doi:10.1038/nature12886

Contents

Supplementary Information 1

Sampling, Library Preparation and Sequencing

Susanna Sawyer*, Bence Viola and Anja Heinze

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

Samples

The Altai Neandertal data was produced from the proximal pedal phalanx. The toe bone was excavated in 2010 from Square B-3, Subsquare D, in layer 11.4 of the East Gallery of Denisova Cave (51.409° N; 84.689° E) in the Altai mountains in Siberia, Russian Federation. It is a proximal phalanx, as evidenced by the single articular facet on the proximal epiphysis, and based on shape and size it likely derives from the 4th or 5th ray. It cannot be sided with certainty. The bone is rather complete, but the majority of the trochlea is missing. There is minor damage on the dorsal aspect of the proximal epiphysis. The shaft is broad and robust, with a marked waist. The base is robust, and tall as well as broad. For a more detailed description and comparative analysis see Mednikova, 2011[1].

The Denisovan finger bone, from which we have previously determined a 30-fold coverage genome, was found in the same gallery of Denisova cave, slightly higher in the stratigraphic sequence in layer 11.2[2]. None of the hominid bones in Denisova cave have been directly dated. However, layer 11 of the east gallery contained animal bones which have been ^{14}C dated to >48–30 kyr ago (see Supplementary Information 12 from Reich et al.[3]).

The Mezmaiskaya Neandertal data was generated from a rib of a Neandertal neonate found in Mezmaiskaya cave, Russia, as described in Golovanova et al.[4]. The neonate ribs are part of a partial skeleton found in quadrant M-26 of layer 3 of Mezmaiskaya cave (ref). Animal teeth from layer 3 have been dated by ^{14}C dating, estimating a date greater than 45kya[4], and ESR dating, which estimates the Mezmaiskaya neonate layer to be 60-70 ky old[5].

DNA Extraction

The Altai Neandertal toe phalanx and the Mezmaiskaya Neandertal rib were sampled using a dentistry drill under clean-room conditions to produce 38 and 45 mg of bone powder, respectively. This bone powder was used to prepare 100uL of DNA extract per Neandertal individual (extract E956 for the Altai Neandertal and extract E733 for Mezmaiskaya Neandertal) as described in Rohland et al.[6].

Library preparation

Overview

Five libraries were prepared from E956, the Altai Neandertal DNA extract (see Table S1.1, Figure S1.1). The first library (L9105) was prepared using a double stranded library preparation method[7]. The other four libraries (L9199, L9198, L9302 and L9303) are independent amplification products of

two libraries that were prepared with a single stranded library preparation method recently used to produce a 30-fold coverage Denisova genome[2]. After library preparation and prior to amplification by PCR, each of the two libraries was split equally in two parts. Library amplification was performed using a double-indexing scheme described elsewhere[7], resulting in five libraries carrying unique combinations of indices. Since a substantial proportion of library molecules generated with the single-stranded method is very short, a size fractionation step using acrylamide gels (gel cut) was carried out to remove library molecules with inserts shorter than approximately 35 bp from the four libraries prepared with this method.

The Mezmaiskaya Neandertal extract E733 was used to prepare five libraries (L4533, L4740 L4741, L4677 and L4678) using the same double stranded library preparation method and amplification scheme as used for the L9105 library from the Altai Neandertal (Table S1.1, Figure S1.1). Like the Altai Neandertal four of the libraries come from two original libraries (L4740/L4677 from one, and L4741/L4678 from another); however unlike the Altai Neandertal, the libraries were split after the first amplification (the indexing amplification), meaning that the split libraries have the same indices, as well as the same amplified molecule base for further amplification (see Figure 1.1 for clarification).

Sample	Library	Library type	Sequencing Lanes
Altai Neandertal	L9105	Double stranded+UDG	3 HiSeq
Altai Neandertal	L9198	Single stranded+UDG	8.5 HiSeq
Altai Neandertal	L9199	Single stranded+UDG	8.5 HiSeq
Altai Neandertal	L9302	Single stranded+UDG	6.5 HiSeq
Altai Neandertal	L9303	Single stranded+UDG	6.5 HiSeq
Mezmaiskaya	L4533	Double stranded+UDG	4 HiSeq + 7.2 GAII
Mezmaiskaya	L4677	Double stranded+UDG	0.15 GAII
Mezmaiskaya	L4678	Double stranded+UDG	0.15 GAII
Mezmaiskaya	L4740	Double stranded+UDG	4 HiSeq
Mezmaiskaya	L4741	Double stranded+UDG	4 HiSeq

Table S1.1: Samples and Libraries sequenced for this project.

Double stranded libraries (L9105, L4533, L4740, L4741, L4677 and L4678):

L9105 was produced from 20uL of E956 Altai Neandertal DNA extract using a double stranded library preparation method and uracil-DNA-glycosylase / endonuclease VIII treatment to remove uracils[7,8]. Illumina Multiplex adapters were extended by clean-room specific four basepair keys to minimize the risk of contamination by libraries generated from other sources during downstream amplifications and sequencing[9].

All Mezmaiskaya libraries were produced from extract E733. L4533 was produced from 19uL of E733. L4740 and L4677 are derived from one library, as are L4741 and L4678. These two original libraries were produced from 20 uL of E733 Mezmaiskaya DNA extract, each. All five libraries were made, in the exact same manner as L9105. Extraction negative controls as well as water controls were carried alongside with each set of library preparations. To determine the number of molecules in each library prior to amplification, qPCR measurements were taken using 1 uL of a 1:10 dilution from each

library[10,11]. Using the qPCR results we could determine that all libraries had at least three orders of magnitude more molecules than any of the negative controls, indicating success of library preparation.

All six libraries from the two Neandertals were amplified in two successive rounds of PCR. The first amplification was performed for 10 cycles using AmpliTaq Gold DNA Polymerase and two index primers[7]. The indexed PCR products were then purified using the MinElute PCR purification kit (Qiagen) and eluted in 30 uL (L9105) and 20uL (L4533, the precursor library to L4677/L4740, and the precursor library to L4678/L4741) of Elution Buffer (EB), respectively.

10uL from the indexed L9105 library was used as template for a second amplification using Herculase II Fusion DNA polymerase (Agilent) and the primer pair IS5 and IS6[10] under PCR conditions described elsewhere[2,12]. L4677/L4740 and L4678/L4741 were split up: L4677 and L4678 were made from 10uL of template while L4740 and L4741 were made from 4 uL of template. All amplifications used Herculase II Fusion DNA polymerase under identical conditions as for L9105. 5 uL of L4533 was amplified with Phusion Hot Start II High-Fidelity DNA polymerase as described in Kircher et al.[7] using the same primer pair as the other libraries.

For the second round of amplification, cycle numbers were adjusted to avoid PCR plateau. The amplified libraries were again purified using the MinElute PCR purification kit and eluted in 15uL of EB. The concentration of all six libraries was measured using a DNA-1000 chip on the Bioanalyzer 2100 (Agilent).

Single stranded libraries (L9198, L9199, L9302 and L9303):

The single stranded libraries were prepared as follows. Two libraries were prepared from 28.5 uL each of the E956 DNA extract using single stranded library preparation[2]. Water controls were also included. The final elution volume for both libraries was 40uL of EB. To measure library concentration, 1 uL of a 1:40 dilution of each library was measured with qPCR. According to this measurement, the two single stranded sample libraries have an order of magnitude more molecules than the control libraries prepared from water. When comparing the qPCR molecule counts obtained from single-stranded and double-stranded library preparation and normalizing for the volume of extract used for library preparation, the molecule counts from both single-stranded libraries are approximately 11 times higher than that of the double-stranded library L9105.

After library preparation and prior to amplification, both libraries were split into two equal parts. L9198 and L9199 both come from the same original library, as do L9302 and L9303. Four PCR amplifications were performed with AccuPrime *Pfx* DNA polymerase (Life Technologies) and different index primer combinations in each reaction[2,12]. Cycle numbers that avoid PCR plateau were determined from the qPCR amplification plots. PCR products were purified using the MinElute PCR purification kit and eluted in 30uL EB. 4 uL of L9198 and L9199 were further amplified using Herculase II Fusion DNA polymerase (Agilent) and the primer pair IS5 and IS6. Amplification products were purified using the MinElute PCR purification kit and eluted in 40uL of EB. L9302 and L9303 were size fractioned directly after the first amplification step, L9198 and L9199 after the second amplification.

Size fractionation was performed as described in Meyer et al.[2]. L9199 was not amplified after size fractionation as this library retained sufficiently high concentration. Concentrations of the final libraries were determined using a DNA-1000 chip on the Bioanalyzer 2100.

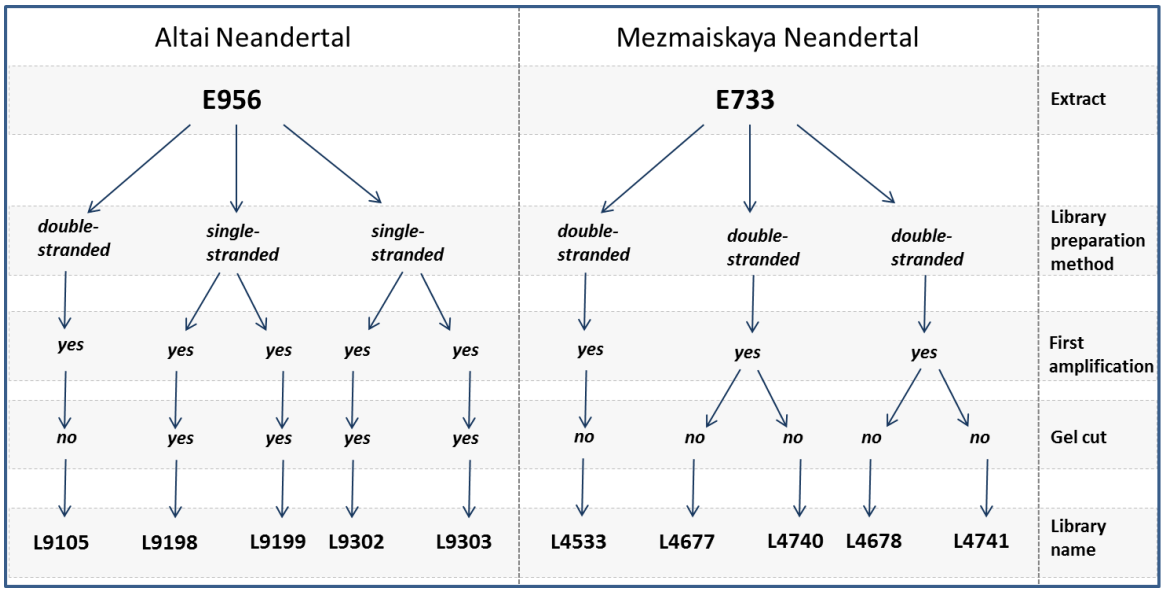


Figure S1.1. An overview of the Altai and Mezmaiskaya Neandertal library preparation schema. The main events that differentiate the libraries are shown, such as at what point libraries were split into two parts.

Sequencing

Sequencing of L9105, L4533, L4740 and L4741 was performed on fifteen lanes of Illumina HiSeq 2000 flowcells (see Supplementary Table S1.1). L9105 occupied 3 lanes, while the other three libraries each had four lanes. Sequences were produced using 101-bp paired-end reads using the CR2 forward (5' – TCTTTCCCTACACGACGCTCTTCCGATCTGTCT) and CR2 reverse (5' – GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTCT) custom sequencing primers. In addition, seven cycles were sequenced for a P7 index using the P7 Illumina Mulitplex primer. The P5 index was not sequenced. The instructions from the manufacturers were followed for multiplex sequencing on the HiSeq platform with a TruSeq PE Cluster Kit v3 - cBot – HS cluster generation kit and a TruSeq SBS Kit v3 sequencing chemistry, with modifications for double-indexed sequencing described elsewhere[7]. An indexed control library of ϕX 174 was spiked into each library prior to sequencing, contributing to 0.5% of the sequences from each lane[7].

L4533 was additionally sequenced on seven full Illumina GAII lanes, as well as 20% of another lane. L4677 and L4678 were each sequenced on 15% of an Illumina GAII lane. For all Illumina GAII sequencing, a 76- cycle paired-end as well as two 7-cycle index reads were performed[2] with the same CR2 custom sequencing primers mentioned above. The instructions from the manufacturers were followed for multiplex sequencing on the GAII platform with vs5 sequencing chemistry with modifications for double-indexed sequencing[7].

Sequencing of L9198 and L9199 was performed on one flow cell each per library on the Illumina HiSeq 2000 platform. In addition L9198 and L9199 were pooled together in equimolar amounts and sequenced as a pool on one additional HiSeq lane. L9302 and L9303 were pooled together in equimolar amounts and the pool was sequenced on 13 lanes spread over two HiSeq flowcells. 95-cycle paired-end reads as well as two 7-cycle index reads were generated for these libraries using the double-index sequencing scheme described elsewhere[2].

References

1 Mednikova, M. B. A proximal pedal phalanx of a paleolithic hominin from Denisova cave, Altai. *Archaeology Ethnology & Anthropology of Eurasia* **39**, 129-138, doi:http://dx.doi.org/10.1016/j.aeae.2011.06.017 (2011).

2 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).

3 Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060, doi:10.1038/nature09710 (2010).

4 Golovanova, L. V., Hoffecker, J. F., Kharitonov, V. M. & Romanova, G. P. Mezmaiskaya cave: A Neanderthal occupation in the Northern Caucasus. *Curr Anthropol* **40**, 77-86, doi:Doi 10.1086/515805 (1999).

5 Skinner, A. R. *et al.* ESR dating at Mezmaiskaya Cave, Russia. *Appl Radiat Isotopes* **62**, 219-224, doi:DOI 10.1016/j.apradiso.2004.08.008 (2005).

6 Rohland, N. & Hofreiter, M. Comparison and optimization of ancient DNA extraction. *BioTechniques* **42**, 343-352 (2007).

7 Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic acids research* **40**, e3, doi:10.1093/nar/gkr771 (2012).

8 Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic acids research* **38**, e87, doi:10.1093/nar/gkp1163 (2010).

9 Fu, Q. *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 2223-2227, doi:10.1073/pnas.1221359110 (2013).

10 Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols* **2010**, pdb prot5448, doi:10.1101/pdb.prot5448 (2010).

11 Meyer, M. *et al.* From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic acids research* **36**, e5, doi:10.1093/nar/gkm1095 (2008).

12 Dabney, J. & Meyer, M. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques* **52**, 87-+, doi:Doi 10.2144/000113809 (2012).

Supplementary Information 2a
Processing and Mapping

Gabriel Renaud, Fernando Racimo, Kay Prüfer*
* To whom correspondence should be addressed (pruefer@eva.mpg.de)

This section describes the processing of the Altai and Mezmaiskaya Neandertal sequence data from the raw nucleotide intensities to the final aligned sequences that were used for genotyping. This processing includes basecalling, alignment to the human and chimpanzee reference genomes, duplicate removal and indel realignment. We also describe the reprocessing of the previously released low-coverage data from Vindija 33.16, 33.25 and 33.26, El Sidrón 1253, Feldhofer 1 and Feldhofer 2. We report the number of aligned sequences for each sample.

Data availability
All sequence data have been submitted to the European Nucleotide Archive (ENA) and are available under the following accessions: Altai Neandertal: ERP002097, Mezmaiskaya Neandertal: ERP002447.

Base calling and raw sequence processing
The 8 libraries from the Altai and Mezmaiskaya Neandertals described in SI 1 were sequenced with 2×95 cycles insert reads and 2x7 cycles index reads on the Illumina HiSeq 2000, processed with RTA 1.13.48 and base-called from the raw intensity files using Ibis[1] 1.1.6. Reads with the correct library index sequence combinations[2] were identified. Any read for which either index sequence has one or more bases with a base quality score of 10 or less is excluded in downstream analysis[3].

For paired-end reads exceeding the length of the library insert, adapters were trimmed and the overlapping sequences were merged into a single sequence by calling a consensus as described in ref.[4]. Partly overlapping paired-end reads were also merged into a single sequence if the overlap spanned at least 11 nucleotides, and a consensus is called on the overlapping stretch[4]. Paired-end reads with less than 11 bases overlap were kept as separate reads. After merging, sequences having at least 5 bases with a quality score < 15 were flagged as poor quality and excluded from further analysis.

Reducing the effects of cytosine deamination on genotyping
The Altai Neandertal sequences produced from libraries prepared using the single-stranded protocol (ie: L9198, L9199, L9302 and L9303) still show signals of cytosine deamination at the first base or last two bases of sequenced reads[5]. Cytosine deamination creates deoxyuracils, which are read as thymines during sequencing and cannot be readily identified and excluded using base quality information. They can therefore influence variant calling. To address this we reduced to 2 the base quality of any 'T' in the first base or last two bases of sequences from all libraries.

Mapping and Duplicate Removal
Reads were mapped to the human genome (GRCh37/1000 Genomes release) and the chimpanzee genome (CGSC 2.1/pantro2) using BWA[6] version 0.5.10 with parameters "–n 0.01 –o 2 –l 16500". These parameters deactivate seeding, and allow more substitutions and up to two gaps. Identical parameters were used for the analysis of the Denisova genome[5]. After mapping, the following sequence reads were discarded: unmapped single reads, completely unmapped pairs, QC-failed single reads and reads shorter than 35bp.

For each library, sequences mapping to the same outer reference coordinates were replaced by a consensus sequence to collapse duplicate reads. Since reads were aligned individually (rather than combined as in a multiple sequence alignment), some reads may differ in the placement of insertions and deletions to the reference. We therefore call the consensus from reads that show the most

common insertion/deletion placement pattern (according to the cigar line in the BAM file). The consensus at each position was inferred as the base with the highest sum of base quality scores; the base quality score was calculated as the difference between the sum-of-qualities for the highest ranking base and the sum-of-qualities for the second highest ranking base, and then limited to 60. After duplicate removal, the sequences from all libraries were combined with *samtools merge* into one BAM file for each chromosome.

Indel Realignment

We used the Genome Analysis Toolkit[7] (GATK) v1.3-14-g348f2db to identify segments of our sequencing reads that contain an implausibly large number of differences to the reference genome (RealignerTargetCreator). We then realigned sequences in these regions using the GATK IndelRealigner. After realignment, the NM/MD fields were recalculated using *samtools fillmd*, and sequences that had an edit distance of more than 20% of the sequence length were removed. The resulting number of reads is shown in Table S2a.1.

Table S2a.1. Number of aligned reads for the two Neandertals sequenced for this study

	Altai Neandertal		Mezmaiskaya Neandertal	
Reference	**Aligned reads**	**With MAPQ ≥ 30**	**Aligned reads**	**With MAPQ ≥ 30**
GRCh37	2,278,039,997	1,927,490,046	32,909,631	23,589,975
pantro2	1,826,919,328	1,505,502,977	32,622,697	21,095,355

Table S2a.2. Library information and per library mapping statistics for the Altai and Mezmaiskaya Neandertals

Library	Sample	Average insert size (bp)	Library type	Mapped sequences
L9105	Altai Neandertal	99.3	Double stranded+UDG	108,702,535
L9198	Altai Neandertal	66.2	Single stranded	737,839,272
L9199	Altai Neandertal	76.7	Single stranded	209,166,129
L9302	Altai Neandertal	71.1	Single stranded	571,800,334
L9303	Altai Neandertal	70.6	Single stranded	583,673,827
L4533	Mezmaiskaya	47.4	Double stranded+UDG	10,474,074
L4677	Mezmaiskaya	47.4	Double stranded+UDG	135,420
L4678	Mezmaiskaya	46.9	Double stranded+UDG	208,782
L4740	Mezmaiskaya	46.8	Double stranded+UDG	14,256,839
L4741	Mezmaiskaya	46.8	Double stranded+UDG	14,514,275

"Mapped sequences" refers to mapped (to hg19), merged sequences, or properly paired mates (counted as 1 sequence per pair).

Low-coverage Neandertal data from previous publications

For downstream analyses, we also re-mapped shotgun Neandertal sequences from ref.[8]: Feldhofer 1, Feldhofer 2, Mezmaiskaya 1, Sidron 1253, Vindija 33.16, Vindija 33.25 and Vindija 33.26. The previously released Mezmaiskaya 1 data stems from the same individual as the additional Mezmaiskaya data introduced in this study. All sequences were aligned to GRCh37/1000 Genomes Release and CGSC 2.1/pantro2 using BWA and subjected to the same post-mapping processing as the Altai Neandertal and Mezmaiskaya data described above. We did not adjust the quality scores of thymines at the beginning and end of reads, since the original libraries for these samples were not

UDG-treated and therefore show cytosine deamination beyond the first base and last two bases of reads. Table S2a.3 summarizes the number of reads for each of the archaic samples for which we have shotgun sequence data.

Table S2a.3. Summary of archaic data used in this study

Sample	**Reads mapping to *GRCh37***	**Reads mapping to *pantro2***
Feldhofer 2	3,032	2,431
Sidron 1253	35,047	31,955
Feldhofer1	36,494	33,450
Mezmaiskaya 1 (Green et al. 2010)	934,027	870,798
Vindija 33.25	20,200,797	18,820,256
Vindija 33.25	20,244,154	18,981,331
Vindija 33.16	24,018,978	22,248,906
Mezmaiskaya 1 (this study)	32,909,631	32,622,697
Altai Neandertal	2,278,039,997	1,826,919,328
Denisova	1,418,957,698	1,355,814,532

Note: Libraries for the samples in the last three rows of the table (marked in green) were UDG-treated to remove deoxyuracils (see SI1 and Meyer et al. 2012 for details of the library preparation protocol). All other samples are from Green et al. 2010 and contain more extensive deamination since they are from libraries that were not UDG-treated.

References

1 Kircher, M., Stenzel, U. & Kelso, J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome biology* **10**, R83, doi:10.1186/gb-2009-10-8-r83 (2009).

2 Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols* **2010**, pdb prot5448, doi:10.1101/pdb.prot5448 (2010).

3 Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic acids research* **40**, e3, doi:10.1093/nar/gkr771 (2012).

4 Kircher, M. Analysis of high-throughput ancient DNA sequencing data. *Methods Mol Biol* **840**, 197-228, doi:10.1007/978-1-61779-516-9_23 (2012).

5 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).

6 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

7 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

8 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722, doi:10.1126/science.1188021 (2010).

Supplementary Information 2b

Altai Neandertal Mitochondrial Genome Sequence

Susanna Sawyer*

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

Mitochondrial sequence determination

In order to assure that the correct mtDNA sequence has been determined for the Altai Neandertal, we aligned four of the HiSeq lanes (lanes 5 to 8) that were sequenced for L9198 (see SI1) to the Vindija 33.16 mitochondrial genome (AM948965). Since we used BWA to align the sequences to the reference genome, and BWA does not map sequences successfully to the beginning and end of a circular genome, we added 240 of the first base pairs to the end of the Vindija 33.16 genome to assure equal coverage of the sequences across the mtDNA genome.

The four HiSeq lanes were processed as described in SI2. Only merged sequences, sequences flagged as mapped and with a length greater than or equal to 35 were considered for this analysis. The filtered sequences were then aligned to an unmodified Vindija 33.16 mitochondrial genome using MIA[1] (https://github.com/udo-stenzel/mapping-iterative-assembler; parameters: -H 1 -i -c). A total of 268,551 sequences aligned, resulting in an average coverage of 1088 fold (Figure S2b.1A). MIA was used to call an mtDNA consensus, which was used for the subsequent analyses. The support for the consensus is high, with no position having a major allele frequency below 0.93 (Figure S2b.1B).

Assessment of the mtDNA relationship between Neandertals

The consensus sequences of the Altai Neandertal and the mtDNA sequences of six previously published Neandertals (Mezmaiskaya – FM865411.1, Feldhofer 1 – FM865407.1, Feldhofer 2 – FM865408.1, Vindija 33.16 – AM948965, Vindija 33.25 – FM865410.1 and Sidron 1253 – FM865409.1) as well as five modern humans (San – AF347008, Yoruba – AF347014, Han Chinese – AF346972, French – AF346981 and Papuan – AF347004), the Denisova finger bone (NC_013993.1) and Chimpanzee (X93335.1) were aligned to each other using the software MAFFT[2,3] v7.017b. The phylogenetic relationship of the aligned sequences was then computed using a Bayesian framework with MrBayes[4,5] 3.2. First a model test was implemented with jModelTest[6] 2.1.3, which indicated that the GTR+G+I substitution model is most suited to the data. MrBayes was then run with default MCMC parameters with the above substitution model for 5,000,000 generations sampling every 1000 generations with a burn-in of 1,000,000 generations. The 4000 resulting trees were then combined into a consensus tree using TreeAnnotator v1.6.2 (http://beast.bio.ed.ac.uk/TreeAnnotator) from the BEAST package [7]. The combined tree was visualized using Figtree v1.3.1. (http://tree.bio.ed.ac.uk/software/figtree/) from the BEAST package[7] (Figure S2b.2). Pairwise mtDNA nucleotide differences between the seven Neandertals, five present-day humans, Denisova finger bone and Chimpanzee were calculated using MEGA[8] 5.05 after the sequences were aligned with MAFFT[2,3] v7.017b (Table S2b.1).

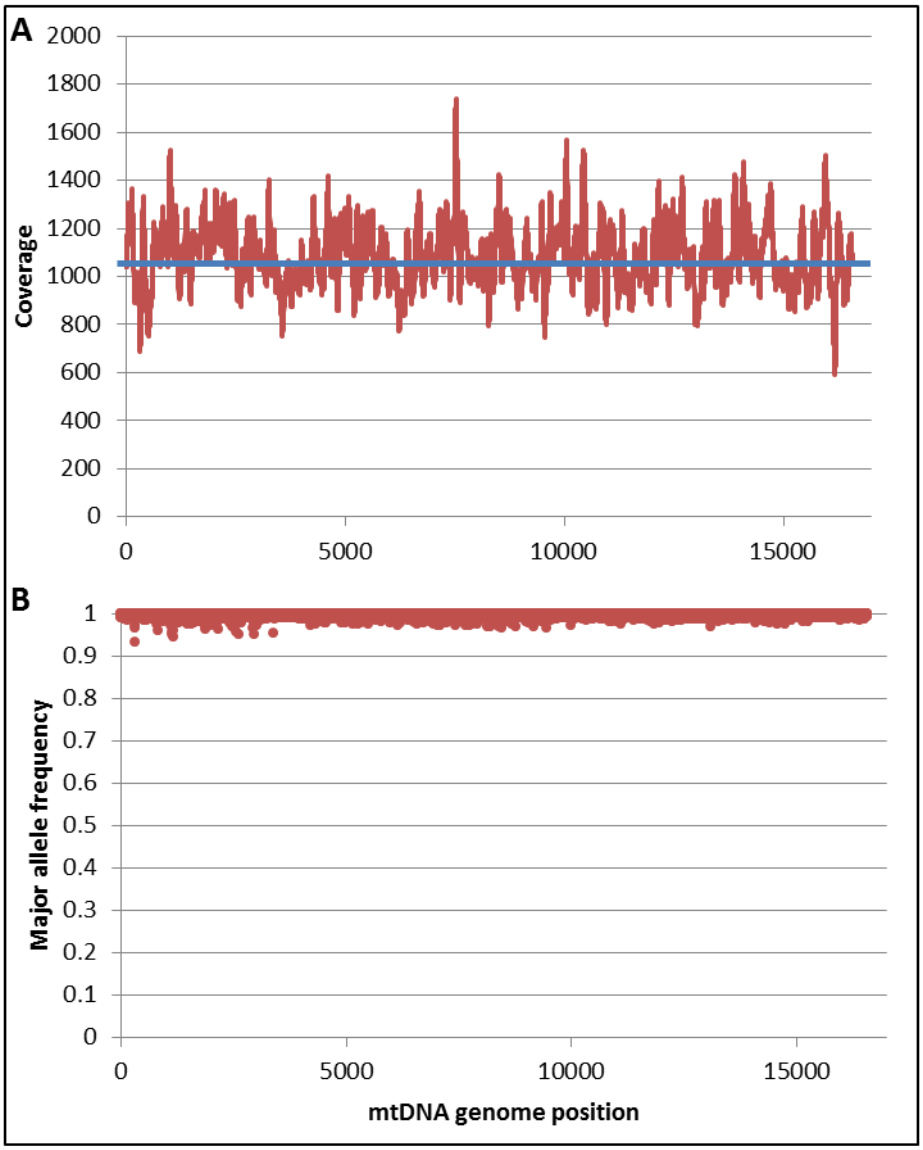


Figure S2b.1. A, The coverage of Altai Neandertal sequences that aligned to the Vindija 33.16 Neandertal genome across the mtDNA genome. The average coverage of 1088-fold is indicated with a blue line. B, The frequency of the major allele at each position across the mtDNA genome. The lowest support for a position is position 302, with a major allele frequency of 0.93.

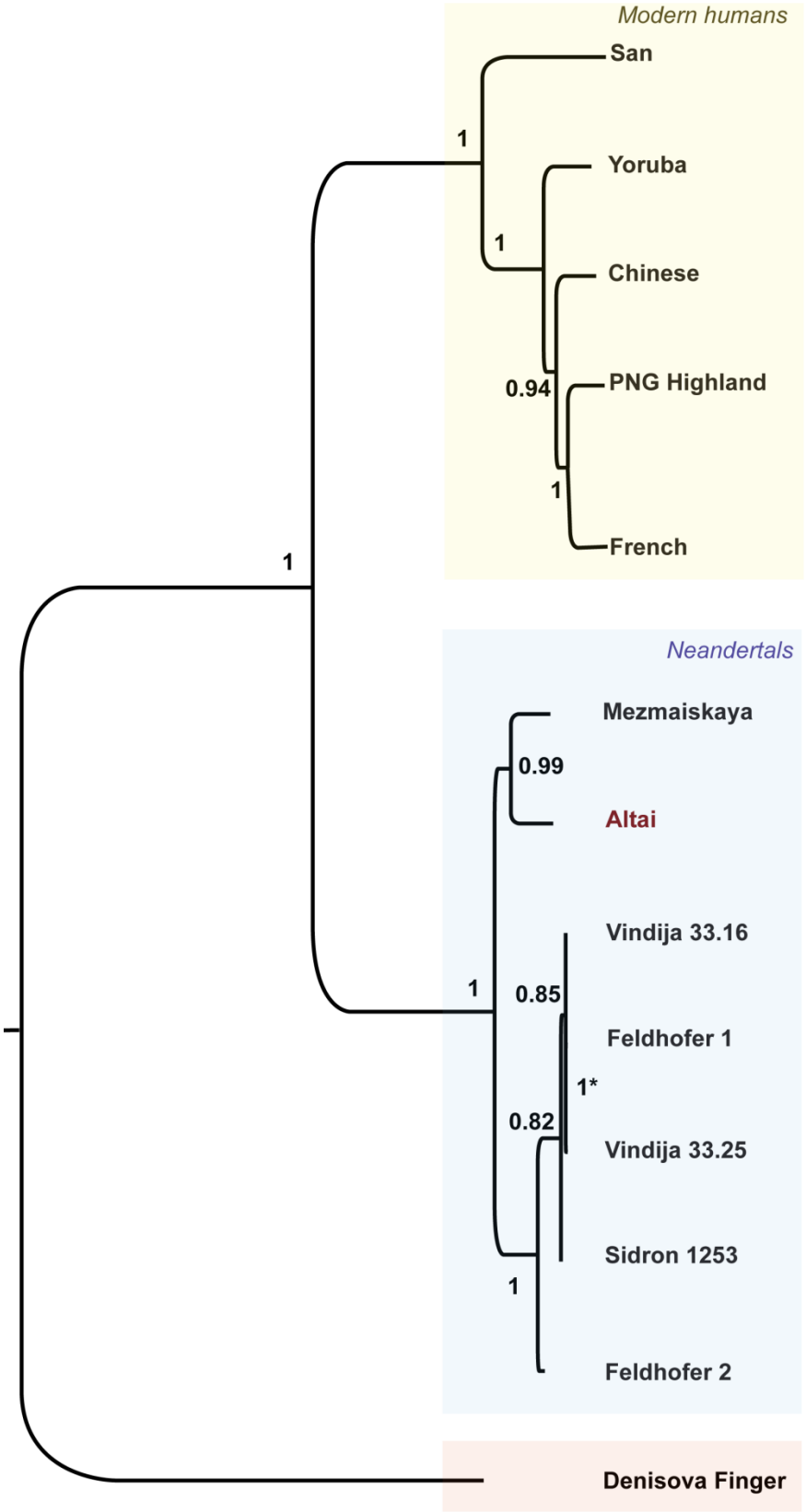


Figure S2b.2. A bayesian mtDNA tree with posterior probabilities. Vindija 33.25 and Feldhofer 1 form a clade with the posterior probability of 1 (shown as 1*). The Chimpanzee was used as an outgroup but is not shown.

doi:10.1038/nature12886

Table S2b.1. Pairwise nucleotide differences between the mtDNA sequences of six previously published Neandertals (in brown), the Altai Neandertal (in red), five present-day humans (in blue), the Denisova finger bone (in green) and the Chimpanzee (in black).

