (FC-104-400x v4 sequencing chemistry and PE-203-4001 cluster generation kit v4). The protocol was followed except that an indexed control PhiX 174 library was spiked into each lane, yielding 2-3% control reads in each lane.

The run was analyzed starting from QSEQ sequence files and CIF intensity files from the Illumina Genome Analyzer RTA 1.6 software. The raw reads were aligned to the corresponding PhiX 174 reference sequence to obtain a training data set for the base caller *Ibis*[2], which was then used to call bases and quality scores. The raw paired-end reads were merged (including adapter removal) by checking for at least 11nt overlap between the first and the second read. For bases in the overlapping sequence that disagreed, the base with the highest quality score was used.

This resulted in two sets of reads for each sample: paired-end reads and merged reads. The paired-end reads were aligned using BWA[3] to the human (NCBI36/*hg18*) and chimpanzee (CGSC 2.1/*pantro2*) genomes using default parameters. Using BWA's *sampe* command, the alignments of the first and second read were combined and converted to SAM/BAM format. Merged reads were aligned separately to these genomes using BWA with default parameters. Using BWA's *samse* command, these alignments were also converted to SAM/BAM format[4]. Subsequently, the BAM output files for paired-end and merged reads were merged using samtools[4] and the alignments to *hg18* and *pantro2* were filtered as follows: (a) Non-mapped merged reads and paired-end reads missing at least one alignment were removed. (b) A mapping quality of at least 30 was required. (c) "Duplicated" reads, *i.e.* read pairs for which another read pair of higher or equal quality had boundaries that map to the same outer coordinates, were

removed. (d) Reads with sequence entropy <1.0 were removed, where entropy (a measure of sequence complexity) is calculated by summing $-p*\log_2(p)$ for each of the four nucleotides. Table S9.1 summarizes the number of raw reads, the fraction of aligned reads, as well as the fraction of reads passing the filters for each library.

The index reads used for the sequencing runs were not further evaluated for downstream analysis. However, they were used to validate the correct assignment of samples to lanes.

Table S9.1: Summary statistics for new sequence data from seven present-day humans

Human (*hg18*)	**Mbuti**	**Karitiana**	**Sardinian**	**Bougainville**	**Cambodian**	**Mongolian**	**Papuan2**
	HGDP00456	**HGDP00998**	**HGDP00665**	**HGDP00491**	**HGDP00711**	**HGDP01224**	**HGDP00551**
Raw clusters	30,562,322	31,188,058	34,994,524	37,133,303	39,665,263	35,608,002	36,576,315
Aligned Merged	5,807,398	8,249,296	11,913,805	7,600,477	5,062,341	5,722,286	5,903,815
Aligned reads PE	36,632,890	31,426,808	33,644,705	45,586,586	54,670,715	46,671,621	45,927,452
Fraction aligned	78.93%	76.83%	82.12%	81.85%	81.68%	81.61%	78.92%
Aligned reads (PF)	37,750,055	34,981,052	40,548,399	46,109,678	52,372,445	46,002,098	45,667,754
PE reads	32,395,570	27,389,156	29,638,756	39,323,910	47,784,930	40,830,772	40,308,448
Proper pairs	32,332,660	27,048,544	29,545,348	39,248,370	47,704,528	40,780,536	40,238,502
Other chromosome	43,664	90,248	34,730	51,008	55,204	31,476	45,012
Merged reads	5,354,485	7,591,896	10,909,643	6,785,768	4,587,515	5,171,326	5,359,306
Fraction (Aligned PF)	70.52%	68.25%	73.52%	71.22%	71.8%	71.86%	69.75%
Giga bases (approx.)	3.81	3.53	4.10	4.66	5.29	4.65	4.61
Coverage (div by 2.8 Gb)	1.4	1.3	1.5	1.7	1.9	1.7	1.6

Chimpanzee (*pantro2*)	**Mbuti**	**Karitiana**	**Sardinian**	**Bougainville**	**Cambodian**	**Mongolian**	**Papuan2**
	HGDP00456	**HGDP00998**	**HGDP00665**	**HGDP00491**	**HGDP00711**	**HGDP01224**	**HGDP00551**
Raw clusters	30,562,322	31,188,058	34,994,524	37,133,303	39,665,263	35,608,002	36,576,315
Aligned Merged	5,069,463	7,327,903	10,467,640	6,626,843	4,370,301	4,917,658	5,060,398
Aligned reads PE	33,163,906	28,221,508	30,407,559	41,698,350	49,732,818	42,309,556	42,058,939
Fraction aligned	70.84%	68.74%	73.36%	73.99%	73.71%	73.22%	71.33%
Aligned reads (PF)	32,720,286	30,352,672	34,935,239	40,254,059	45,770,600	39,951,244	39,944,616
PE reads	28,155,518	23,741,424	25,592,806	34,464,458	41,876,372	35,598,738	35,454,732
Proper pairs	28,081,040	23,488,914	25,511,098	34,365,680	41,776,742	35,533,110	35,365,764
Other chromosome	41,650	62,980	33,638	52,414	54,566	34,172	46,964
Merged reads	4,564,768	6,611,248	9,342,433	5,789,601	3,894,228	4,352,506	4,489,884
Fraction (Aligned PF)	61.0%	59.26%	63.26%	62.0%	62.6%	62.21%	60.74%
Giga bases (approx.)	3.30	3.07	3.53	4.07	4.62	4.04	4.03
Coverage (div by 2.8 Gb)	1.2	1.1	1.3	1.5	1.7	1.4	1.4

Note: For mapping to both (top) the human reference genome *hg18*, and (bottom) the chimpanzee reference genome *pantro2*, this table reports the number of raw reads for each library, and various other metrics including the fraction of paired end reads obtained after filtering (PF).

Access to the raw sequence data

The alignments of reads to *hg18* and *pantro2* are available in BAM format from http://genome.ucsc.edu/Denisova.

References for SI 9

1. Meyer, M. and Kircher, M., Ilumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010**, pdb.prot5448 (2010).
2. Kircher, M., Stenzel, U. and Kelso J., Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* **10**, R83 (2009).
3. Li, H. and Durbin, R., Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754 (2009).
4. Li, H., et al., The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (2009).

Supplementary Information 10

Robustness of inferences about population history from *D*-statistics.

David Reich*, Eric Y. Durand, Richard E. Green, Montgomery Slatkin and Nick Patterson

* To whom correspondence should be addressed (reich@genetics.med.harvard.edu)

Our "*D*-statistics" measure the extent to which derived alleles are shared across different sets of hominins. They are important for our comparisons of present-day and archaic hominins, as we used them in ref. 1 to demonstrate that Neandertals share more derived alleles with non-Africans than with Africans, and in this study to demonstrate that Denisovans share more derived alleles with Melanesians than with other non-Africans.

The first line of evidence for the robustness of our *D*-statistics is that we estimate consistent values for them across a large number of present-day humans. This implies that our *D*-statistics are likely to reflect true historical relationships, instead of experimental or data processing artifacts specific to particular samples. In particular, we observe quantitatively consistent *D*-statistics (within about two standard deviations) for all comparisons involving two humans from the same geographic grouping when we analyze 3 Africans, 3 Melanesians, and 6 other non-African and compare them to archaic hominins (Table 1 of the main text). The consistency is also evident in an analysis of genotyping data from 938 individuals from the CEPH-HGDP diversity panel, where we use a less error prone technology but continue to find that present-day humans separate into three discontinuous groups in their relationships to Neandertals and Denisovans (Figure 2 of the main text).

To further test whether the *D*-statistics can support reliable inferences about history, in this Supplementary Note we explore their robustness to: (i) analyses of various subsets of the data, (ii) sequencing error in modern and ancient DNA, and (iii) others metrics of data quality.

Table S10.1: Consistency of D-statistics across transversion substitution classes

Base substitution class	***D*(Eurasian, Melanesian, Denisova, Chimp)**		***D*(Eurasian, African, Neandertal, Chimp)**	
	***D*-statistic**	**Std. Err.**	***D*-statistic**	**Std. Err.**
TA or AT	-3.6%	0.7%	4.4%	0.6%
GC or CG	-5.2%	0.8%	5.8%	0.7%
TG or AC	-6.5%	0.7%	4.9%	0.7%
GT or CA	-3.1%	0.8%	5.7%	0.7%
*Heterogeneity**	*P=0.004*		*P=0.36*	

Note: To maximize precision, we pooled all reads from Eurasians (n=6), Africans (n=3), and Melanesians (n=3), and used the frequencies of each observed allele to compute expected values. Standard errors are from a Block Jackknife (100 blocks).

***** To test for heterogeneity across the 4 transversion substitution classes, we perform a χ^2 test with 3 degrees of freedom.

Consistency of *D*-statistics across base-substitution classes and chromosomes

We began by exploring the robustness of our *D*-statistics across each of the possible transversion substitution classes, not including transitions because of their known susceptibility to deamination-induced damage in ancient DNA[1]. We computed two *D*-statistics: *D*(Eurasian, Melanesian, Denisova, Chimp), which supports a history of gene flow between Denisovans and

Melanesians, and *D*(Non-African, African, Neandertal, Chimp), which represents one of our three lines of evidence that Neandertals are more closely related to non-Africans than to Africans[1]. We pool all reads from the 6 Eurasians, the 3 Melanesians, and the 3 Africans to increase the precision of our *D*-statistics, a procedure that does not bias our results (SI 8).

The *D*-statistic involving Denisova is slightly heterogeneous across the four possible transversion substitution classes (P=0.004), a phenomenon that we have observed before and that might be due to systematic differences in the probability of recurrent mutation[1] (Table S10.1) Encouragingly, however, the sign of the statistic is consistent (-3.1% to -6.5%), and thus consistently supports the findings about population history. The *D*-statistic involving Neandertal is fully consistent across base substitution classes (P=0.36 in a test for heterogeneity).

We also explored the consistency of the *D*-statistics across chromosomes. Table S10.2 shows that the signals are consistent, with formal tests for heterogeneity being non-significant across the 22 autosomes (P=0.11 for the *D*-statistic involving Denisova and P=0.74 for the *D*-statistic involving Neandertal). Interestingly on chromosome X, the *D*-statistic involving Neandertal is somewhat weaker than on the autosomes.

Table S10.2: Consistency of D-statistics across chromosomes

Base substi-tution class	***D*(Eurasian, Melanesian, Denisova, Chimp)**		***D*(Non-Afr., African, Neandertal, Chimp)**	
	***D*-statistic**	**Std. Err.**	***D*-statistic**	**Std. Err.**
1	-2.5%	1.4%	6.1%	1.4%
2	-2.3%	1.6%	4.8%	1.4%
3	-8.1%	2.1%	5.1%	1.7%
4	-4.8%	1.5%	5.8%	1.5%
5	-7.7%	1.8%	4.%	1.7%
6	-9.1%	2.0%	7.9%	1.8%
7	-1.7%	1.8%	2.6%	1.5%
8	-3.5%	1.9%	4.4%	1.9%
9	-1.3%	1.7%	7.5%	2.2%
10	-1.7%	1.9%	4.8%	1.8%
11	-0.8%	1.6%	4.9%	1.8%
12	-4.0%	2.3%	7.6%	2.1%
13	-2.1%	2.3%	1.7%	2.4%
14	-7.8%	2.7%	6.2%	2.5%
15	-4.4%	2.2%	3.%	1.6%
16	-5.1%	2.5%	6.8%	2.1%
17	-10.0%	2.2%	5.7%	2.%
18	-8.3%	2.4%	3.7%	2.3%
19	-2.4%	2.9%	8.3%	2.8%
20	-8.4%	2.5%	3.5%	2.6%
21	-6.9%	2.9%	4.0%	2.8%
22	-8.4%	3.2%	3.3%	2.9%
*Heterogeneity**	*P=0.11*		*P=0.74*	
Autosomes	-4.7%	0.6%	5.2%	0.5%
X	-2.1%	2.3%	0.5%	1.9%

***** To test for heterogeneity across all the autosomes (excluding X), we perform a χ^2 test with 21 degrees of freedom.

Robustness of *D*-statistics to sequencing error

In this section, we explore whether the levels of sequencing error in our data are sufficient to generate the *D*-statistics in Table 1 of the main text that support a history of gene flow. We used *D*(French, Papuan1, Denisova, Chimp = -4.0% ± 0.7%, and *D*(Han, Yoruba, Neandertal, Chimpanzee) = 5.7% ± 0.7% to represent these *D*-statistics in the analyses that follow. Our exploration of the effect of sequencing error is related to that in Appendix I of SOM 15 of ref. 1, and here we extend it to *D*-statistics involving Denisova.

To compute the expected value of the *D*-statistic *D*(H_1, H_2, H_3, Chimpanzee), we need to distinguish between the observed counts, which include the effect of sequencing error, and the counts that we would observe if there were no sequencing error. To do this, we define n_{BABA} as the observed number of biallelic substitutions that cluster H_1 and H_3 (to the exclusion of H_2 and chimpanzee) and similarly define n_{ABBA} as the observed number of biallelic substitutions that cluster H_2 and H_3. We recall that the *D*-statistic is defined as the difference in these quantities normalized by their sum. We use the following notations for expected counts and error rates:

(i) *Notation for the true counts*: We denote the count of each base substitution class in the absence of sequencing error as n_{abcd}, using lower case letters as a contrast to the upper case letters for observed counts. Under the null hypothesis that H_3 is equally closely related to (H_1, H_2) and that mutation rates have been equal in H_1 and H_2 since population divergence, we expect symmetry in the true counts under transposition of samples H_1 and H_2: $n_{abcd} = n_{bacd}$.

(ii) *Notation for sequencing error rates:* We denote the probability of misreading nucleotide "*a*" as "*b*" in individuals H_1, H_2, and H_3 respectively as $e_1^{a\to b}$, $e_2^{a\to b}$, $e_3^{a\to b}$, and we assume that the chimpanzee error rate is negligible: $e_{\text{chimp}}^{a\to b}=0$. As a shorthand, we use $e_k^{a\to *}$ to refer to the probability of misreading nucleotide *a* in individual *k* as one of the 3 alternative alleles.

To estimate the sequencing error rates for each of the 12 possible substitution patterns, we measure the excess branch length compared with the human reference sequence *hg19* since divergence of both from their common genetic ancestors (we use chimpanzee as an outgroup for these analyses). The error rate estimates, which we present in Table S10.3, are substantially lower than those in SI 2 (Table S2.4), because here we remove the 50% of the data with the lowest sequencing base quality (instead of the lowest 5% as in Table S2.4), to match the stringent filters that we developed for the *D*-statistic analysis (SI 6). The only difference to the filtering of SI 6 is that we include nucleotides within 5 bp of the end of the Neandertal read, 1 bp of the end of the Denisova read, and position 34 of the Papuan1 individual (41 on the reverse strand). Thus our estimates of sequencing errors in Table S10.3 are likely to be overestimates compared to the error rates that apply to real data. This means that our estimate of the bias due to sequencing error is conservatively too large.

$$
\begin{aligned}
E[n_{BABA}] \approx\ & n_{baba}\left(1-e_1^{b\to *}-e_2^{a\to *}-e_3^{b\to *}\right)+\left(e_1^{a\to b}e_3^{a\to b}n_{aaaa}\right)+\left(e_1^{a\to b}n_{aaba}+e_1^{c\to b}n_{caba}+e_1^{d\to b}n_{daba}\right)\\
&+\left(e_2^{b\to a}n_{bbba}+e_2^{c\to a}n_{bcba}+e_2^{d\to a}n_{bdba}\right)+\left(e_3^{a\to b}n_{baaa}+e_3^{c\to b}n_{baca}+e_3^{d\to b}n_{bada}\right)\\
E[n_{ABBA}] \approx\ & n_{abba}\left(1-e_1^{a\to *}-e_2^{b\to *}-e_3^{b\to *}\right)+\left(e_2^{a\to b}e_3^{a\to b}n_{aaaa}\right)+\left(e_1^{b\to a}n_{bbba}+e_1^{c\to a}n_{cbba}+e_1^{d\to a}n_{dbba}\right)\\
&+\left(e_2^{a\to b}n_{aaba}+e_2^{c\to b}n_{acba}+e_2^{d\to b}n_{adba}\right)+\left(e_3^{a\to b}n_{abaa}+e_3^{c\to b}n_{abca}+e_3^{d\to b}n_{abda}\right)
\end{aligned}
\qquad \text{(S10.1)}
$$

Table S10.3: Error rates estimates for each substitution class after filtering out 50% of sites

	Transitions				Transversions							
	$e^{A\to G}$	$e^{G\to A}$	$e^{C\to T}$	$e^{T\to C}$	$e^{A\to C}$	$e^{A\to T}$	$e^{G\to C}$	$e^{G\to T}$	$e^{C\to A}$	$e^{C\to G}$	$e^{T\to A}$	$e^{T\to G}$
Denisova * †	.029%	.029%	.029%	.028%	.009%	.004%	.009%	.003%	.003%	.009%	.004%	.009%
Vindija * †	2.45%	.110%	.110%	2.38%	.024%	.023%	.021%	.023%	.022%	.022%	.023%	.023%
Han	.010%	.049%	.048%	.010%	.015%	.010%	.012%	.015%	.015%	.012%	.010%	.014%
Papuan1 *	.013%	.048%	.048%	.013%	.018%	.019%	.010%	.010%	.010%	.010%	.018%	.018%
Yoruba	.009%	.043%	.043%	.009%	.013%	.009%	.010%	.011%	.011%	.009%	.009%	.013%
San	.008%	.044%	.043%	.008%	.009%	.007%	.010%	.010%	.010%	.010%	.007%	.009%
French	.010%	.045%	.045%	.010%	.012%	.008%	.014%	.014%	.014%	.014%	.007%	.012%
Mbuti	.008%	.016%	.016%	.008%	.012%	.004%	.009%	.007%	.007%	.009%	.004%	.012%
Karitiana	.005%	.021%	.021%	.004%	.010%	.003%	.008%	.032%	.032%	.008%	.003%	.010%
Sardinian	.007%	.009%	.009%	.007%	.013%	.002%	.005%	.001%	.002%	.005%	.002%	.013%
Bougainville	.006%	.010%	.010%	.006%	.013%	.002%	.006%	.001%	.001%	.006%	.001%	.013%
Cambodian	.005%	.011%	.010%	.005%	.008%	.001%	.005%	.001%	.001%	.005%	.001%	.007%
Mongolian	.005%	.009%	.009%	.005%	.010%	.001%	.006%	.001%	.001%	.006%	.001%	.010%
Papuan2	.006%	.012%	.012%	.006%	.012%	.002%	.005%	.004%	.004%	.005%	.002%	.012%

Note: To match the data quality filtering used to compute our *D*-statistics, we remove the 50% of nucleotides of lowest quality from the data set in a base- and individual-specific manner (SI 6). The only exception to the filtering of SI 6 is that we do not implement the additional filters based on position in the read. Specifically, for Neandertal we include sites within 5bp of the end of the read, for Denisova within 1 bp of the end, and for Papuan1 we include data from nucleotide 34 (41 on the reverse strand). This means that our estimates of the error rates for these three samples (and thus our estimates of the bias due to sequencing error) is likely to be conservatively too high.

* The error rate estimates for the Vindija and Denisova samples are in theory slightly too small due to us not taking into account the fact that the archaic hominins were interred tens of thousands of years ago, and thus there has been less time for mutations to accumulate on these lineages than on *hg19* since their divergence. When we estimate the effect of this by assuming that the Denisova branch is shortened by 50,000 years and the Neandertal branch by 40,000 years (so that the mutations that are estimated to have accumulated in this time are now interpreted as sequencing error), we find that it only changes the estimates of the transversion error rates by 10% for Denisova and 2% for Vindija, which has a negligible effect on the expected bias due to sequencing error (the expected biases in this table change by $\leq$0.001%).

It is of particular interest that the estimated error rates for the Denisova data are remarkably low, and indeed are quite comparable to our estimated error rates for moder humans.

We now write the expected values of $E[n_{BABA}]$ and $E[n_{ABBA}]$ in terms of the counts multiplied by the error rates (Equation S10.1). All of the terms correspond to at most a single sequencing error, with the exception that we also include a term to represent sites where there are two independent errors from a true *aaaa* pattern (causing it to be misread as either *baba* or *abba*). Monomorphic *aaaa* patterns occur at such a high rate in real data that we wished to explore whether double-errors at such sites could contribute to the expected bias even when multiplied by the (very low) probability of two sequencing errors.

We used Equation S10.1 to compute the expected value of the numerator $E[n_{BABA}-n_{ABBA}]$ and denominator $E[n_{BABA}+n_{ABBA}]$, thus obtaining Equation S10.2. The expectation for the denominator is $2n_{baba} \sim n_{BABA}+n_{ABBA}$ under the approximation that sequencing error makes a relatively small contribution to the sum of these counts. The expectation for the numerator has 8 terms once we take into account the fact that $n_{abcd}=n_{bacd}$ under the null hypothesis of no gene flow. Using the notation $\Delta^{a\to b} \equiv (e_1{}^{a\to b} - e_2{}^{a\to b})$ to denote the difference in the error rates for H_1 and H_2 for a base substitution class, we obtain the following:

$$E[D(H_1, H_2, H_3, Chimp)] \approx \frac{E[n_{BABA} - n_{ABBA}]}{E[n_{BABA} + n_{ABBA}]} \approx \frac{1}{2n_{baba}} \begin{bmatrix} n_{baba}\left(\Delta^{a\to *} - \Delta^{b\to *}\right) + n_{aaaa}\left(e_3{}^{a\to b}\right)\Delta^{a\to b} \\ + n_{aaba}\Delta^{a\to b} + n_{caba}\Delta^{c\to b} + n_{daba}\Delta^{d\to b} \\ - n_{bbba}\Delta^{b\to a} - n_{bcba}\Delta^{c\to a} - n_{bdba}\Delta^{d\to a} \end{bmatrix} \quad \text{(S10.2)}$$

To understand the relative contributions of each of the 8 terms in the brackets of Equation S10.2, we counted all 256 = 4×4×4×4 base patterns that pass the filters of SI 6 for the ordered sets {French, Papuan1, Denisova, Chimpanzee} and {Han, Yoruba, Neandertal, Chimpanzee}. The first ordered set of samples is a "worst case" scenario: the error rate for Papuan1 is higher than for any of the other samples (Table S10.3), and thus any bias due to sequencing error is expected to be worse for this comparison than for comparisons not involving Papuan1.

Table S10.4 presents the expected bias based on all 8 possible *BABA* and *ABBA* base patterns where *A* and *B* are related by a transversion substitution. We observe that the expected bias is dominated by Term 6, reflecting the effect of single errors on the pattern *bbba*. The relative contributions of all the other terms is almost negligible. Importantly, the expected bias for both of the representative *D*-statistics in Table S10.4 is far less than the observed *D*-statistic. For *D*(French, Papuan1, Denisova, Chimpanzee), the expected bias is 0.09% compared with the empirical observation of -4.0 ± 0.7%, and thus is insufficient to generate the observed skew. For *D*(Han, Yoruba, Neandertal, Chimpanzee), the expected bias is -0.09% compared with the observation of 5.7 ± 0.7%, and is thus also insufficient to generate the observed skew.

Table S10.4: Sequencing error is expected to produce negligible bias for two key D-statistics

Term	***D*(French, Papuan1, Denisova, Chimp)**	***D*(Han, Yoruba, Neandertal, Chimp)**
(1) $n_{baba}(\Delta^{a\to *} - \Delta^{b\to *})$	-	-
(2) $n_{aaaa}(e_3{}^{a\to b})\Delta^{a\to b}$	-0.01%	0.02%
(3) $n_{aaba}\Delta^{a\to b}$	-0.01%	0.02%
(4) $n_{caba}\Delta^{c\to b}$	-	-
(5) $n_{daba}\Delta^{d\to b}$	-	-
(6) $-n_{bbba}\Delta^{b\to a}$	0.10%	-0.12%
(7) $-n_{bcba}\Delta^{c\to a}$	-	-
(8) $-n_{bdba}\Delta^{d\to a}$	-	-
Expected bias in D-statistic	0.09%	-0.09%
Observed D-statistic	-4.0 ± 0.7%	5.7 ± 0.7%

Note: This analysis sums over all *BABA* and *ABBA* base patterns where *A* and *B* are related by a transversion. We use "-" to indicate terms that are expected to contribute a negligible bias (<0.005%).

Robustness of *D*-statistics to read alignment bias

We were concerned that *D*-statistics and other important statistics in this study might be affected by alignment bias; differences in the accuracy of mapping of reads from different hominins to the chimpanzee reference sequence. This could cause some reads to appear closer to archaic hominins than others, leading to *D*-statistics that incorrectly suggest gene flow.

Consider a *D*-statistic of the form *D*(H_1, H_2, H_3, Chimpanzee). When the data for H_1 and H_2 are generated using different experimental procedures and subjected to different bioinformatic processing, it is possible that artifacts due to differences in processing will cause one of them to appear more closely related to H_3. However suppose that they are processed identically. In this case, any apparent difference between H_1 and H_2 should be uncorrelated to that between H_3 and

chimpanzee. On this basis, we argue that if (H_1, H_2) use the same sequencing and alignment strategy, biases in the *D*-statistic should be minimal or not arise. Thus, to be very conservative and filter out processing artifacts aggressively, in Table 1 we only report *D*-statistics in which both H_1 and H_2 are compared among the 5 males sequenced in ref. 1, or the 7 males sequenced in this study. (This is perhaps more conservative than is necessary: Table S8.2 presents results for all possible pairwise comparisons and obtains results that are entirely consistent with Table 1.)

To further test whether read alignment bias might be affecting our conclusions about population history based on *D*-statistics, we stratified the data based on the number of sequencing reads covering each site, since it is known that loci with unusual coverage are prone to read-alignment errors resulting from, for example, segmental duplications. We use this as a validation tool, to explore how sensitive our key inferences to alignment bias. We focused on 3 alignments:

Alignment of {French, Yoruba, Neandertal, Chimpanzee}.
Alignment of {any Eurasian, any Melanesian, Denisova, Chimpanzee}.
Alignment of {Neandertal, Denisova, Yoruba, Chimpanzee}.

We explored the sensitivity of two statistics:

(a) $D(H_1, H_2, H_3, \text{Chimpanzee}) = (n_{BABA}-n_{ABBA})/(n_{BABA}+n_{ABBA})$.
If this is significantly different from 0, H_3 shares more derived alleles with H_1 or H_2.

(b) $R(H_1, H_2, H_3, \text{Chimpanzee}) = 2n_{BBAA}/(n_{BABA}+n_{ABBA})$.
This is sensitive to whether H_1 or H_2 are sister groups relative to H_3. If they are sister groups, *R* should be significantly greater than 1 (n_{BBAA} is much greater than n_{BABA} or n_{ABBA}). (This is similar to the *E*-statistic of SI 7, where $E>0$ also shows that H_1 and H_2 are sister groups.)