	Feldhofer 1	Vindija 33.25	Vindija 33.16	Feldhofer 2	Sidron 1253	Mezmaiskaya	Altai	French	PNG Highland	Chinese	Yoruba	San	Denisova finger bone
Vindija 33.25	0												
Vindija 33.16	10	10											
Feldhofer 2	10	10	8										
Sidron 1253	10	10	10	6									
Mezmaiskaya	46	46	46	40	44								
Altai	**49**	**49**	**51**	**45**	**47**	**35**							
French	196	196	198	192	192	193	187						
PNG Highland	205	205	207	201	201	204	198	25					
Chinese	197	197	199	193	193	194	188	33	40				
Yoruba	194	194	196	190	190	193	189	38	43	39			
San	210	210	212	206	206	207	199	92	99	97	96		
Denisova finger bone	378	378	378	370	376	370	368	384	385	387	384	381	
Chimpanzee	1433	1433	1433	1429	1433	1434	1433	1452	1457	1453	1456	1463	1459

References

1 Briggs, A. W. *et al*. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**, 318-321, doi:10.1126/science.1174462 (2009).

2 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066 (2002).

3 Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol*, doi:10.1093/molbev/mst010 (2013).

4 Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310-2314, doi:10.1126/science.1065889 (2001).

5 Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574 (2003).

6 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**, 772, doi:10.1038/nmeth.2109 (2012).

7 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**, 1969-1973, doi:10.1093/molbev/mss075 (2012).

8 Tamura K, P. D., Peterson N, Stecher G, Nei M, and Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* **28**, 2731-2739 (2011).

Supplementary Information 3

Extended VCF processing

Fernando Racimo* and Martin Kircher

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

GATK genotype calls

We used GATK's Unified Genotyper[1] to call genotypes for all genomic sites (--output_mode EMIT_ALL_SITES --genotype_likelihoods_model BOTH). Genotype calls were made separately for the Altai Neandertal and each of the Panel B modern humans each aligned to both the human (GRCh37, version of the 1000 Genomes project[2]) and the chimpanzee (panTro2) reference genomes. In order to keep the processing identical to the previously analyzed high-coverage data of the Denisovan and 11 modern human genomes (Panel A), we used GATK version 1.3 (v1.3-14-g348f2b) and performed the "re-call" procedure described in Supplementary Note 6 of Meyer et al. 2012 (ref.[3]). This allows for an individual to have sites where both alleles differ from the reference. Genotypes were also produced for the six low coverage Neandertal genomes (Feldhofer 1 and 2, El Sidrón 1253, Vindija 33.16, 33.25 and 33.26) and the Mezmaiskaya Neandertal (MezE733) used in SOM6a and SOM7, but since the majority of sites in these genomes have very low coverage, the latter part of the processing (the "re-call") was skipped.

Updated pipeline for Extended VCF creation

After calling genotypes, we incorporated further information to the Variant Call Format (VCF) files such as mapability scores, conservation scores and inferred ancestor alleles, as previously described in Meyer et al. 2012 (ref.[3]). We created extended VCFs for the newly sequenced Altai Neandertal genome, the Denisovan individual from Meyer et al. 2012, as well as the 25 present-day human genome sequences generated as described in Meyer et al. 2012 (11 individuals from "panel A") and in SI 4 (14 individuals from "panel B"). We updated the extended VCF creation pipeline and incorporated the following additional information:

1) Non-reference alleles obtained from 1000 Genomes Project (1000G) data[2] 20110521 release that do not appear in the called individual no longer appear in the ALT field; instead, the ALT field only describes non-reference alleles that are present in the reads used for calling the individual. Alternative alleles from 1000G can still be found in the 1000gALT subfield of the INFO field.

2) Genomic Evolutionary Rate Profiling (GERP) scores were incorporated into the subfield GRP of the INFO field. This data was obtained from the UCSC genome browser[4]. The scores are per-base estimates of evolutionary constraint using maximum likelihood evolutionary rate estimation on a 35-mammal alignment. They are indicative of putative functional elements[5,6].

3) An additional systematic error flag was incorporated into the subfield SysErrHCB of the INFO field. This flag marks positions identified as systematic errors based on shared SNPs with high strand bias in human, chimpanzee and bonobo exomes (Castellano et al. in submission).

4) All reference sequence bases from the Ensembl Compara EPO 6 primate alignments[7,8] (Ensembl release 64) at a given position are now present in the "TSseq" subfield of the INFO field (comma separated and in the same order as the TS subfield).

References

1 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

2 Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:nature09534 [pii]
10.1038/nature09534 (2010).

3 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).

4 Fujita, P. A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic acids research* **39**, D876-882, doi:10.1093/nar/gkq963 (2011).

5 Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-913, doi:gr.3577405 [pii]
10.1101/gr.3577405 (2005).

6 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).

7 Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814-1828, doi:gr.076554.108 [pii]
10.1101/gr.076554.108 (2008).

8 Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res* **18**, 1829-1843, doi:gr.076521.108 [pii]
10.1101/gr.076521.108 (2008).

Supplementary Information 4

25 deep genomes from present-day humans of which 13 are experimentally phased

Jacob O. Kitzman, Heng Li, Swapan Mallick, Arti Tandon, Hélène Blanche, Howard Cann, Jay Shendure and David Reich*

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

(i) Sample preparation and shotgun sequencing

We previously reported deep whole genome sequences (WGS) for 11 individuals from diverse populations[1] (10 from the CEPH-Human Genome Diversity Panel[2] and 1 Dinka individual from Sudan[3]). We call these individuals "Panel A" in this paper.

In this new study, we supplement this dataset with 14 new individuals: "Panel B". These are:

- 11 individuals that are different from those in Panel A but are from exactly the same set of populations and have the same provenance. The informed consent procedure for these samples was previously discussed in ref. 1. We thank Michael Hammer for sharing this sample.
- 1 Mixe Native American for which there is informed consent for genetic studies of population history and for which HLA, microsatellite, and SNP genotypes have previously been reported[4,5,6]. We thank Cheryl Winkler and William Klitz for sharing this sample.
- 2 aboriginal Australians from a diversity panel maintained at the European Collection of Cell Cultures (ECCAC). These samples were collected with informed consent for genetic analysis of population diversity, and genome-wide SNP data for these samples were previously reported[7,8]. We do not know the geographic source of these samples within Australia; however, our genetic analyses indicate that they do not have recent admixture with non-Australian populations, for example Europeans or East Asians, and that they form a deep clade with Papuans[7,8]. At our request, the ECCAC carried out an independent re-review to determine if these samples were appropriate for whole genome sequencing and dissemination of data, and approved their use for this purpose.

For the 14 Panel B individuals, we shipped DNA to Illumina Inc. (San Diego, USA). We shipped 3-5μg for all samples made from cell lines, and less for the DNK07 sample. Illumina prepared TruSeq libraries with a tight size selection (average insert length of 328 bp and average standard deviation of 37 bp) using their in-house protocol and sequenced them on HiSeq2000 instruments for 2×100 cycles.

(ii) Mapping and generation of BAM files

We used BWA[9] version bwa-0.5.9 to map the reads from each of the Panel B individuals to the human reference genome (*hg19/GRCh37* which we extended by adding the Epstein Barr virus) and to chimpanzee (*panTro2*). We used the command "bwa aln -q15", which removes the low-quality ends of reads. We marked potential PCR duplicates with Picard (http://picard.sourceforge.net/). The mappings of reads to both *hg19/GRCh37* and *panTro2* have been deposited in BAM format as a Public Data Sets on Amazon Web Services (http://aws.amazon.com/datasets), and are freely and publicly available there along with VCF files containing the genotype calls we used in this study (and also the Altai and Denisova data). Table S4.1 reports summary statistics for both the 11 Panel A individuals reported in ref. 1, and the 14 Panel B individuals newly reported in this study.

It is important to highlight the differences in the data for Panels A and B. For Panel A, we prepared four barcoded libraries for each of the 11 individuals using the method of ref. 10, and then combined all 44 into a single pool in approximately equimolar amounts. Since the libraries for all Panel A samples were sequenced together, the possibility of artifactual differences across samples due to differences in sequencing machines or reagents was minimized. In contrast, for the Panel B set of 14 individuals, we simply shipped individual tubes of DNA to Illumina. The Panel A and Panel B sequences are also different in terms of GC bias and insert length distribution.

doi:10.1038/nature12886

Table S4.1. Summary statistics on shotgun sequencing of 25 present-day human samples

Population	Panel	Sample ID	Sex	Mapping to human reference *hg19*				Mapping to chimpanzee reference *PanTro2*			
				Mapped reads	Reads part of proper pairs	After duplicate removal	% of reads used	Mapped reads	Reads part of proper pairs	After duplicate removal	% of reads used
Dinka	A	DNK02	M	947,105,021	930,831,666	851,171,031	76.37%	916,549,057	884,628,712	801,553,975	71.91%
Mbuti*	A	HGDP00456	M	886,228,397	866,931,212	746,142,578	79.36%	848,074,664	814,965,242	686,143,163	72.98%
French*	A	HGDP00521	M	993,077,880	971,254,974	822,589,731	78.34%	940,705,102	903,150,342	760,166,624	72.40%
Papuan*	A	HGDP00542	M	957,209,275	935,417,074	802,115,913	79.00%	915,884,235	878,398,804	734,275,412	72.32%
Sardinian*	A	HGDP00665	M	916,389,106	892,615,658	767,127,809	79.15%	873,543,296	835,455,770	700,043,844	72.23%
Han*	A	HGDP00778	M	1,005,969,464	978,790,234	862,350,953	80.80%	967,050,867	924,225,448	796,567,857	74.74%
Yoruba*	A	HGDP00927	M	1,184,897,821	1,153,642,906	1,000,910,182	80.30%	1,124,940,368	1,075,787,652	906,738,040	72.75%
Karitiana*	A	HGDP00998	M	996,500,295	968,156,064	813,991,887	76.59%	952,528,353	909,023,542	742,750,359	69.89%
San*	A	HGDP01029	M	1,276,748,296	1,230,278,942	1,056,952,329	77.62%	1,174,425,666	1,109,986,162	890,881,364	65.42%
Mandenka*	A	HGDP01284	M	937,936,405	908,846,748	775,089,405	77.23%	897,038,052	853,579,424	701,087,589	69.85%
Dai*	A	HGDP01307	M	1,002,157,271	983,337,904	880,009,435	83.27%	961,796,375	924,978,454	814,542,161	77.08%
Dinka	B	DNK07	M	1,158,499,813	1,107,910,972	1,049,704,758	72.30%	1,095,360,543	1,029,017,632	998,584,960	68.78%
Mbuti	B	HGDP00982	M	1,169,454,752	1,145,773,878	1,100,916,262	88.44%	1,103,137,610	1,057,524,550	1,046,440,758	84.06%
French	B	HGDP00533	M	1,307,405,760	1,285,999,396	1,219,794,176	88.53%	1,242,743,065	1,196,513,030	1,167,797,135	84.76%
Papuan	B	HGDP00546	M	1,342,439,485	1,317,456,890	1,243,879,150	87.62%	1,259,606,421	1,210,049,074	1,182,374,925	83.28%
Sardinian	B	HGDP01076	M	1,185,446,548	1,167,004,292	1,104,809,095	88.26%	1,122,604,612	1,081,706,380	1,054,826,921	84.26%
Han	B	HGDP00775	M	1,128,858,502	1,105,675,046	1,056,349,592	88.21%	1,064,951,424	1,020,478,846	1,006,227,060	84.02%
Yoruba	B	HGDP00936	M	1,244,868,836	1,216,598,668	1,145,344,324	86.31%	1,177,758,888	1,127,691,870	1,092,629,351	82.34%
Karitiana	B	HGDP01015	M	1,115,249,791	1,090,385,210	1,038,851,335	87.48%	1,050,045,069	1,005,464,890	987,833,318	83.18%
San	B	HGDP01036	M	1,200,160,761	1,175,098,710	1,126,237,368	87.84%	1,132,074,063	1,086,103,658	1,074,773,400	83.82%
Mandenka	B	HGDP01286	M	1,172,063,761	1,145,826,934	1,091,587,340	87.82%	1,100,277,526	1,053,489,210	1,037,551,934	83.48%
Dai	B	HGDP01308	M	1,170,307,516	1,148,730,494	1,100,561,917	88.80%	1,107,295,709	1,063,500,330	1,049,421,241	84.68%
Australian*	B	WON,M	M	1,268,632,894	1,251,777,542	1,214,027,921	91.12%	1,201,522,870	1,159,797,770	1,157,904,245	86.91%
Australian*	B	BUR,E	F	1,317,339,767	1,294,841,294	1,246,624,085	89.88%	1,250,529,869	1,201,657,988	1,190,061,965	85.80%
Mixe*	B	MIXE0007	F	1,279,650,592	1,244,992,900	1,113,486,997	81.40%	1,207,250,661	1,148,786,420	1,062,827,309	77.69%

Notes: Data from the 11 Panel A individuals were previously published[1], while data from the 14 Panel B individuals are new.
* Data for the 13 individuals indicated with an asterisk (10 Panel A and 3 Panel B) were also experimentally phased.

(iii) Fosmid libraries and sequencing

We generated experimentally phased genomes for 13 individuals: 10 from Panel A and 3 from Panel B. For the phasing, we followed the method of ref. 11 with minor modifications.

Briefly, frozen pellets of lymphoblastoid cells were prepared and were the sources for genomic DNA used for fosmid construction. For the ten HGDP-CEPH samples, approximately twenty million packed, PBS-washed and dried cells were prepared at CEPH in Paris, France and then shipped frozen to Seattle, USA. For the three other samples, the cell lines were processed in Seattle.

High molecular weight genomic DNA was extracted with the Gentra Puregene kit (Qiagen) and sheared to approximately 40 kb using a Hydroshear instrument (Digilab) for 20 cycles at speed code 16. The sheared DNA was size selected to 38-40 kb after pulsed field electrophoresis (18 hours at 170V, initial A=1, final A=6, 1% low-melt agarose in 0.5X TBE buffer) and visualized with SYBR Gold stain (Invitrogen). Excised gel slices were melted at 70ºC for 10 minutes in a water bath, spun at 15,000g for 1 minute, and cooled to 47ºC for 1 minute. Agarose was digested using beta-agarase (Promega, ½ unit per 200mg molten agarose) for 1 hour at 47ºC. Digests were split into 500ul fractions, cooled on ice for 4 minutes, and the remaining gel was centrifuged at 15,000g for 20 minutes. DNA precipitation was carried out on ice for 1 hour by addition of 50ul 3M NaOAc and 1ml 100% EtOH, and pelleted by centrifugation at 15,000g and at 4ºC for 45 minutes. Pellets were rinsed twice with cold 70% EtOH and resuspended in 20ul LoTE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8). The DNA was end-repaired using the End-IT kit (Epicentre), followed by cleanup using standard phenol-chloroform extraction and ethanol precipitation. End-repaired DNA was ligated to the pCC1Fos fosmid vector in a 10ul reaction, and incubated overnight at 16ºC with 2000 U T4 DNA Ligase (New England Biolabs). Clone packaging, infection, and titration were performed using the CopyControl fosmid cloning kit (Epicentre) following the manufacturer's instructions. For each fosmid library, a single bulk infection culture was split by dilution into pools of 1,000 - 5,000 clones each in 1.5ml media, and grown out in deep 96-well plates. Fosmid clone DNA was then isolated in sets of 96 by standard alkaline lysis miniprep. Clone pools were converted to shotgun libraries for Illumina sequencing with the "version 1" Nextera library preparation kit (Epicentre), using 0.1ul transposase enzyme per reaction. The resulting sheared libraries were amplified and tagged by PCR using a set of 96 distinct barcoding primers as previously described[12].

Each set of 96 barcoded clone pool sequencing libraries was combined and sequenced with a 9 base pair index read to allow separation by pool after sequencing. The median number of clones per pool was 1,885, corresponding to an average of 2.2% clone coverage per genome. This clone coverage is sparse enough that for the great majority of sites, the coverage is 0 or 1 clone per fosmid pool. When only 1 clone covers a region, there is no ambiguity as to the haplotype to which a read belongs so that sequences are effectively phased over the ~35 kb of the fosmids. On average, we had 7.6-fold fosmid coverage per position in the genome, with a range of 3.9 to 17.8 across the 13 samples (Table S4.2).

Shotgun sequencing of the fosmids pools was performed using 75bp paired-end reads for the 10 Panel A individuals (at the Beijing Genome Institute, Shenzhen, China) or 50bp paired-end reads for the 3 Panel B individuals (at the Harvard Medical School Biopolymers Facility, Boston, USA). Summary statistics on the amount of sequencing performed and the phasing results are reported in Table S4.2.

(iv) Computational phasing

We used BWA to align reads sequenced in the fosmid pools to the human reference genome (*hg19*/*GRCh37*). We identified fosmids by using a simple score-based Hidden Markov Model (HMM). This HMM models the per-base coverage along the reference genome. It has two hidden states: *F* to emit a site on the fosmids and *N* to emit a site not on the fosmids. By definition, state *F* emits non-zero coverage while *N* emits zero coverage. We use a dynamic programming algorithm to find the optimal path where *F* states identify the fosmid regions. We manually set the transition and emission scores such that the identified fosmid segments are ~35kb and not too fragmented or

misjoined. The consensus of the fosmids is then generated automatically by the HMM. The output of this step is a SAM file in which each "read" represents an aligned fosmid sequence.

Our phasing algorithm follows ref. 13. Briefly, we begin by merging the fosmid alignment and the short-read alignment. We identify potential heterozygous sites using SAMtools[14] and implicitly build a weighted graph where a vertex at position i is a binary vector of size k (15 by default) representing alleles at k contiguous heterozygous sites ending at i. An edge (v_i,v_{i+1}) is present if there exists a read (a short read, a read pair or a fosmid) consistent with both vertices v_i and v_{i+1}. We set the weight of an edge to be the number of supporting reads. We use dynamic programming to identify the best weighted path which represents the pair of complementary haplotypes supported by the most reads. This resolves the phase. If no reads contain a gap consisting of k or more uncalled heterozygotes, the algorithm is guaranteed to find the optimal phasing. However, due to uneven coverage across fosmids (which arises due to the fact that fosmids in the pool are not all represented at equal molarity), fosmid sequences may contain gaps. In this case, there is a possibility of switch errors when the phase across the gap cannot be inferred for any other reads.

After the initial phasing, we align each read back to the two haplotypes to identify its phase. We mark a read as chimeric if it is clearly joining the two haplotypes (which may happen if we misjoin two fosmids from opposite haplotypes), and we break such reads into two. We mark a read as ambiguous if there is no clear switch point and it is about equally distant to both haplotypes in terms of hamming distance. Ambiguous reads are dropped. By aligning reads to the phased haplotypes, we also identify regions that strongly violate the 2-haplotype assumption, which may reflect mismapping, copy number changes, or reference genome misassembly. We flag these and ignore them for analyses.

The output of the phasing pipeline is a master file showing the phased regions and the haplotypes composed of the alleles at candidate heterozygous sites. We perform haploid SNP calling later on these BAMs to derive the haploid consensus sequence. Table S4.2 gives summary statistics on the phasing, which was variable in its quality with a mean N50 contig size of 471 kb over all samples and a range from 222 kb (for Yoruba individual HGDP00927) to 839 kb (for Australian individual WON,M). The 13 phased genomes are available as an Amazon Web Services Public Data Set (http://aws.amazon.com/datasets) along with the BAM and VCF files for A and B panel individuals.

doi:10.1038/nature12886

Table S4.2. Summary statistics on the 13 phased genomes

Sample	Panel	Sample ID	Gb of reads from short inserts*	Gb of reads from fosmids*	Number of fosmid pools	Mean fosmids per pool	Fosmid clone coverage of genome†	Mean ± std. dev. read coverage per fosmid	Length of phased blocks (Gb)	N50 size of phased contigs (kb)
Mbuti	A	HGDP00456	72	54	192	2,044	4.9	3.15 ± 3.12	2.40	324
French	A	HGDP00521	80	63	288	1,528	5.5	3.41 ± 4.20	2.49	331
Papuan	A	HGDP00542	78	75	480	1,055	6.3	3.57 ± 3.52	2.56	687
Sardinian	A	HGDP00665	74	51	192	3,316	8.0	1.75 ± 1.50	2.52	538
Han	A	HGDP00778	84	49	288	1,237	4.5	3.38 ± 3.48	2.46	335
Yoruba	A	HGDP00927	97	28	192	1,610	3.9	2.20 ± 2.45	2.42	222
Karitiana	A	HGDP00998	79	54	192	3,920	9.4	1.38 ± 1.46	2.46	532
San	A	HGDP01029	103	51	192	1,974	4.7	3.20 ± 2.79	2.56	392
Mandenka	A	HGDP01284	75	59	384	1,180	5.7	3.30 ± 3.17	2.62	608
Dai	A	HGDP01307	85	51	192	2,047	4.9	2.86 ± 2.49	2.38	273
Australian1	B	WON,M	118	58	288	4,952	17.8	0.67 ± 0.35	2.53	839
Australian2	B	BUR,E	121	44	192	5,146	12.4	0.80 ± 0.23	2.51	700
Mixe	B	MIXE0007	108	38	192	4,701	11.3	0.67 ± 0.19	2.26	345

* After duplicate removal.

† (Fosmid clone coverage of genome) = (Number of fosmids pools) x (Mean fosmids per pool) x (35,000 bp) / (2,800,000,000 bp / genome)

References

[1] Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222-6.

[2] Cann HM, de Toma C, Cazes L, Legrand Marie-Fernande, Morel V, Piouffre L, Bodmer J, et al.(2002) A human genome diversity cell line panel. Science 296, 261-2.

[3] Cox MP, Mendez FL, Karafet TM, Pilkington MM, Kingan SB, Destro-Bisol G, Strassmann BI, Hammer MF (2008) Testing for archaic hominin admixture on the X chromosome: model likelihoods for the modern human RRM2P4 region from summaries of genealogical topology under the structured coalescent. Genetics 178, 427-437.

[4] Hollenbach JA, Thomson G, Cao K, Fernandez-Vina M, Erlich HA, Bugawan TL, Winkler C, Winter M, Klitz W (2001) HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives. Hum Immunol. 62, 378-90.

[5] Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C, Mazzotti G, Poletti G, Hill K, Hurtado AM, Labuda D, Klitz W, Barrantes R, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Llop E, Rothhammer F, Excoffier L, Feldman MW, Rosenberg NA, Ruiz-Linares A (2007) Genetic variation and population structure in Native Americans. PLoS Genet. 3, e185.

[6] Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, García LF, Triana O, Blair S, Maestre A, Dib JC, Bravi CM, Bailliet G, Corach D, Hünemeier T, Bortolini MC, Salzano FM, Petzl-Erler ML, Acuña-Alonzo V, Aguilar-Salinas C, Canizales-Quinteros S, Tusié-Luna T, Riba L, Rodríguez-Cruz M, Lopez-Alarcón M, Coral-Vazquez R, Canto-Cetina T, Silva-Zolezzi I, Fernandez-Lopez JC, Contreras AV, Jimenez-Sanchez G, Gómez-Vázquez MJ, Molina J, Carracedo A, Salas A, Gallo C, Poletti G, Witonsky DB, Alkorta-Aranburu G, Sukernik RI, Osipova L, Fedorova SA, Vasquez R, Villena M, Moreau C, Barrantes R, Pauls D, Excoffier L, Bedoya G, Rothhammer F, Dugoujon JM, Larrouy G, Klitz W, Labuda D, Kidd J, Kidd K, Di Rienzo A, Freimer NB, Price AL, Ruiz-Linares A (2012) Reconstructing Native American population history. Nature 488, 370-4.

[7] Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A (2011) Adaptations to climate-mediated selective pressures in humans. PLoS Genet. 7, e1001375.

[8] Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AM, Ko YC, Jinam TA, Phipps ME, Saitou N, Wollstein A, Kayser M, Pääbo S, Stoneking M (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet. 89, 516-28.

[9] Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-60.

[10] Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res. 22, 939-46.

[11] Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE, Shendure J (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol. 29, 59-63.

[12] Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, Shendure J (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 11, R119.

[13] He D, Choi A, Pipatsrisawat K, Darwiche A, Eskin E (2010) Optimal algorithms for haplotype assembly from whole-genome sequence data. Bioinformatics 26, 183-90.

[14] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.

Supplementary Information 5a

Genome Data Quality

Fernando Racimo*, Susanna Sawyer, Cesare de Filippo, Michael Siebauer, Matthias Meyer, Philip L.F. Johnson, Michael Lachmann, Kay Prüfer*

* To whom correspondence should be addressed (fernando_racimo@eva.mpg.de, pruefer@eva.mpg.de)

We calculated basic statistics and estimated contamination for sequences from Altai Neandertal and Mezmaiskaya libraries. Altai Neandertal libraries showed a consistently high percentage of endogenous sequences (~70%) while Mesmaiskaya libraries yielded around 4% endogenous molecules. The alignment of Altai Neandertal sequences yield an average coverage of 52×, while Mezmaiskaya covered on average only half the genomic positions (0.5×).

We used several methods to estimate the fraction of contaminating human sequences in the Altai and Mezmaiskaya Neandertal data. Based on contamination by human mitochondrial sequences we estimate a contamination rate of 0.6% and 0.8% for Mezmaiskaya and Altai Neandertal respectively. When testing for male Y-chromosomal sequences among the sequences of the female Mezmaiskaya and Altai samples, we estimate a rate of contamination by male sequences of ~0.5% for both Mezmaiskaya and Altai Neandertal. Autosomal contamination was estimated using the reads covering positions where humans carry a fixed derived variant compared to great apes. The test gave an estimate of 0.8% contamination for Altai and 0.001% for Mezmaiskaya. A last test makes use of the the deep coverage and the inbred tracks in Altai Neandertal. In these regions of homozygosity, reads not matching the Altai Neandertal consensus must be sequencing error or contamination. This analysis gave a point estimate of 1.2% autosomal contamination.

DNA library characteristics

We first looked at four aspects of data quality for all of the ancient DNA libraries prepared: the percent of sequences that mapped to the human genome (GRCh37), the fragment length distribution, the damage patterns and the base composition. These four aspects were assessed independently per library using subsets of the available data; one HiSeq lane was used per single-stranded library for L9198 and L9199. L9302 and L9303 were pooled together before sequencing, and one HiSeq lane was used from the pool. All available data were used for the double-stranded libraries (two lanes, L9105).

Proportion of endogenous DNA and fragment length distribution

The Altai Neandertal shows a strikingly high proportion of mapped sequences (~ 70% in all libraries) when compared to Mezmaiskaya (~ 4%), as well as the previously sequenced Neandertals from Croatia and Spain (less than 5%)[1]. To date, the only other Pleistocene hominin sample for which a similar proportion of mapped sequences has been reported is the distal manual phalanx that was excavated at the same site and yielded the high-coverage Denisova genome[2]. It is tempting to speculate that the extraordinary DNA preservation and absence of excessive microbial contamination in the two samples may either be due to favorable environmental conditions in this cave or the type of bone that was sampled (both are phalanxes).

Since both single and double-stranded libraries were prepared for the Altai Neandertal, we can evaluate the effects of library preparation. We find that the double-stranded library (L9105) shows a slightly lower proportion of mapped sequences than the libraries made with the single-stranded method (Figure S5a.1). Subtle differences between single and double-stranded library preparation have been observed before[2]. In addition, the single-stranded library method yielded a larger proportion of small fragments compared to the double-stranded method, possibly due to the single-stranded method being more efficient at transforming short fragments into library sequences[2].

Compared to the lanes of sequence that were analyzed for all other libraries, the lane for L9199 has a strikingly non-uniform pattern in its fragment length distribution. We have included the fragment length distribution of all sequences to show that mapped and unmapped sequences follow the same size trajectories and that this signal is not due to human contamination by molecules of specific sizes. To further investigate this, we analyzed data from another lane where L9199 and L9198 were sequenced in a pool and did not find the pattern in the sequences of either library. The non-uniform fragment size distribution is therefore not a property of the library but must be an artifact of the sequencing process affecting eight lanes, which corresponds to 8% of the total sequence data in the Altai Neandertal genome.

Base Composition and Damage Patterns

The base composition plots in Figure S5a.2 show that the double-stranded libraries have a slight bias towards GC-rich sequences, especially in shorter fragments, with GC bias being more pronounced in the Altai Neandertal (L9105) than in the Mezmaiskaya libraries. In contrast, the single-stranded libraries show an AT-rich bias (Figure S5a.2). Both observations are in agreement with those made in the Denisova genome paper[2]. L9199 shows a non-uniform base composition that may be caused by the uneven fragment length distribution described above.

All libraries were treated with uracil-DNA-glycosylase and endonuclease VIII to remove deoxyuracils that arose from cytosine deamination, which would otherwise manifest as C-to-T and G-to-A changes (the latter only occurring with double-stranded library preparation[3]). In double-stranded libraries, only a small residual signal of damage-derived C-to-T and G-to-A substitutions is detected at the 5' and 3' terminal positions (3.5% or lower; see Fig S5a.3: panels Mezmaiskaya and L9105). As shown previously[2],

the single-stranded library method is less effective at removing uracils from terminal nucleotides, especially at the 3' end where we observe C-to-T changes with a frequency of up to 30%.

Genomic coverage and GC Dependence

Using the VCF files generated from alignments to the human reference sequence (GRCh37), we computed coverage statistics for the uniquely mapable regions of the genome for the Altai Neandertal, the Denisova individual and each of the present-day humans from panel A and panel B (Table S5a.1). We observe that coverage is highest for the Altai Neandertal, followed by panel B individuals (Fig. S5a.4), then Denisova and the panel A individuals. The mean coverage for Altai Neandertal is 52×. In comparison, confidently mapped reads (MQ >= 30) to unique regions (map35_100%, see SI5b) of the human chromosome 21 yielded an estimate of 0.48× coverage for Mezmaiskaya.

Next, we used the number of GC bases in the 50 bases flanking each position (25 on each side) to divide the genome into bins of different GC content. The coverage distributions in these classes show a strong GC dependence for both Altai Neandertal and Denisova (Fig. S5a.5). The coverage is highest for AT rich regions of the genome and lowest for GC rich regions.

Previous analyses of the Denisova genome showed that regions of exceedingly high and low coverage in the Denisova and present-day human genomes were associated with high divergence to other genomes, and we observe a similar trend in the new data from this study (Fig. S5a.6). Part of this pattern may be explained by duplications in the genomes that are only present once in the human reference, so that reads from several locations in the sequenced genome are aligned to only one location in the reference genome. This is true even after correcting for local GC content, which as discussed above affects local coverage; Fig. S5a.6 shows that the high divergence artifact occurs at different coverage levels depending on local GC content. We therefore introduced filters, discussed in detail in SI 5b, which remove the lowest 2.5% and highest 97.5% of the ancient genome data based on their GC-corrected coverage.

Contamination estimates

We obtained four complementary estimates of contamination from present-day human DNA.

Mitochondrial contamination estimate

We estimated mitochondrial contamination as described in (1) with 4 changes. First, all sequences shorter than 35 bp in length were removed before fastq conversion and MIA re-alignment. Second, the re-alignment with MIA was done to the revised Cambridge Reference Sequence (rCRS, NC_012920.1). Third, diagnostic positions were defined as those where all 311 human mtDNA sequences differ from the Neandertal mtDNA consensus (instead of 95% of the 311 humans as used in ref.[1]). Fourth, the bases were not clipped ahead of alignment; thus, damage-derived C-to-T changes at the first and last two base pairs of the sequences are frequently observe in the Altai Neandertal data (Figure S5a.3). Contamination was checked using diagnostic

positions as described in ref.[4] with the following change specifically for the Altai Neandertal: a read is not counted as potential contamination (and is instead ignored) if the following four conditions are met—(i) the read overlaps one of the diagnostic positions, (ii) the diagnostic position is a transition, (iii) the nucleotide in the read is a T or A consistent with ancient DNA deamination, and (iv) the sample sequence matches the 311 human mtDNA. The Mezmaiskaya Neandertal has low levels of cytosine deamination (below 3.5%, Figure S5a.2) and therefore did not have the extra damage filter applied for the contamination estimate. The Altai Neandertal contamination estimate used 84 diagnostic positions, while the Mezmaiskaya Neandertal used 89 diagnostic positions. The mitochondrial contamination estimate for all Altai Neandertal libraries combined is 0.78% (95% C.I. = 0.75-0.82), while it is 0.57% (95% C.I. = 0.49-0.65) for the Mezmaiskaya Neandertal. The mtDNA contamination estimates for each library separately is presented in Table S5a.2.