Table S10.5: Effect of read coverage on a {French, Yoruba, Neandertal, Chimp} alignment

Coverage	$\mathbf{n_{BBAA}}$	$\mathbf{n_{BABA}}$	$\mathbf{n_{ABBA}}$	$\boldsymbol{D = (n_{BABA}-n_{ABBA})/(n_{BABA}+n_{ABBA})}$	**Std. Err.**	**Z-score**	$\boldsymbol{R = 2(n_{BBAA})/(n_{BABA}+n_{ABBA})}$	**% of data**
All data	71,222	23,586	21,499	4.6%	0.6%	8.0	3.16	100%
Stratified by French read coverage								
1	17,418	5,534	5,003	5.0%	1.1%	4.7	3.31	24%
2	17,366	5,637	5,017	5.8%	1.0%	6.0	3.26	24%
3	14,204	4,687	4,318	4.1%	0.8%	5.1	3.15	20%
4	9,959	3,306	2,991	5.0%	1.2%	4.3	3.16	14%
5	5,884	2,032	1,868	4.2%	1.4%	3.0	3.02	8%
6	3,263	1,063	1,067	-0.2%	2.2%	-0.1	3.06	5%
7	1,584	556	529	2.5%	2.4%	1.1	2.92	2%
Stratified by Yoruba read coverage								
1	20,265	6,318	5,875	3.6%	0.9%	4.0	3.32	28%
2	19,154	6,277	5,667	5.1%	1.0%	5.4	3.21	27%
3	14,249	4,726	4,217	5.7%	1.0%	5.6	3.19	20%
4	8,722	2,942	2,749	3.4%	1.1%	3.0	3.07	12%
5	4,738	1,665	1,443	7.1%	1.5%	4.8	3.05	7%
6	2,223	797	735	4.0%	2.0%	2.0	2.90	3%
7	966	389	356	4.5%	2.6%	1.7	2.60	1%
Stratified by Neandertal read coverage								
1	43,047	14,589	13,374	4.3%	0.7%	6.4	3.08	61%
2	18,189	5,851	5,319	4.8%	0.9%	5.1	3.26	25%
3	6,785	2,118	1,841	7.0%	1.5%	4.8	3.43	9%
4	2,172	658	609	3.9%	2.4%	1.6	3.43	3%
5	671	198	186	3.0%	3.7%	0.8	3.49	1%

Neandertals share more derived alleles with non-Africans than with Africans

Table S10.5 shows *D*- and *R*-statistics for the {French, Yoruba, Neandertal, Chimpanzee} alignment. The *D*-statistic, whose significantly positive value is one of the lines of evidence for gene flow from Neandertals into modern humans[1], is within roughly two standard deviations of the genome average regardless of the read coverage of French, Yoruba, or Neandertal. We do observe some sensitivity to alignment errors in the *R*-statistic, which decreases slightly from low coverage (R~3.3) to high coverage (R~3.0) of present-day humans. Whatever the coverage, however, there is strong evidence for French and Yoruba being more closely related to each other than either is to Neandertal (*R*>>1).

Denisovans share more derived alleles with Melanesians than with other Eurasians

Table S10.6 shows the *D*-statistics for all possible {Eurasian, Melanesian, Denisova, Chimpanzee} alignments where the Eurasian and Melanesian samples are sequenced at the same time and using the same protocol (the same comparisons as in Table 1). The *D*-statistic has a consistent negative sign for all 10 comparisons. No consistent trend in the magnitude of the statistics is observed with increasing coverage, showing that the excess sharing of derived alleles between the Denisova individual and Melanesians is not an artifact of bias in read mapping.

Table S10.6: Effect of read coverage on D(Eurasian, Melanesian, Denisova, Chimpanzee)

	French / Pap.1	**Han / Pap.1**	**Karitia. / Pap.2**	**Karitia. / Boug.**	**Sardin. / Pap.2**	**Sardin. / Boug.**	**Cambo. / Pap.2**	**Cambo. / Boug.**	**Mongol. / Pap.2**	**Mongol. / Boug.**
No. sites	177,611	177,590	41,372	43,118	59,952	63,103	70,082	72,884	61,170	63,924
***D* (all data)**	-4.2±0.8	-4.6±0.8	-6.8±1.2	-7.3±1.0	-5.8±0.9	-5.2±0.9	-6.0±1.1	-4.4±0.9	-5.3±0.9	-5.2±0.8
Eurasian coverage (*D*-statistic as %)										
1	-2.8±1.0	-3.6±1.0	-6.8±1.3	-7.3±1.1	-6.0±1.1	-5.5±1.1	-5.5±1.3	-4.3±1.3	-4.7±1.2	-5.2±1.1
2	-4.0±1.0	-4.1±1.1	-7.7±2.1	-8.2±1.9	-5.8±1.4	-5.3±1.4	-6.5±1.2	-5.4±1.3	-6.0±1.3	-5.6±.3
3	-5.3±1.2	-7.0±1.2	-5.6±4.1	-7.3±3.5	-4.8±2.3	-2.6±1.9	-8.0±2.0	-3.1±1.8	-6.5±2.2	-4.4±2.2
4	-5.0±1.3	-5.2±1.4			-6.0±3.5	-8.2±3.3	-7.2±3.4	-4.4±3.0	-9.1±3.5	-2.3±4.0
5	-6.3±1.6	-4.7±1.5								
6	-4.4±1.7	-4.3±2.2								
7	-4.9±2.3									
8	-8.1±2.8									
Melanesian coverage (*D*-statistic as %)										
1	-5.6±0.9	-6.5±0.9	-6.5±1.3	-7.5±1.3	-6.3±1.1	-5.0±1.2	-6.9±1.2	-4.1±1.3	-5.3±1.1	-6.4 ± 1.0
2	-3.4±0.9	-4.5±1.1	-7.6±1.8	-7.1±1.4	-7.3±1.3	-6.4±1.3	-6.0±1.5	-6.1±1.4	-5.2±1.5	-4.7 ± 1.4
3	-4.5±1.4	-4.1±1.3	-9.3±2.7	-8.7±2.6	1.6±2.2	-3.3±2.1	-2.0±2.2	-3.7±2.1	-6.9±2.0	-2.3 ± 2.1
4	-2.6±1.5	-0.8±1.4	-2.9±3.4	-1.0±4.2	-5.2±3.3	0.0±3.0	-2.1±3.3	0.3±2.5	-3.7±3.1	-1.5 ± 3.0
5	-1.5±1.8	1.8 ± 2.1								
6	0.7 ± 3.0	-2.6±2.9								
Denisova coverage (*D*-statistic as %)										
1	-3.8±0.9	-4.0±0.9	-8.1±1.9	-9.4±1.7	-6.4 ± 1.6	-5.6±1.4	-6.8±1.6	-5.1±1.4	-4.5 ± 1.6	-4.1 ± 1.2
2	-4.2±1.0	-5.7±1.0	-4.6±2.0	-6.4±1.9	-3.9 ± 1.4	-4.6±1.7	-5.4±1.5	-4.6±1.6	-6.8 ± 1.6	-7.5 ± 1.6
3	-5.3±1.1	-5.8±1.2	-6.7±2.0	-6.3±1.9	-5.8 ± 1.9	-6.1±1.6	-7.3±1.7	-4.1±1.6	-5.3 ± 1.8	-5.0 ± 1.7
4	-6.0±1.4	-4.7±1.7	-9.7±2.7	-5.7±2.5	-6.8 ± 2.2	-6.8±2.0	-6.2±2.3	-6.8±2.0	-1.7 ± 2.2	-4.5 ± 2.2
5	-1.3±1.9	-3.2±1.8	-7.0±3.6	-8.7±3.6	-6.8 ± 2.7	-0.2±2.7	-5.6±2.3	-3.3±2.5	-6.4 ± 3.0	-3.0 ± 3.1
6	-5.4±2.5	-1.5±2.9	-2.3±4.7	-4.8±4.3	-5.1 ± 3.6	-5.3±3.9	-0.4±3.6	0.5±3.5	-9.0 ± 4.0	-5.9 ± 3.8
7			-7.9±5.9	-9.8±5.8	-14.1 ± 5.1	-6.7±5.1	-2.2±5.2	-3.0±4.7	11.3 ± 4.8	-13.5 ± 4.5

Note: Cells with <1% of data are blank. The "*D* (all data)" line does not exactly match the *D*-statistics reported in Table 1, since there we randomly sample a read from each individual to represent each site, and here we average across all possible samplings to maximize precision (Equation S8.1).

Denisovans and Neandertals are sister groups

Table S10.7 shows the results for the {Denisova, Neandertal, Yoruba, Chimpanzee} alignment. In contrast to what is observed for Table S10.5 and Table S10.6, the *D*-statistics are not

qualitatively consistent. In particular, when the data are stratified by the coverage of Denisova reads, *D* is positive for low coverage and negative for high coverage, and when the data are stratified by the coverage of Neandertal reads, the sign is negative for low coverage and positive for high coverage. There is no reason that demographic history could cause a dependence of the sign of the *D*-statistic on read coverage. Thus, Table S10.7 shows that the *D*(Denisova, Neandertal, Yoruba, Chimpanzee) statistic is not robust to alignment bias or mapping error. This further highlights the fact that to be confident in our inferences based on *D*-statistics, the first two samples used to compute the statistic (H_1, H_2) should be experimentally processed in the same way. This is not the case for Denisova and Neandertal sequences, as they differ in read length, sequence error profiles, and the way that they were mapped.

While the *D*-statistics in Table S10.7 are not robust to coverage, the *R*-statistics are positive regardless of read coverage. Thus, the *R*-statistic is adequate for showing that Denisovans and Neandertals are sister groups relative to modern humans, supporting the validity of the analyses presented in SI 7. (In SI 6, we also present an independent analysis, based on sequence divergence, showing that Neandertals and Denisovans are sister groups in the nuclear genome.)

Table S10.7: Effect of read coverage on a {Denisova, Neandertal, Yoruba, Chimp} alignment

Coverage	**n_{BBAA}**	**n_{BABA}**	**n_{ABBA}**	***D* = (n_{BABA}-(n_{BABA}+n_{ABBA})**	**Std. Err.**	**Z-score**	***R* = 2(n_{BBAA})/(n_{BABA}+n_{ABBA})**	**% of data**
All data	48531	22694	24551	-3.9%	0.7%	-5.4	2.05	100%
Stratified by Denisova read coverage								
1	17136	8156	7854	1.9%	0.9%	2	2.14	35%
2	13654	6099	6668	-4.5%	1.0%	-4.5	2.14	28%
3	8689	4008	4638	-7.3%	1.2%	-6.2	2.01	18%
4	4745	2224	2626	-8.3%	1.5%	-5.6	1.96	10%
5	2369	1085	1399	-12.7%	1.8%	-7	1.91	5%
6	1049	526	661	-11.3%	2.6%	-4.3	1.77	2%
7	415	235	309	-13.6%	3.8%	-3.6	1.52	1%
Stratified by Neandertal read coverage								
1	29882	13348	15097	-6.2%	0.8%	-7.5	2.10	61%
2	12386	5854	6152	-2.5%	1.0%	-2.5	2.06	25%
3	4281	2252	2190	1.4%	1.4%	1	1.93	9%
4	1348	792	709	5.5%	2.1%	2.6	1.80	3%
5	380	263	220	8.9%	3.5%	2.6	1.57	1%
6	117	92	80	7.2%	5.3%	1.4	1.36	0%
7	53	40	34	7.3%	7.9%	0.9	1.44	0%
Stratified by Yoruba read coverage								
1	12989	6623	6956	-2.5%	0.9%	-2.6	1.91	28%
2	13105	6025	6491	-3.7%	0.9%	-4.1	2.09	27%
3	9893	4357	4870	-5.6%	1.2%	-4.7	2.14	20%
4	6177	2712	2996	-5.0%	1.2%	-4.1	2.16	12%
5	3383	1448	1559	-3.7%	1.8%	-2.1	2.25	7%
6	1599	747	791	-2.9%	2.1%	-1.4	2.08	3%
7	730	318	378	-8.7%	3.2%	-2.7	2.10	1%

References for SI 10

[1] Green, R. E. et al., A draft sequence of the Neandertal genome. *Science* **328**, 710 (2010).

Supplementary Information 11

A population genetic model fit to the data.

Eric Y. Durand and Montgomery Slatkin*

* To whom correspondence should be addressed (slatkin@berkeley.edu)

In this section we develop and analyze a mathematical model of population history compatible with the relationships observed among the Denisova individual, Vindija Neandertals and present-day humans. In accordance with SI 6, we assume that Denisova (D) and Vindija (V) are sister groups. After the divergence of Denisova and Vindija, there was gene flow between Vindija Neandertals and the ancestor of non-African populations represented by French (F), Han (H) and Melanesians (M), and between Denisovans and the ancestors of Melanesians. This model is represented in Figure 3 of the main text. In what follows, we first derive the analytical expectation of different *D* statistics under our model, and use the observed values of these *D* statistics to estimate the parameters of our model. Next, we consider alternative models of population history that could fit the data and illustrate why we favor our first model. Finally, we show that the data do not enable us to distinguish between two alternative explanations of the discordance between the mtDNA and the nuclear histories, namely incomplete lineage sorting and admixture between Denisovans and a more diverged hominin.

Analytical expectation of *D* statistics for a given set of parameters

We assume that we have one nucleotide site sampled from three populations (P_1, P_2 and P_3), and from a chimpanzee (*C*). The three populations may represent present-day humans, Vindija Neandertals (V) or Denisovans (D). By assumption, the chimpanzee carries the ancestral nucleotide (denoted A), and we restrict our analysis to biallelic polymorphisms where P_3 carries the derived nucleotide (denoted B). There are two cases of interest: P_1 has the ancestral nucleotide and P_2 has the derived (denoted ABBA) or the reverse (denoted BABA). We then focus on the test statistic $D(P_1, P_2, P_3, C)$=[Pr(ABBA)–Pr(BABA)]/[Pr(ABBA)+Pr(BABA)]. In particular, we are interested in deriving the analytical expectations of *D*(Afr, F, D, C), *D*(Afr, F, V, C) and *D*(Afr, M, V, C), where Afr is San or Yoruba. We note that replacing French by Han does not change the above *D* statistics in our model.