Autosomal contamination estimate

We estimated contamination rates on the autosomes using the same method presented in[2]. This method is a maximum likelihood based co-estimation of sequence error, contamination and two population parameters. The method is based on the assumption that contamination of the sample will contribute human derived alleles for which the ancient individual is ancestral (here, we use human sites that appear fixed derived as compared to great ape outgroups). Contamination and sequence error can thus be inferred from low frequency allele counts at homozygous positions or overrepresented derived alleles at heterozygous sites. We apply the method to the Altai Neandertal and Mezmaiskaya datasets. Reads were required to have a minimal length of 35 (a filter used throughout this paper) and a mapping quality of at least 30. Only regions of high uniqueness were considered (map35_100%, see SI5b) and bases with a quality score below 30 were excluded. The method estimates low contamination in both samples; the estimated contamination was 0.8% (CI: 0.79-0.83%) for Altai Neandertal and 0.0013% (0-1.12%) for Mezmaiskaya.

Autosomal contamination estimate (taking advantage of recent inbreeding)

Third, we took advantage of the fact that the Altai Neandertal appears to be highly inbred (SI 10) to obtain an autosomal contamination estimate without the need to co-estimate heterozygosity. Briefly, we selected regions of the Altai Neandertal genome that are 95% composed of strict runs of homozygosity, defined to be segments >50kb that have no heterozygous sites (SI 10). Treating the few heterozygotes found in these regions as missing data, we calculated the error rates for each possible base change at sites where most humans and Altai Neandertal have the same genotype. We then calculated a likelihood for contamination L(c) using formula S5a.1 for all sites in homozygous regions.

$$L(c) = \prod_{B1->B2} (1-\varepsilon_{B1,B2}-c)^{N_{B1,B2}-k_{B1,B2}} (\varepsilon_{B1,B2}+c)^{k_{B1,B2}} \quad \text{(S5a.1)}$$

where c is the contamination level, $\varepsilon_{B1,B2}$ is the error rate from base B1 to base B2, $N_{B1,B2}$ is the total number of reads overlapping sites we look at where Altai has base B1 and present-day humans are fixed for base B2, and $k_{B1,B2}$ is the number of times we observe base B2 in those reads. Using this simplified method, we obtain a contamination estimate of 1.16% (95%CI: 1.15% - 1.18%) for the Altai Neandertal, which is close to the other estimates.

Y chromosome estimate of contamination (only sensitive to contamination from males)

We calculated a Y-chromosome contamination estimate for both the Altai and Mezmaiskaya Neandertals. Our method relies on the same idea as that in ref.[1]. Briefly, we compute the number of unique Y-chromosome fragments we would expect to observe if the individual were a male. If we see a significant depletion in these fragments, we can reject the hypothesis that the individual is a male and assume that such fragments come from male contamination. We can then use the number of unique Y-chromosome fragments to obtain a contamination estimate, with the caveat that the method can only detect male contamination.

To find unique Y-chromosome fragments, we used the UCSC mapability track of 40-mers (CGR Align 40 (wgEncodeCrgMapabilityAlign40mer) in the UCSC table browser) to filter for positions that had a mapability score of 1 and fall inside regions of a size of at least 500 basepairs of uniquely mapable sequence on the Y chromosome. We then removed regions that overlap with sequences from the four females from the 1000 genomes trio data as was done in ref.[2]. This gives us 744 regions in the Y-chromosome.

To compute the number of Y chromosome fragments we would expect if the Neandertal is a male we used the following formula: (Number of aligned reads in the whole genome of the Neandertal) × (the number positions in the Y-chromosome / (Genome size). The genome size was calculated as 2×(autosomal positions) + (X-chromosome positions) + (Y-chromosome positions). This means that for the Altai Neandertal we expect 1,382,781,988 × 653,451 / 3,443,138,544 = 262,430 Y fragments, and for the Mezmaiskaya Neandertal we expect 17,350,675 × 653,451/ 3,443,138,544 = 3,293 Y fragments.

We then determined the actual number of Y fragments seen in our Neandertals that fall completely within the same Y-chromosome regions. We see 1,450 Y fragments in the Altai Neandertal and 16 Y fragments in the Mezmaiskaya Neandertal. As both Neandertals are most likely female we can use the number of Y fragments seen to calculate contamination by dividing the number of fragments seen by the number of fragments expected if the Neandertal were male. This gives us a contamination estimate of 0.55% (95% CI = 0.52-0.58) for the Altai Neandertal and 0.49% (95% CI = 0.28-0.79) for the Mezmaiskaya Neandertal.

Samples	IDs	Autosomes		X chromosome	
		mean	median	mean	median
Denisova	Denisova	30.7	31	31.3	31
Altai Neandertal	AltaiNea	52.5	51	53.4	52
Dinka (A)	DNK02	28.2	28	14.6	14
Dinka (B)	DNK07	35.4	35	18.0	18
Mandenka (A)	HGDP01284	24.8	25	12.6	12
Mandenka (B)	HGDP01286	37.0	37	18.4	18
Mbuti (A)	HGDP00456	24.6	25	12.6	12
Mbuti (B)	HGDP00982	37.4	37	18.3	18
San (A)	HGDP01029	33.2	33	17.0	17
San (B)	HGDP01036	38.6	38	19.4	19
Yoruba (A)	HGDP00927	32.4	32	16.6	16
Yoruba (B)	HGDP00936	39.0	39	19.4	19
Karitiana (A)	HGDP00998	26.3	26	13.4	13
Karitiana (B)	HGDP01015	35.3	35	17.6	17
Dai (A)	HGDP01307	28.6	28	14.1	14
Dai (B)	HGDP01308	37.3	37	18.8	18
Han (A)	HGDP00778	28.0	28	14.4	14
Han (B)	HGDP00775	35.5	35	17.7	17
French (A)	HGDP00521	27.0	27	13.8	13
French (B)	HGDP00533	42.6	42	21.6	21
Sardinian (A)	HGDP00665	24.9	25	12.8	13
Sardinian (B)	HGDP01076	38.3	38	19.1	19
Papuan (A)	HGDP00542	26.2	26	13.6	13
Papuan (B)	HGDP00546	42.8	43	21.5	21
Mixe (B)	MIXE0007	37.1	37	36.4	36
Australian (B)	WON,M	42.1	42	20.8	20
Australian (B)	BUR,E	42.3	42	41.1	41

Table S5a.1: Mean and median coverage for Denisova, Altai Neandertal, Panel A and B data (labeled (A) and (B) in Samples column). Values are given separately for the autosomes and the X chromosome. All possible 35 basepair windows overlapping a given position are required to align only to this position with up to two mismatches to ensure uniquely mapable regions. The female samples with higher chromosome X coverage are Denisova, Altai, Mixe, and one of the Australian individuals (BUR,E). All other samples are male.

Library	Average coverage	Contamination estimate	95% confidence intervals	Neandertal sequences	Human sequences
L9105	467	0.5%	0.4%-0.6%	17,400	87
L9198	1598	0.8%	0.7%-0.8%	63,611	493
L9199	627	0.7%	0.6%-0.8%	24,849	175
L9302	1419	0.7%	0.6%-0.7%	56,982	386
L9303	1419	0.9%	0.9%-1%	86,588	829
Altai Neandertal Combined	**5530**	**0.78%**	**0.75%-0.82%**	**249,430**	**1970**
L4533	129.6	0.6%	0.5%-0.8%	8556	51
L4740	205.9	0.5%	0.4%-0.7%	13,351	69
L4741	214.8	0.6%	0.5%-0.7%	14,022	85
Mezmaiskaya Combined	**550.3**	**0.57%**	**0.49%-0.65%**	**35,929**	**205**

Table S5a.2: MtDNA contamination estimates of the Altai Neandertal and the Mezmaiskaya Neandertal by library. The percent contamination is shown with 95% confidence intervals.

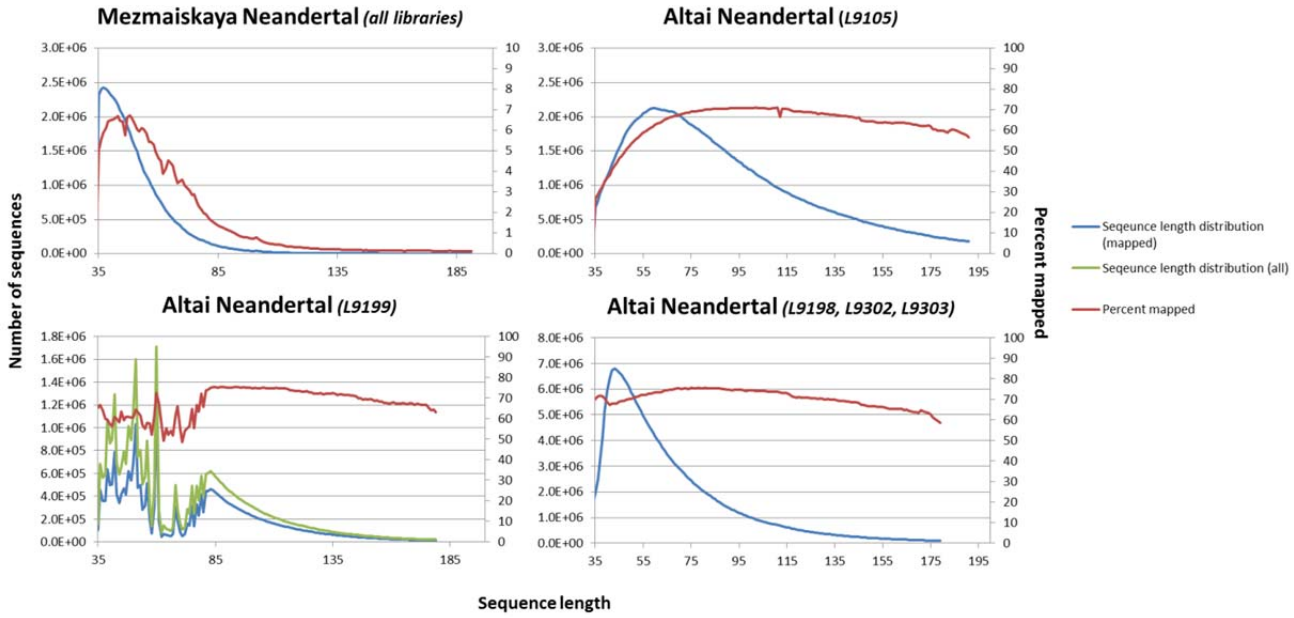


Figure S5a.1: Fragment length distribution and proportion of mapped sequences. A comparison of fragment length distributions and the percent of sequences that mapped over sequence length between the two Neandertals as well as between library preparation methods. Libraries from the same individual and library preparation method were combined. Analyses were performed without a map quality filter and restricted to merged sequences only (corresponding to full-length sequences of the fragments). All Mezmaiskaya Neandertal libraries were prepared using the double-stranded method, as was library L9105 of the Altai Neandertal. L9198, L9199, L9302 and L9303 of the Altai Neandertal were made with a single-stranded library method. The L9199 library is shown separately, because the fragment size distribution and proportion of mapped reads is non-uniform. We also show the fragment length distribution of all sequences (including unmapped) for this library.

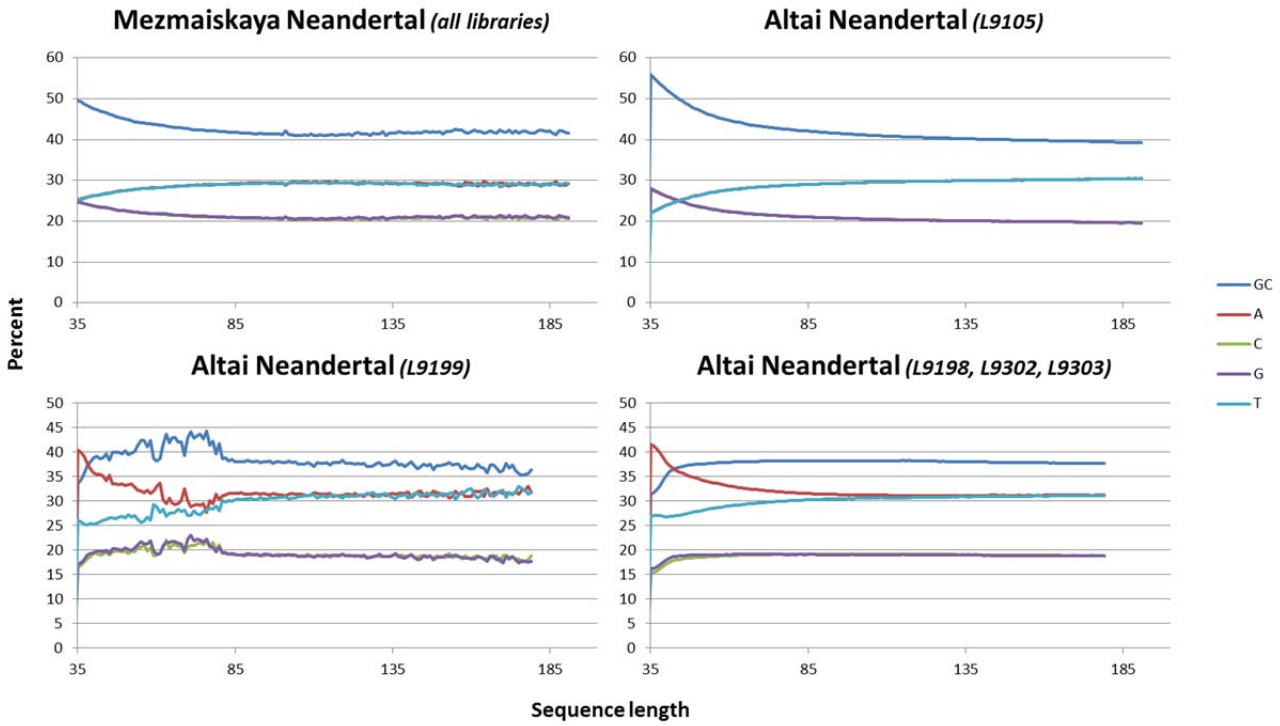


Figure S5a.2: Base composition of aligned sequences as a function of fragment length. The top two graphs show double-stranded libraries (note that the orientation of the template strand cannot be inferred in these libraries) while the bottom ones come from single-stranded libraries. L9199 is shown separately from L9198, L9302 and L9303 for the reasons discussed in the text.

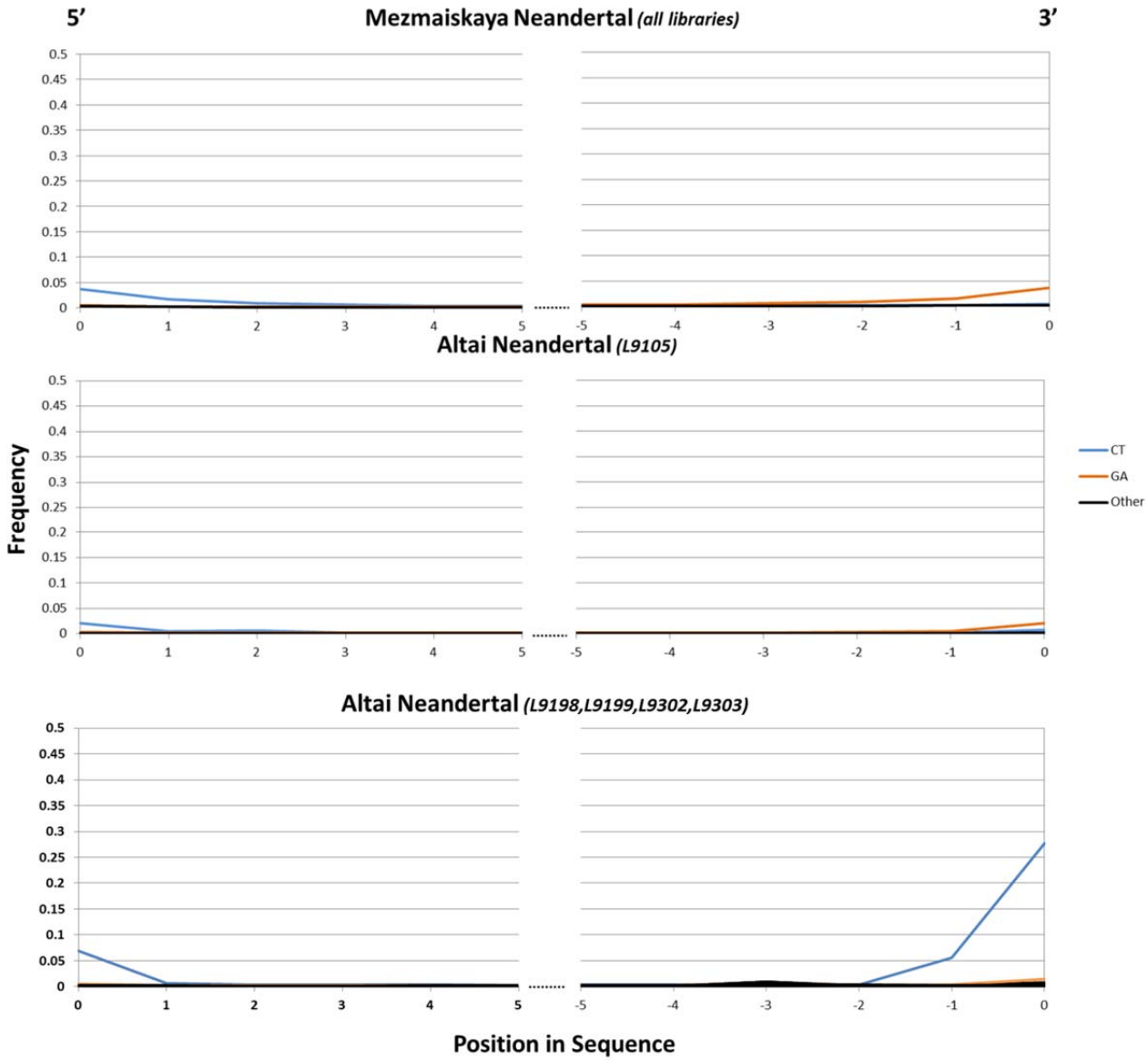


Figure S5a.3: Damage-derived substitution patterns. The frequency of C-to-T and G-to-A differences to the human reference sequence is shown as a function of the position in the alignment. The three Mezmaiskaya libraries were combined as well as the single stranded Altai Neandertal libraries as they had the same damage pattern. The top two graphs are double-stranded libraries while the bottom graph shows the damage patterns from the single-stranded libraries. All base changes other than C to T and G to A are shown as 'other'.

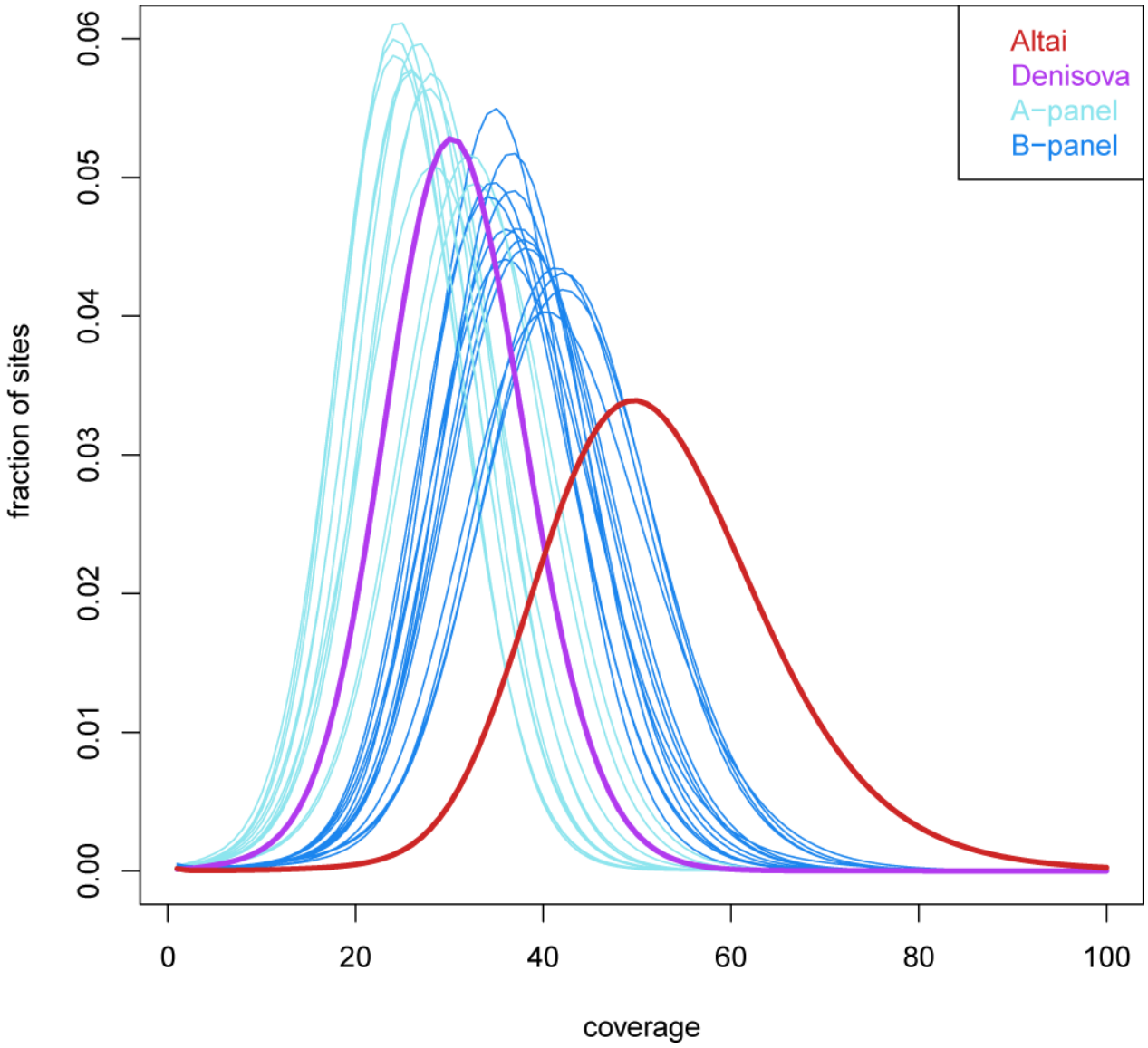


Figure S5a.4: Coverage distributions for Altai Neandertal, Denisova and present-day humans from panels A and B (autosomes only).

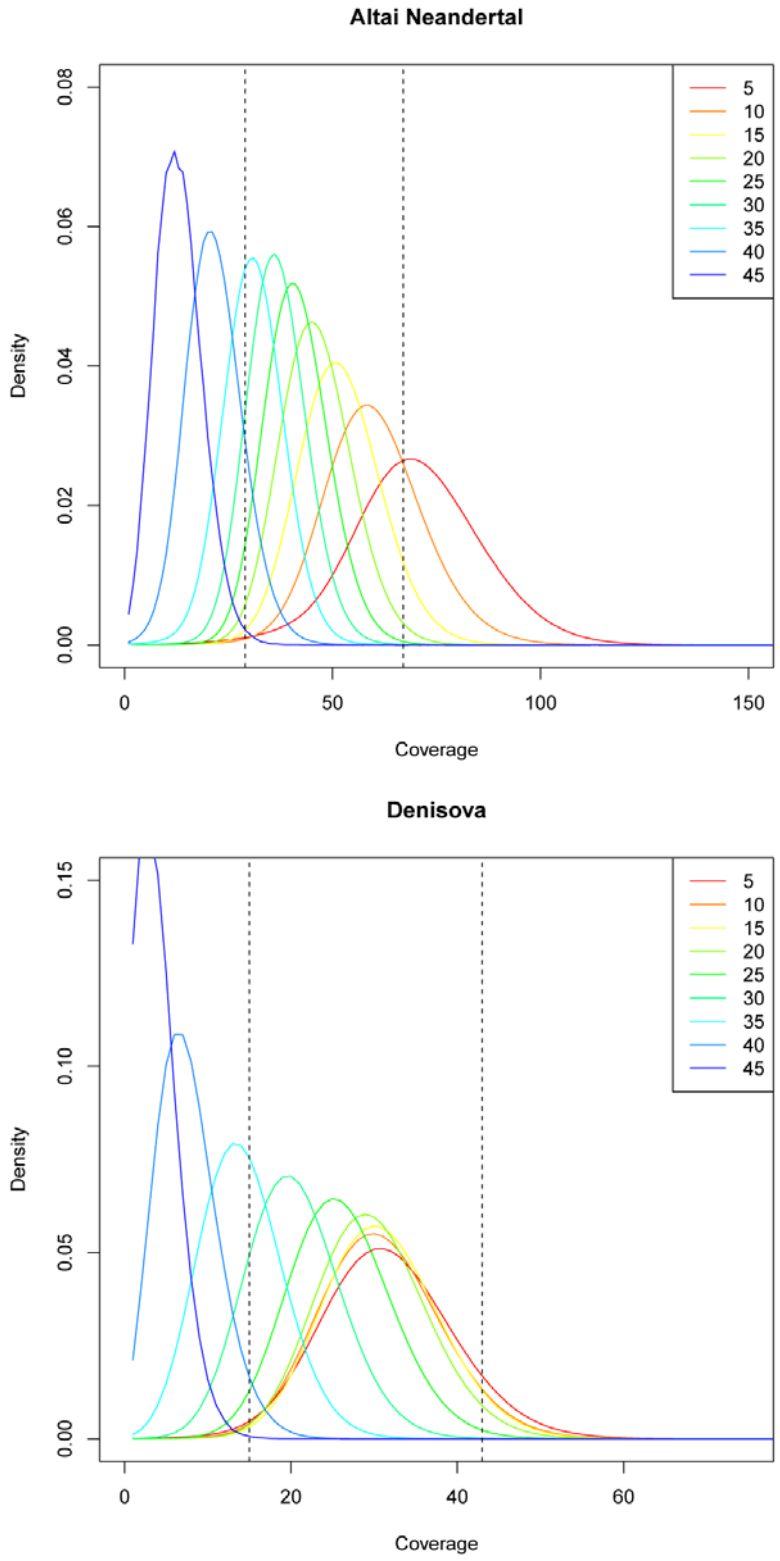


Figure S5a.5: Coverage in different GC bins (given as number of bases in a 51 basepair window centered on the analyzed nucleotide) for Altai Neandertal (top) and Denisova (bottom). Each bin sums regions of 5 different GC values. The first value is 5 and contains regions with at least 5 GC bases and at most 9 GC bases. Dashed lines delimit the range that includes 95% of reads overall.

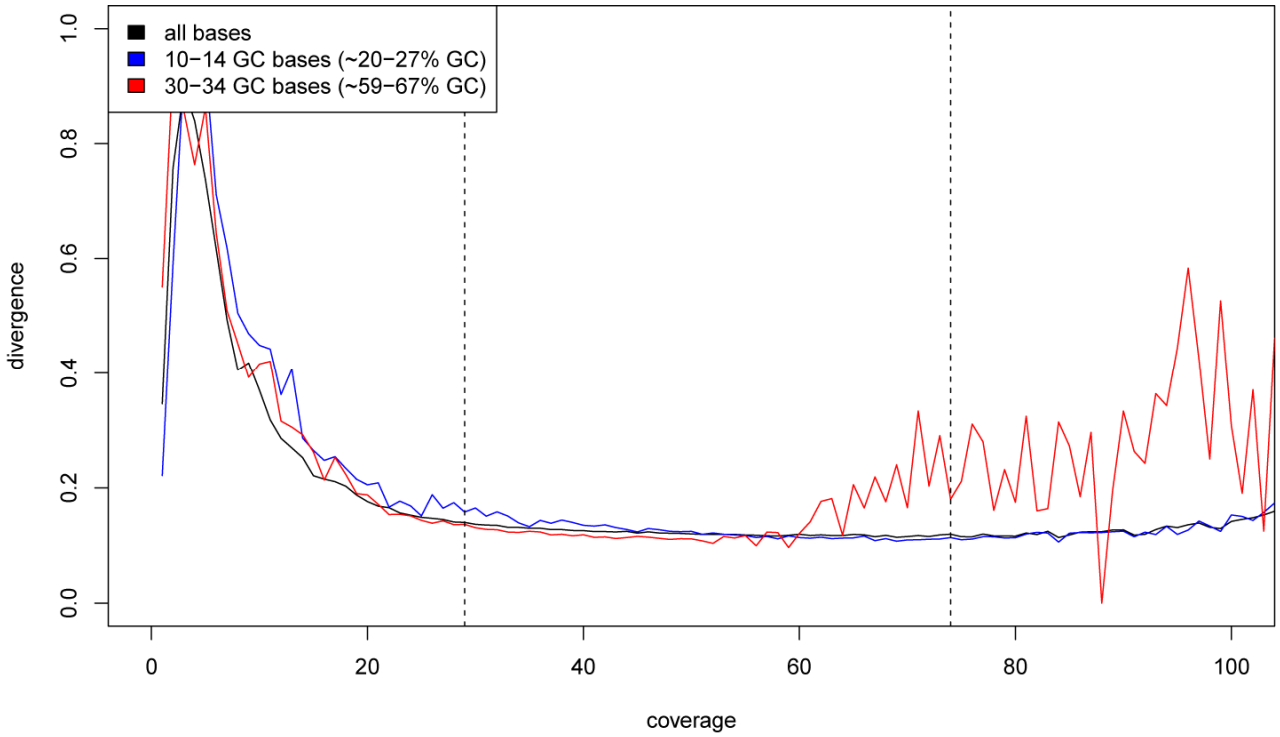


Figure S5a.6: Divergence between the Altai Neandertal and the human reference genome sequence measured as a fraction of human-chimpanzee divergence. Results for all data are given as a black solid line, for a very low GC bin as a blue line, and for a very high GC bin as a red line. The vertical black dashed lines indicate the 2.5th and 97.5th percentile coverage cutoff for the entire distribution; it is evident that these cutoffs are not appropriate for the very low (blue) or very high GC (red) sites, motivating our decision to apply GC-corrected coverage cutoffs for the ancient samples (see SI 5b).

References

1 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722, doi:10.1126/science.1188021 (2010).

2 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).

3 Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 14616-14621, doi:10.1073/pnas.0704665104 (2007).

4 Green, R. E. *et al.* A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**, 416-426, doi:10.1016/j.cell.2008.06.021 (2008).

Supplementary Information 5b

A Minimal Set of Filters for All Analyses

Cesare de Filippo*, Martin Kircher, Janet Kelso, Heng Li, David Reich, Kay Prüfer*

* To whom correspondence should be addressed (cesare_filippo@eva.mpg.de, pruefer@eva.mpg.de)

We describe a minimal set of filters for the Neandertal, Denisova and modern human genome data that are applied throughout all analyses presented in this paper. These filters primarily aim at reducing the influence of mapping errors. Briefly, using published genome annotation and alignability tracks we restrict our analysis to unique regions in the genome that are not in tandem repeats. We further require that reads align confidently based on map-quality scores, and exclude genomic positions that are outliers in the genome-wide distribution of coverage depth. We tested the effect of these filters on estimates of divergence and heterozygosity as well as their influence on D-statistics as described in the respective supplementary notes SI 6a and 9.

Filters

In the following, we describe four filters that we apply on Neandertal, Denisova and modern human data. These filters are applied after sequence based filtering described in SI2, such as removal of reads with too many low quality bases and duplicate removal. In addition, sites that were called as indels by GATK (SI3) were removed from analyses.

Tandem Repeat Filter (TRF)

We downloaded the Tandem Repeat Finder annotation for hg19[1] and pantro2[2] from the UCSC genome browser. Regions identified as repetitive in these tracks were excluded from further analyses.

Mapping Quality Filter (MQ30)

We use the root-mean-square mapping quality entry (MQ field in VCFs) for each position in the pantro2 and hg19 VCFs produced by GATK (see SI 3). For a position to be included in our analysis, the MQ must be at least 30 (Phred scale).