Expected values of ABBA and BABA for populations (Afr, F, V, C)

Here we detail the derivation of the expected value of *D*(Afr, F, V, C). The derivation is similar to that in SOM 19 of ref 1, and is presented here for completeness. We then provide the analytical expectations of the three other *D* statistics of interest.

There are three genealogical scenarios under the model of Figure 3 that can create either ABBA or BABA sites:

1. *The F lineage traces its ancestry through the modern human side of the phylogeny (probability 1-f_1), and between t_{Afr} and t_{V}, the Afr and F lineages do not coalesce (probability $\left(1-1/(2N)\right)^{t_{\mathrm{V}}-t_{\mathrm{Afr}}}$).*

   In this case, the Afr and F lineages trace their ancestry back to the modern human-Vindija ancestral population without coalescing, so that all three lineages (Afr, F and V) coalesce in the ancestral population. The expected time between the first and second

coalescent events (which can produce ABBA or BABA sites) is 2*N*. With probability 1/3, the first coalescence is between Afr and V or F and V, producing the two configurations of interest. Thus:

$$\Pr_1(\text{ABBA}) = \Pr_1(\text{BABA}) = \left(1 - \frac{1}{2N}\right)^{t_V - t_{Afr}} \frac{2N\mu}{3}\left(1 - f_1\right). \quad \text{(S11.1)}$$

2. *The F lineage traces its ancestry through the Vindija side of the phylogeny (probability f_1), and between t_{GF1} and t_V the two Vindija lineages do not coalesce (probability $\left(1-1/(2N)\right)^{t_V - t_{GF1}}$).*

In this case, the Afr and F lineages trace their ancestry back to the modern human-Vindija ancestral population without coalescing, so the three lineages (Afr, F and V) coalesce in the ancestral population. The expected time between the first and second coalescent events (which can produce ABBA or BABA sites) is 2*N*. With probability 1/3 the first coalescence is between Afr and V or F and V, producing the two sites of interest. Thus:

$$\Pr_2(\text{ABBA}) = \Pr_2(\text{BABA}) = \frac{2N\mu f_1}{3}\left(1 - \frac{1}{2N}\right)^{t_V - t_{GF1}}. \quad \text{(S11.2)}$$

3. *The F lineage traces its ancestry through the Vindija side of the phylogeny (probability f_1), and between t_{GF1} and t_V the two Vindija lineages coalesce.* This history results only in ABBA sites (never BABA sites in the absence of recurrent mutation in the same genealogy). The probability that there is coalescence before t_V is $\left[1-\left(1-1/(2N)\right)^{t_V - t_{GF1}}\right]$. Once the coalescence occurs, the ancestral lineage cannot coalesce with Afr before t_V. After t_V, the average coalescence time is 2*N*. Therefore, the expected length of the internal branch is $t_V + 2N - \left(t_{GF1} + \bar{t}\right)$, where $\bar{t}$ is the expected coalescence time in the Vindija lineage, given that the coalescence occurs before t_V. A little analysis shows that:

$$\bar{t} = 2N - \frac{(t_V - t_{GF1})\left(1 - \frac{1}{2N}\right)^{t_V - t_{GF1}}}{1 - \left(1 - \frac{1}{2N}\right)^{t_V - t_{GF1}}} \quad \text{(S11.3)}$$

Thus:

$$\Pr_3(\text{ABBA}) = \mu f_1\left(t_V - t_{GF1}\right). \quad \text{(S11.4)}$$

The overall probability of ABBA and BABA is obtained by adding. The mutation rate term μ cancels and we obtain,

$$E\left[D(\text{Afr, F, V, C})\right] = \frac{\Pr(\text{ABBA}) - \Pr(\text{BABA})}{\Pr(\text{ABBA}) + \Pr(\text{BABA})}$$

$$= \frac{3f_1\left(t_V - t_{GF1}\right)}{3f_1\left(t_V - t_{GF1}\right) + 4N(1-f_1)\left(1-\frac{1}{2N}\right)^{t_V - t_{Afr}} + 4Nf_1\left(1-\frac{1}{2N}\right)^{t_V - t_{GF1}}}. \quad \text{(S11.5)}$$

If there is no gene flow, f_1=0, $E(D)$=0.

The expectations of the other D statistics are obtained using a similar reasoning. The *D* statistic involving one African lineage, the French (or Han) lineage and the Denisova lineage is obtained by substituting t_{GF1} by t_D in Equation S11.5. Indeed, in the case where the F lineage traces its ancestry though the Vindija side of the phylogeny, the F and D lineages cannot coalesce more recently than t_D. Thus,

$$E\left[D(\text{Afr, F, D, C})\right] = \frac{3f_1\left(t_V - t_D\right)}{3f_1\left(t_V - t_D\right) + 4Nf_1\left(1-\frac{1}{2N}\right)^{t_V - t_D} + 4N\left(1-f_1\right)\left(1-\frac{1}{2N}\right)^{t_V - t_{Afr}}}. \quad \text{(S11.6)}$$

The *D* statistic involving one African lineage, a Melanesian lineage and the Vindija lineage is more complicated because the Melanesian population undergoes two successive events of gene flow. Its expectation is equal to

$$E\left[D(\text{Afr, M, V, C})\right] = \frac{X}{Y}, \text{ where} \quad \text{(S11.7)}$$

$$X = 3f_2\left(t_V - t_D\right) + 3f_1\left(1-f_2\right)\left(t_V - t_{GF1}\right) \text{ and}$$

$$Y = 3f_2\left(t_V - t_D\right) + 3f_1\left(1-f_2\right)\left(t_V - t_{GF1}\right) + 4Nf_1\left(1-f_2\right)\left(1-\frac{1}{2N}\right)^{t_V - t_{GF1}}$$
$$+4N\left(1-f_1\right)\left(1-f_2\right)\left(1-\frac{1}{2N}\right)^{t_V - t_{Afr}} + 4Nf_2\left(1-\frac{1}{2N}\right)^{t_V - t_D}.$$

Estimating the proportions of gene flow

In this section, we derive the expected value of the *R* statistics (defined in SI 8) under our model. *R* statistics are calculated as the ratio between the numerators of two *D* statistics. We denote

$S(P_1, P_2, P_3, C)$=Pr(ABBA)-Pr(BABA) the numerator of $D(P_1, P_2, P_3, C)$. Here, we use S and Y to denote the San and Yoruba lineages, respectively. We first derive the expected value of

$$R_{\text{Neandertal}} = \frac{S(\text{S, F, D, C})}{S(\text{S, V, D, C})}. \tag{S11.8}$$

In the previous section, we showed that $S(\text{S, F, D, C}) = \mu f_1 \left(t_V + \tau_V - \left(t_D + \bar{t}\right)\right)$, where $\bar{t}$ is the expected coalescence time in the Vindija lineage, given that the coalescence occurs before t_V. τ_V is the expected coalescence time in the population ancestral to both modern humans and archaic hominins. When we derived D statistics, we took $\tau_V = 2N$ because we assumed a constant population size. To derive S(S, V, D, C), it is sufficient to take f_1=1 in S(S, F, D, C) (indeed, F and V are identical populations for f_1=1). Thus, in agreement with SI 8, we find that

$$R_{\text{Neandertal}} = f_1. \tag{S11.9}$$

This statistics does not depend on any other parameter of the model. We also note that this result does not require an assumption of a constant population size in ancestral populations, which increases the generality of the estimate of f_1 as noted in SI 8.

Now, we derive the expected value of

$$R_{\text{Denisova}} = \frac{S(\text{F, M, S, C})}{S(\text{Y, V, S, C})}. \tag{S11.10}$$

There are six classes of event that produce the patterns ABBA and BABA for the populations F, M, S, and C.

1. *The M lineage originated from the Denisova population (probability f_2), and the F lineage originated from the Vindija population (probability f_1).* This event produces ABBA and BABA with equal probability.

   $$\Pr_1(\text{ABBA}) = \Pr_1(\text{BABA}) = f_1 f_2 \left(1 - \frac{1}{2N}\right)^{t_V - t_D} \frac{2N\mu}{3}. \tag{S11.11}$$

   This corresponds to the case where both the M and F lineages originated in archaic populations but did not coalesce before t_V.

2. *The M lineage originated from the Denisova population (probability f_2), and the F lineage traces its ancestry through the modern human side of the phylogeny (probability $1-f_1$).* This event produces ABBA and BABA with probabilities:

$$\Pr_2(\text{ABBA}) = f_2(1-f_1)\mu\left(1-\frac{1}{2N}\right)^{t_V-t_{Afr}}\frac{2N\mu}{3},$$
$$\Pr_2(\text{BABA}) = f_2(1-f_1)\mu\left(t_V - t_{Afr} + 2N - t^* + \left(1-\frac{1}{2N}\right)^{t_V-t_{Afr}}\frac{2N\mu}{3}\right). \quad \text{(S11.12)}$$

The extra term in BABA corresponds to the case where the F and S lineages coalesced between times t_{Afr} and t_V. We denote by t^* the expected coalescence time of two lineages in the population ancestral to modern humans, given that they coalesced before time t_V.

3. *The M lineage traces its ancestry through the modern human side of the phylogeny (probability $(1-f_2)(1-f_1)$), and the F lineage traces its ancestry through the modern human side of the phylogeny (probability $1-f_1$).* This event produces ABBA and BABA with equal probability:

$$\Pr_3(\text{ABBA}) = \Pr_3(\text{BABA}) = (1-f_1)^2(1-f_2)\left(1-\frac{1}{2N}\right)^{t_{Afr}-t_M}\frac{2N\mu}{3}. \quad \text{(S11.13)}$$

This corresponds to the case where both M and F originated from the population ancestral to all non-African modern humans and did not coalesce before t_{Afr}.

4. *The M lineage traces its ancestry through the modern human side of the phylogeny (probability $(1-f_2)(1-f_1)$), and the F lineage originated from the Vindija side (probability f_1).* This event produces ABBA and BABA with probabilities:

$$\Pr_4(\text{ABBA}) = f_1(1-f_1)(1-f_2)\mu\left(t_V - t_{Afr} + 2N - t^* + \left(1-\frac{1}{2N}\right)^{t_V-t_{Afr}}\frac{2N\mu}{3}\right),$$
$$\Pr_4(\text{BABA}) = f_1(1-f_2)\mu\left(1-\frac{1}{2N}\right)^{t_V-t_{Afr}}\frac{2N\mu}{3}. \quad \text{(S11.14)}$$

The extra term in ABBA corresponds to the case where the Melanesian and San lineages coalesced between times t_{Afr} and t_V.

5. *The M and the F lineages originated from the Vindija population (probability $(1-f_2)f_1^2$).* This event contributes to ABBA and BABA equally:

$$\Pr_5(\text{ABBA}) = \Pr_5(\text{BABA}) = f_1^2(1-f_2)\left(1-\frac{1}{2N}\right)^{t_V-t_D}\frac{2N\mu}{3}. \quad \text{(S11.15)}$$

6. *The M lineage originated from the Vindija population (probability $(1-f_2)f_1$), and the F lineage traces its ancestry through the modern human side of the phylogeny (probability $1-f_1$).* This event produces ABBA and BABA with probabilities:

$$\Pr_6(\text{BABA}) = f_1(1-f_1)(1-f_2)\mu\left(t_V - t_{Afr} + 2N - t^* + \left(1-\frac{1}{2N}\right)^{t_V - t_{Afr}}\frac{2N\mu}{3}\right),$$

$$\Pr_6(\text{ABBA}) = f_1(1-f_2)\mu\left(1-\frac{1}{2N}\right)^{t_V - t_{Afr}}\frac{2N\mu}{3}. \quad \text{(S11.16)}$$

Thus we have:

$$S(\text{F, M, Afr, C}) = \Pr(\text{ABBA}) - \Pr(\text{BABA}) = \mu f_2(1-f_1)\mu(t_V - t_{Afr} + 2N - t^*). \quad \text{(S11.17)}$$

To derive the denominator of R_{Denisova}, we note that the extra opportunity for the Yoruba and San lineages to coalesce between times t_{Afr} and t_V creates an excess of BABA for populations Y, V, S, C. Therefore we have :

$$S(\text{Y, V, S, C}) = \Pr(\text{ABBA}) - \Pr(\text{BABA}) = \mu(t_V - t_{Afr} + 2N - t^*). \quad \text{(S11.18)}$$

Finally, we find that

$$R_{\text{Denisova}} = f_2(1-f_1). \quad \text{(S11.19)}$$

Although we assumed constant population size to simplify some of the derivations, this result also fully holds for arbitrary varying population size. Using the same numerical values as in SI 8, we find that observed R statistics are consistent with f_1 = 1.3 – 3.7% and f_2 = 3.8 – 5.8%, in agreement with the findings of SI 8.

Estimating other parameters from D statistics

The D statistics derived above depend on the following parameters: N, t_V, t_{Afr}, t_D, t_{GF1}, t_{GF2}, f_1, and f_2. From a sensitivity analysis (not shown), we concluded that the parameters that have the strongest effect on the D statistics are: N, t_V, t_D, f_1 and f_2. The ancestral population size cannot be reliably estimated from D statistics. Thus, we will estimate other parameters for different values of N. For the remaining parameters, we used the same values of t_{Afr}=3,000, t_{GF1}=2,500 and t_{GF2}=1000 (in generations) that were used in SOM 19 of ref. 1. Changes in those parameters have little impact on the D statistics of interest. We already estimated f_1 = 1.3-3.7% and f_2 = 3.8-5.8%.