[1] http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/simpleRepeat.txt.gz

[2] http://hgdownload.soe.ucsc.edu/goldenPath/panTro2/database/simpleRepeat.txt.gz

Genome Alignability Filter

We produced two tracks for the genome that aim at excluding regions that may lead to ambiguous alignments of short sequences.

1. *map35_50%:* This filter requires that at least 50% of all possible 35mers overlapping a position do not find a match to any other position in the genome allowing for up to one mismatch[1].
2. *map35_100%:* This filter requires that all possible 35mers overlapping a position do not find a match to any other position in the genome allowing for up to one mismatch.

The window size of 35 basepairs was chosen since this size corresponds to the minimum read length used in the Neandertal and Denisova dataset. Unique regions according to this definition are expected to be comparable between ancient and present-day humans despite differences in sequence length. Whether the more stringent *map35_100%* or less stringent *map35_50%* filter is used is specified in the respective Supplementary Information notes.

Coverage Filter

One possible reason for high or low sequence coverage in some regions of the genome are segmental duplications. An excess-of-coverage approach is for instance used in SI 8 to detect the presence of segmental duplications in archaic and modern human genomes and to estimate their copy number. If the duplication state does not match that of the human or chimpanzee references used for alignment, they may lead to an over-collapsing of sequences and erroneous calls of heterozygotes or differences to the reference (see also SI 5). We therefore exclude sites that fall within the 2.5% and 97.5% quartiles of the coverage distribution from all modern human genomes (according to the DP field in the 'FORMAT' column of the hg19 or pantro2 VCFs). The coverage distributions are calculated separately within *map35_50%* and *map35_100%* regions and for the autosomes and chromosome X. Table S5b.1 shows the cutoffs used for the modern human genomes aligned to hg19.

For the ancient genomes, we have observed a strong dependence of coverage on GC-content (SI 5). To account for this, we partition the reference sequence in bins of GC content according to the number of G and C bases occurring within 25 basepairs to each side of a position. A total of 11 bins were defined based on GC counts of 0-4, 5-9, 10-14, …, 45-49, 50-51 bases. The coverage distribution in each bin was determined using reads with a mapping quality ≥30. The coverage cutoffs were then applied for each bin separately within the *map35_50%* and *map35_100%* regions. We do not apply a GC-corrected coverage cutoff to present-day humans.

Samples (Panel)	IDs	Map35_100%				Map35_50%			
		Autosomes		Chromosome X		Autosomes		Chromosome X	
		2.5%	97.5%	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
Dinka (A)	DNK02	14	42	6	24	14	42	6	24
Dinka (B)	DNK07	21	51	9	29	20	51	9	29
Mandenka (A)	HGDP01284	12	39	5	22	11	39	5	22
Mandenka (B)	HGDP01286	20	55	9	31	20	55	8	31
Mbuti (A)	HGDP00456	11	39	5	22	11	39	5	22
Mbuti (B)	HGDP00982	20	58	8	30	19	58	8	30
San (A)	HGDP01029	17	50	7	29	16	50	7	29
San (B)	HGDP01036	21	57	9	32	21	57	9	32
Yoruba (A)	HGDP00927	16	48	7	27	16	48	7	28
Yoruba (B)	HGDP00936	22	58	9	32	21	58	9	32
Karitiana (A)	HGDP00998	13	41	5	23	12	41	5	23
Karitiana (B)	HGDP01015	19	53	8	29	19	53	8	29
Dai (A)	HGDP01307	13	45	5	25	13	45	5	25
Dai (B)	HGDP01308	21	54	9	30	20	54	9	31
Han (A)	HGDP00778	14	43	6	24	13	43	6	25
Han (B)	HGDP00775	20	53	8	29	19	53	8	29
French (A)	HGDP00521	13	41	6	23	13	41	6	23
French (B)	HGDP00533	23	62	11	35	22	62	10	35
Sardinian (A)	HGDP00665	12	38	5	22	12	38	5	22
Sardinian (B)	HGDP01076	21	57	9	31	21	57	9	31
Papuan (A)	HGDP00542	12	40	5	23	12	40	5	23
Papuan (B)	HGDP00546	23	63	10	35	22	63	10	35
Mixe (B)	MIXE 0007	21	53	22	52	21	53	21	52
Australian (B)	WON,M	23	64	10	34	22	64	10	34
Australian (B)	BUR,E	24	62	24	60	23	62	23	59

Table SX.1: 2.5% and 97.5% quartiles of the coverage distributions using alignability filters map35_50% and map35_100% for the autosomes and chromosome X. The values were used as cutoffs for the present-day human Panel A and B data.

Bases Passing Filters

The autosomes of the human genome assembly (GRCh37/1000 Genomes release) encompass 2.68 gigabases (Gb) for which a sequence has been determined (this excludes N's). Of these, between 2.00 and 2.04 Gb remain in all individuals after applying all filters (TRF, MQ30, Coverage) in the *map35_50%* alignability regions, and between 1.64 and 1.66 Gb remain when applying all filters in *map35_100%* regions. When restricting to the positions covered in all individuals, we retain 0.90 and 0.70 Gb for *map35_50%* and *map35_100%*, respectively.

Availability

The here described filters are available in the form of *.bed* files from http://bioinf.eva.mpg.de/altai_minimal_filters/.

References

1 Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496, doi:10.1038/nature10231 (2011).

Supplementary Information 6a

Divergence Estimation

Gabriel Renaud*, Fernando Racimo, Kay Prüfer*, Janet Kelso

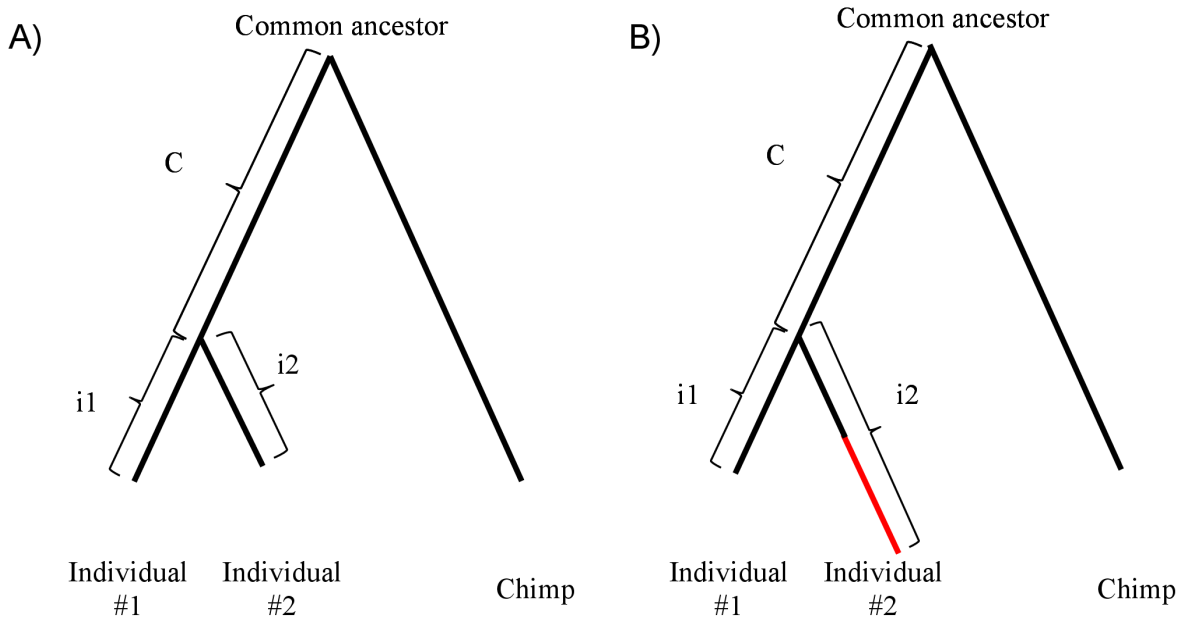
* To whom correspondence should be addressed (gabriel_renaud@eva.mpg.de, pruefer@eva.mpg.de)

This section presents divergence estimates between present-day and archaic hominins. The divergence estimates are consistent with the known relationship among the populations. However, comparing Denisova and Altai Neandertal divergence to present-day humans, we find that Denisova gives a consistently deeper divergence. This signal is observed in all present-day human populations, including African, and does not fit a simple scenario in which Denisova and Altai Neandertal are sister groups and present-day human genomes constitute an equidistant outgroup. An analysis of divergence in short non-overlapping windows along the genome shows that this signal manifests as a shift towards higher divergence over the entire distribution.

Divergence Triangulation

We calculated divergence using the triangulation method previously applied to archaic genomes[1-3]. Let individual #1 and individual #2 be the two individuals for which we seek to estimate divergence. For a given site that differs between these two individuals and a common ancestor sequence, we can parsimoniously assign the change to a lineage when two sequences agree and the third is different. A change can be assigned either to the lineage leading to the ancestor, or the lineages leading to individual #1 and individual #2.

Figure S6a.1: *Method for calculating divergence between 2 individuals. If both have the same error rates, divergence computed on either branch should be the same. A) High-quality ancient genomes can show a shortened branch. B) Individuals with a high error rate (e.g. due to low coverage) will have a lengthened branch (red). These issues can be avoided by calculating divergence only using the changes assigned to an extant, high-quality genome (individual #1 in the figure).*



One of the two individuals may be ancient, so its branch to the common ancestor will be shortened relative to an extant lineage (Fig. S6a.1). Alternatively, one of the individuals may show more genotype errors, which would artificially inflate divergence. To avoid these issues, divergence is calculated by only considering changes assigned to individual #1 ($i1$) and the ancestor (c) as:

$$div(ind_1, ind_2) = \frac{i1}{i1+c} \quad \text{S6a.1}$$

As a general rule, we use extant, high-coverage samples as individual #1 in our comparisons when available. For comparisons between archaic individuals, we always choose the higher quality genome (Altai Neandertal or Denisova) for individual #1.

An estimate of variability is given for each divergence estimate by calculating the 95% confidence interval for a binomial distribution with the probability parameter set to the estimated divergence.

Datasets

We assigned changes using the human-chimpanzee ancestor inferred from the 6-primate EPO alignments[4,5]. The raw alignments were parsed and formatted as a tab-delimited file using a custom computer program[a]. Only the alignment blocks where a single human and a single chimpanzee base were present in the EPO block were retained for this analysis, thus excluding regions that were ambiguously aligned or duplicated in either lineage.

Genotypes were extracted from VCF files (SI 3). We used the sites retained after the filtering described in Supplementary 5b. To infer the genotype, we used the PHRED-likelihood (PL) values generated by GATK. The PL field provides the likelihood of all possible genotypes given the observed data. It contains 3 subfields, the likelihood of being homozygous for the reference allele, homozygous for an alternative allele, or heterozygous. If one genotype is at least ~2000-times more likely than the remaining two (phred-scale difference of at least 33) then this genotype was used to infer an allele: The homozygous genotype was selected as allele for at homozygous sites, and one random allele was selected at heterozygous sites. However, sufficient statistical power to distinguish between homozygous sites and heterozygous ones is only available with sufficient coverage. To call allele in low-coverage samples or for low-coverage sites, we also consider sites where the difference in phred-scaled likelihood between the two homozygous calls exceeds 33, thereby ignoring the likelihood of the heterozygous genotype. The more likely homozygous genotype was then chosen for the allele.

Divergence Estimates

When comparing an individual from panel A with another individual from the same population from panel B, we observed that A panel members gave consistently deeper average divergence to the archaic hominins than B panel members (divergence for Altai Neandertal: A panel: 11.62%, B panel 11.33%; divergence for Denisova A panel: 11.83% B panel: 11.57%). This observation can be explained by a higher genotyping error in A panel individuals, possibly due to the lower coverage compared to B panel individuals (see SI5). For the following analyses in this section, we will treat A and B panels separately to reflect the difference in error between these two datasets.

[a] "grenaud/epoParser · GitHub." 2012. 13 Nov. 2012 https://github.com/grenaud/epoParser

We estimated divergence on all autosomes using panel A, panel B, and both high-coverage archaic hominins as individual #1 and various ancient humans as individual #2. Specifically, the ancient humans were the Feldhofer 1 Neandertal (Feldhofer 2 was excluded due to limited data), a Neandertal from Mezmaiskaya (Mez1), a Neandertal from El Sidron (Sid1253), three Neandertals from Vindija (Vindija 33.16, 33.25 and 33.26), and the high-coverage Altai Neandertal and Denisova.

Table S6a.1: Percent divergences seen between Neandertals & each group of present-day humans

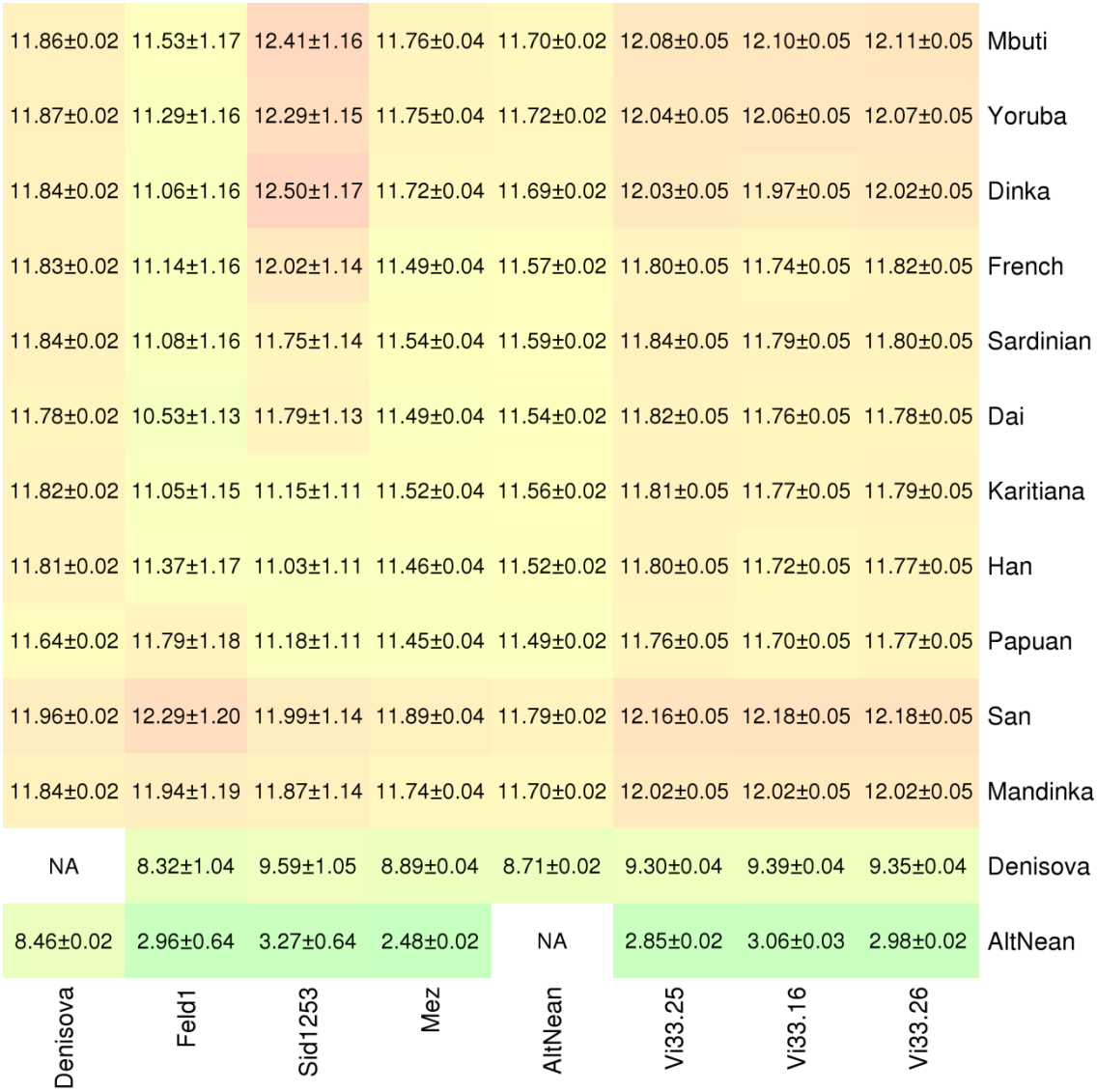
	A panel				B panel			
	Non-Africans		Africans		Non-Africans		Africans	
	Lowest	Highest	Lowest	Highest	Lowest	Highest	Lowest	Highest
Altai	11.49	11.59	11.69	11.79	11.18	11.38	11.40	11.43
Mezmaiskaya	11.46	11.53	11.69	11.86	11.08	11.28	11.36	11.40
Vindija 33.16	11.70	11.79	11.97	12.18	11.41	11.58	11.69	11.78
Vindija 33.25	11.76	11.84	12.02	12.16	11.47	11.65	11.72	11.78
Vindija 33.26	11.77	11.82	12.02	12.18	11.46	11.64	11.76	11.79

The results of these pairwise divergence estimates are shown as a heatmap in Figure S6a.2. We find that Non-Africans have consistently lower divergence to each Neandertal described in this study compared with Africans (see Table S6a.1). However, Sidron (Sid1253) and Feldhofer (Feld1) show no consistent difference, likely due to the very limited amount of data available for these samples as indicated by the wide confidence intervals. The lower divergence to non-Africans compared to African individuals is likely due to the admixture of Neandertals into the ancestors of non-Africans.

For Denisova, the present-day humans with the lowest divergence are the Papuans and Australians, consistent with the previously reported signal of gene flow into the ancestors of these present-day humans[3]. Interestingly, the divergence of African individuals to Denisova (A panel: 11.87% ± 0.03%, B panel: 11.62% ± 0.03%) was consistently higher than to the Altai Neandertal (A panel: 11.72% ± 0.03%, B panel: 11.41% ± 0.03). Similarly, the divergence for each of the non-African individuals to Denisova compared to Altai Neandertal is between 0.15-0.31 greater. Those differences could for example be due to contamination or to an unknown archaic component in the Denisovan ancestry.

Figure S6a.2: *Heat maps of divergence for each pair of individuals for both the A and B panels. Each cell represents the divergence between the individuals on the vertical and horizontal axes. The vertical axis is individual #1 whereas the horizontal axis is individual #2.*

A panel individuals



B panel individuals:

Denisova	Feld1	Sid1253	Mez	AltNean	Vi33.16	Vi33.25	Vi33.26	
11.62±0.02	10.59±1.14	11.24±1.12	11.37±0.04	11.42±0.02	11.72±0.05	11.72±0.05	11.76±0.05	Yoruba
11.68±0.02	10.15±1.12	11.31±1.12	11.32±0.04	11.38±0.02	11.58±0.05	11.65±0.05	11.64±0.05	Han
11.65±0.02	10.52±1.14	11.10±1.12	11.26±0.04	11.37±0.02	11.57±0.05	11.62±0.05	11.62±0.05	Dai
11.60±0.02	10.63±1.14	10.94±1.10	11.18±0.04	11.31±0.02	11.48±0.05	11.54±0.05	11.54±0.05	Mixe
11.42±0.02	10.58±1.13	11.75±1.14	11.12±0.04	11.24±0.02	11.44±0.05	11.53±0.04	11.51±0.05	Papuan
11.42±0.02	10.39±1.13	11.46±1.13	11.14±0.04	11.23±0.02	11.42±0.05	11.53±0.04	11.48±0.05	Australian2
11.61±0.02	9.80±1.10	10.97±1.11	11.18±0.04	11.30±0.02	11.52±0.05	11.54±0.04	11.56±0.05	Sardinian
11.59±0.02	10.06±1.12	10.88±1.11	11.15±0.04	11.28±0.02	11.45±0.05	11.51±0.04	11.51±0.05	Karitiana
11.38±0.02	10.03±1.12	11.05±1.11	11.08±0.04	11.18±0.02	11.40±0.05	11.46±0.04	11.46±0.05	Australian1
11.63±0.02	10.81±1.15	11.74±1.14	11.41±0.04	11.43±0.02	11.74±0.05	11.78±0.05	11.79±0.05	Mbuti
11.64±0.02	10.86±1.16	11.55±1.13	11.38±0.04	11.43±0.02	11.71±0.05	11.75±0.05	11.77±0.05	Dinka
11.63±0.02	10.77±1.15	12.14±1.15	11.36±0.04	11.41±0.02	11.69±0.05	11.75±0.05	11.76±0.05	Mandinka
11.61±0.02	11.29±1.16	11.75±1.14	11.20±0.04	11.31±0.02	11.49±0.05	11.57±0.05	11.56±0.05	French
11.60±0.02	11.94±1.19	11.99±1.15	11.41±0.04	11.40±0.02	11.78±0.05	11.78±0.05	11.79±0.05	San
NA	8.32±1.04	9.59±1.05	8.89±0.04	8.71±0.02	9.39±0.04	9.30±0.04	9.35±0.04	Denisova
8.46±0.02	2.96±0.64	3.27±0.64	2.48±0.02	NA	3.06±0.03	2.85±0.02	2.98±0.02	AltNean

Is a difference in contamination responsible for the deeper divergence in Denisova?

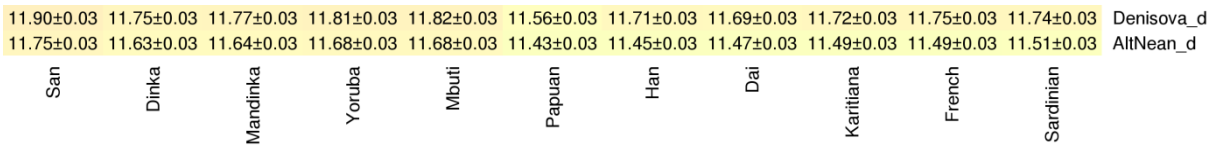
To test the effect of contamination on the signal, we recalculate African-Altai and African-Denisova divergence using reads that show deamination patterns. Deamination has been found to increase with time, so reads containing deamination are more likely endogenous[6].

We selected reads with a cytosine to thymine change at the positions where residual deamination is found after uracil removal (last base 5' end + last two bases 3' end; see SI1 and SI5). Bases at these positions were required to have a minimum quality score of 30 (corresponding less than 1 error in a 1000 basepairs). The cytosine residue at the corresponding genomic position was required to be present in the consensus sequence (genotype in the VCF). In order to avoid contaminating human molecules to be falsely classified as deaminated read, we further required that no human from the 1000 Genomes Project phase1 data shows the thymine at the position under consideration. With this methodology, new BAM files containing putatively deaminated reads for the Altai Neandertal and the Denisova were obtained and genotyping was performed according to the method of SI 3. The deaminated reads from Altai Neandertal provided an average coverage of 2.4 while those from Denisova gave an average coverage of 2.0.

*S6a*Using only the deaminated reads, we recalculate divergence (see Figure S6A.3). We find that Denisova continues to show deeper divergence to African individuals compared to the Altai Neandertal. We conclude that the difference is not caused by a difference in contamination.

***Figure S6a.3**: Divergence for deaminated reads from Denisova (Denisova_d) and Altai (Altai_d).*

A panel individuals:



B panel individuals:

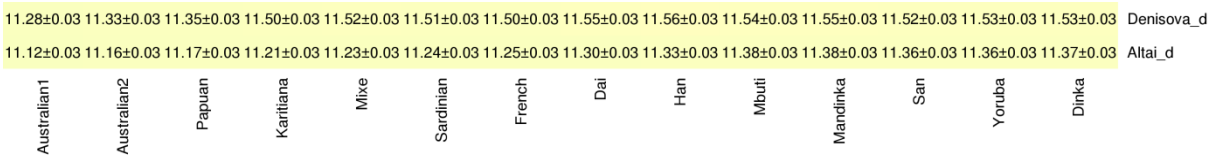
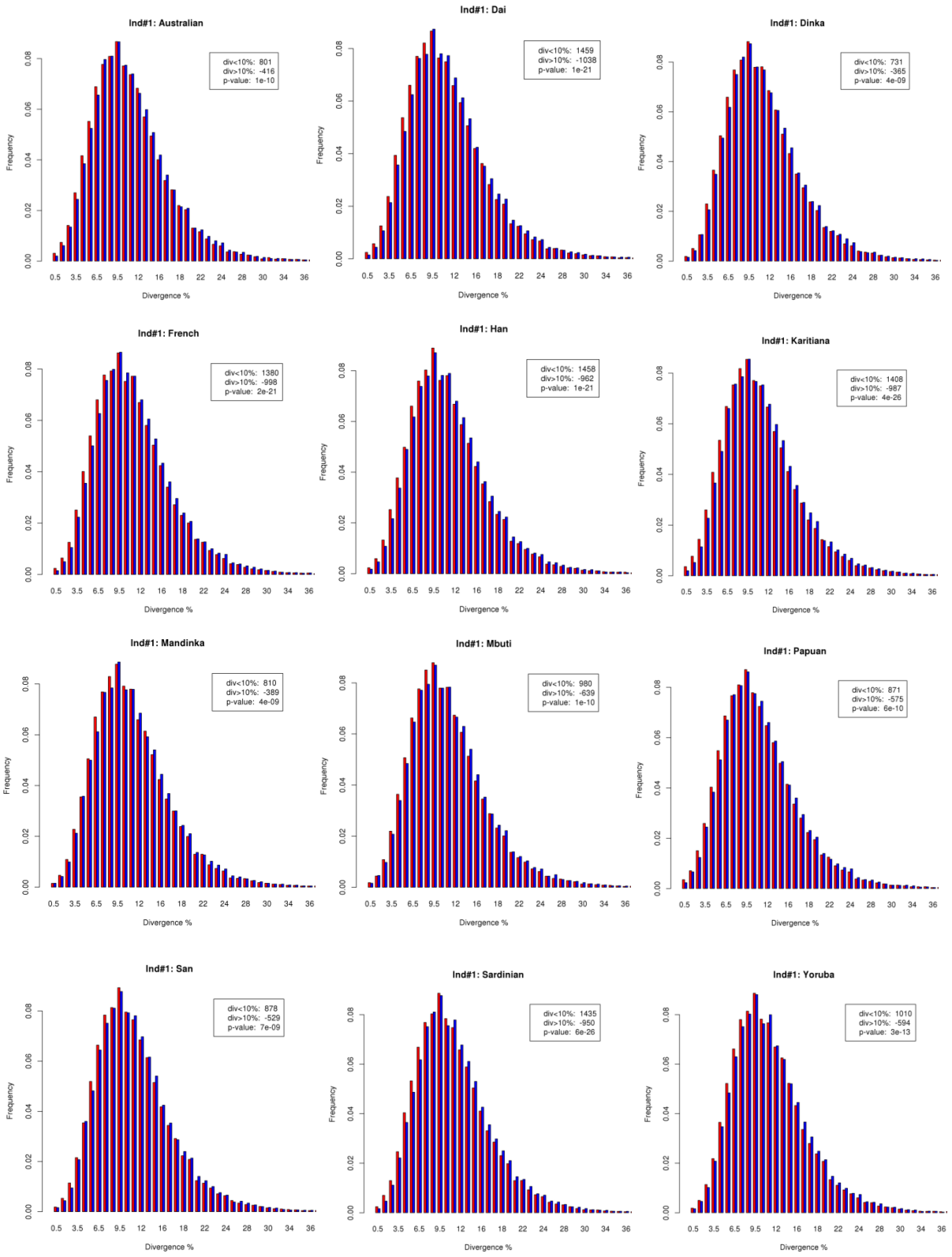


Figure S6a.4: *Distribution of divergence in windows of 40kb for Denisova (blue) and Neandertal (red) to 12 modern humans from the B panel. The legend contains the number of windows in Altai Neandertal minus the ones in Denisova that gave a low divergence estimate (<10% of the human-chimpanzee divergence) and a higher divergence (between 10% and 20% of the human-chimpanzee divergence). The p-value was obtained using a two sided Wilcoxon rank test on the divergence values for Denisova versus the Altai Neandertal for all genomic windows.*



Divergence by windows

To test whether the deeper Denisova-African divergence is a caused by a small fraction of genomic regions or whether the signal is distributed over the genome, we calculated divergence in 40kb windows. In each case, a genomic window was retained if at least 25% of its sites had a resolved genotype. The resulting distributions for both archaic samples were plotted side by side (Figure S6a.4). We computed the difference in windows of low divergence (between 0% and 10%) and high divergence (between 10% and 20%) between Denisova and Altai Neandertal. The divergence distribution for Altai Neandertal and Denisova was significantly different in all comparisons to modern humans (Wilcoxon rank test, Fig. S6a.4). However, we observe no outlier regions in Denisova. We test several scenarios for explaining this shift in SI 16a,b.

Testing the effect of filtering

Throughout this section, we used the set of minimal filters described in SI 5b. Here, we tested how these filters affect our analysis by comparing four individuals ($Dinka_A$, $Sardinian_B$, Altai Neandertal and Denisova) to the human reference and the human-chimpanzee common ancestor under different filtering settings. In each comparison, we assigned differences to either the human reference or the individuals' lineage. Keeping with our previous nomenclature, we then calculated the ratio:

$$r(ind_1, ind_2) = \frac{i2}{i1} \qquad \text{S6a.2}$$

Under the assumption of equal mutation rates on all lineages, this ratio is expected to be close to 1 since the same time passed on each lineage since the split of human reference and the tested individual. The archaic genomes may give slightly smaller ratios since the archaic genomes did not experience new mutations from the moment of death (see SI 6b for a detailed analysis of this branch shortening). However, with more error in the tested individuals' sequence, this ratio is expected to increase since error will most likely be assigned to this individuals' lineage. Thus, with increasing stringency on filtering, we expect a ratio close to 1 while failing to remove error will result in larger values. The following combinations of filters were tested:

- No filters
- map35_100%
- map35_100% + MQ30
- map35_100% + coverage + tandem repeat
- map35_50%
- map35_50% + MQ30
- map35_50% + coverage + tandem repeat

Our results (Figure S6a.5) show that all filters decrease the ratio compared to no filtering and that the ratio moves closer to the expected value of 1 when filters are applied. The more stringent mappability filter map35_100% gave generally smaller values than map35_50%, even when other filters were applied in combination, indicating that more error remains with the latter mappability filter. When applying map35_100% with all additional filters, we observe the smallest ratios, including branch shorting by 2.8% for Altai Neandertal and a slight excess of lineage length of 0.39%, 8.33% and 5.75% for the Denisova, Dinka and Sardinian individuals, respectively.

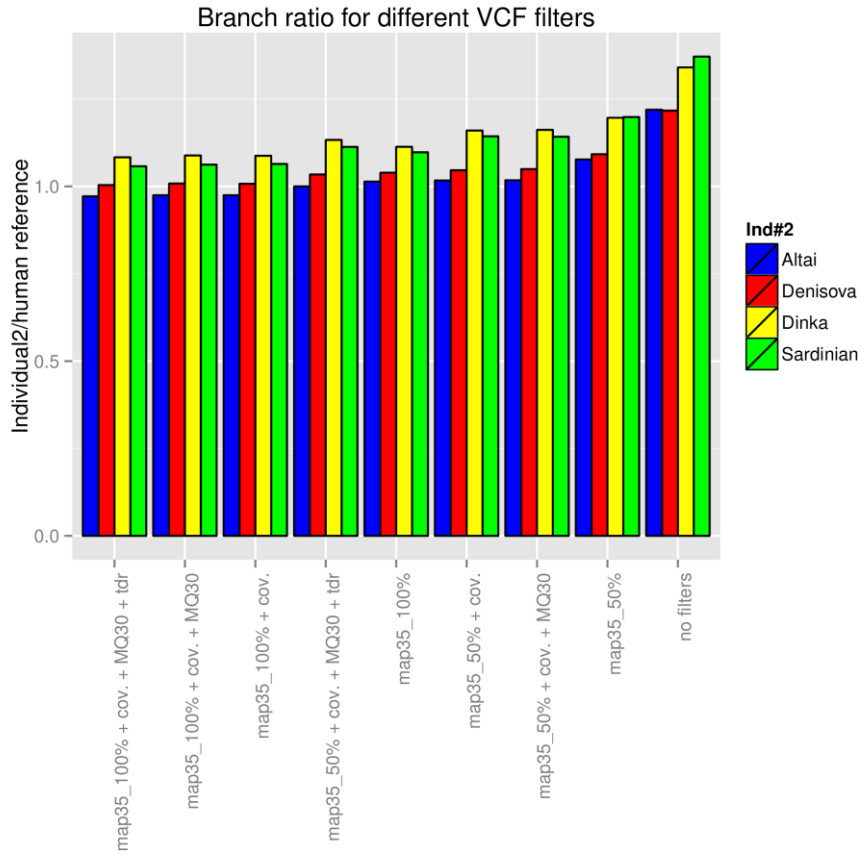


***Figure S6a.5**: Bar plot of the ratio of the mutations in individual #2 (Altai Neandertal, Denisova, Dinka from the A panel and Sardinian from the B panel) and the human reference for different filter settings (mappability, mapping quality greater than 30 (MQ30), coverage (cov.) and tandem repeats (tdr)). The filters have been sorted according to the branch ratio for the Altai Neandertal.*

A Neighbor joining tree of archaic and modern individuals

We use identical filters to our divergence triangulation calculation to extract alleles from all individuals. The extracted information is used to calculate a pairwise rate of transversions substitutions between all individuals and the chimpanzee-human common ancestor.