Table S11.1: Range of parameters values compatible with the observed D-statistics

Parameters	Accepted range
N	6000 – 16500
t_V	9500 – 26000
t_D	6500 – 21500

We explored all permutations for a broad range of values for the three important parameters. We accepted every combination that was within the confidence intervals of the four *D* statistics (observed value ± twice the standard error). Table S11.1 summarizes the accepted values for the 3 parameters. We also present the best fit parameters for different values of the ancestral population size *N* (Table S11.2).

Table S11.2: Point estimates of key parameters using D-statistics

Parameters	N = 8,000	N = 10,000	N = 12,000	N = 15,000
t_V	13,450	16,150	18,800	22,850
t_D	9,350	11,000	12,650	15,200

Note: To constrain our parameter estimates, we use the empirical observations *D*(S, F, D, C) = 0.018, *D*(S, F, V, C) = 0.047, *D*(S, P, V, C) = 0.065 and *D*(V, D, S, C) = -0.040.

Remarks on ancestral population size

In the section above, we assumed that the ancestral population size was constant at all epochs. Here we make a few remarks about the effects of relaxing this assumption. Denote N_H as the ancestral population size of modern humans, N_D as the size of the population ancestral to Vindija and Denisova and N_V as the size ancestral to modern humans, Vindija and Denisova.

- Varying N_V has a symmetrical effect on ABBA and BABA. Thus it does not affect the numerator of the *D* statistics. However, when N_V grows, the branch lengths of the gene tree increase, increasing both ABBA and BABA counts. Therefore, a larger value of N_V leads to a lower value of *D*, requiring a larger value of t_V and t_D to match the data. Overestimating N_V leads to overestimating the times t_V and t_D.
- N_H has the opposite effect of N_D. If t_H and t_D are close, which is suggested by the very similar genetic divergence estimates shown in SI 2 and SI 6, then N_H and N_D will have effects of the same magnitude on *D* statistics

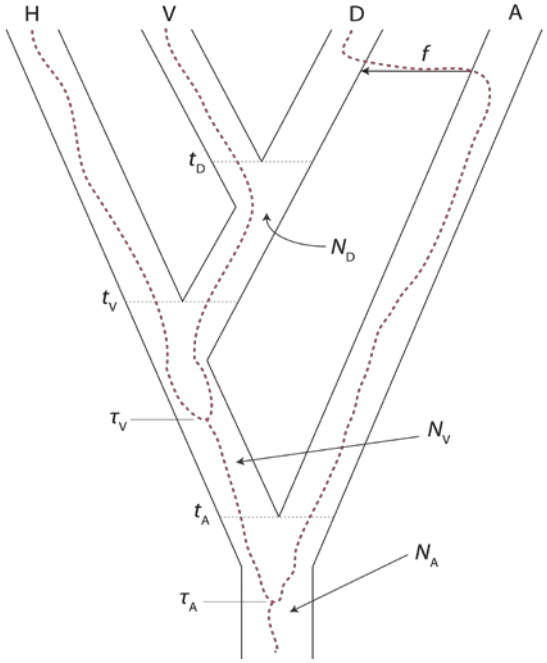


Figure S11.1 Does the discordance of mtDNA and nuclear histories occur due to incomplete lineage sorting or admixture with a more diverged hominin? The red dashed line represents the case of admixture.

Discordance of mtDNA and nuclear histories

If the model illustrated in Figure 3 is correct, there is still the question of why the mtDNA gene tree is discordant with the population tree inferred from the nuclear data. Here we ask

whether the deep divergence of the mtDNA lineage might be a result of admixture with an ancient hominin, for which we have no data, or of incomplete lineage sorting in the common ancestral population. We show that given reasonable assumptions about the size of the ancestral populations, the discordance can be explained either by admixture or incomplete lineage sorting. Therefore, the discordance of the mtDNA and nuclear gene genealogies does not necessarily imply that there was gene flow from an ancient hominin into Denisovans, although it is consistent with that hypothesis.

We are interested in computing the probability that Denisovan mtDNA lineage is derived from an extinct, more ancient hominin (A), given the observed mtDNA pattern. Define τ_V as the coalescence time between H and V, and τ_A as the coalescence of the resulting ancestral lineage with D. We assume that the ancestral population size is constant and equal to N_A. The notation is shown in Figure S11.1. In the section that follows, we allow for the possibility that the ancestral effective population sizes may be unequal (relaxing the assumption depicted in Figure 3).

We denote the probability that the observed mtDNA pattern occurred because of admixture by $\Pr(\text{Ad.} \mid \text{Data})$. Bayes' formula shows that

$$\Pr(\text{Ad.} \mid \text{Data}) = \frac{f \Pr(\text{Data} \mid \text{Ad.})}{f \Pr(\text{Data} \mid \text{Ad.}) + (1-f)\Pr(\text{Data} \mid \text{No Ad.})}, \tag{S11.20}$$

where f is the probability of admixture from A to D.

First, we derive the probability of data in the case of no admixture. It is the product of the probabilities that V and D did not coalesce between t_D and t_V, that V and H coalesced first at time τ_V and that the remaining coalescence occurred at time τ_A:

$$\Pr(\text{Data} \mid \text{No Ad.}) = \frac{1}{N_A^2} \exp\left(-\frac{t_V - t_D}{N_D}\right) \exp\left(-\frac{3(\tau_V - t_V)}{N_A}\right) \exp\left(-\frac{\tau_A - \tau_V}{N_A}\right). \tag{S11.21}$$

We now derive the probability of data when admixture occurs. Given that V and H already coalesced, the probability that D coalesced at time τ_A is equal to

$$\Pr(\tau_A \mid \tau_V, N_A, \text{Ad.}) = \frac{1}{N_A} \exp\left(-\frac{\tau_A - \tau_V}{N_A}\right). \tag{S11.22}$$

The probability that V and H coalesced at time τ_V is the sum of two terms, depending whether the first coalescence event occurred after or before t_A:

$$\Pr\left(\tau_V \mid N_A, t_V, t_A, \text{Ad.}\right) = \Pr\left(\tau_V, \text{coal}_1 < t_A \mid N_A, t_V, t_A, \text{Ad.}\right) + \Pr\left(\tau_V, \text{coal}_1 > t_A \mid N_A, t_V, t_A, \text{Ad.}\right), \tag{S11.23}$$

where coal_1 is the first coalescence event. Before t_A, V and H are the only two lineages present in the same population. Thus,

$$\begin{aligned}\Pr\left(\tau_V, \text{coal}_1 < t_A \mid N_A, t_V, t_A, \text{Ad.}\right) &= \Pr\left(\tau_V \mid N_A, t_V, t_A, \text{Ad.}\right) \mathbb{I}\left(\tau_V < t_A\right) \\ &= \frac{1}{N_A} \exp\left(-\frac{\tau_V - t_V}{N_A}\right) \mathbb{I}\left(\tau_V < t_A\right).\end{aligned} \tag{S11.24}$$

If no coalescence event occurred before t_A, there are three lineages (V, H and D) present in the same population. Thus, the probability that V and H coalesced at time τ_V has to be multiplied by the probability that V and H coalesced first. Therefore, we find that

$$\begin{aligned}\Pr\left(\tau_V, \text{coal}_1 > t_A \mid N_A, t_V, t_A, \text{Ad.}\right) &= \frac{1}{3} \times \Pr\left(\tau_V \mid N_A, t_V, t_A, \text{Ad.}\right) \mathbb{I}\left(\tau_V > t_A\right) \\ &= \frac{1}{3} \times \frac{3}{N_A} \exp\left(-\frac{3\left(\tau_V - t_A\right)}{N_A}\right) \mathbb{I}\left(\tau_V > t_A\right).\end{aligned} \tag{S11.25}$$

The probability that V and H coalesced at time τ_V is then:

$$\Pr\left(\tau_V \mid N_A, t_V, t_A\right) = \frac{1}{N_A} \exp\left(-\frac{\tau_V - t_V}{N_A}\right) \mathbb{I}\left(\tau_V < t_A\right) + \frac{1}{N_A} \exp\left(-\frac{3\left(\tau_V - t_A\right)}{N_A}\right) \mathbb{I}\left(\tau_V > t_A\right). \tag{S11.26}$$

Finally we have

$$\Pr\left(\text{Data} \mid \text{Ad.}\right) = \Pr\left(\tau_A \mid \tau_V, N_A\right) \times \Pr\left(\tau_V \mid N_V, N_A, t_V, t_A\right). \tag{S11.27}$$

Remark. Interestingly, the probability of admixture given the observed pattern does not depend on the exact value of τ_A. All the information is contained in t_A, and the terms containing τ_A

cancel in the numerator and the denominator. While this result is counterintuitive, it is explained by the fact that, if there was admixture, coalescence cannot occur before t_A. Also, this result is no longer true if we assume that the population size of the ancestral population that lived between t_V and t_A is different from N_A.

Numerical example.
We use the following numerical values for our parameters (times in generations):
t_D = 9,150, t_V = 12,500, t_A = 30,000, τ_V = 40,000, τ_A = 24,000, f = 0.01.

We note that N_A, the effective number of females, is expected to be half of the effective population size, assuming a freely mixing population where males and females have the same variance in their number of offspring.

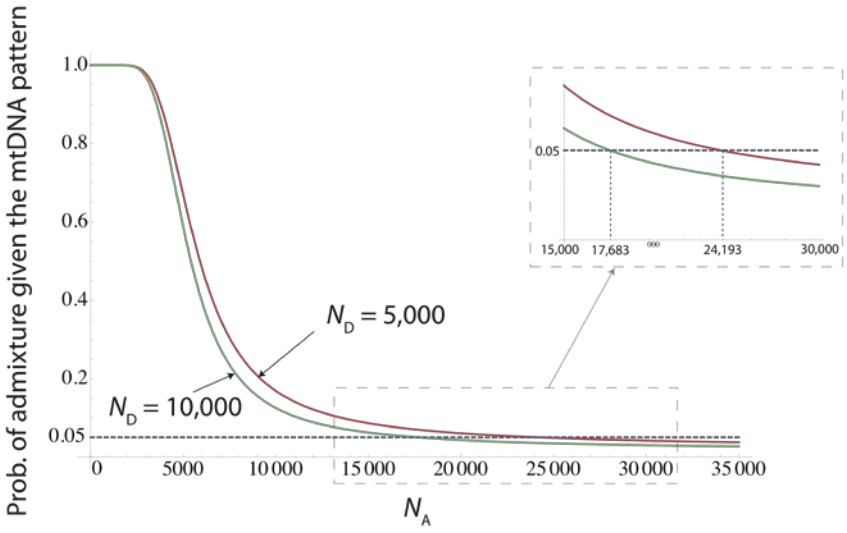


Figure S11.2: Probability of admixture between a more diverged hominid and Denisovans given the mtDNA pattern, as a function of the effective number of females in the ancestral population.

We now plot the probability that the mtDNA pattern occurred because of admixture as a function of the effective number of females in the ancestral population, N_A. Using N_D=5,000, we find that the probability that the mtDNA pattern arose by admixture is below 5% for an effective number of females N_A>24,193. Increasing N_D to 10,000, the probability is below 5% for N_A>17,683. Rejecting at the 5% level the hypothesis that the observed mtDNA pattern occurred because of admixture from a more diverged hominid requires very large ancestral population sizes, even for a small admixture fraction. Figure S11.2 presents the curves underlying these inferences.

We conclude that a small amount of gene flow (f=0.01) from an ancient hominin into Denisovans could create the observed discordance between the mtDNA and nuclear gene genealogies but it remains possible that the discordance could be the result of incomplete lineage sorting in the common ancestral population. Higher levels of gene flow (larger f) would make it more likely that the discordance is the result of gene flow, but at present we do not have data that allow us to obtain an independent estimate of f.

Gene flow must have come from a population closely related to Denisovans
Here we show that the admixture signal seen in Melanesians must have originated from a close relative of Denisovans to explain the observed patterns. Figure S11.3 illustrates a model in which

Melanesians mixed with an unsampled population Z that diverged from Denisova at time t_Z. This model slightly generalizes that of Figure 3.

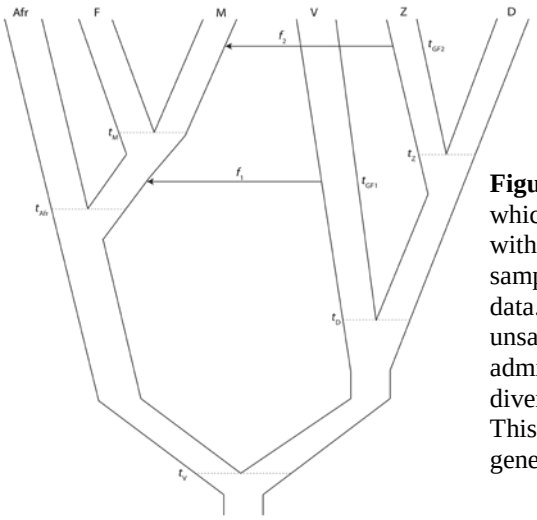


Figure S11.3: A model in which Melanesians admixed with a close relative of the sampled Denisovan fits the data. We use Z to denote the unsampled population that admixed with M, which diverged from D at time t_Z. This is a slight generalization of Figure 3.