We observe that the low-coverage individuals in our comparison (Mezmaiskaya and Vindija), give a substantially higher rate of transversions compared to other individuals, likely due to higher error in these sequences (Table S6a.2). Similar to the method presented in Reich et al (2010) (Supplementary Information, Section 6, p. 34), we use this excess of transversions to chimpanzee to estimate an error rate. Since rates are highly similar between humans, we use the modern human with the lowest rate as a reference point for the expected rate of transversions if a sequence is mostly error free and treated the excess in the Mezmaiskaya and Vindija samples as error. These individual error rates are then subtracted from all comparisons involving low-coverage samples.

To estimate the reliability of the tree, we generate 1000 bootstrap replicas over the rate of transversions in windows of 5Mb size. For each individual replica, we recalculate the error rates for the low-coverage Neandertals using the lowest modern human transversion rate and correct the values before calculating the neighbor joining tree.

Figure S6a.6 shows the neighbor tree calculated after error correction and with bootstrap support values. The exact relationship of the three Vindija individuals is unresolved. However, the Altai Neandertal falls confidently with other Neandertals.

Table S6a.2: *Rate of transversions between the chimpanzee common ancestor and archaic and present-day human individuals.*

Ancient Individual	**% transversion**
Mezmaiskaya	0.420%
Vindija25	0.417%
Vindija16	0.395%
Vindija26	0.392%
Denisova	0.373%
Altai	0.371%

Panel A Individual	**% transversion**
San_A	0.375%
$Mandinka_A$	0.375%
$Mbuti_A$	0.374%
Han_A	0.374%
$French_A$	0.374%
$Yoruba_A$	0.374%
$Karitiana_A$	0.373%
$Sardinian_A$	0.373%
$Papuan_A$	0.373%
Dai_A	0.373%
$Dinka_A$	0.373%

Panel B Individual	**% transversion**
San_B	0.373%
$Mandinka_B$	0.373%
$Mbuti_B$	0.374%
Han_B	0.373%
$French_B$	0.373%
$Yoruba_B$	0.373%
$Karitiana_B$	0.373%
$Sardinian_B$	0.373%
$Papuan_B$	0.373%
Dai_B	0.373%
$Dinka_B$	0.373%
$Mixe_B$	0.373%
$Australian_{B1}$	0.372%
$Australian_{B2}$	0.373%

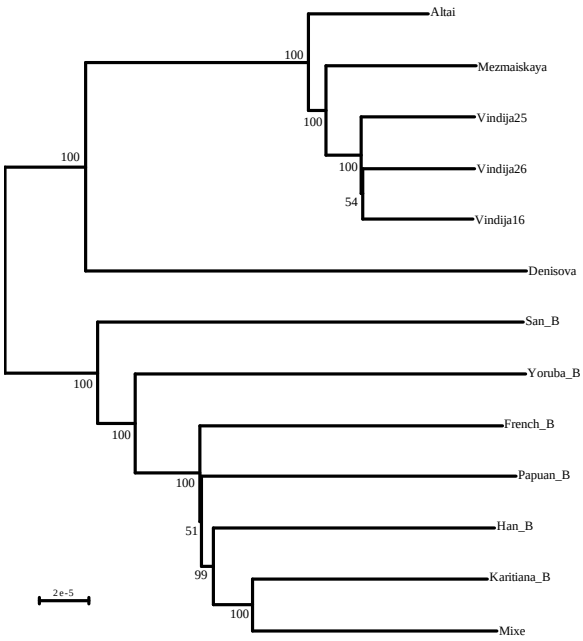


Figure S6a.6: *Neighbor joining tree of archaic and present-day human individuals (B-panel). The tree was calculated using the method by Saitou and Nei (1987)[7] as implemented in the R-package phangorn[8].*

References

1 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722, doi:10.1126/science.1188021 (2010).

2 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).

3 Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060, doi:10.1038/nature09710 (2010).

4 Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome research* **18**, 1814-1828, doi:10.1101/gr.076554.108 (2008).

5 Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome research* **18**, 1829-1843, doi:10.1101/gr.076521.108 (2008).

6 Sawyer, S., Krause, J., Guschanski, K., Savolainen, V. & Paabo, S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PloS one* **7**, e34131, doi:10.1371/journal.pone.0034131 (2012).

7 Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* **4**, 406-425 (1987).

8 Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593, doi:10.1093/bioinformatics/btq706 (2011).

Supplementary Information 6b

Branch shortening

Martin Kircher*

* To whom correspondence should be addressed (mkircher@uw.edu)

Over time, lineages accumulate changes due to germline mutations. While lineages of extant samples continue to accumulate these changes, an individual that died many generations ago will have stopped accumulating changes in the past. This section examines the discrepancy between lineage-specific mutations leading to extinct archaic hominids and present-day modern humans.

We evaluated whether a shorter branch is observed for the Neandertal and Denisova archaic genomes as compared to present-day modern humans following the methodology described by Meyer et al. 2011 (ref.[1]). Briefly, we use genotype calls (described in SI 3) (sampling a random allele for heterozygote sites) and test, at each genomic position, for differences between two individuals and the inferred ancestor of human and chimpanzee (provided by the Ensembl Compara v64 six primate alignments[2,3]). The first sample is denoted "individual#1", the second "individual#2". We count the following three types of substitutions at sites for which every individual (Denisova genome, Altai Neandertal genome, A-panel and B-panel human genomes) passes the filters described in SI 5b:

individual#1-specific (i1)	:=	(individual#2 != individual#1) and (individual#1 == anc.)
individual#2-specific (i2)	:=	(individual#2 != individual#1) and (individual#2 == anc.)
common (c)	:=	(individual#2 == individual#1) and (individual#2 != anc.)

While divergence can be measured as i1 / (i1 + c) (see SI 6a), branch shortening is measured by the ratio of the individual#2-specific and individual#1-specific counts (i2/i1). To put branch shortening on the same scale as divergence, we normalize branch shortening by the length of the extant lineage to the common ancestor: $(1-i2/i1)\cdot(i1/(i1+c))$. Years are obtained by multiplying these numbers with an ancestor divergence of 6.5 million years.

We have previously shown differences in divergence and branch-shortening estimates when comparing modern human samples from whole-genome shotgun data to the human reference sequence (GRCh37)[1]. These differences are probably the result of two confounding factors: (i) Sequences generated from whole genome shotgun sequencing and mapping assembly are of lower quality than the finished human reference genome. Thus, if genotypes from one of the shotgun genomes are compared to the human reference, sequence errors will increase the inferred length of the shotgun lineage. This will, in turn, lead to an increased divergence estimate and a decreased estimate of branch shortening for the lineage with more errors. (ii) An alignment bias to GRCh37 may lead to a preferential loss of non-reference genome alleles, which reduces the length of the lineage. This causes a decreased estimate of divergence and an increased estimate branch shortening since lineage specific alleles are lost. Since these two counter-acting biases are difficult to disentangle, we try to limit divergence and branch-shortening calculations to genomes of similar quality. Conversely, comparison of samples from the same population but different

sequencing batches to GRCh37 as the reference allows us to identify differences in quality. The divergence and branch-shortening estimates for using GRCh37 as reference and different minimum genotype quality cutoffs in the shotgun sequencing sample are available in Table S6b.1.

As outlined above, the estimation of branch shortening is sensitive to differences in sequence and alignment quality. Unfortunately, we cannot easily control for these differences. For example, we have indication from other analyses (e.g. coverage, heterozygosity and divergence) that the more recent sequencing of the B-panel produced a present-day human data set that is of higher quality than the A-panel (analyzed in Meyer et al.[1]) despite identical processing of both datasets down to identical software versions. Differences in branch length [(A-panel -B-panel)/mean] are more pronounced when increasing the minimum genotype quality cutoffs (GQ field in the VCF file), indicating that they are linked to genotyping problems. Based on the fact that divergence and branch-shortening results show considerably lower variation for different genotype quality cutoffs in the B-panel, these genotyping problems seem to be more prevalent in the A-panel. Individuals from the same populations that were sequenced in A-panel and B-panel give consistently shorter lineages for the B-panel individuals, indicating a higher quality in the B-panel. We note that the difference between the two Australian samples from the same sequencing batch (B-panel) is considerably smaller than the difference between sequencing batches. This difference also decreases with increasing genotype quality, suggesting that higher genotype quality cutoffs make these two individuals more comparable instead of pronouncing differences as has been seen in comparisons between panels. Due to the higher quality, estimates that are less affected by a genotype quality cutoff and the consistent difference between the two data sets, we argue that branch shortening should be analyzed based on the B-panel results.

We have previously reported a branch shortening of 1.16% [1.13-1.27%] for the high coverage Denisovan genome[1]. With the modified set of filters used in this study (SI 5b), we now obtain 1.06% (1.04%-1.09%) for the A-panel and 0.82% (0.74%-0.93%; 48ka-60ka) for the B-panel (Table S6b.2). This suggests that we were able to reduce errors on the human branches to a larger extent by the new filter set and further confirms that there are fewer errors in the B-panel than in the A-panel. When determining branch-shortening for the high-coverage Altai Neandertal, we measure a shortening of 1.03% (0.96%-1.14%; 62ka-74ka) compared to the B-panel. When comparing to the high-coverage Denisovan genome, the Altai Neandertal branch is shortened by 0.22%. Thus, both, measuring branch shortening separately to the present-day humans as well as measuring between the two archaic genomes, give an almost identical point estimate of the Altai Neandertal bone being older than the Denisova finger bone (0.22%; ~14ka).

For high GQ cutoffs, the branch-shortening results for the different human reference samples vary considerably (Table S6b.2), as also seen for the GRCh37-based results before. When we plot the median coverage in each sample versus the branch shortening result, we see that coverage and branch shortening are negatively correlated (Figure S6b.1). We see that for high coverage human samples, those results seem to stabilize. Samples with coverage above 38x might therefore produce the most reliable results. If we limit the analysis to those samples (BUR,E, HGDP00533, HGDP00546, HGDP00936, HGDP01036, HGDP01076, WON,M), branch shortening

for the Denisova bone is 0.81% (0.77%-0.84%; 50-54ka) and 1.02% (0.99%-1.05%; 64-68ka) for the Altai Neandertal bone.

Again, we caution that the estimated branch shortening is sensitive to differences in error rate. This analysis examines less than 0.08% of all genomic sites (the average sequence divergence of each of the archaic genomes from a human genome) and measures differences in less than 0.0065% of all genomic sites, i.e. 1% branch shortening. To put the reported numbers into perspective to an analysis using all B-panel samples, the maximum branch difference observed when comparing only among the B-panel humans is 0.19% (average 0.05%). For those B-panel samples with a median coverage of 38x and higher, the maximum branch difference is 0.09% (average 0.03%). We are therefore confident that both Neandertal and Denisova show branch shortening. An accurate estimate of the extent, however, will require further advances. Higher coverage and/or an improved data analysis pipeline for the present-day humans (e.g. including realignment of insertion/deletions as done for the archaic humans) should increase the stability of this analysis.

References

1 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).

2 Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814-1828, doi:gr.076554.108 [pii] 10.1101/gr.076554.108 (2008).

3 Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res* **18**, 1829-1843, doi:gr.076521.108 [pii] 10.1101/gr.076521.108 (2008).

doi:10.1038/nature12886

Table S6b.1: Divergence and branch length ratios when using GRCh37 as reference and different minimum genotype quality cutoffs for the shotgun sequencing sample. Quality differences are apparent between samples of the same population sequenced in the A-panel and B-panel (*italic*) for both divergence and branch-shortening estimates. Differences in branch shortening estimates are calculated as (A-panel - B-panel)/mean. Differences are more pronounced when increasing the minimum genotype quality cutoffs. Comparison between two Australian individuals from panel B are shown in grey and do not increase in difference with more stringent genotype quality.

		Divergence				Branch length ratio (ind.#2-specific/ind.#1 -specific)							
Sample ID	**Population**	**GQ0**	**GQ30**	**GQ60**	**GQ90**	**GQ0**	**GQ30**	**GQ60**	**GQ90**	**$Diff_{GQ0}$**	**$Diff_{GQ30}$**	**$Diff_{GQ60}$**	**$Diff_{GQ90}$**
Denisova	Archaic	0.120802	0.121424	0.122514	0.122408	0.927896	0.923326	0.912806	0.982881				
Neandertal	Archaic	0.117621	0.117527	0.116889	0.114689	0.907667	0.904480	0.894579	0.870719				
WON,M	*Australian*	0.075126	0.075130	0.075154	0.076465	0.988213	0.988173	0.983210	0.999616				
BUR,E	*Australian*	0.075004	0.075006	0.075028	0.076056	0.992378	0.992389	0.987716	0.998746	-0.4%	-0.4%	-0.5%	0.1%
HGDP01307	Dai	0.072695	0.072677	0.074433	0.097869	1.027813	1.027305	1.032005	1.293903				
HGDP01308	*Dai*	0.072527	0.072529	0.072595	0.076100	0.999923	0.999874	0.990863	1.035670	2.8%	2.7%	4.1%	22.2%
DNK02	Dinka	0.081779	0.081736	0.083735	0.116044	1.020939	1.021275	1.036260	1.436623				
DNK07	*Dinka*	0.081977	0.081987	0.082263	0.089219	0.991088	0.991140	0.986561	1.091355	3.0%	3.0%	4.9%	27.3%
HGDP00521	French	0.070940	0.070876	0.073271	0.112842	1.024126	1.023391	1.031882	1.387509				
HGDP00533	*French*	0.070637	0.070640	0.070655	0.071752	0.995186	0.995177	0.990282	1.000429	2.9%	2.8%	4.1%	32.4%
HGDP00778	Han	0.073472	0.073347	0.074820	0.102661	1.023920	1.022425	1.018636	1.322595				
HGDP00775	*Han*	0.072855	0.072862	0.072938	0.078544	1.008057	1.008050	0.996501	1.068226	1.6%	1.4%	2.2%	21.3%
HGDP00998	Karitiana	0.072463	0.072238	0.074167	0.110231	1.026433	1.022249	1.025410	1.374892				
HGDP01015	*Karitiana*	0.072314	0.072314	0.072355	0.076757	0.994597	0.994453	0.988280	1.046568	3.2%	2.8%	3.7%	27.1%
HGDP01284	Mandenka	0.082836	0.082367	0.087515	0.169014	1.022657	1.024419	1.101441	1.778018				
HGDP01286	*Mandenka*	0.082642	0.082596	0.082257	0.086826	0.991969	0.992133	0.988869	1.074518	3.0%	3.2%	10.8%	49.3%
HGDP00456	Mbuti	0.089154	0.088771	0.093717	0.167298	1.022521	1.023812	1.110119	1.885118				
HGDP00982	*Mbuti*	0.089064	0.089064	0.089052	0.093989	0.985161	0.985122	0.980676	1.070122	3.7%	3.9%	12.4%	55.2%
MIXE0007	*Mixe*	0.072086	0.072090	0.072189	0.075131	0.994629	0.994608	0.987766	1.023420				
HGDP00542	Papuan	0.074821	0.074652	0.076747	0.114814	1.027744	1.025150	1.039690	1.434155				
HGDP00546	*Papuan*	0.075067	0.075069	0.075046	0.075702	0.995607	0.995575	0.989399	0.995839	3.2%	2.9%	5.0%	36.1%
HGDP01029	San	0.091457	0.091334	0.091171	0.100638	1.024617	1.025122	1.013437	1.205206				
HGDP01036	*San*	0.091880	0.091881	0.091864	0.094691	0.989000	0.988998	0.984034	1.042336	3.5%	3.6%	2.9%	14.5%
HGDP00665	Sardinian	0.071026	0.070937	0.075380	0.145415	1.030735	1.029293	1.061472	1.529499				
HGDP01076	*Sardinian*	*0.070859*	0.070863	0.070905	0.073863	0.992186	0.992156	0.987169	1.026479	3.8%	3.7%	7.3%	39.4%
HGDP00927	Yoruba	0.082862	0.082769	0.082984	0.095096	1.024417	1.024759	1.015776	1.202793				
HGDP00936	*Yoruba*	0.083026	0.083029	0.083046	0.086234	0.989347	0.989339	0.984759	1.034271	3.5%	3.5%	3.1%	15.1%
									Average	3.1%	3.0%	5.5%	30.8%
									Min	2.2%	2.1%	2.5%	11.5%
									Max	4.2%	4.2%	11.4%	61.3%

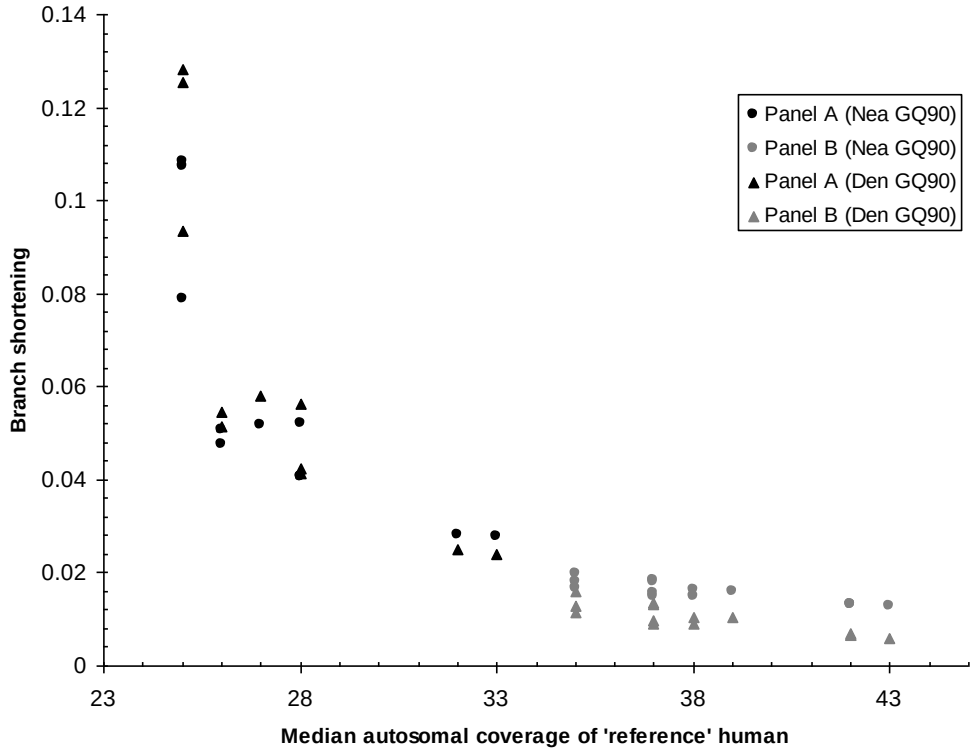
doi:10.1038/nature12886

Table S6b.2: Branch shortening for the Denisova and Altai Neandertal samples when compared among each other and to both panels of humans.

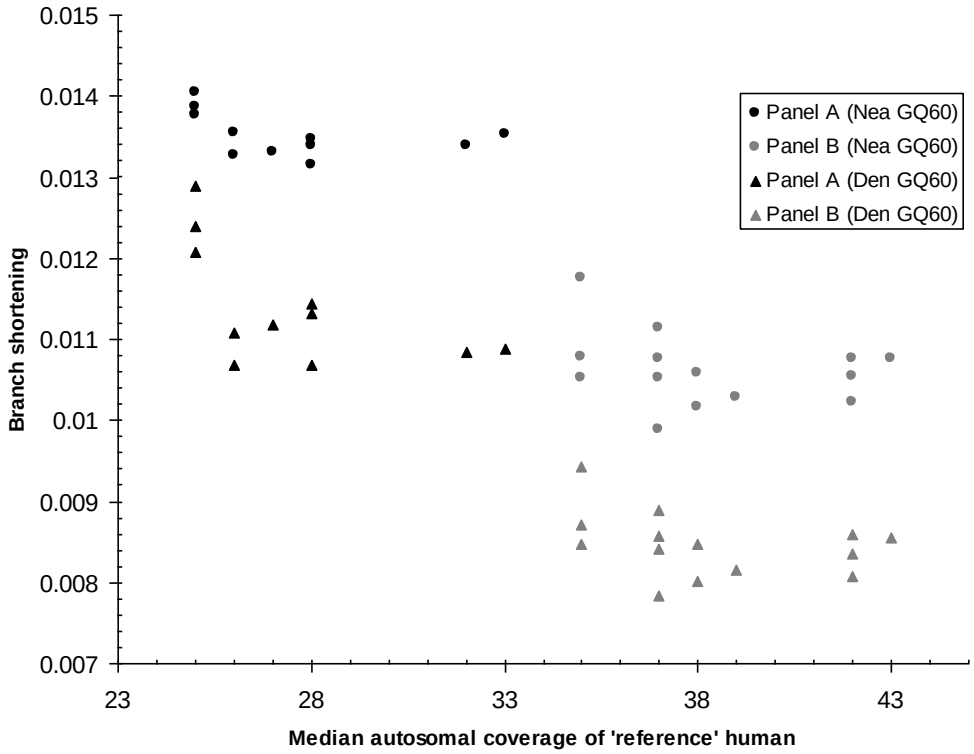
		Denisova				Neandertal			
Sample	**Population**	**GQ0**	**GQ30**	**GQ60**	**GQ90**	**GQ0**	**GQ30**	**GQ60**	**GQ90**
Neandertal/Denisova		-0.22%	-0.22%	-0.26%	-0.78%	0.22%	0.22%	0.26%	0.77%
DNK02	**Dinka**	1.04%	1.04%	1.14%	5.62%	1.26%	1.26%	1.34%	5.22%
HGDP00456	**Mbuti**	1.07%	1.07%	1.29%	12.55%	1.28%	1.29%	1.40%	10.75%
HGDP00521	**French**	1.04%	1.04%	1.12%	5.79%	1.26%	1.26%	1.33%	5.17%
HGDP00542	**Papuan**	1.08%	1.08%	1.11%	5.44%	1.29%	1.29%	1.36%	5.08%
HGDP00665	**Sardinian**	1.09%	1.09%	1.21%	9.33%	1.30%	1.30%	1.39%	7.88%
HGDP00778	**Han**	1.04%	1.04%	1.07%	4.23%	1.26%	1.26%	1.31%	4.06%
HGDP00927	**Yoruba**	1.07%	1.07%	1.08%	2.50%	1.29%	1.29%	1.34%	2.82%
HGDP00998	**Karitiana**	1.06%	1.06%	1.07%	5.14%	1.28%	1.28%	1.33%	4.76%
HGDP01029	**San**	1.09%	1.09%	1.09%	2.39%	1.31%	1.31%	1.35%	2.77%
HGDP01284	**Mandenka**	1.06%	1.06%	1.24%	12.81%	1.27%	1.27%	1.38%	10.84%
HGDP01307	**Dai**	1.07%	1.07%	1.13%	4.15%	1.29%	1.29%	1.35%	4.05%
HGDP01308	**Dai**	0.87%	0.87%	0.89%	0.99%	1.09%	1.09%	1.11%	1.56%
HGDP00533	**French**	0.84%	0.84%	0.86%	0.69%	1.05%	1.05%	1.08%	1.33%
HGDP00775	**Han**	0.93%	0.93%	0.94%	1.29%	1.14%	1.14%	1.18%	1.80%
HGDP01286	**Mandenka**	0.81%	0.80%	0.84%	1.33%	1.02%	1.02%	1.05%	1.80%
HGDP00982	**Mbuti**	0.74%	0.74%	0.78%	1.35%	0.96%	0.95%	0.99%	1.83%
HGDP00546	**Papuan**	0.84%	0.84%	0.85%	0.59%	1.05%	1.05%	1.08%	1.29%
HGDP01036	**San**	0.77%	0.77%	0.80%	1.05%	0.99%	0.99%	1.02%	1.62%
HGDP01076	**Sardinian**	0.82%	0.82%	0.85%	0.92%	1.03%	1.03%	1.06%	1.49%
HGDP00936	**Yoruba**	0.78%	0.78%	0.81%	1.04%	1.00%	1.00%	1.03%	1.59%
HGDP01015	**Karitiana**	0.83%	0.83%	0.87%	1.14%	1.05%	1.05%	1.08%	1.67%
WON,M	**Australian**	0.78%	0.78%	0.81%	0.66%	1.00%	1.00%	1.02%	1.32%
BUR,E	**Australian**	0.81%	0.81%	0.84%	0.65%	1.03%	1.03%	1.05%	1.33%
MIXE0007	**Mixe**	0.83%	0.83%	0.86%	0.90%	1.05%	1.05%	1.08%	1.49%
DNK07	**Dinka**	0.80%	0.80%	0.85%	1.60%	1.01%	1.01%	1.05%	1.99%

Figure S6b.1: Branch shortening vs. the median autosomal coverage (Table S5.1) of the human sample used as individual#1 for the Denisova and Altai Neandertal samples after applying high GQ cutoffs of 90 (A) and 60 (B) – for enhancing the quality effects seen.

A



B



Supplementary Information 7

A Drift Tree of Archaic and Modern Humans

Qiaomei Fu* and Janet Kelso

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

To further investigate the relationships between the high coverage Altai Neandertal and other Neandertals from Mezmaiskaya and Vindija[1] as well the recently sequenced Denisovan[2], and a set of worldwide present-day human populations we produced a maximum likelihood drift tree of populations using TreeMix[3].

Genotypes from the VCF files (SI 3) were converted to allele frequency counts as follows:

1) Standard filters (map35_100%, Coverage, MQ, Tandem Repeats; see SI5b) were applied to all 25 present-day humans, the Denisovan and the Altai Neandertal. Identical filters to the modern human read data were also applied to shotgun sequencing data of a female Bonobo[4].

2) Only transversions were considered in all individuals. This is necessary because the Vindija individuals were not treated with uracil-DNA-glycosylase (UDG) and endonuclease VIII, and the ancient DNA damage is therefore still present in these samples.

3) Phred-scaled likelihoods (the VCF 'PL' field) were used to identify confidently called homozygous and heterozygous genotypes for the low coverage (less than 0.45 fold coverage) Vindija and Mezmaiskaya individuals. First, we remove all sites where no genotype receives a likelihood of 30 or greater, meaning that there is no sufficient difference between the genotypes to call any genotype reliably. We call diploid genotypes only if the difference between the smallest and the second smallest PL phred-scaled likelihood is at least 30. We call haploid genotypes if the difference is less than 30 and assign the genotype of the homozygous call.

The tree was built by TreeMix and is rooted using bonobo[4]. Bonobo-specific sites were removed to improve the resolution of the tree. There are 15,727 positions that are variable across all individuals (i.e. the three Vindija individuals (Vi33.16, Vi33.25, Vi33.26), Mezmaiskaya, the Altai Neandertal, Denisovan and 25 individuals from 13 present-day human populations). The relationships between all individuals are shown in Figure S7.1. The known population relationships for present-day humans are evident, with an initial split between African and non-African populations, followed by a later split of European and Asian populations.

The Altai Neandertal is confidently grouped with the other Neandertals from Mezmaikaya and Vindija (Figure S7.1). The small number of sites, together with the remaining error in the low coverage Vindija individuals (evidenced by the long branch lengths), does not allow the relationships between the Vindija individuals to be cleanly resolved.

We tested whether combining the three Vindija individuals, to increase the number of sites that can be compared, changes the tree. This yields 190,866 positions that are variable. We find that the tree is stable with a reduced standard error (Figure S7.2).

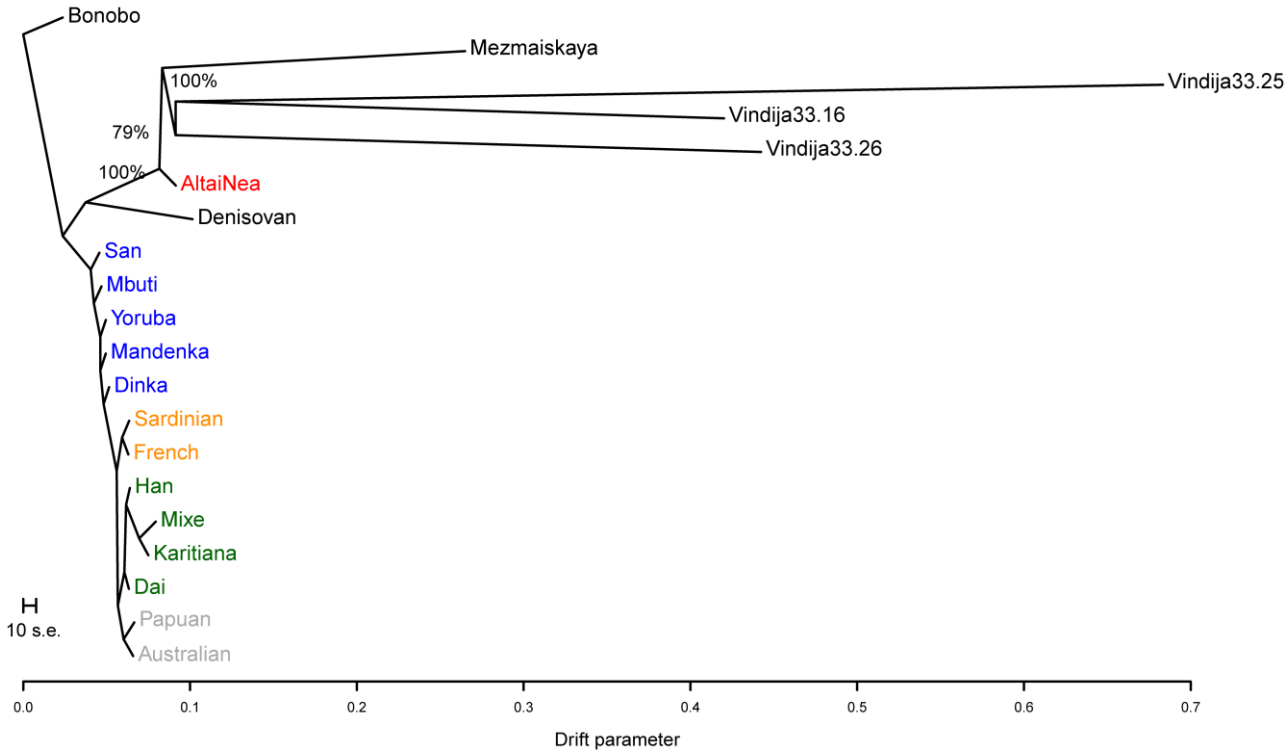


Figure S7.1. Maximum likelihood tree of bonobo, Denisova, Neandertals from Mezmaiskaya, Vindija and the Altai, and 25 individuals from 13 present-day human populations.