To show that t_Z must be recent, we derive D(Afr, F, D, C), D(Afr, F, V, C), D(Afr, M, V, C) and D(Afr, M, D, C) under this model. D(Afr, F, D, C), D(Afr, F, V, C) and D(Afr, M, V, C) have the same expectation as under the model of Figure 3. Using the same methodology as before, it is easy to show that:

$$E\left[D(\text{Afr, M, D, C})\right]=\frac{X'}{Y'}, \text{ where}$$

$$X' = 3f_2(t_V - t_Z) + 3f_1(1-f_2)(t_V - t_D) \text{ and}$$

$$Y' = 3f_2(t_V - t_Z) + 3f_1(1-f_2)(t_V - t_D) + 4Nf_1(1-f_2)\left(1-\frac{1}{2N}\right)^{t_V - t_D}$$
$$+4N(1-f_1)(1-f_2)\left(1-\frac{1}{2N}\right)^{t_V - t_{Afr}} + 4Nf_2\left(1-\frac{1}{2N}\right)^{t_V - t_Z}. \quad \text{(S11.28)}$$

We observed D(Afr, M, D, C)=0.063 and D(Afr, M, D, C)=0.065. Compared to the model of Figure 3, this model requires f_2 to be bigger in order to match D(Afr, M, D, C)=0.063. This is because if the M lineage originated from Z (probability f_2), it cannot coalesce with D until t_Z. The older t_Z is, the larger f_2 needs to be to match D(Afr, M, D, C)=0.063. In turn, the larger f_2 is, the larger D(Afr, M, D, C). By varying t_{Afr}, t_Z, t_D, t_V, f_1 and f_2, we see that t_Z cannot be larger than $t_D/3$. Using t_D=9,000 generations, t_Z is constrained to be lower than 3,000 generations. Therefore, if Melanesians admixed with an archaic hominin, it must have been with one closely related to the population represented by the Denisova individual.

An alternative to gene flow: a model with structured ancestral populations

Here we explore a model in which the ancestral population of modern humans and of Denisova and Vindija was comprised of two subpopulations with gene flow between them. The population subdivision then persisted in the ancestral population, which is comprised of three subpopulations. We assume that subpopulations exchange migrants at rate m per generation. At time T in the past, the ancestral population becomes panmictic. We also assume that Denisovans admixed with Melanesians at rate f_2. This model is a generalization of the kind analyzed by Slatkin and Pollack[2].

To show that ancient structure can fit the observed *D* statistics, we simulated the model represented in Figure S11.4 using Hudson's *ms* software. We report the simulated *D* statistics in Table S11.3. The objective of this section is not to estimate the parameters of the ancient structure model but to reproduce the major features of the data to show that this model cannot be rejected based on *D*-statistics alone.

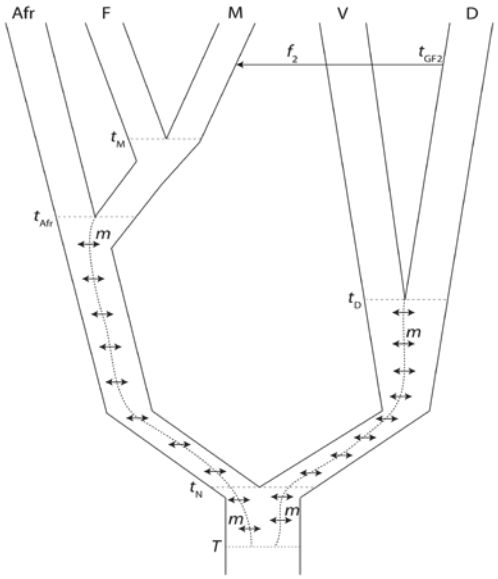


Figure S11.4: Model of ancient structure that fits the genetic data and cannot be ruled out.

By assuming the same migration rate in different ancestral populations, we can obtain a qualitative fit to *D* statistics. An even better fit could be obtained by letting the migration rate *m* vary in different ancestral populations. For simplicity, we use a model that still involves an episode of gene flow. We note that a more complicated model in which the population ancestral to French and Melanesians is subdivided in two, the ancestral population of all present-day humans is subdivided into three, and the ancestral population of present-day humans, Vindija and Denisova into four, could also fit the data.

Although models that assume ancestral subdivision can be made to fit the data, they are not especially plausible on biological grounds. Population subdivision has to persist for hundreds of thousands of years to produce the asymmetries detected by the D-statistics. Such subdivision would require that the geographic barriers that separated local populations persisted for very long times, much longer than seems reasonable for highly mobile and adaptable hominins.

Table S11.3: D-statistics computed from simulations of ancient structure (Figure S11.4 model)

Populations	***D*-stat**	**Comments**
S, F, V, C	0.046	French and Melanesians are closer to Vindija than Africans.
S, M, V, C	0.055	
S, F, D, C	0.015	French are closer to Vindija than to Denisova.
S, M, D, C	0.058	Melanesians are more closely related to Denisovans than Africans and French.
V, D, S, C	-0.051	Vindija is closer to Africans than Denisova.

Note: We used the following parameter values (times in generations): N=10,000; t_{GF2}=1,500; t_M=2,500; t_{Afr}=8,000; t_D=9,000; t_V=12,500; T=25,000; m=5.

References for SI 11

1. Green, R.E. et al., A draft sequence of the Neandertal genome. *Science* **328,** 710 (2010).
2. Slatkin, M. and Pollack, J.L., Subdivision in an ancestral species creates asymmetry in gene trees. *Molecular Biology and Evolution* **25**, 2241 (2008).

Supplementary Information 12

Morphology of the Denisova molar and phalanx. Stratigraphy and dating.

Bence Viola*, Michael Richards, Sahra Talamo, Michael V. Shunkov, Anatoly P. Derevianko, Jean-Jacques Hublin

* To whom correspondence should be addressed (bence.viola@eva.mpg.de)

The Denisova 4 molar

Denisova 4 (Denisova 2000 Г-2/29) is likely to be a left upper third molar. The specimen is well preserved; it is almost complete with only the apical half of the distobuccal root missing. Several cracks traverse the crown, but they do not alter the shape or the dimensions significantly.

In occlusal view, the crown is a rounded trapezoid, strongly tapering distally. The distal half of the crown is slightly displaced lingually, resulting in the hypocone and metacone tips lying lingually of the protocone and the paracone. It does not show the characteristic morphology seen in most first and some second upper molars of Neandertals, where a large hypocone contributes to a strongly rhomboid crown. The protocone is slightly displaced distally relative to the paracone. The trigon and talon basins are large. Decreasing cusp size order is: protocone, followed by subequal metacone and paracone, the hypocone is the smallest.

The mesial marginal ridge is high, and carries several accessory cusps. There is no anterior fovea or Crista transversa anterior. Instead, there is a bifurcated median fissure that encloses the largest of the accessory cusps on the mesial marginal ridge. The Crista obliqua is slightly incised by the median fissure. A marked distal marginal ridge carrying two accessory tubercles encloses the posterior fovea. The larger of the two accessory tubercles is adjacent to the hypocone, while the smaller one is next to the metacone and probably represents the metaconule (cusp 5). Several small finger-like protrusions are present on the lingual face of the protocone, representing a low grade (Grade 3 of ASUDAS[1]) expression of a Tuberculum Carabelli.

The crown is very high, with bulging buccal and especially lingual walls.

The roots are relatively short (length of the lingual root from the cervix: 12.4 of the mesiobuccal root 12.7), but very robust. There is a massive, strongly lingually flaring lingual root and two, probably only slightly separated buccal roots. The roots are not taurodont.

The apices of the lingual and mesiobuccal roots are closed. The wear of the crown is slight, with small wear facets visible on all cusps. Mesially a ~3mm × 3mm interproximal facet is visible.

As the crown is only slightly worn, the absence of a distal interproximal facet and the lack of wear facets on the distal aspects of the protocone, metacone, and hypocone could be consistent with the tooth being an M^2 preceding a non-erupted M^3. However the morphology makes this scenario less likely. The identification as an M^3 is primarily based on the tapering of the crown and large talon basin with a round distal outline. Some *Homo erectus* and Middle Pleistocene fossils have second molars with tapering, trapezoidal crowns but the tapering is less marked, the outline of the distal end is less rounded, and the talon basin is shorter in these specimens. Also, the marked distal marginal ridge with accessory tubercles is more like that seen in third molars.

Denisova 4 is very large. The mesiodistal diameter is 13.1 mm, measured as the greatest distance parallel to the occlusal surface and buccal and lingual walls; while the buccolingual diameter is 14.7, measured at a right angle to the plane of the mesiodistal measurement. If it is an M^3, it is outside the range of variation of all fossil human taxa, with the exception of *Homo habilis* and *Homo rudolfensis*. It falls into the range of variation of *Australopithecus afarensis* and *africanus* (Figure S12.1a, which replicates Figure 4c in the main text). Even as an M^2, it would be at the very upper end of the *Homo erectus* and Middle Pleistocene range of variation, and larger than any Neandertal or early modern human (Figure S12.1b). We explicitly exclude two specimens from these size comparisons, an upper M^2 from Obi-Rakhmat, Uzbekistan[2,3], and the M^3 of Oase 2[4,5]. The Obi-Rakhmat M^2 is significantly larger than even most Australopithecines, but this, and its unusual morphology, either results from a gemination of the tooth germ, or from a fusion of the germ with a supernumerary tooth[6]. The Oase 2 M^3s also show an unusual morphology, with numerous extra cusps[5,7]. As the Oase teeth are unerupted, it is impossible to judge whether the morphology and large size of these teeth also results from pathological processes. Despite Trinkaus' assertions, there are parallels in postcanine tooth size with Oase: Aterian hominins from North Africa show tooth sizes in the same range, so that large postcanine dentition might have been common in African early modern humans before they colonized Western Eurasia[8].

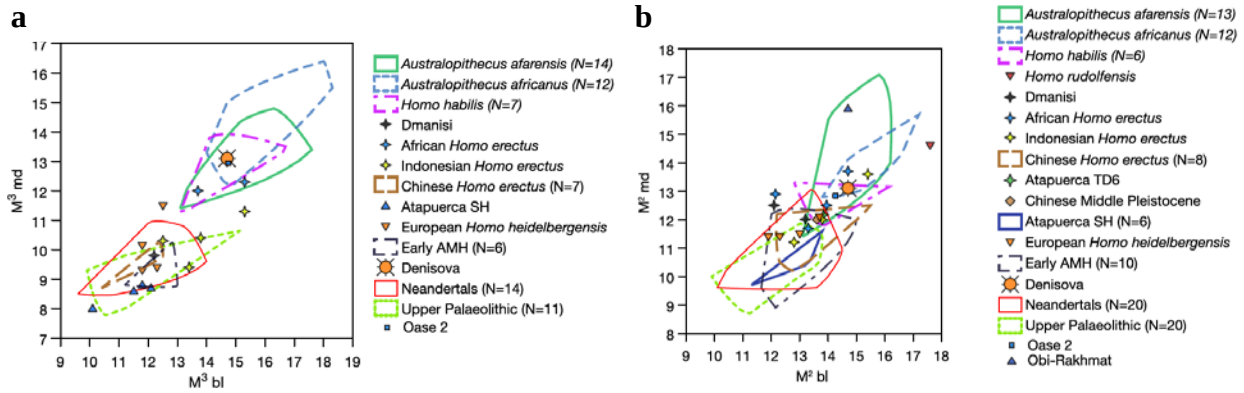


Figure S12.1: Comparison of the size of the Denisova molar to diverse other hominins. We plot mesiodistal length against buccolingual length for (a) third and (b) second molars (panel (a) replicates Figure 4c in the main text). The Denisova molar is an outlier relative to modern humans and Neandertals. Two specimens, Obi-Rakhmat and Oase 2, fall outside their groups (Neandertals and Upper Paleolithic modern humans, respectively). These specimens show unusual morphology or have been shown to be pathological.

Comparative morphology

The most diagnostic upper molars are the first ones, while upper M^2s and especially M^3s are more variable making their taxonomic identification rather problematic.

A general trend in all *Homo* starting with *H. erectus* is the reduction of the M^3. This results in most individuals in a size decrease from M^1 to M^3, unlike in *Australopithecines* or early *Homo*, where tooth sizes increase from M^1 to the M^3. We take the large size of Denisova 4 as an indication that this might have been the largest tooth in the molar row.

Upper M^3s of Homo erectus are reduced. In Chinese Homo erectus this is a general size reduction of the M^3 (e.g. refs. 9 and 10), while in Indonesian Homo erectus this manifests itself mainly in a reduction of the mesio-distal diameter, resulting in a short, but very broad crown[11].

Similarly, Middle Pleistocene hominins from China also show strongly reduced M^3s, for example in the case of Jinniushan[12]. There is significant variability in the reduction of the M^3 though, with the two specimens from Yunxian showing the two opposites of the morphological cline: EV 9002 has extremely reduced, peg-shaped M^3s, while in EV 9001 the M^3 is the largest molar[13]. Homo erectus from the Turkana basin also shows evidence of M^3 reduction[14,15]. LB 1, the holotype of Homo floresiensis, shows very small alveoli for the M^3, which is interpreted by Brown and collaborators[16] as evidence for a reduced M^3.

M^2s of Asian *Homo erectus* and Middle Pleistocene *Homo* are rather similar; they all share a trapezoidal outline where the mesial buccolingual diameter is significantly larger than the distal one. The root morphology of these groups is rather variable, with both well separated and splayed roots and fused, taurodont roots occurring[9]. If considered as an M^2, Denisova 4 fits reasonably well into this group, and is mainly differentiated by its lingually skewed hypocone and metacone and large talon basin. Denisova 4 does not show the mesiodistally narrow molar crowns seen in Early *Homo*, African *Homo erectus*[15] and in Dmanisi[17].