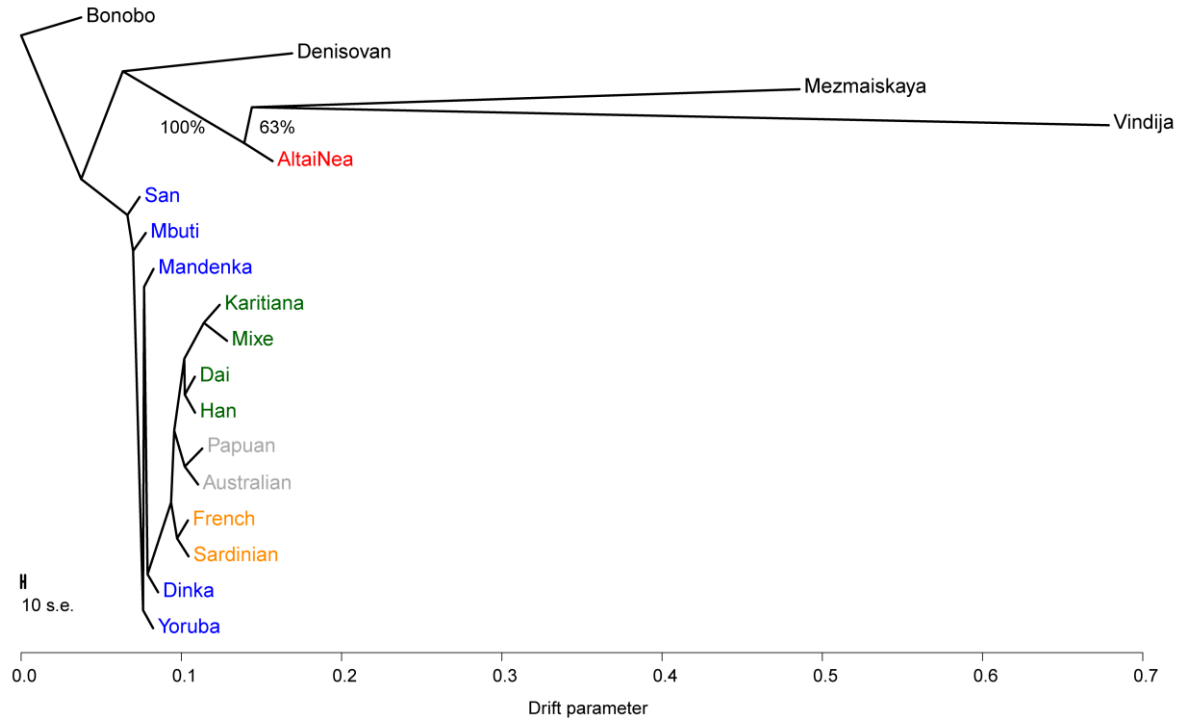


Figure S7.2. TreeMix maximum likelihood tree of Bonobo, Mezmaiskaya, Vindija and Altai Neandertals, Denisova and the 13 present-day populations.

TreeMix provides evidence for admixture events between populations. The most well-supported admixture event is that from the Denisovan to Papuans and is estimated at approximately 5%, which is similar to the estimate of 6% in ref.[2]. Additional admixture events that can be inferred by TreeMix are less well-supported, and of a lesser magnitude than the Denisovan-Papuan admixture. Within the top five predicted events we do not detect the Neandertal to modern human, the super-archaic to Denisovan, or the Neandertal to Denisovan admixtures.

References

1 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722, doi:10.1126/science.1188021 (2010).
2 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).
3 Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *precedings.nature.com* **http://precedings.nature.com/documents/6956/version/1** (2012).
4 Prufer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**, 527-531, doi:10.1038/nature11128 (2012).

Supplementary Information 8

Segmental duplications, copy number and structural diversity in the Altai Neandertal

Peter H. Sudmant*, Kay Prüfer*, and Evan E. Eichler

* To whom correspondence should be addressed (psudmant@gmail.com, pruefer@eva.mpg.de)

Segmental duplications and copy number variation

Methods

To identify copy number variants (CNVs) and assess segmental duplication diversity in the Altai Neandertal genome we constructed a map of copy number estimates across the genome at a scale of 500bp of unmasked sequence. This map was constructed as described in Sudmant *et al* 2010[1] by dividing reads into their 36 base-pair (bp) constituents and mapping them to the repeat masked human reference sequence Build 37 (HG19) using the mrsFAST read aligner[2] (version 2.3.0.2, allowing up to 2 mismatches/36bp). Read depth was then corrected for underlying GC content and copy numbers were estimated using a calibration curve generated from regions of known copy number. In addition to the high coverage Altai Neandertal individual we generated copy number maps for the low coverage Mezmaiskaya Neandertal, the high coverage Denisova individual[3], and 25 diverse high coverage human individuals (**Table S8.1**).

Table S8.1: Individuals assessed for copy number variation and segmental duplication diversity. Correlations (r) are calculated in regions known copy number between the copy number and GC-corrected read depth.

Individual	Sequencing group	Sex	mrsFAST 36-mer fold-coverage	r
Mezmaiskaya Neandertal	Mezmaiskaya Neandertal	F	0.329722	0.717253
Altai Neandertal	Altai Neandertal	F	49.7008	0.918732
Denisova	Denisova	F	21.6454	0.905983
DNK02_Dinka	A_team_11 Human	M	10.8766	0.924249
HGDP_00456_Mbutipygmy	A_team_11 Human	M	10.2303	0.920364
HGDP_00521_French	A_team_11 Human	M	11.3448	0.935537
HGDP_00542_Papuan	A_team_11 Human	M	11.0158	0.930799
HGDP_00665_Sardinian	A_team_11 Human	M	10.5201	0.929221
HGDP_00778_Han	A_team_11 Human	M	11.6409	0.924928
HGDP_00927_Yoruba	A_team_11 Human	M	13.4773	0.933963
HGDP_00998_Karitiana	A_team_11 Human	M	11.4252	0.930952
HGDP_01029_San	A_team_11 Human	M	14.1071	0.933836
HGDP_01284_Mandenka	A_team_11 Human	M	10.6812	0.932112
HGDP_01307_Dai	A_team_11 Human	M	11.0267	0.930635
BUR_E_Australian	B_team_14 Human	F	15.0871	0.927677
DNK07_Dinka	B_team_14 Human	M	12.9279	0.932636
HGDP_00533_French	B_team_14 Human	M	15.2971	0.930403
HGDP_00546_Papuan	B_team_14 Human	M	15.3589	0.933366
HGDP_00775_Han	B_team_14 Human	M	12.6516	0.929352
HGDP_00936_Yoruba	B_team_14 Human	M	14.4002	0.928613
HGDP_00982_Mbuti	B_team_14 Human	M	12.9866	0.930841
HGDP_01015_Karitiana	B_team_14 Human	M	13.0453	0.930927
HGDP_01036_San	B_team_14 Human	M	13.5586	0.930782
HGDP_01076_Sardinian	B_team_14 Human	M	14.1145	0.932511
HGDP_01286_Mandenka	B_team_14 Human	M	13.1231	0.933333
HGDP_01308_Dai	B_team_14 Human	M	13.0375	0.922476
MIXE_0007_Mixe	B_team_14 Human	F	14.1177	0.927027
WON_M_Australian	B_team_14 Human	M	14.8937	0.93116

Quality control

We began by assessing the quality of each of the individual genomes analyzed for copy number variation using two different approaches. We first calculated the correlation coefficient (r) between GC-corrected read depth and copy state in regions of known copy number. This simple metric has a mean of 0.93 among the modern humans assessed and values of 0.91, 0.92 and 0.71 in the Denisova, Altai and Mezmaiskaya individuals respectively, demonstrating similar sensitivity to detect copy number changes in modern humans and the high coverage ancient genomes. The lower coverage of the Mezmaiskaya genome renders it far less comparable to the other genomes assessed. We next assessed the ability to assign a copy number state of 2 to regions of euchromatic sequence unlikely to be subject to copy number variation. We select these 'control' loci by identifying all continuous blocks >100Kb in the genome remaining after subtracting all known modern human structural variants in the database of genomic variants (DGV), segmental duplications, and gaps. As sequencing coverage often varies with GC content as a result of biases in library construction, we assessed our ability to accurately identify the copy number of windows in these control loci in different GC-content bins (**Figure S8.1**). All of the modern humans assessed showed remarkable power to accurately predict copy number with an average of >99% accuracy over all GC bins. The high coverage ancient genomes performed similarly, with 98.9% of regions accurately predicted. However, only 72% of all loci could be accurately predicted in the lower coverage Mezmaiskaya individual. From these analyses we thus determined that copy number prediction in the Mezmaiskaya genome is of poorer quality and we decided to exclude Mezmaiskaya predictions from the comparison with the high-coverage genomes. Mezmaiskaya genotypes are reported for specific candidate regions, but should be interpreted with caution.

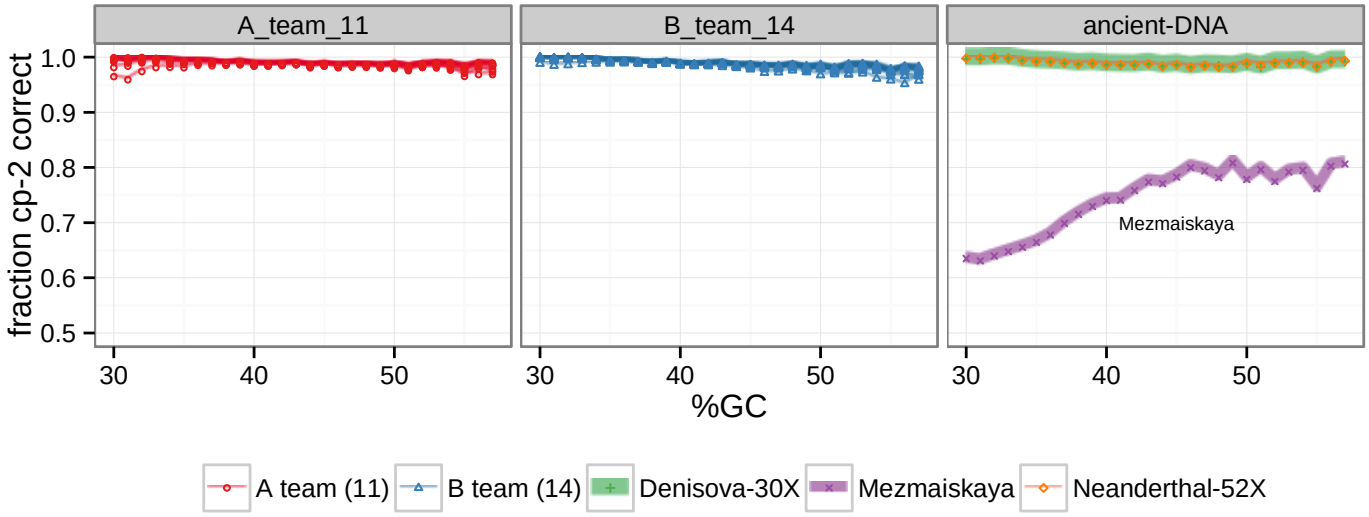


Figure S8.1: Fraction of diploid loci accurately identified as copy number 2 is plotted as a function of different GC% bins. ~99% of loci can be accurately estimated among modern humans and the high coverage Neandertal and Denisova ancient genomes. Additionally, accuracy is robust to varying GC% contents. The accuracy of copy number prediction in the low coverage Mezmaiskaya individual is much lower than the other individuals and varies with GC% content.

Segmental duplication content

Individual segmental duplication blocks were independently identified on a per-individual basis using a scale-space filtering[4] based segmentation algorithm (Sudmant *et al*, *under review*). Briefly, this algorithm transforms the waveform signal of windowed copy number estimates, $f(x)$, into a set of waveforms, $f(x,\sigma)$ parameterized by the *scale* variable σ, which represents the standard deviation of a Gaussian smoothing kernel. This waveform surface is then explored for inflections that persist across multiple scales which correspond to likely copy number variants. Among most humans, the Denisova and Altai Neandertal, the number of duplicated bp was largely consistent (~159-168Mbp for regions > 1kb; see **Table S8.2)**. However, one individual, Mbuti sample HGDP-00982, showed with 220Mbp an aberrantly high number of duplicated base-pairs, ~142Mbp more than any other sample. These excess duplications localized to chromosome 9 (**Figure S8.2**) where HGDP-00982 exhibits a largely triploid chromosome consistent with a heterogeneous population of cells, most of which contain three chromosome 9 copies. As such, chromosome 9 was excluded from analyses of HGDP-00982.

Table S8.2: Total per-individual base-pair count of duplicated loci. The highlighted Mbuti individual showed an aberrantly high number of duplicated base-pairs.

individual	bp of duplication blocks	
	>1kb	>5kb
HGDP_00778_Han	159,096,031	136,111,890
HGDP_00546_Papuan	159,653,313	135,777,753
HGDP_00521_French	159,796,254	136,764,984
HGDP_01076_Sardinian	160,117,937	137,003,195
HGDP_00542_Papuan	160,378,153	136,636,137
HGDP_01015_Karitiana	160,494,458	136,612,512
HGDP_00998_Karitiana	161,020,311	136,855,531
WON_M_Australian	161,155,468	138,012,893
HGDP_00533_French	161,245,194	137,154,676
HGDP_00927_Yoruba	161,247,326	138,324,764
HGDP_00775_Han	161,265,616	137,191,628
HGDP_01286_Mandenka	161,270,461	137,675,939
HGDP_01029_San	161,455,879	137,296,731
HGDP_00665_Sardinian	161,524,439	137,585,663
HGDP_01036_San	161,658,420	137,676,744
HGDP_01308_Dai	161,710,048	137,350,801
HGDP_01284_Mandenka	161,797,486	137,454,281
HGDP_00456_Mbutipygmy	161,895,897	136,956,751
HGDP_00936_Yoruba	161,904,005	138,808,517
Homo_denisova-Denisova_30x	162,316,840	139,490,602
DNK02_Dinka	162,511,615	139,145,109
MIXE_0007_Mixe	162,762,286	139,094,970
BUR_E_Australian	163,378,311	139,141,468
Altai-Neandertal	163,490,404	141,474,129
DNK07_Dinka	163,535,649	140,106,402
HGDP_01307_Dai	168,284,716	141,594,294
HGDP_00982_Mbuti	220,879,971	195,927,474

We next sought to identify lineage specific duplications among the 25 modern humans and two ancient individuals in addition to three high-coverage non-human great apes, the Western chimpanzee Clint, the Western lowland gorilla Kamilah, and the Bornean orangutan KB5404-Billy (**Table S8.3**). As expected, the vast majority of duplications are shared among modern human, Denisova and Neandertal individuals. Interestingly, we find no Neandertal-Denisova specific shared duplications, though we do identify 75.2kbp of human, 76.6kbp Denisova, and 181.1kbp of Neandertal specific duplication, intersecting a total of 13 unique genes (**Table S8.4**).

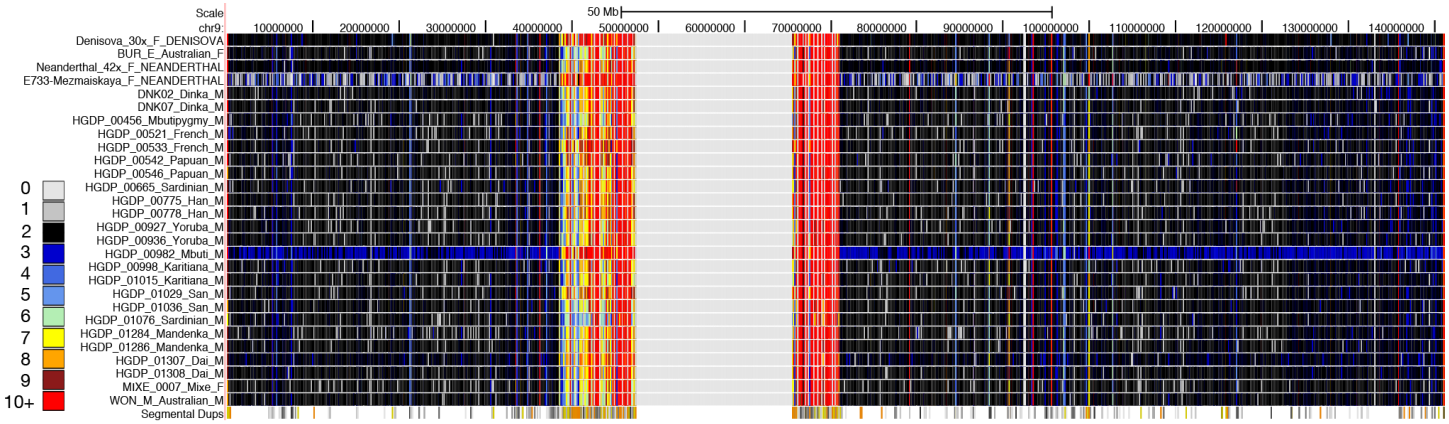


Figure S8.2: Mbuti sample HGDP-00982 demonstrates a complete duplication of chromosome 9, likely the result of a cell-line artifact. The Mezmaiskaya Neandertal is additionally plotted demonstrating that copy number estimations perform poorly on this individual.

Table S8.3: Summary of the number of duplicated base-pairs along and shared between lineages. Copy corrected duplicated bp refers to the number of duplicated bp corrected for over-counting segmental duplications represented in the human reference. H,N,D,C,G and O stand for Human, Neandertal, Denisova, Chimpanzee, Gorilla and Orangutan respectively.

Lineage	Duplicated bp	Copy-corrected duplicated bp	events
H-D	67520	66194	7
H	75207	53388	3
D	76585	139355	8
N	181101	279646	6
H-N	271110	219981	19
C-H-D-N	4791330	4341010	297
H-D-N	6450328	5679030	306
C	8039549	12086082	1751
G	14849699	39623204	1562
O	19458418	33373255	3378
G-C-H-D-N	25275862	30800081	1153
O-G-C-H-D-N	69582673	273666449	5955

Table S8.4: Lineage specific segmental duplications along each of the terminal branches and genes encompassed.

Locus	length	lineage	genes	Genotypes			
				Modern Humans (median)	Denisova	Altai	Mezmaiskaya
chr12:122079832-122087495	**7663**	Altai-Neandertal	ORAI1	2	2	4	3
chr12:132295389-132391442	**96053**	Altai-Neandertal	MMP17,ULK1	2	2	4	2
chr19:9284044-9291195	**7151**	Altai-Neandertal		2	2	4	4
chr20:281880-290717	**8837**	Altai-Neandertal		2	2	10	9
chr3:12639069-12641393	**2324**	Altai-Neandertal	RAF1	2	2	7	3
chr6:95473793-95532866	**59073**	Altai-Neandertal		2	2	3	2
chr11:39901956-39909545	**7589**	Denisova		2	4	2	2
chr1:161272681-161274838	**2157**	Denisova	MPZ	2	4	2	2
chr12:49894191-49897733	**3542**	Denisova	SPATS2	2	4	2	2
chr19:55302094-55315197	**13103**	Denisova	KIR3DP1,KIR2DL4	2	4	2	2
chr2:48781187-48787915	**6728**	Denisova		2	3	2	2
chr4:68542692-68577288	**34596**	Denisova	UBA6,LOC550112	2	3	2	2
chr4:68579206-68581585	**2379**	Denisova	LOC550112	2	3	2	2
chr7:140872574-140879065	**6491**	Denisova	LOC100131199	2	6	2	2
chr1:108924526-108990191	**65665**	Modern Human		4	2	2	2
chr16:30200098-30206185	**6087**	Modern Human	CORO1A ,LOC606724,BOLA2	6	2	2	2
chr2:87417089-87420544	**3455**	Modern Human		4	2	2	2

Of particular interest is the modern human-specific duplication on 16p11.2 which encompasses the *BOLA2* gene. This locus is the breakpoint of the 16p11.2 micro-deletion, which results in developmental delay, intellectual disability, and autism[5,6]. We genotyped the *BOLA2* gene in 675 diverse human individuals sequenced to low coverage as part of the 1000 Genome Project Phase I[7] to assess the population distribution of copy numbers in homo-sapiens (**Figure S8.3**). While both the Altai Neandertal and Denisova individual exhibit the ancestral diploid copy number as seen in all the non-human great apes, only a single human individual exhibits this diploid copy number state.

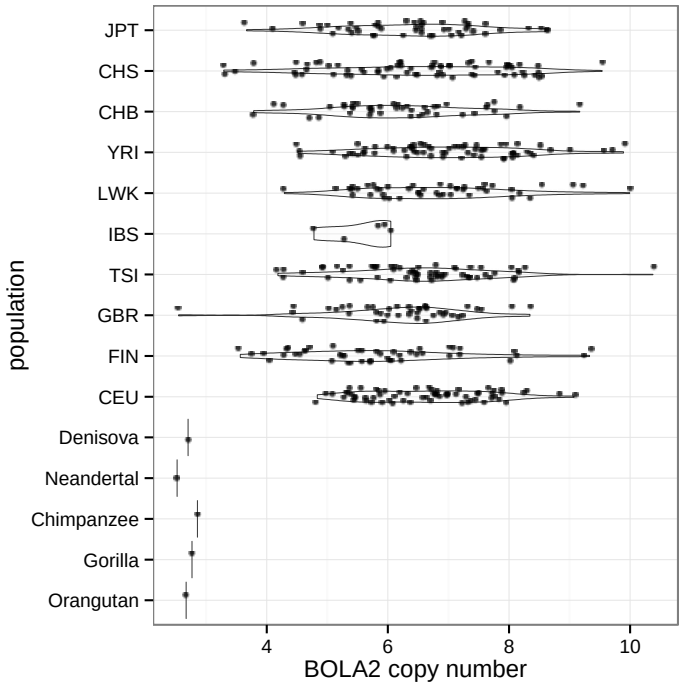


Figure S8.3: Population distribution of BOLA2 copy number in 675 humans from the 1000 genomes project in comparison to the Denisova and Neandertal individuals and non-human great apes. Individual copy-number genotypes are superimposed over violin plots of the distribution of copy-number states in individual populations.

We have previously reported that the chromosome 18 *ROCK1* duplication present in all humans is absent from the Denisova individual[3], suggesting perhaps that this too is a human specific duplication. This duplication was however found in the Altai Neandertal individual suggesting instead that the *ROCK1* locus has been specifically deleted in the Denisova.

We assayed next for lineage specific gene expansions along the terminal Denisova, Neandertal and present-day human lineages, identifying genes where one of these lineages showed expansion over the ancestral state. Two modern human specific gene expansions were identified, *TPTE2*, which has additionally undergone expansion in Gorillas, and the *Amy1* and *Amy2* genes (**Figure S8.4**). *Amy1* and *Amy2* encode for amylase enzymes. *Amy1* has been suggested to be selected for higher copy number in populations with starch rich diets[8]. Genotyping an additional 675 humans we find that *Amy1* ranges from 2-18 copies among individuals, with only 11 individuals of Finnish and CEU descent sharing the same ancestral copy number states as the Denisova and Altai Neandertal. Consistent with previous observations, we find Asian populations have significantly higher copies of *Amy1* than non-Asian populations ($P=1.07x10^{-12}$).

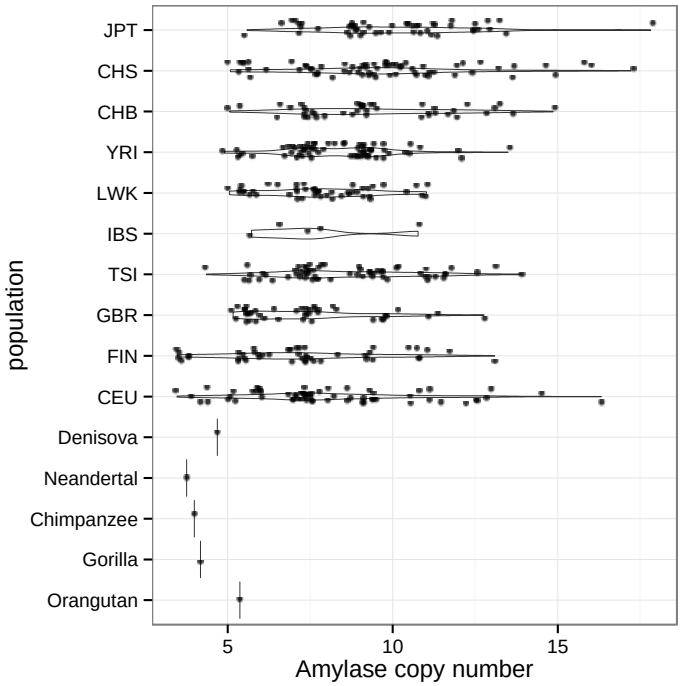


Figure S8.4: *Amy1* copy number genotypes in 675 humans sequenced to low coverage by the 1000 Genomes Project in addition to the Altai Neandertal, Denisova and three non-human great apes. The Altai Neandertal and Denisova share the ancestral copy number of the Amylase gene along with 11 Finnish and CEU Europeans.

Deletions

We next assessed lineage specific fixed/homozygous deletions in the Altai Neandertal and Denisova individual identifying 212 regions encompassing a total of 1.5Mbp of sequence lost along these lineages (**Table S8.5** attached, **Table S8.6**). Among these loci, 17 exon deletions were identified on the archaic lineages (**Table S8.7**), including complete loss of the *GSTT1*, *MRGPRG*, and *C11orf36* genes and partial loss of the *GSTTP2* and *SPINK14* genes along the Denisova-Neandertal ancestral lineage. Each of *GSTT1*, *MRGPRG* and *C11orf36* show copy number polymorphism among modern humans and *GSTT1* shows strong continental population stratification with Asian populations having higher frequency of loss (**Figure S8.5**).

Table S8.6: Summary of lineage specific homozygous deletions identified along the Neandertal-Denisova branch.

lineage	bp	loci
Neandertal-Denisova	141723	35
Neandertal	646519	77
Denisova	723582	100

Table S8.7: Summary of gene exon-deletions along the Neandertal-Denisova lineage

Lineage	Gene	Gene coordinate	Exons deleted	Total Exons
Denisova	LCE3C/LCEB	chr1:152573137-152573562	1	1
Denisova	GUCY2GP	chr10:114067935-114116353	2	19
Denisova	PLEKHA5	chr12:19282625-19526602	1	28
Denisova	FABP6	chr5:159614373-159665729	1	7
Neandertal	C1orf227	chr1:213003484-213020991	2	3
Neandertal	TACC2	chr10:123748688-124014057	1	23
Neandertal	FLI1	chr11:128562388-128683162	1	9
Neandertal	C12orf70	chr12:27619742-27655118	2	9
Neandertal	TFAMP1	chr7:1654105-1656328	1	1
Neandertal	GSDMD	chr8:144635556-144645231	1	14
Neandertal	PLEKHA2	chr8:38758752-38831430	1	12
Neandertal-Denisova	MRGPRG	chr11:3239173-3240043	1	1
Neandertal-Denisova	C11orf36	chr11:3239561-3244361	2	2
Neandertal-Denisova	LOC391322	chr22:24373116-24374043	2	2
Neandertal-Denisova	GSTT1	chr22:24376138-24384284	5	5
Neandertal-Denisova	GSTTP2	chr22:24385937-24401899	1	5
Neandertal-Denisova	SPINK14	chr5:147549295-147554961	1	4

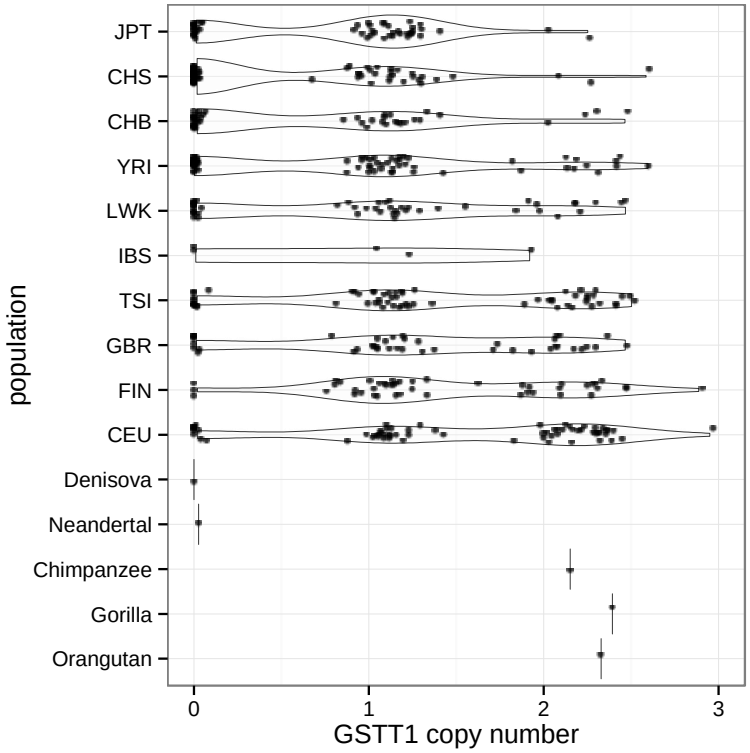


Figure S8.5: Copy number genotypes of the GSTT1 gene in 675 diverse humans from the 1000 Genomes Project, the Denisova, Neandertal and three non-human great ape genomes. Asian populations show a higher frequency of homozygous and hemizygous deletions than African or European populations.

Structural variation

The chromosome two fusion is present in Neandertals

The modern human genome differs from the genomes of great apes by a fusion of two chromosomes. This fusion event formed the human chromosome two[9]. As a remnant of the fusion of the two chromosomes, chromosome two shows a short stretch of telomeric repeat sequence in forward and reverse direction in the interior of the chromosome[10]. This sequence was also found in the Denisovan genome, showing that the

fusion event predates the split of modern humans and Denisovans[3]. We repeat this analysis with the Altai Neandertal sequence and find 5 reads spanning the fusion site.

Processing and Results

We use the Altai Neandertal sequences aligned to the human reference genome hg19 and extract reads that overlap the fusion region on chromosome two. A total of five reads overlap the region and all reads have a mapping quality score of 37. Figure S8.6 shows the alignment of the reads to the human reference.

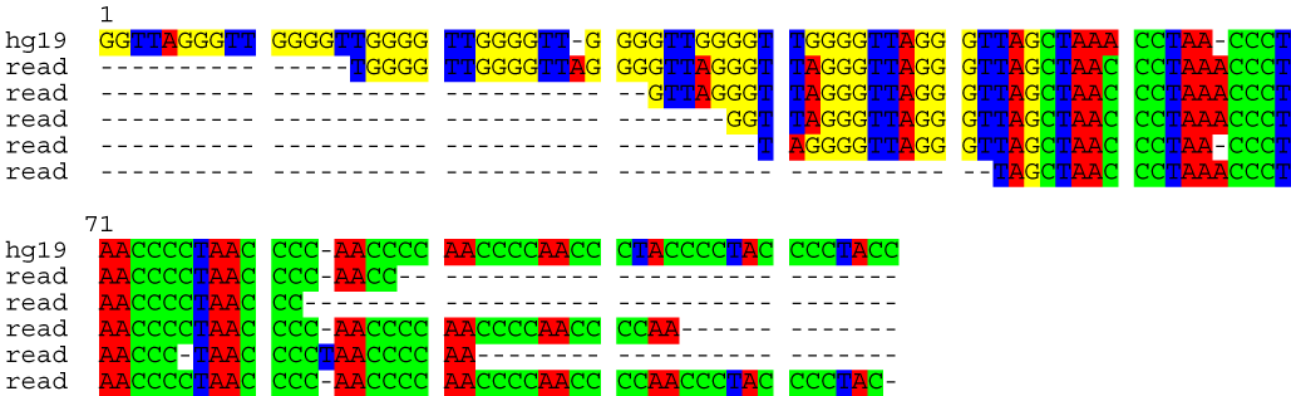


Figure S8.6: Human reference (hg19; chr2:114,360,452-114,360,565) and five Altai Neandertal reads aligning to the region.

References

1. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330,** 641–646 (2010).
2. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods* **7,** 576–577 (2010).
3. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338,** 222–226 (2012).
4. Witkin, A. Scale-space filtering: A new approach to multi-scale description. **9,** 150–153 (1984).
5. Kumar, R. A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17,** 628–638 (2008).
6. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358,** 667–675 (2008).
7. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
8. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39,** 1256–1260 (2007).
9. Yunis, J. J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215,** 1525–1530 (1982).
10. IJdo, J. W., Baldini, A., Ward, D. C., Reeders, S. T. & Wells, R. A. Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci. U.S.A.* **88,** 9051–9055 (1991).

Supplementary Information 9

Neandertal genetic heterozygosity

Kay Prüfer* and Cesare de Filippo

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

We compare the heterozygosity over all autosomes in the Altai Neandertal genome to the heterozygosity in Denisova and modern Humans using two different approaches. Both approaches consistently show that the Neandertal genome has a substantially lower heterozygosity than modern humans and a slightly lower heterozygosity than the Denisova genome. Specifically, we find an average of less than two heterozygous sites every 10,000 basepairs and this heterozygosity corresponds to 16-32% of present-day humans and 81% of Denisova.

Estimating Heterozygosity from Genotype Calls

We estimated the autosomal heterozygosity from genotype calls using GATK (see SI 3) by dividing the number of heterozygous genotypes by the total number of genotypes that passed the filtering thresholds described in SI 5b. Specifically, we removed indels and positions based on the mappability tracks (using the more stringent threshold *Map35_100%*), coverage, mapping quality, and the simple repeats tracks. To obtain a comparable heterozygosity estimate for all the genomes, we restricted to positions that passed filtering in all 27 individuals. This left almost 700 million positions (roughly 26% of the genome), and we report the GATK heterozygosity estimates in Table S9.1 and Figure S9.1. This analysis suggests that the heterozygosity of Neandertal is 16-32% of present-day humans and is also reduced compared with Denisova (~81%). On average, over every 10,000 nucleotides on the autosomes we observed ~1.8 heterozygous sites in Neandertal and ~2.1 in Denisova. However, after removing long homozygous regions that are the result of recent inbreeding within Altai Neandertal (regions >2.5cM; cutoff: 92.5% in regions of strict run of homozygosity: see SI 10), the heterozygosity level of Altai is very similar to Denisova and also less reduced to those of modern humans (Table S9.1).

We also tested the effect of different settings for a set of minimal filters (SI5b) on the heterozygosity estimates. We find that Altai Neandertal remains, independent of filtering, the individual with the lowest heterozygosity. The filters for mapability (*Map35_100%)* and coverage had a noticeable effect on relative and absolute heterozygosity estimates, while further filtering of tandem repeats and by mapping quality had little effect. The less stringent mappability track *Map35_50%*, gave slightly higher estimates than the more stringent *Map35_100%* track. However, the relative heterozygosity estimates remained stable with a maximum difference of 3.5% compared between the two mapability tracks.

Estimating Heterozygosity using *mlRho*

To estimate heterozygosity, we used *mlrho*[1], a maximum likelihood estimator of population mutation rate (θ) from high-coverage sequencing data of a single individual. Heterozygosity is approximated by θ under the infinite sites model and when the value of θ is small.

We ran *mlrho* on all the data from present-day humans (see SI 4), Denisova[2] and the Altai Neandertal. All data were filtered as described in the previous paragraph and in SI 5b. Table S9.1 shows the estimated heterozygosity for all individuals. In agreement with the estimate based on genotype calls, we observe a consistently lower heterozygosity in Neandertal as compared to Denisova and all modern humans. The Altai heterozygosity increases to levels similar to Denisova when inbred regions are excluded (regions >2.5cM; cutoff: 92.5% in regions of strict run of homozygosity: see SI 10).

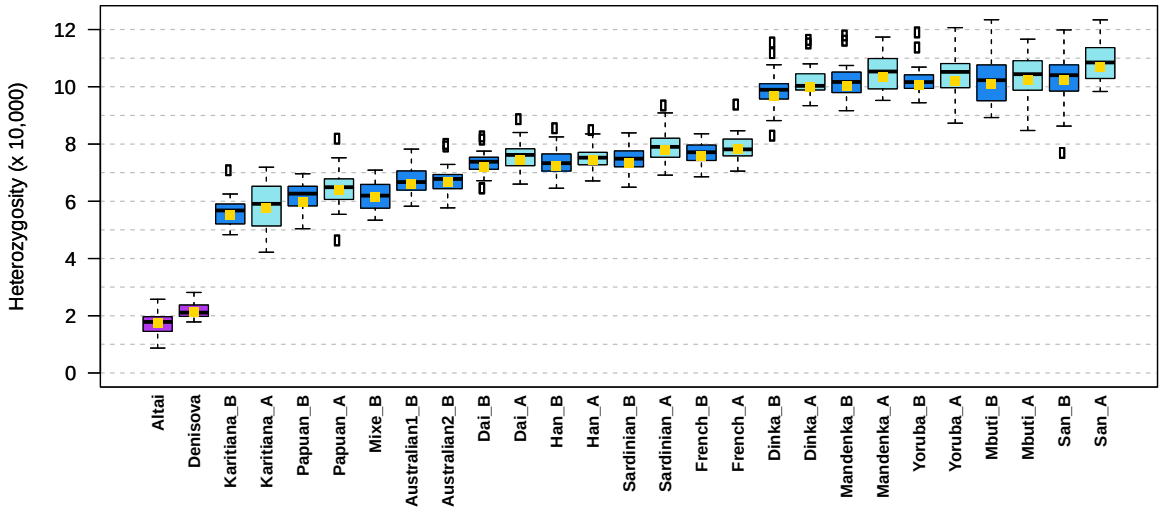


Figure S9.1: Heterozygosity estimates and autosomal distributions. The yellow squares are the average genomic heterozygosity estimates (i.e. those in Table S9.1) for 10,000 sites and the box-and-whiskers represent the distributions across 22 autosomes given by the R-function 'boxplot' with default parameters[3]. Archaic samples are in purple; modern human from A- and B-panels are in light and dark blue, respectively. The samples 'Australian1_B' and 'Australian2_B' are 'WON,M' and 'BUR,E', respectively (see Table S9.1).

Table S9.1: Comparisons of heterozygosity estimates using genotype calls and mlrho with the filters described in SI 5b. The values are for every 10,000 sites, except for the relative estimates, which are in percentage.

		Genotype calls				*mlrho*			
		Het.		Het.rel. (%) [#]		θ		θ.rel. (%) [#]	
SAMPLE	**POP**	**all**	**no inb.**	**all**	**no inb.**	**all**	**no inb.**	**all**	**no inb.**
Altai	Neandertal	1.76	2.21	100	100	1.67	2.12	100	100
Denisova	Denisova	2.15	2.18	82	101	1.88	1.91	89	111
A-Panel									
HGDP00998	Karitiana	5.77	5.89	30	38	5.65	5.76	30	37
HGDP00542	Papuan	6.39	6.42	27	34	6.28	6.31	27	34
HGDP01307	Dai	7.45	7.48	24	30	7.31	7.34	23	29
HGDP00778	Han	7.45	7.55	24	29	7.31	7.4	23	29
HGDP00665	Sardinian	7.81	7.87	22	28	7.68	7.74	22	27
HGDP00521	French	7.82	7.87	22	28	7.68	7.73	22	27
DNK02	Dinka	10.00	10.06	18	22	9.86	9.92	17	21
HGDP01284	Mandenka	10.33	10.44	17	21	10.20	10.3	16	21
HGDP00927	Yoruba	10.20	10.29	17	21	10.00	10.1	17	21
HGDP00456	Mbuti	10.24	10.43	17	21	10.10	10.3	17	21
HGDP01029	San	10.70	10.74	16	21	10.50	10.5	16	20
B-Panel									
HGDP01015	Karitiana	5.53	5.56	32	40	5.52	5.55	30	38
HGDP00546	Papuan	5.99	6.04	29	37	5.98	6.04	28	35
MIXE_007	Mixe	6.13	6.24	29	35	6.10	6.2	27	34
WON,M	Australian	6.60	6.64	27	33	6.60	6.64	25	32
BUR,E	Australian	6.67	6.73	26	33	6.66	6.73	25	32
HGDP01308	Dai	7.20	7.30	24	30	7.15	7.26	23	29
HGDP00775	Han	7.24	7.32	24	30	7.18	7.25	23	29
HGDP01076	Sardinian	7.35	7.45	24	30	7.34	7.44	23	28
HGDP00533	French	7.58	7.68	23	29	7.57	7.67	22	28
DNK07	Dinka	9.68	9.73	18	23	9.64	9.69	17	22
HGDP01286	Mandenka	10.02	10.10	18	22	10.00	10.1	17	21
HGDP00936	Yoruba	10.07	10.16	17	22	10.10	10.2	17	21
HGDP00982	Mbuti	10.10	10.14	17	22	10.10	10.1	17	21
HGDP01036	San	10.22	10.30	17	21	10.20	10.3	16	21

[#]: relative (rel.) values report the ratio of heterozygosity in Altai Neandertal to the heterozygosity of the individual in each row;
'all' refers to all positions (~698.3 Mbp) that passed minimal filters (see SI 5b) in all individuals;
'no inb.' exclude the inbred segments of Altai Neandertal (see SI 10) from all individuals and span a total of ~543.3 Mbp.

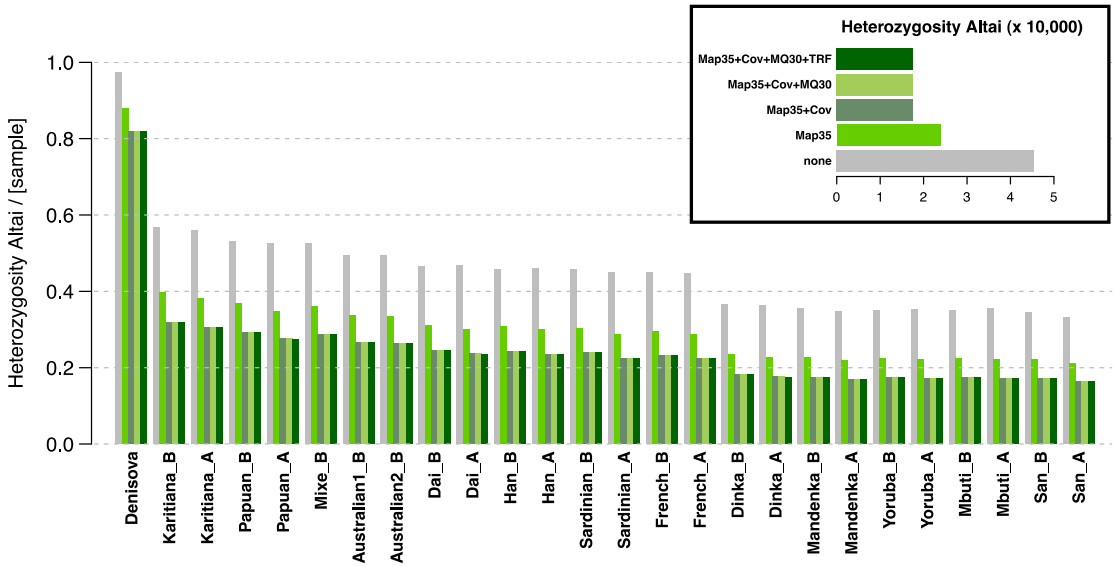


Figure S9.2: The filtering effects on the relative heterozygosity of Altai. The bars are colored according to the filters used and labeled in the top-right: 'Map35' stands for mappability *Map35_100%*, 'Cov' for coverage, 'MQ30' for mapping quality ≥ 30, and 'TRF' for simple repeats (see SI 5b for more details). The samples 'Australian1_B' and 'Australian2_B' are 'WON,M' and 'BUR,E', respectively (see Table S9.1).

References

1 Haubold, B., Pfaffelhuber, P. & Lynch, M. mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Molecular ecology* **19 Suppl 1**, 277-284, doi:10.1111/j.1365-294X.2009.04482.x (2010).

2 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).

3 Becker, R. A., Chambers, J. M. & Wilks, A. R. *The new S language : a programming environment for data analysis and graphics*. (Wadsworth & Brooks/Cole Advanced Books & Software, 1988).

Supplementary Information 10

Recent Inbreeding and Background Inbreeding

Flora Jay* and Montgomery Slatkin

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

We detect long homozygous segments in the Altai Neanderthal genome that suggest that this individual was inbred. We compare the number of homozygous segments and the fraction of the genome in homozygous segments (coverage) to those found when simulating different inbreeding scenarios. We find evidence for recent inbreeding, with a probable inbreeding coefficient of 1/8. The number of shorter homozygous segments and their coverage indicates that there is background inbreeding as well: the common ancestor or ancestors of the parents may have been themselves somewhat inbred. We also use the length distribution of homozygous segments on the X chromosome to narrow the range of possible scenarios for the recent inbreeding.

Methods

Filtering

We filtered the data using the minimal filters described in SI 5b (with map35_100%).
Additionally we do not consider a site as heterozygous if it:

- has genotype quality lower than 40
- surrounds an indel (-5bp/+5bp)
- is not well balanced (ie one allele is found in more than 70% of the reads). We filtered out these sites after observing that regions of homozygosity were enriched for unbalanced heterozygous.

Scan

We call a segment a "strict run of homozygosity" (strict ROH) if it is longer than 50kb, has no heterozygous sites, has less that 50% missing data and less than 70% of missing plus filtered data. We use the term homozygous by descent (HBD) for tracts that are sufficiently long to be most likely identical by descent (IBD). HBD tracts might contain a few heterozygous sites due to sequencing errors or mutations, so we searched for tracts that have a high fraction of sites in a strict ROH (i. e. we tolerate some heterozygous sites). To do so, we slide a window of size 1Mb across the genome, with a step of size 100kb. If the window contains at least *p%* of sites in a strict ROH, it is considered as a putative HBD tract. Consecutive HBD tracts are merged. HBD tracts detected this way are at least 1Mb long. To convert to centimorgans we use the uniform, genome-wide average rate of 1.3 cM/Mb.

Identification of HBD tracts

To identify putative HBD tracts in the Altai Neandertal, Denisova, and B-panel genomes, we applied the filters and the scan described in Methods.

The detection of tracts is sensitive to the parameter *p*. A larger value for *p* will increase the stringency of the scan. A more stringent scan, in turn, will lead to the fragmentation of detected tracts, so that a true HBD tract is often represented by several smaller, neighbouring tracts. We use this observation to find an appropriate value for *p* by testing for clustering of detected tracts along the genome. We run our scan for different values of *p*, calculating the minimum distance between detected tracts each time. For each chromosome and tested *p* parameter, we then shuffle the detected tracts, and again calculate the minimum distance between tracts. By repeating the procedure 100 times we obtain the neutral distribution of the minimum distances and test if the observed minimum is smaller than expected, computing a p-value for each chromosome and each value of the parameter *p*. Figure S10.1 shows for each scan the percentage of chromosomes on which tracts longer than 2.5 cM were significantly clustered (p-value<0.05). In order to be conservative, we chose the largest *p* for which this percentage of chromosomes is less than 5%. We applied the clustering test separately to all individuals, as the parameter *p* should depend on the average heterozygosity (for a given *p* the risk of false positive is higher for an individual harboring a smaller heterozygosity).

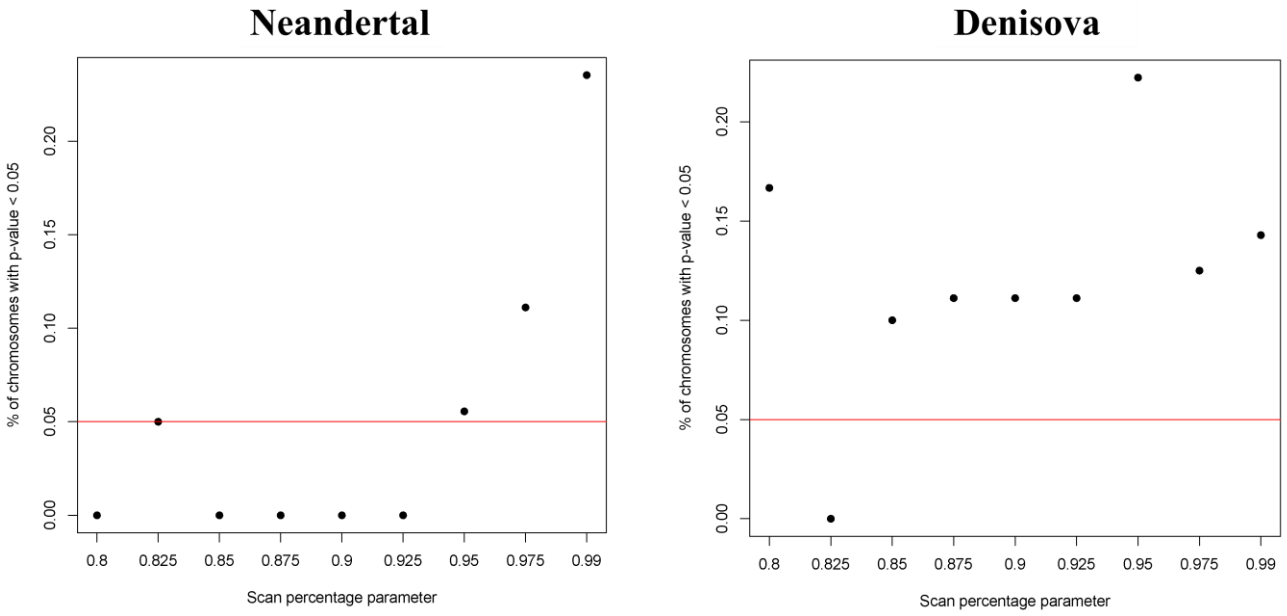


Figure S10.1: For each value of parameter *p*, percentage of chromosomes on which HBD tracts longer than 2.5 cM are significantly clustered (p-value<0.05). The red line indicates 5%. The largest *p* for which the percentage of chromosomes is less than 5% is 0.925 for Neanderthal (left) and 0.825 for Denisova (right).

Figure S10.2 shows the position of HBD tracts found for Altai, Denisova, and a Papuan individual for chromosomes 8 and 14.

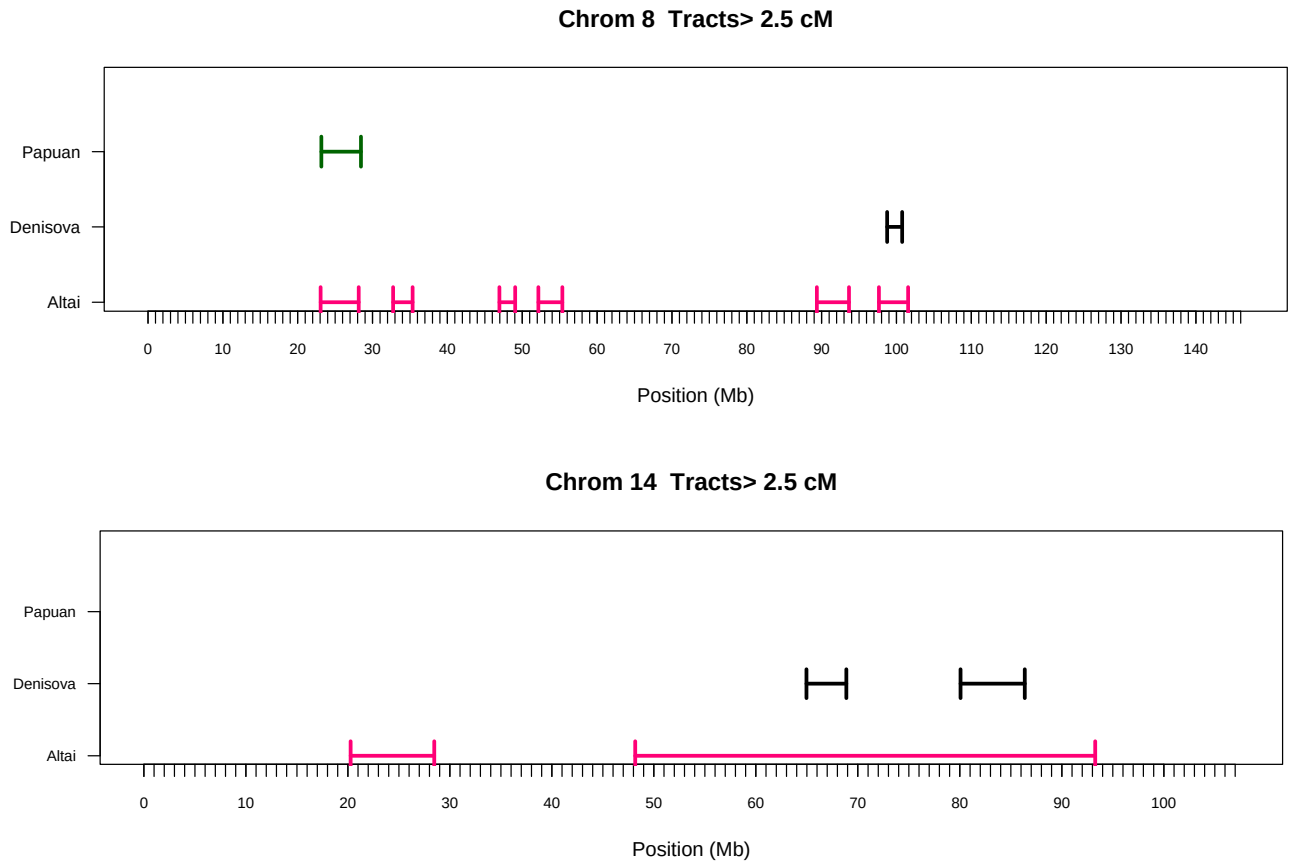


Figure S10.1: HBD tracts identified in chromosomes 8 and 14 for Papuan (top line, green), Denisova (middle line, black), and Altai (bottom line, pink).

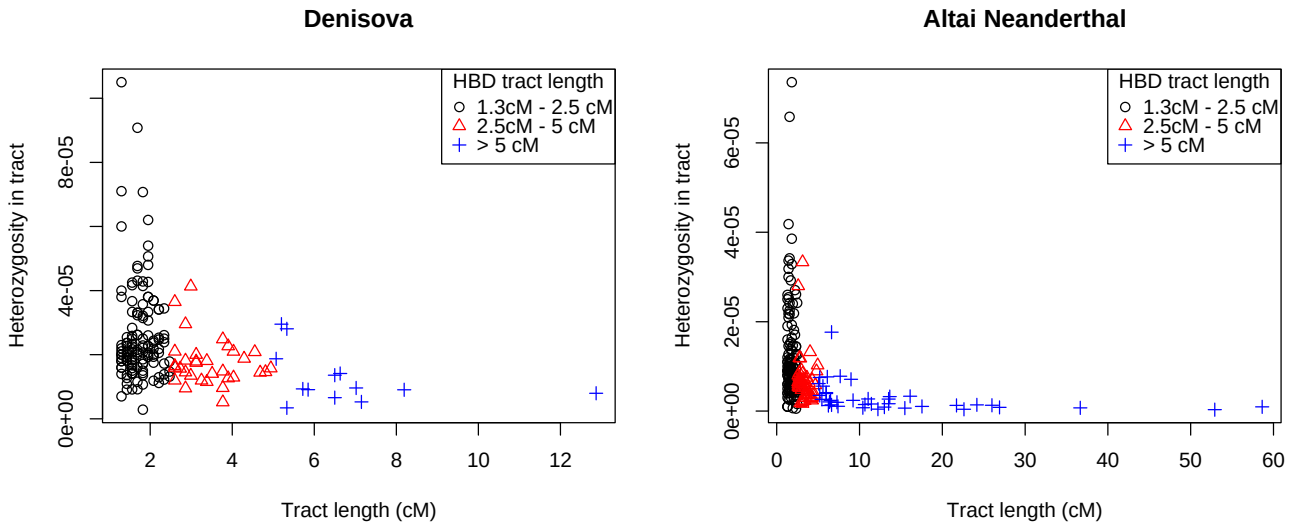


Figure S10.3: Heterozygosity in HBD tracts detected by the scan for Denisova and Altai Neanderthal as a function of the tract length.

Heterozygosity of HBD tracts

Heterozygosity in tracts shorter than 5cM is significantly higher than in tracts longer than 5cM (Figure S10.3, t-test p-value=2.4×10^{-15}). This pattern can not be explained by sequencing errors in HBD tracts, because there is no reason to assume that the sequencing error rate is higher in short tracts. However, this pattern could indicate that either there is a high false positive rate of the HBD scan for detecting short tracts, or there is background inbreeding, or both. If there is background inbreeding, older IBD tracts would have had time to accumulate recent mutations and thus have a higher heterozygosity. Note that this argument does not hold if there is only recent inbreeding with common ancestors living in the same generation, because all tracts experienced the same number of meioses.

To infer recent inbreeding ancestry we will thus focus on tracts longer than 10 cM to eliminate both false positive tracts and tracts created by background inbreeding.

Simulation of inbreeding scenarios

We simulated complete sequences of the 22 autosomes for 700 individuals according to 7 scenarios of recent inbreeding. The mutation rate per bp was set to 1.3×10^{-8}. Simulated individuals are the offspring of closely related parents which we group into three categories based on the inbreeding coefficient:

Group A (inbreeding coefficient = 12.5%):

- double first cousins
- grandfather granddaughter (or grandmother grandson)
- half siblings
- uncle and niece (or aunt and nephew)

Group B (inbreeding coefficient = 25%):

- full siblings

Group C (inbreeding coefficient = 6.25%):

- half uncle and niece (or half aunt and nephew)
- first cousins

Identification of inbreeding scenario in simulated data

We extract HBD tracts from the simulated data. (All of these tracts are due to inbreeding and have well-defined coordinates since they are simulated). For a given simulated individual, we keep tracts longer than a threshold *t*, and compare the number and coverage of these tracts to what is observed in the remaining simulated individuals. The probability that the test individual corresponds to one of the 7 simulated inbreeding scenarios is given by the probability of observing values as extreme as the ones calculated for the test individual in each scenario. This was repeated for all simulated individuals to calculate the probability of correctly identifying a scenario. Figure S10.4 shows the false discovery rate as a function *t*. The false discovery rate (fdr) between the groups A, B, and C of inbreeding scenarios increases with the length threshold used, because we remove more and more information. For *t* between 10 cM and 30 cM, the fdr between groups is around 20%. The fdr within group is higher than 30% whatever the threshold used.

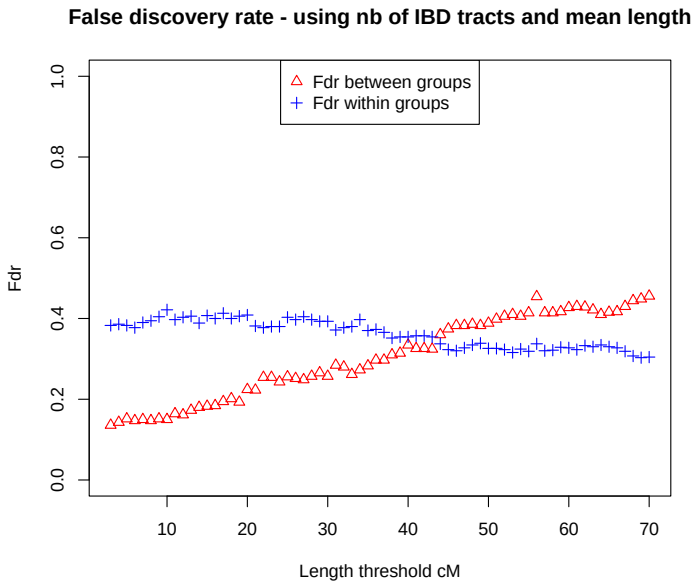


Figure S10.4: Probability of falsely identifying the inbreeding scenario for simulated individuals. The scenario is identified using the number of HBD tracts and the coverage for different length threshold. Two rates are calculated: for individuals falsely assigned to an inbred scenario in a different group (fdr between, red triangle), or in the same group (fdr within, blue plus). Group A (double first cousins, grandfather granddaughter, half siblings, uncle and niece), Group B (siblings), Group C (half uncle and niece, first cousins).

Recent inbreeding in Altai Neanderthal

Figure S10.5 shows the number of HBD tracts longer than 10 cM (top) found for the Altai Neanderthal, Denisova and simulated individuals, their fraction of the genome sequence in HBD tracts (coverage, middle panel) and the length of the longest tract (bottom).

To identify the group of inbreeding scenarios (A, B, or C) that best explains the Altai tracts, we focus on thresholds larger than 10 cM, but smaller than 30 cM, as it was found that the false discovery rate is smaller for those (Figure S10.4). For different length thresholds the probability for the Altai Neanderthal to be from the 7 different scenarios is shown in Figure S10.6. Group A is clearly the most likely, but depending on the threshold used the most likely scenario within group A changes. Thus it is impossible to distinguish between the four inbreeding scenarios in group A. According to simulations, the probability of wrongly identifying a scenario as group A using tract length thresholds between 10 and 30 CM is 22% (see Table S10.1), whereas the same probability for Group C is 36%.

We then varied the length thresholds to determine whether we can discriminate between scenarios within Group A. The false discovery rate within this group is around 68 % which is very high. The grandfather-granddaughter and half-siblings scenarios are similar since they both correspond to sharing one common ancestor, with 4 meioses occurring. However, even ignoring one of those two scenarios only decreases the false discovery rate to 56%. Figure S10.7 shows how the probability of the inbreeding scenario for the Altai

individual depends on the length threshold used, and Table S10.2 specifies the false discovery rates for each scenario.

Conclusion. The inbreeding coefficient of the Altai individual is likely to be 1/8, which implies that her parents were double first cousins, grandfather and granddaughter, grandmother and grandson, half siblings, uncle and niece, or aunt and nephew, but that we cannot distinguish among these possibilities using runs of homozygosity on the autosomes. This number provides an upper bound for the inbreeding coefficient as a smaller false positive rate or unobserved heterozygous sites (due to missing data) might decrease the total length of homozygous tracts.

Table S10.1: False discovery rates for simulated data. Probabilities are given for a scenario being misidentified as Group y although it is from Group x, using length thresholds between 10 cM and 30 cM. Group A (double first cousins, grandfather granddaughter, half siblings, uncle and niece), Group B (siblings), Group C (half uncle and niece, first cousins).

Given assignment to / Probability of truth being	Group A	Group B	Group C
Group A	0.78	0.29	0.35
Group B	0.06	0.71	0.01
Group C	0.16	0.00	0.64

Table S10.2: False discovery rates for simulated data. Probability of a scenario x given that it has been identified as y, using length thresholds between 10 cM and 60 cM. * denotes scenarios for which gender could be switched (eg. grandfather-granddaughter or grandmother-granddaughter).

Given assignment to / Probability of truth being	Double 1st cousins	Grandfather granddaughter*	Half siblings	Uncle and niece*
Double 1st cousins	0.37	0.19	0.11	0.30
Grandfather granddaughter*	0.18	0.28	0.34	0.23
Half siblings	0.17	0.30	0.35	0.21
Uncle and niece*	0.28	0.24	0.21	0.26

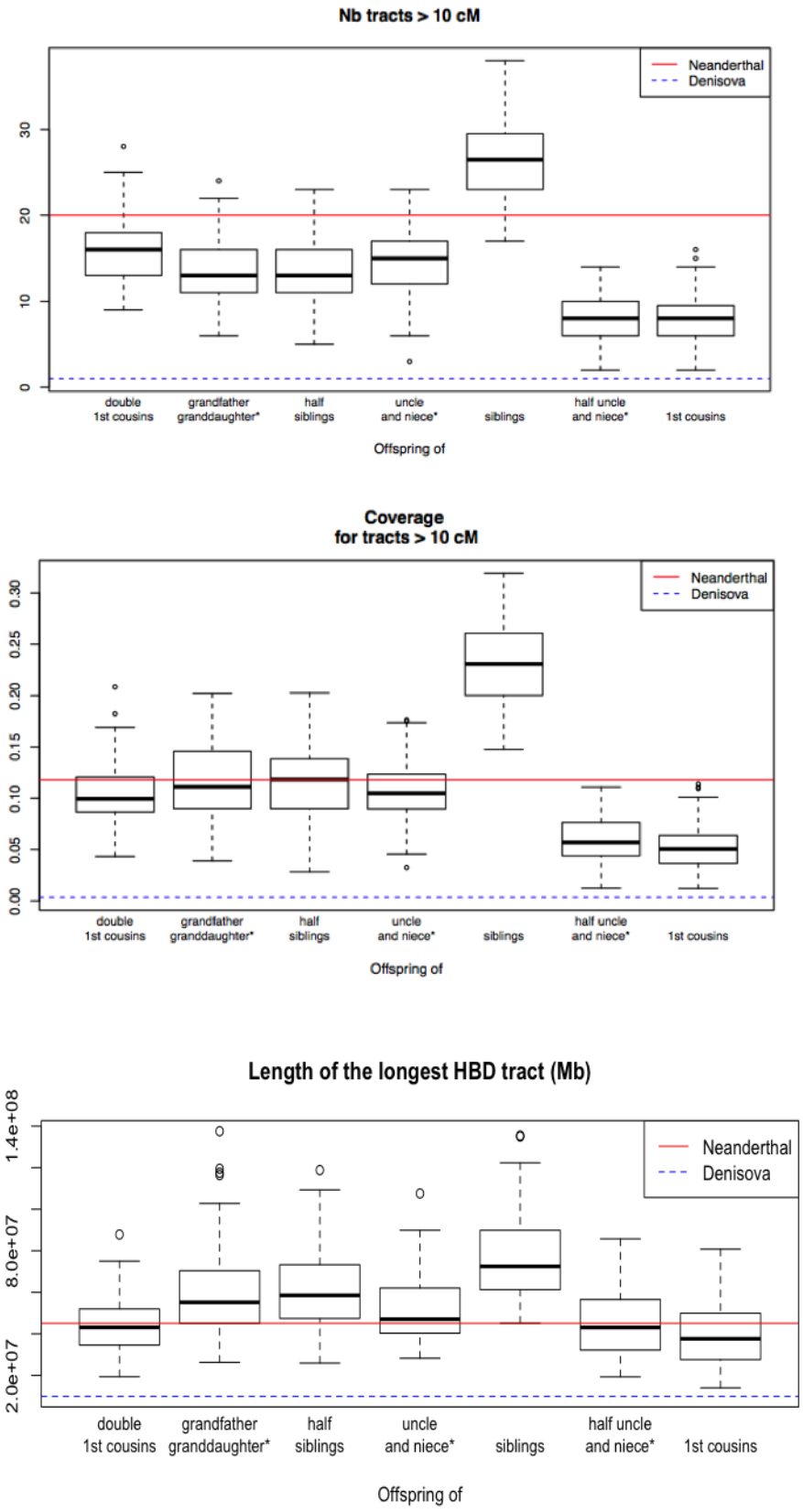


Figure S10.5: Number (top), and coverage (middle) of HBD tracts longer than 10 cM, and length of the longest HBD tract (bottom), for Neanderthal (red line), Denisova (blue dotted line), and simulations under 7 inbreeding scenarios (boxes). * denotes scenarios for which gender could be switched (eg. grandfather-granddaughter or grandmother-granddaughter).