Recently, Bailey[18,19,20] described a suite of derived characteristics of Neandertal upper molars. These include an enlarged hypocone, a small metacone, centralized cusp tips and a constricted and rhomboid crown outline. This morphology is well expressed in Neandertal M^1s, much less so in M^2s, while the M^3s are rather variable and frequently reduced. The M^3 reduction is mainly a result of the reduction or even lack of the hypocone (68.6 % of Neandertal M^3s have reduced hypocones[16]). Only the roots of Neandertal lower molars have been studied in detail[21], but according to our observations upper molars follow a similar morphological pattern with long, and frequently pyramidal roots ("taurodontism").

Denisova 4 shows a trapezoidal crown outline, with a large metacone and small hypocone that are unlike M^1s and in lesser degree M^2s of Neandertals. This could be explained by the tooth being an M^3, but several other traits, such as the very robust, splaying separate roots, the strong flare of the crown and the large talon basin are also in stark contrast to the Neandertal morphology, thus making the determination of the tooth as M^2 or M^3 irrelevant.

Dental characteristics typical of Neandertals appear relatively early in the Neandertal lineage. Dental remains from Tautavel and the Sima de los Huesos of Atapuerca already display clear derived conditions in the Neandertal direction[20,22]. This suggests that these features evolved at least 350 thousand years ago[23] and possibly before 530 thousand years ago[24]. The Denisova phalanx does not allow any morphological comparisons but the Denisova molar shows a generalized archaic morphology reminiscent of earlier Homo, completely unlike Neandertals.

The Denisova 3 phalanx

Denisova 3 (Denisova 2008 Д-2/ 91), is the proximal epiphysis of a juvenile manual phalanx, preserving the proximal articular surface and the bone surrounding it. It is broken about 2 mm distally of the unfused proximal epiphyseal line. Fusion of the proximal epiphysis of the distal phalanges commences between 13.5 years (females) and 16 years (males)[25], and thus we can assume that its age is younger than this. The exact age cannot be determined, as the position, i.e. to which ray the phalanx belongs, is unclear, but an age of at least 6-7 years seems to be probable based on size (maximum radioulnar breadth of the proximal epiphysis: 7.5 mm, maximum dorsopalmar height of the proximal epiphysis: 5.1 mm).

Stratigraphic position and dating

The Denisova 4 tooth was recovered from the South Gallery of Denisova Cave, Square Г-2, Layer 11.1, during the 2000 excavation campaign. The excavation zone of the South Gallery is not directly connected to that of the Central Chamber, or the Eastern Gallery from which the phalanx derives. The correlation of the main stratigraphic levels is based on the geological and archaeological evidence. Thus Layer 11 in the main chamber and galleries is comparable, but the subhorizons such as 11.1 and 11.2 are not necessarily equivalent in different parts of the cave.

The dates previously published[26,27] from Layer 11 are from the Central Chamber and the South Gallery, and thus are not stratigraphically connected to the hominin phalanx. We thus undertook a radiocarbon dating programme to clarify the age of Layer 11 in the East Gallery, and obtained additional dates from the South Gallery, analyzing human modified bones where possible.

As the phalanx and tooth were too small for direct radiocarbon dating we selected bone fragments that were spatially very close to the human remains. Two of the dated specimens in close proximity of the phalanx are culturally clearly attributable to an Upper Palaeolithic tool industry. One is a rib with regular incisions (Denisova 2008 Д-2/22), while the other is a bone projectile point blank (Denisova 2008 Д-2/104). The rib was separated by about 15 cm laterally, and 16 cm vertically above the phalanx, while the blank was 40 cm laterally, and 4 cm vertically below the phalanx. We also dated three other animal bones from the undisturbed part of Layer 11.2 and 11.3 of the East Gallery (Table S12.1).

Table S12.1: Radiocarbon dates from layer 11 of Denisova Cave

Lab number	Site	Taxon	Layer	Age (^{14}C BP)	δ^{13}C	δ^{15}N	%C	%N	C:N
SOAN-2504	Central Hall	Unidentified bone	11	>37,235	n.d.	n.d.	n.d.	n.d.	n.d.
KIA-25285	South Gallery	Hyena bone	11.2	48,650 ± 2,380	n.d.	n.d.	n.d.	n.d.	n.d.
AA-35321†	South Gallery	Charcoal *	10/11	29,200 ± 360	n.d.	n.d.	n.d.	n.d.	n.d.
OxA-V-2359-16	East Gallery	Ovis/Capra	11.2	>50,000	-18.8	5.0	45.0	16.5	3.2
OxA-V-2359-15	East Gallery	Ovis/Capra (w/cutmarks) †,‡	11	15,740 ± 65	-18.6	3.9	46.1	17.1	3.2
OxA-V-2359-14	East Gallery	Bison (w/cutmarks) †	11.3	>50,000	-19.6	6.2	46.3	17.0	3.2
OxA-V-2359-20	East Gallery	Rib (w/regular markings) †,‡	11	30,100 ± 210	-17.9	7.3	44.5	16.5	3.1
OxA-V-2359-21	East Gallery	Bone tool blank †,‡	11	23,170 ± 110	-19.2	3.7	45.7	16.9	3.2
OxA-V-2359-17	South Gallery	Ovis/Capra	11.2	>50,000	-19.2	6.2	46.1	16.9	3.2
OxA-V-2359-18	South Gallery	Bison	11.2	>50,000	-19.4	5.8	45.8	16.9	3.2

Notes: The first three dates are from ref. 26 and 27, while the others are new dates from the East and South Galleries. Bone collagen δ^{13}C values are reported relative to the vPDB standard, and δ^{15}N values relative to the AIR standard. Errors on the isotope measurements are ±0.1%.

* This sample comes from the erosion surface between layers 10 and 11 in the Central Hall, and thus gives a minimum age for the deposits of Layer 11 in this part of the cave.

† Denotes bones modified by humans.

‡ These samples derive from the disturbed part of Layer 11.

Cutmarked, or herbivore bones were not available from layer 11.1 of the South gallery, from where the molar derives. Therefore, to get a terminus post quem for this specimen, we dated two herbivore bone fragments from the underlying Layer 11.2.

Bone samples were pretreated in the Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology in Leipzig. Pre-treatment procedures follow the procedures developed in collaboration with the Oxford Radiocarbon Accelerator Unit (ORAU) and interlab

comparisons of dates obtained from samples prepared in both labs shows that the results are comparable. Approximately 500 mg of bone was first cleaned and then demineralised in 0.5M HCl at room temperature until no CO_2 effervescence is observed (2 to 5 days). The samples were rinsed and 0.1M NaOH were then added for 30 minutes to remove humics. The NaOH step was followed by further rinsing and then the sample was gelatinized in a pH3 solution at 75°C for 20h. The resulting gelatine was first filtered in a Eeze-Filter™' (Elkay Laboratory Products (UK) Ltd.) to remove larger particles and then through a 30 kDa ultrafilter (Sartorius "Vivaspin 15"). The samples were then lyophilized for 48 hours. The resulting collagen was then converted to graphite and the radiocarbon age measured at ORAU. The resulting dates were then blank corrected using dates obtained on infinite age (i.e. radiocarbon dead) bone collagen prepared in the Leipzig laboratory and measured at the ORAU.

The seven new dates obtained cluster into two age groups (Table S12.1). Four are infinite, older than 50,000 years BP, while the rest are younger than 30,000 radiocarbon years BP. Both of the specimens attributable to the Upper Palaeolithic had ages younger than 30,000 radiocarbon years BP (Denisova 2008 Д-2/22, 27,930 ± 210 BP and Denisova 2008 Д-2/104, 23,010 ± BP). Based on these dates, we now assume that the wedge shaped, discoloured area from which these specimens derive, which until now was assumed to be a result of chemical leaching, is actually a disturbance in Layer 11. Thus, despite the spatial proximity to the phalanx, the Upper Palaeolithic material is not necessarily contemporary with it, as the phalanx derives from supposedly undisturbed deposits of Layer 11.2.

We have evidence for human activity in both of the time periods, as a bison petrosal with cutmarks is dated to >50,000 BP. Both of the radiocarbon dates from herbivore bones from the South Gallery taken from the layer underneath the molar (Layer 11.2) belong to the earlier (>50,000 BP) cluster, and a previously obtained date from this horizon is of similar age.

As all ^{14}C dates from Layer 11 in the South Gallery fall into the earlier cluster, it is likely that the hominin molar derives from the earlier occupation. The dating of the phalanx is more problematic. As evidenced by several ^{14}C dates younger than 30ka from Layer 11 in the East Gallery, it is clear that there has been major post-depositional mixing in this part of the cave. Based on the available evidence, we cannot exclude that the phalanx derives from the later occupation, but we think this is unlikely due to several factors. First, this would either require the Denisovans to independently develop Upper Paleolithic technology, or it would assume the simultaneous presence of modern humans who produced the UP industry and of Denisovans. Second, the mtDNA evidence suggests a relatively recent divergence between the individuals represented by the phalanx and the tooth (95% HDP less than 16,000 years ago, see SI 13), which is incompatible with them being separated by more than 20,000 years.

Thus, we propose the following scenario: a first hominin occupation of the cave more than 50,000 radiocarbon years ago by the Denisova hominins, and a second occupation during the Upper Palaeolithic, at 30,000 years BP or later, probably by modern humans. As the hominin remains are not dated directly, we cannot exclude the possibility that they belong to the second, later, occupation, and may have therefore contributed to the production of Upper Palaeolithic artifacts. However, we think that this is unlikely.

References for SI 12

1. Turner, C., Nichol, C. and Scott, G., In *Advances in Dental Anthropology* (eds.) Kelley, M. and Larsen, C. 13-31 (Wiley Liss, 1991).
2. Glantz, M. et al., New hominin remains from Uzbekistan, *Journal of Human Evolution* **55**, 223 (2008).
3. Bailey, S., Glantz, M., Weaver, T.D. and Viola, B., The affinity of the dental remains from Obi-Rakhmat Grotto, Uzbekistan. *Journal of Human Evolution* **55**, 238 (2008).
4. Trinkaus, E., Milota, S., Rodrigo, R., Mircea, G. and Moldovan, O., Early modern human cranial remains from the Peştera cu Oase, Romania. *Journal of Human Evolution* **45,** 245 (2003).
5. Trinkaus, E., Denisova Cave, Pestera cu Oase, and Human Divergence in the Late Pleistocene, *Paleoanthropology* **2010**, 196 (2010).
6. Smith, T.M., et al., Dental development and age at death of a Middle Paleolithic juvenile hominin from Obi-Rakhmat Grotto, Uzbekistan. In press in *150 Years of Neanderthals Discoveries* (eds.) Condemi, S. and Weniger, G. (Springer, Dordrecht, 2010).
7. Rougier, H. et al., Peştera cu Oase 2 and the cranial morphology of early modern Europeans, *Proc Natl Acad Sci USA*, **104**, 1165 (2007).
8. Hublin, J.-J. et al., Dental evidence for the Aterian human populations of Morocco. In press in *Modern Origins: a North African Perspective* (eds). Hublin, J.-J. and McPherron, S. (Springer, Dordrecht, 2010).
9. Weidenreich, F., The dentition of Sinanthropus pekinensis: a comparative odontography of the hominids. *Palaeontologia Sinica, New Series D* **1**,1 (Peiping (Beijing), 1937).
10. Woo, J.K., Preliminary report on a skull of Sinanthropus lantianensis of Lantian, Shensi. *Scientia Sinica* **14**, 1032 (1965).
11. Indriati, E. and Antón, S.C., Earliest Indonesian facial and dental remains from Sangiran, Java: a description of Sangiran 27. *Anthropological Science* **116**, 219 (2008).
12. Jianing, H., Preliminary study on the teeth of Jinniushan archaic Homo sapiens. *Acta Anthropologica Sinica* **19**, 216 (2000).
13. Li ,T. and Etler, D., New Middle Pleistocene hominid crania form Yunxian in China. Nature **357,** 404 (1992).
14. Brown, B. and Walker, A., In *The Nariokotome Homo erectus Skeleton.* (eds.) Walker, A.C. and Leakey, R.E.F. 161-192 (Harvard University Press, 1993).
15. Wood, B. A., *Koobi Fora Research Project. Volume 4. Hominid cranial remains*. (Clarendon Press, 1991).
16. Brown, P. et al., A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature* **431**, 1055 (2004).
17. Martinón-Torres, M., *et al.*, Dental remains from Dmanisi (Republic of Georgia): Morphological analysis and comparative study. *Journal of Human Evolution* **55,** 249 (2008).
18. Bailey, S., Beyond shovel-shaped incisors: Neandertal dental morphology in a comparative context. *Periodicum Biologorum* **108**, 253 (2006).
19. Bailey, S.E., A morphometric analysis of maxillary molar crowns of Middle-Late Pleistocene hominins. *Journal of Human Evolution* **47**, 183 (2004).
20. Bailey, S.E., *Neandertal dental morphology: Implications for modern human origins* (Arizona State U., 2002).
21. Kupczik, K. and Hublin, J.-J., Mandibular molar root morphology in Neanderthals and Late Pleistocene and recent Homo sapiens. *Journal of Human Evolution* in press (doi.org/10.1016/j.jhevol.2010.05.009) (2010).
22. Martinon-Torres, M. *et al.*, In *Dental Perspectives on Human Evolution: State of the Art Research in Dental Paleoanthropology* (eds.) Bailey, S.E. and Hublin, J.-J. (Springer, 2007).
23. Falguères, C. et al., New U-series dates at the Caune de l'Arago, France. *Journal of Archaeological Science* **31**, 941 (2004).
24. Bischoff, J.L. *et al.*, High-resolution U-series dates from the Sima de los Huesos hominids yields 600+∞-66 kyrs: implications for the evolution of the early Neanderthal lineage. *Journal of Archaeological Science* **34**, 763 (2007).
25. Scheuer, L., and Black, S., *Developmental Juvenile Osteology* (Academic Press, 2000).
26. Derevianko, A.P., Shunkov, M.V., and Volkov, P.V., A paleolithic bracelet from Denisova Cave. *Archaeology, Ethnology and Anthropology of Eurasia* **34**, 13 (2008).
27. Derevianko, A.P. *et al.*, *Arkheologiya, geologiya i paleogeografiya pleistotsena i golotsena Gornogo Altaya* [Archaeology, geology, and the Pleistocene and Holocene palaeogeography of the Mountainous Altai]. (Nauka, Novosibirsk, 1998).

Supplementary Information 13

DNA extraction, library preparation and mtDNA analysis of the Denisova tooth

Johannes Krause*, Qiaomei Fu, Tomislav Maricic and Martin Kircher

* To whom correspondence should be addressed (krause@eva.mpg.de)

DNA extraction and library preparation

A total of 50 mg of bone was removed from the internal parts of one of the Denisova molar roots using a sterile dentistry drill in our clean room facility, where procedures that minimize contamination from present-day human DNA are rigorously implemented[1]. In this facility, DNA was extracted as described in ref. 2. A sequencing library was produced with prior UDG and Endo VIII treatment as described in ref. 3 using the same modified Illumina multiplex protocol[4] as for the Denisova phalanx (SI 1). A 7nt-index (5'- AATCTTC -3') that is not available outside of the clean room was added by a PCR reaction that was set up in the clean room. Libraries were then removed from the clean room facility and PCR cycles were performed (adding the index and the outer adapter sequences required for sequencing).

Illumina Sequencing and primary data processing

After indexing PCR, the library was amplified an additional 10 cycles in a 100µl reaction containing 50µl Phusion™ High-Fidelity Master Mix and 500nM of primers sitting at the outer P5 and P7 *Illumina* library grafting sequences. The annealing temperature was 60°C. The amplified product was spin column purified and quantified on an Agilent 2100 Bioanalyzer DNA 1000 chip.

To enrich for mitochondrial DNA, we used a recently described DNA capture on beads protocol[5]. For this protocol, two overlapping long-range PCR products encompassing the whole mitochondrial genome were produced[6], using DNA extracted from the saliva of a European individual as template. The PCR products were purified using Qiagen spin columns and quantified by NanoDrop. The two products were pooled in equimolar amounts to a total amount of 3 µg. The pooled products were then sonicated (Bioruptor, Diogenode, Liege, Belgium), producing fragments of 150 to 700 bases as observed on a 2% agarose gel. The products were biotinylated by ligation to a biotin-carrying adapter and immobilized on streptavidin-coated magnetic beads. The amplified tooth library was made into single-stranded DNA by incubating it at 95ºC for 3 min and then adding streptavidin beads coated with the fragmented long range PCR product. The mixture was incubated under rotation at 65ºC in a hybridization oven (SciGene, Model 700, Sunnyvale, CA, USA). After 48 hours, the beads were washed and library molecules were eluted by heating for 3 minutes at 95ºC. The DNA concentration was measured by qPCR (Mx3005P Real Time PCR System, Stratagene, La Jolla, CA), and the eluted library was further amplified for 15 cycles using primers complementary to the outer P5 and P7 *Illumina* library grafting sequences.

The enriched library was sequenced on the *Illumina Genome Analyzer IIx* platform using 2 x 76 + 7 cycles on half a sequencing lane according to the manufacturer's instructions for multiplex sequencing (FC-104-400x v4 sequencing chemistry and PE-203-4001 cluster generation kit v4). The manufacturer's protocol was followed except that an indexed control PhiX 174 library

(index 5'- TTGCCGC-3') was spiked into each lane, yielding a fraction of 2-3% control reads in all lanes of the run.

The sequencing data were analyzed starting from QSEQ sequence files and CIF intensity files from the Illumina Genome Analyzer RTA 1.6 software. The raw reads were aligned to the PhiX 174 reference sequence to obtain a training data set for the base caller Ibis[7]. Raw sequences called by Ibis 1.1.1 were filtered for the 'AATCTTC' index as described[4]. The paired-end reads were subjected to a fusion process (including removal of adapter sequences and adaptor dimers) by requiring at least an 11nt overlap between the two reads. In the overlapping sequence, quality scores were combined and the base with the highest base quality score was called. Only sequences merged in this way were used for further analysis. The small proportion of molecules longer than 191nt was thus discarded. As the library was sequenced to high PCR duplicate redundancy (21,739,162 merged clusters; 12,561,359 different sequences; 7,974,163 singletons), only sequences observed at least three times (1,827,232) were kept for analysis.

Assembly procedure

The 1,827,232 merged sequences that occurred at least 3 times were used as input for a custom iterative mapping assembler[8]. In the first assembly round, sequences were aligned to the revised Cambridge Reference Sequence (rCRS)[9]. The mapper uses a position-specific scoring matrix designed to capture the most relevant features of nucleotide misincorporations affecting ancient DNA sequences. Since several amplification steps were performed, the aligned sequences were filtered for uniqueness by grouping sequences with the same direction, start, and end coordinates. From each such cluster a consensus sequence was generated by taking, for each position, the base with the highest quality score. This resulted in a total of 15,094 distinct aligned sequences that aligned to the reference mtDNA. The average fragment length of the mapped sequences was 64.3 bp. The average coverage of the complete mtDNA was 58.5-fold (minimum 8- fold and maximum 144- fold). Among the 16,570 positions in the circular mtDNA genome, the average fraction of sequences that agree with the assembled mtDNA was 99.95% (lowest 91%, highest 100%). This is different from the phalanx mtDNA, where some positions showed a minor base frequency of up to 30% (all 8 were consistent with cytosine deamination[8]). We note that the UDG + Endo VIII treatment of the Denisova tooth mtDNA is expected to remove deaminated cytosines, resulting in a reduced number of positions with appreciable minor base frequencies.

Table S13.1: Pairwise nucleotide differences among six mtDNA genomes

	Phalanx	**Tooth**	**Neandertal**	**Human**	**Chimpanzee**
Denisova Phalanx (FN673705)					
Denisova Tooth (FR695060)	2				
Neandertal (Vindija 33.16; NC_011137)	380	380			
Modern human (rCRS; AC_000021)	386	386	202		
Chimpanzee (NC_001643)	1462	1462	1434	1451	
Bonobo (NC_001643)	1454	1454	1419	1433	678

Phylogenetic and divergence time analysis

In addition to the Denisovan tooth (FR695060) mtDNA and the Denisovan phalanx mtDNA (FN673705), we chose a single mtDNA genome to represent Neandertals (Vindija 33.16; NC_011137), modern humans (rCRS; AC_000021), chimpanzees (NC_001643), and bonobos (NC_001644). These six hominin mtDNAs were aligned using the software Muscle[10]. Pairwise nucleotide differences between mtDNAs were calculated using MEGA 4.1[11] (Table S13.1).

We estimated a phylogenetic tree and mtDNA divergence times using the Bayesian algorithm of BEAST v1.5.3[12]. The alignment was analyzed using two molecular clock models: a strict molecular clock and a relaxed uncorrelated log-normal molecular clock[13]. We set the prior distribution for the mtDNA divergence time between chimpanzees and humans to follow a normal distribution with a mean of 6 million years (±500,000 years). For each analysis, we used a model that assumes a constant population size across the phylogeny and ran 20,000,000 generations of the Markov Chain Monte Carlo with the first 2,000,000 generations discarded as burn-in. We chose the General Time Reversible sequence evolution model with a fraction of invariable sites (GTR+I) determined by the best-fit model approach of Modeltest[14] and PAUP*[15].

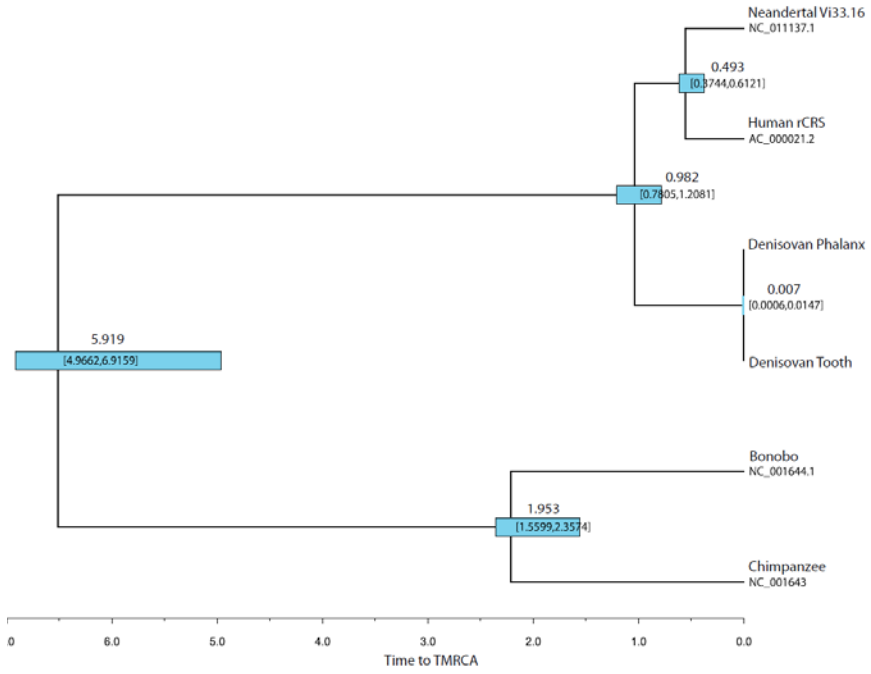


Figure S13.1 – Maximum clade probability tree displayed as a chronogram from BEAST analysis of the unpartitioned mtDNA alignment. The evolution of all lineages are compatible with a strict molecular clock and the GTR+I substitution model. Node bars illustrate the width of the 95% highest posterior density for the divergence estimates. Numbers in bold indicate the posterior mean estimates of divergence times and the 95% highest posterior density values below. Time to MRCA is displayed in millions of years.

A consensus tree of all 20,000 trees was inferred using TreeAnnotater V.1.4.8[12] (Figure S13.1). The Denisovan tooth and phalanx mtDNA differ at two positions (3,600 and 16,399 in the rCRS), which are covered by 24 and 41 independent reads in the Denisova tooth and 155 and 169 times in the phalanx, respectively. The minor base frequency in both mtDNA alignments for the two positions is at most 1.8%. These two differences thus suggest that the tooth and the phalanx belong to two different individuals.

Using both a strict and a relaxed clock, the mean divergence time of the two Denisova bones was inferred to be less than 7,500 years with a 95% highest probability density (HPD) between 500 and 16,000 years (Table S13.2), with the relaxed molecular clock yielding slightly higher values.

We can also give an estimate of the effective population size of the Denisovans, which is obviously limited by the fact that we have a sample size of only two. A sequencing of 311 full human mtDNA[16] shows only a 0.06% chance to have two differences or less between two individuals. The median number of differences is 42, and the 5% quantile is 25. Since coalescence time scales with effective population size, as does the number of differences between individuals, our point estimate is that the population from which the two Denisovans

were sampled has an effective population size that is 20 times smaller than that of humans (with a 95% confidence interval of 5 to 80 fold smaller). A similar estimate is obtained by asking what effective population size would result in an expected coalesce time of 7,500 years.

Table S13.2: TMRCAs of phalanx and tooth, and human and Neandertal mtDNAs

Method	**Tooth (D_1) - Phalanx (D_2) TMRCA (95% HPD)**	**Human (H)- Neandertal (N) TMRCA (95% HPD)**	**HN–D_1D_2 TMRCA (95% HPD)**
Strict clock	6,860 (624 - 14,656)	491,300 (374,400 - 612,100)	982,500 (780,500 - 1,208,100)
Relaxed clock	7,117 (567-15,589)	496,600 (304,200 - 686,000)	995,900 (634,100 - 1,323,300)

Note: We calibrate the dates against an assumed human-chimpanzee mtDNA genetic divergence time of 6 Mya. The analysis is performed on all 16,586 sites simultaneously using a GTR+I substitution model.

References for SI 13

1. Green, R. E. et al., The Neandertal genome and ancient DNA authenticity. *EMBO J* **28**, 2494 (2009).
2. Rohland, N. and Hofreiter, M., Ancient DNA extraction from bones and teeth. *Nat Protoc* **2**, 1756 (2007).
3. Briggs, A.W. et al., Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38,** e87 (2010).
4. Meyer, M. and Kircher, M., Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*, **2010**, pdb.prot5448 (2010).
5. Maricic, T., Whitten, M. and Paabo, S., Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *Plos ONE* **5,** e14004.
6. Meyer, M., Stenzel, U., Myles, S., Prufer, K. and Hofreiter, M., Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res* **35**, e97 (2007).
7. Kircher, M., Stenzel, U. and Kelso, J., Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* **10**, R83 (2009).
8. Briggs, A.W. et al., Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**, 318 (2009).
9. Andrews, R.M. et al., Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**, 147 (1999).
10. Edgar, R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792 (2004).
11. Kumar, S., Tamura, K. and Nei, M., MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**, 150 (2004).
12. Drummond, A. J. and Rambaut, A., BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**, 214 (2007).
13. Drummond, A. J., Ho, S. Y., Phillips, M. J. and Rambaut, A., Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**, e88 (2006).
14. Posada, D., Using MODELTEST and PAUP* to select a model of nucleotide substitution. *Curr Protoc Bioinformatics* **Chapter 6**, Unit 6 5 (2003).
15. PAUP* beta version. Phylogenetic analysis using parsimony (*and other methods). (Sinauer Associates, Sunderland, MA, 2002).
16. Green, R. E. et al., A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**, 416 (2008).