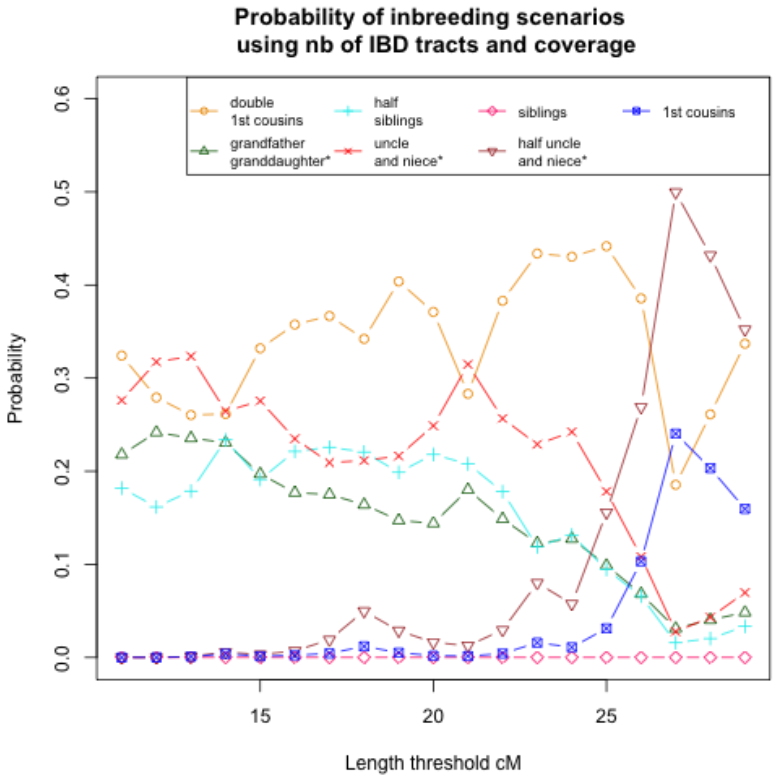


Figure S10.6: Respective probability of each inbreeding scenario given the HBD tracts observed in Altai Neanderthal, as a function of the minimum length threshold used.

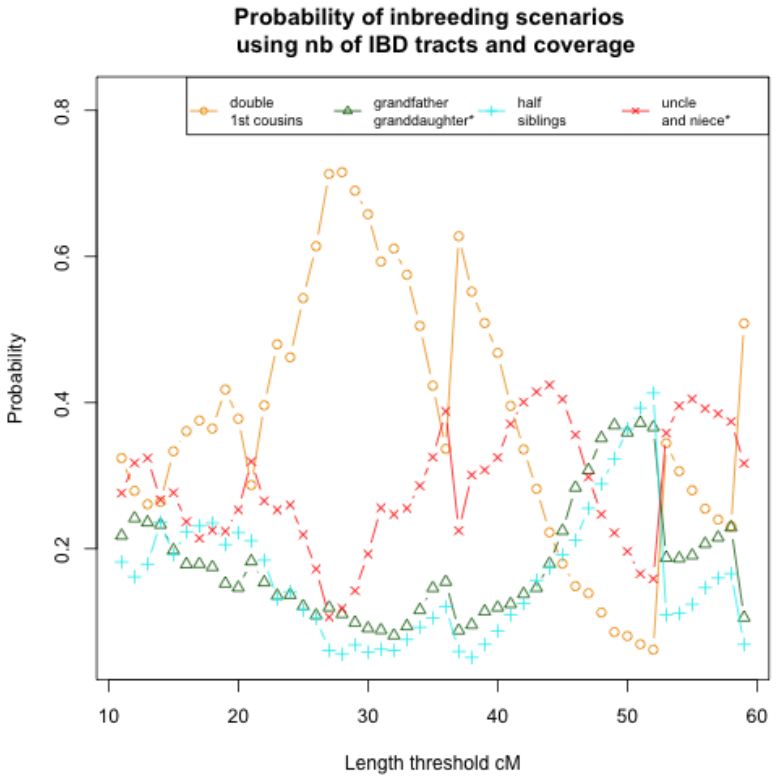


Figure S10.7: Respective probability of each inbreeding scenario in Group A given the HBD tracts observed in Altai Neanderthal, as a function of the minimum length threshold used.

Recent inbreeding in X chromosome?

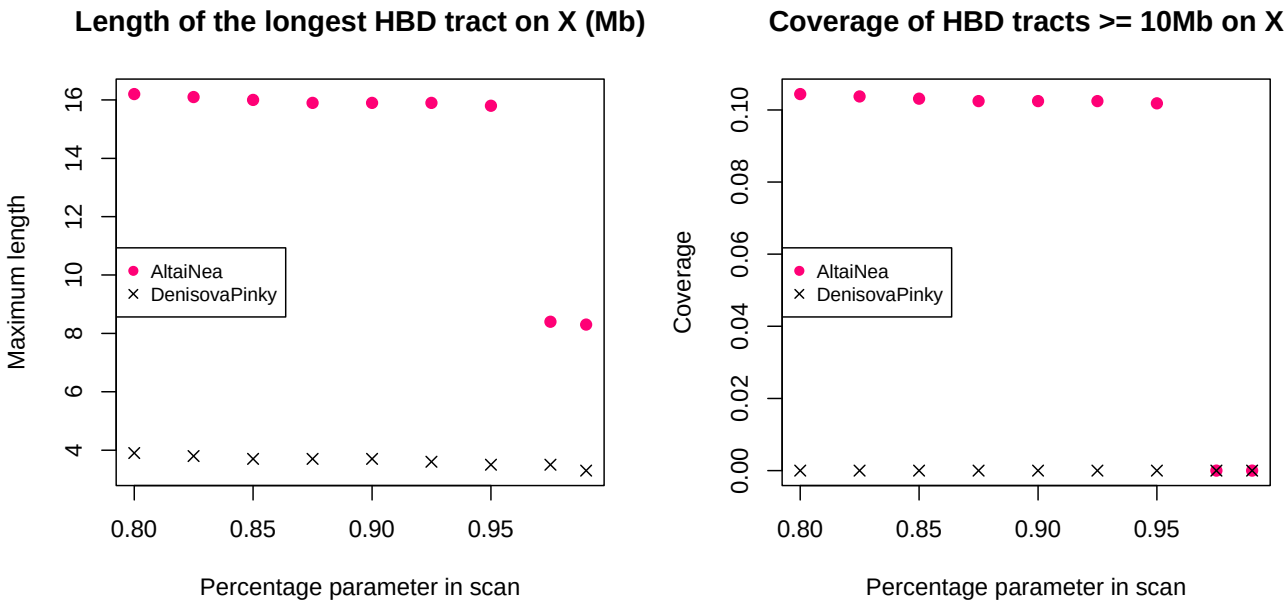


Figure S10.8: Length of the longest HBD tact (left) and coverage of HBD tracts longer than 10Mb (right) for the X chromosome. Pink circles: Altai, black cross: Denisova.

Because diversity for X chromosome is smaller than for the autosomes, we expect more false positive HBD tracts. When using scans with *p<=95%*, we detected one HBD tract ~16 Mb long. For p>=0.975 the longest HBD tract is ~8.3 Mb long and thus not necessarily due to recent inbreeding. (Figure S10.8, left) Whatever the percentage used, the HBD tracts identified were not found to be significantly clustered. However for all percentages the coverage ranges between 18% and 30% for tracts longer than 5Mb which is larger than what was found for autosomes (13.5%), and for p<=95% the coverage of tracts longer than 10Mb is ~10% which is similar to autosomes.

If we assume recent inbreeding on X, all scenarios for which the pedigree has 2 successive males can be excluded since the X sequences of the common ancestor(s) would be lost in such cases and the inbreeding coefficient for X would be 0. Therefore, we can exclude four scenarios (Figure S10.9A-D). Two pedigrees corresponding to a same scenario might have different positive inbreeding coefficients for X, depending on the number of females in the pedigree (eg. 1/8 and 3/16 for the child of an uncle and his niece). Pedigrees for which the expected inbreeding coefficient on X is identical or similar to the one on autosome belong to the following scenarios: uncle-niece, double-first-cousins, first-cousins, half uncle-niece, and half aunt-nephew. However, the variance in inbreeding coverage for two chromosomes under the same scenario is large, and it is thus difficult to pinpoint a pedigree by using only the X chromosome.

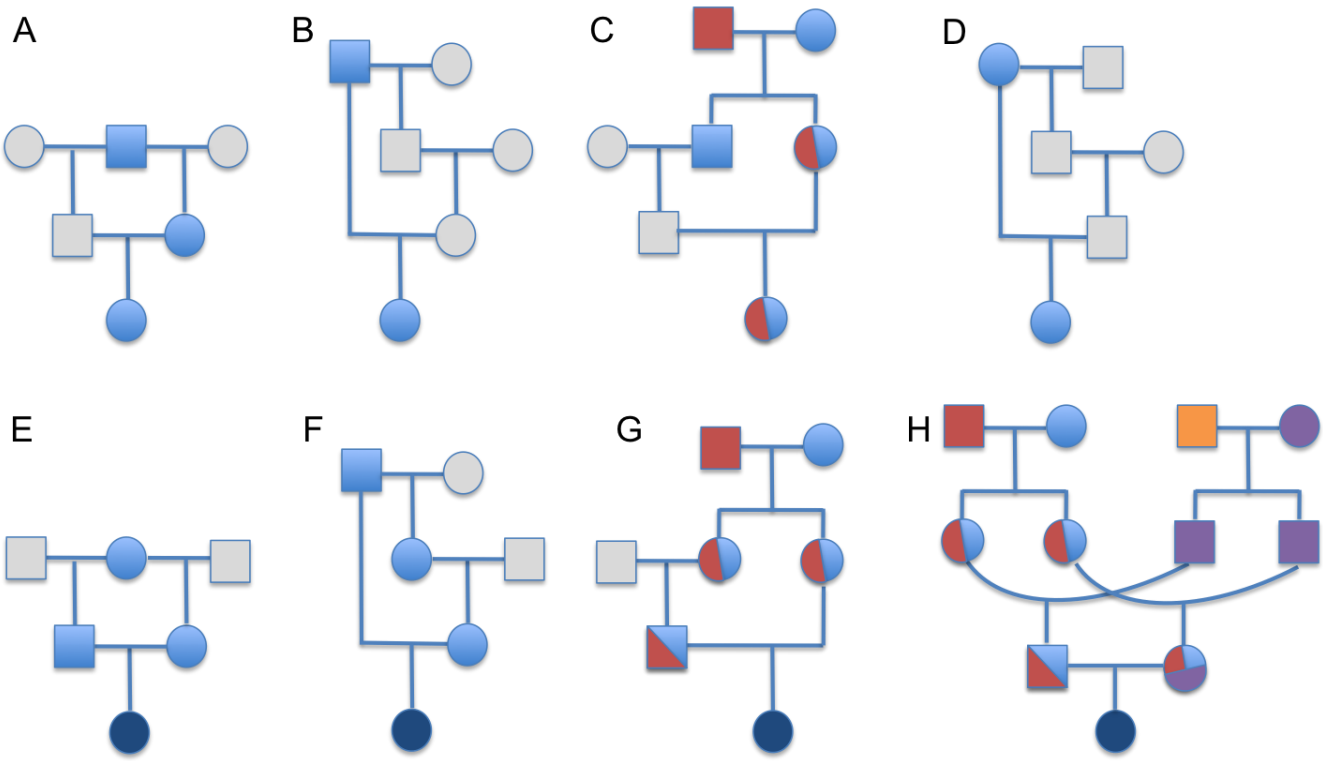


Figure S10.9: Non-exhaustive illustration of pedigrees that can be excluded (top, A-D) or not excluded (bottom, E-H), using X chromosome information. Gray denotes the absence of X sequence coming from the recent common ancestor(s). Other colors denote the potential presence of X sequence coming from the common ancestor(s). Dark blue indicates that both parents might carry X chunks inherited from the same recent common ancestor, thus the individual might be inbred for X. The pedigrees depict cases of the following scenarios: offspring of half-siblings (A,E), grandfather-granddaughter (B, F), aunt-nephew (C,G), grandmother-grandson (D), double-first-cousins (H)

Background inbreeding

Background inbreeding is the additional identity be descent created by common ancestors in the more distant past. For example, in Figure S10.9 E, there would be background inbreeding if the two males who mated with the female were themselves closely related or if the female were somewhat inbred. Background inbreeding will create tracts of IBD but they will be shorter than tracts created by recent inbreeding. We define background coverage to be the excess of coverage of HBD tracts that cannot be explained by recent inbreeding. To estimate the background coverage, we use tracts longer than 2.5 cM (to reduce false positives) and shorter than 10 cM, which is the lower limit of tract length we used to infer recent inbreeding, and calculate the total coverage minus the coverage found in simulated data for each inbreeding scenario.

Figure S10.10 shows the background coverage for Neanderthal under the 4 different Group A inbreeding scenarios. Values range from 4.9 % to 8.0% with a mean 6.9 %. Note that this is an upper bound because false positive HBD tracts would artificially increase the coverage. Assuming any of these inbreeding scenarios, the background coverage is significantly larger in Altai than in Denisova (p-value < 2.2e-16). All individuals from the B-panel have a smaller background coverage than Denisova. Among present-day populations, Native American populations (Karitiana and Mixe) and the Papuan population were found to be

the most inbred, which was already known[1,2]. Note that several individuals also have HBD tracts longer than 10 cM that are not taken into account in the computation of the background coverage.

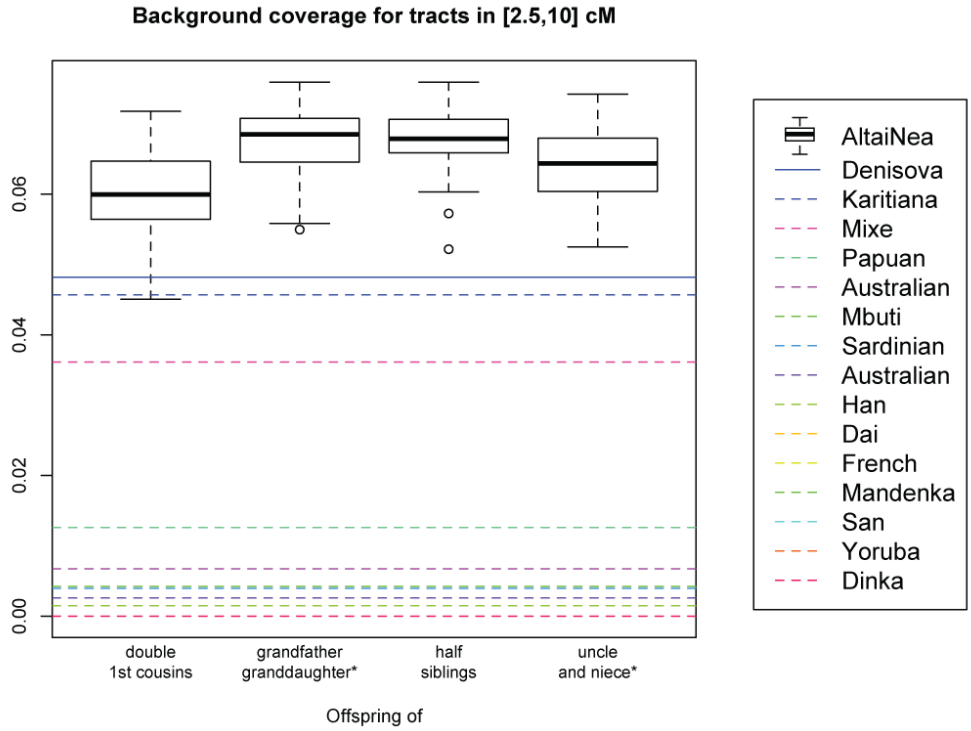


Figure S10.10: Background coverage for tracts between 2.5 and 10 cM. Boxes show the remaining Neandertal coverage by HBD tracts after subtracting the expected proportion of HBD tracts under the four different scenarios given on the x-axis. Lines denote the background coverage for Denisova and the B-panel individuals. The legend gives the individual populations ranked by their coverage value (highest value for the Altai Neandertal, lowest for the Dinka individual). We identified 36 tracts longer than 2.5 and smaller than 10cM for the Karitiana individual, 44 for Denisova, 69 for Neandertal whereas the maximum number of tracts due to recent inbreeding was 14 in the simulated data.

Bottleneck scenarios

Because the heterozygosity is overall quite low in Altai, we investigate the hypothesis of one or successive bottlenecks as an alternative for background inbreeding (or an explanation for why inbreeding occurred when the bottleneck produce extremely small population sizes). Using MS we simulated sequences under 3 types of scenarios:

(A) Ten successive bottlenecks starting *t* generations ago and uniformly spaced in time, with an initial population size of 15000, and a population size of 3000 at time of sampling (Figure S10.11A). This mimics a smooth decrease of the population size. *t* varies between 1,000 and 200,000 generations.

(B) Ten successive bottlenecks starting 12000 generations ago and uniformly spaced in time, with an initial population size of 8000, and a population size that varies at time of sampling (Figure S10.11B). This mimics a smooth decrease of the population size starting right after the split from

modern humans. Population size at sampling varies between 100 and 2000, results are shown for intermediate values.

(C) Only one very strong and recent bottleneck starting between 5 and 50 generations before sampling; the size after bottleneck varies from 0.25% to 10% of the initial size (Figure S10.11C shows results for 1%).

(D) One very strong and recent bottleneck, starting between 5 and 50 generations before sampling, that lasts for only 10 or 20 generations (Figure S10.11D).

For each set of parameters we simulated 1000 independent diploid tracts of length 5 cM (assuming the 2 randomly sampled haploid tracts belong to individuals that mate). We compared the distribution of the heterozygosity for simulated tracts to the distribution for tracts randomly chosen in Altai. We removed inbred tracts that can be explained by one of the recent inbreeding scenarios. Figure S10.11 shows that one or several bottlenecks can lead to some regions having a heterozygosity as low as the one found in the Altai. For Models D this requires extremely drastic bottlenecks (population size reduced to 0.2% of initial size during 10 generations) which are too extreme to be likely. Moreover, none of the simulated distributions for scenarios A,C, and D(gray) match the observed distribution (pink). Scenarios of type B provide a better fit to the data (eg. successive bottlenecks reducing the population size from 8000 individuals 12000 generations ago to ~600 individuals at time of sampling). However, they are still unable to fit both the lower and the upper tails at the same time. The upper tail could potentially be explained by some gene flow from another population, as this could create a longer tail by increasing heterozygosity in some part of the genome.

All simulated bottleneck scenarios failed to explain the whole pattern of heterozygosity observed in Altai. We additionally investigated scenarios with one very strong bottleneck after the split from modern humans, and scenarios roughly mimicking the demography inferred by PSMC (only 4 different phases). They did not provide a good fit to the data. Scenarios with more complex changes in population size were not investigated.

Conclusion: The observed background coverage of HBD tracts could be explained by the presence of background inbreeding in the population. Alternatively, a demographic scenario of random mating with successive bottlenecks starting after the split from modern humans that induce a very small population size at time of sampling (~600 individuals) also provides a reasonable fit to the data. Note that when a population is very small for a long time the chance of mating between distant cousins is not negligible even in case of random mating.

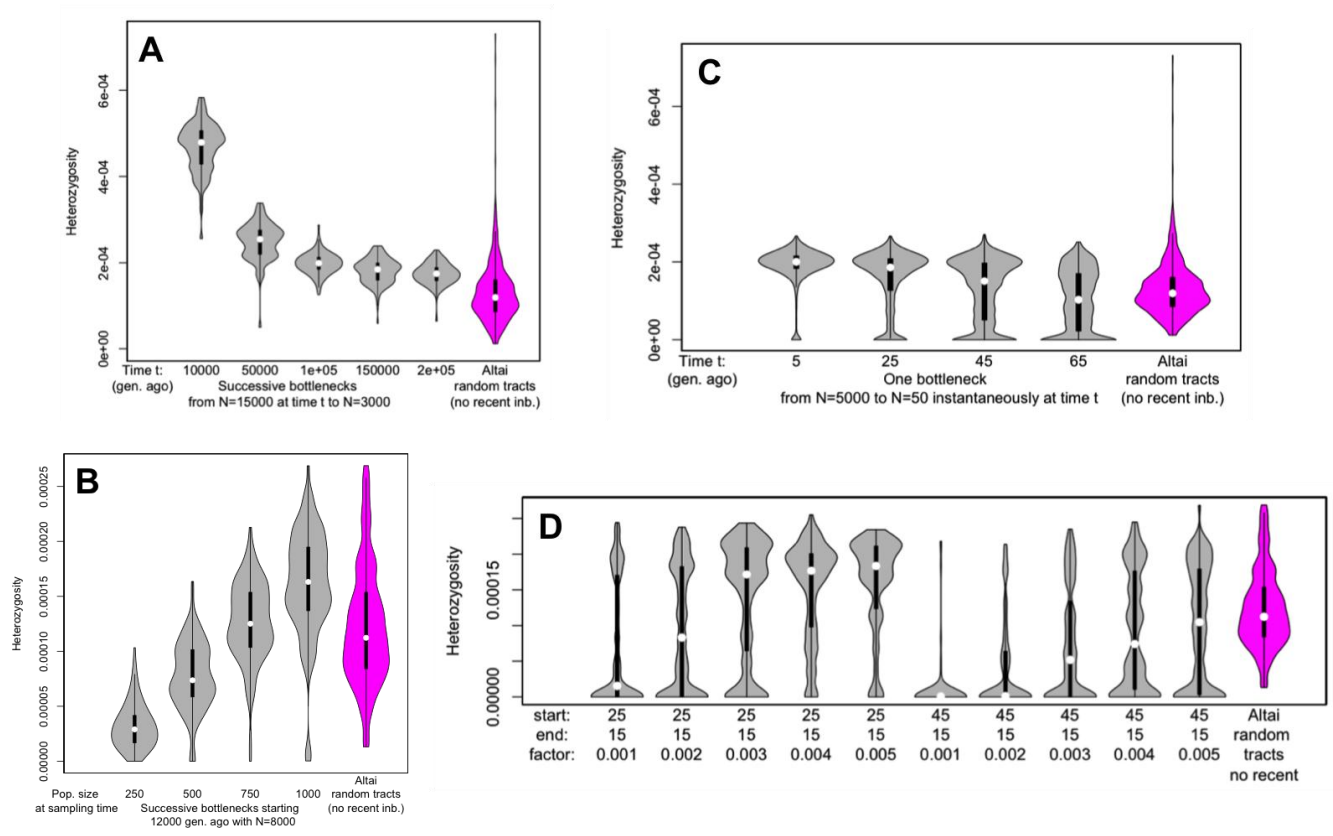


Figure S10.11: Magenta: Distribution of heterozygosity from 1000 tracts 5 cM long randomly chosen from the Altai sequence (after removing recently inbred tracts). Gray: Distribution of heterozygosity from 1000 tracts simulated under different bottleneck scenarios. A: 10 successive bottlenecks starting *t* generations ago. B: 10 successive bottlenecks starting 12k generations ago; population size at time of sampling varies. C: One recent bottleneck. *t* generations before sampling the population size is reduced to 1/100th. D: One recent and short bottleneck; *Start* and *End* denote the starting and ending times of the bottleneck in generations, *factor* denotes the reduction percentage (ie for *factor* = 0.005 the population size after bottleneck is 0.005 * 5000 = 25).

References

1 Kirin, M. *et al.* Genomic runs of homozygosity record population history and consanguinity. *PloS one* **5**, e13996, doi:10.1371/journal.pone.0013996 (2010).

2 Pemberton, T. J. *et al.* Genomic patterns of homozygosity in worldwide human populations. *American journal of human genetics* **91**, 275-292, doi:10.1016/j.ajhg.2012.06.014 (2012).

Supplemental Online Material 11
Comparison of X- and Autosome diversity

Ines Hellmann* and Kay Prüfer
To whom correspondence should be addressed (ines.hellmann@bio.lmu.de)

The ratio of X vs. Autosome diversity is influenced by many factors: sex-biased mutation, recent demographic events and differences in reproductive variance between the sexes. The last effect can reflect social structure, *i.e.* whether the population is polyandrous or polygynous, or matri- or patrilocal. In practice, these effects are hard to tease apart, but we believe that comparisons between closely related populations, such as Denisovans and Neandertals, can still provide insight.

Estimating diversity
We used the alignments described in SI 2a (Altai Neandertal) and Meyer et al. 2012 (Denisova). We restricted analyses to bases with coverage ranging from 15X to 85X for the Altai Neandertal and from 12X to 50X for the Denisovan individual. These cutoffs were chosen by looking at the coverage distributions to make them approximately normal.
Furthermore, we used the 'manifesto' filters (SI 5b) so as to look only at sites with high mapability and genotype quality. We also used annotations from the human genome (hg19) downloaded from the UCSC Browser (http://genome.ucsc.edu/) to exclude gaps in the hg19 build as well as exons (knownGenes), repeats (rmsk), genomic duplications (genomicSuperDups) and first introns. To remove the pseudoautosomal regions we used the locations provided by the HapMap recombination Map (http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/) (chrX:150,118-2,695,340 and chrX:154,969,038-155,235,078). Furthermore, we excluded runs of homozygosity in the Altai individual of >2Mb that are likely due to recent inbreeding (SI 10). Note that the putative Altai inbreeding tracts were also removed from the Denisovan genome, but not from Bonobo[1] or present-day humans[2] in Table S11.1.
Next, we used the program mlRho[3] (version 1.5) to obtain estimates of nucleotide diversity within our data.

Mutation rate correction
In order to correct for male mutation bias, we used the divergence on the human lineage since the Bonobo (panpan1) split using orangutan[4] (ponabe2) as the outgroup. Tri-allelic sites (where all 3 species differ) were excluded. Furthermore, only sites that were used for diversity calculations in the archaic human genomes were used to estimate divergence for the mutation rate correction.

X-Autosome effective population sizes
The Altai Neandertal has the lowest Nx/Na ratio compared to Bonobo, Europeans (CEPH, NA12878), African (Yoruba, NA19240) and Denisovan. At least in the 5 populations examined here, there seems to be a trend of low Nx/Na ratio associated with low overall Ne, indicating that one of the causes of variation in Nx/Na ratios is recent changes in population sizes rather than sex-differences in reproductive variance[5].

For the modern human populations, the demography can be reasonably well approximated. We used an Out-of-Africa model to simulate the diversity within one European and one African individual using the software msms[6]. For the autosomes the parameters used were as described in Gutenkunst et al. (2009)[7], supplementary section 5.2. For the X chromosome, the diversity (theta) and recombination rate (rho) where multiplied by the inheritance factor β and the coalescent times were re-scaled by dividing by β. If the same number of females and males reproduce (i.e. Nf/Nm=1), β=¾ for the X-chromosome. We determine the best fitting value of β by simulating Nx/Na over a grid corresponding to a 100-fold difference between male and female population sizes (Nf vs. Nm) and comparing those values to the mutation-rate-corrected observed values of Nx/Na (Figure S11.2).

For the Africans, demography does not distort the Nx/Na ratio, as can be seen by the perfect correspondence of β and Nx/Na in Figure 2A. In contrast, the complex demography of Europeans (the out-of-Africa bottleneck followed by the agricultural expansion) leads to a strong distortion of the Nx/Na ratios. While the observed value of 0.78 implies a smaller Nf/Nm compared to Africans, the corrected value of 0.91 is the highest among the analyzed populations. We note that we assumed for these estimates that the β value corresponding to Africans has little impact on the estimates of β for the Europeans.

The demography of archaic human populations is known with less certainty. Some aspects are described in Supplementary Sections SI 17 and SI 18, amongst which the one with the biggest potential to impact Nx/Na is an abrupt 10x decline in population size approximately 8000 generations ago, which corresponds to 2 in coalescent time if N_0=1000 (msms 2 1000 -t 60 -r 60 -eN 2*β 10). Assuming that this demography is true for both the Altai Neandertal and the Denisovan individual, β is 0.81 and 0.83, respectively.

Because of the uncertainties in the demography of archaic humans, we also took an analytic approach[5] to estimate the expected β over a range of fold population size changes and times (Figure S11.3). From this we conclude that the most realistic parameters lead to Nx/Na being smaller than 0.75 and hence would cause an underestimate of the ratio of female to male reproductive variance (Nf/Nm).
In summary, we cannot detect any significant differences in Nf/Nm in bonobos, archaic and modern human populations. If anything, breeding success in females appears to have a lower variance than in males, as indicated by Nf/Nm >1.

Table S11.1: Composite Maximum Likelihood estimates of Ne and Nx/Na. The ancestral Ne of bonobos and humans was assumed to be 45,000, the split time 4.5Mya., a generation time of 20 years. Average human background selection values of a window had to be >0.9. β is the demography corrected Nx/Na, which can be converted into the ratio of female to male effective population sizes (Nf/Nm).

	Ne	Nx/Na	β	Nf/Nm	Liklihood x 10^{-3}	# of A-windows	# of X-windows
Altai	2800	0.74 (0.59-0.99)	0.81 (0.69-1.0)	1.54 (0.58-7.37)	-24	2997	85
Denisova	2600	0.77 (0.62-0.97)	0.83 (0.71-0.99)	1.83 (0.73-6.28)	-24	2965	84
Bonobo	12000	0.87 (0.76-1.02)	same as Nx/Na*	2.45 (1.05-8.32)	-20	2918	177
Yoruba	11000	0.85 (0.73-0.98)	0.84 (0.72-0.97)	1.97 (0.77-5.5)	-20	3971	179
CEPH	9000	0.78 (0.68-0.91)	0.91 (0.88-0.96)	3.31 (2.53-4.83)	-19	2812	181

*assuming no recent population size changes

Figure S11.1: Likelihood surface for Ne and Nx/Na ratio for the Altai Neandertal and the Denisovan. This Likelihood surface was calculated using the Method as described in Hammer et al. (2009)[8] whereas diversity was estimated using mlRho and mutation rates from the substitutions on the lineage from human to the most recent common ancestor of humans and Bonobos, using orangutan as an outgroup.

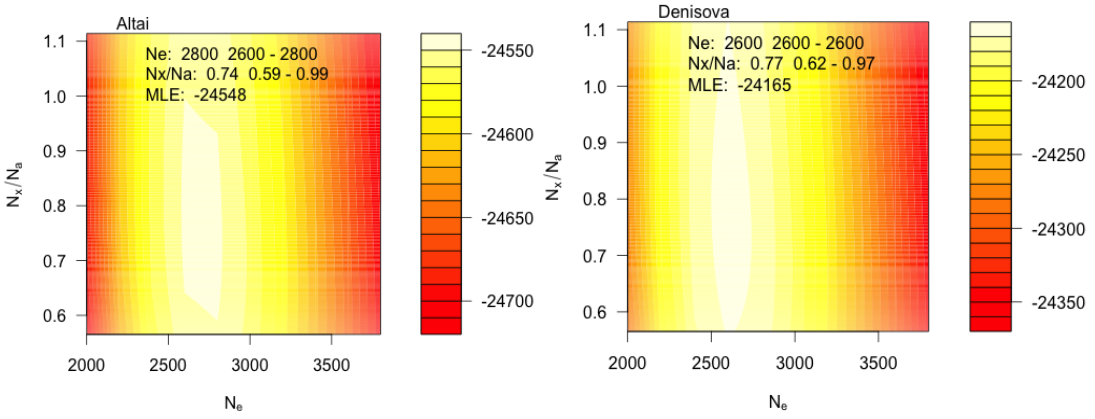


Figure S11.2: In panels A & B the dots represent simulations of Nx/Na for the Out of Africa Model from Gutenkunst et al (2009) for varying inheritance factors (β). The green lines correspond to the estimates from one Yoruba (A) and one CEPH (B) individual including the confidence intervals. In panel (C) a population decline (see text) is simulated, from which the β-estimates for Altai (green) and Denisovan (blue) are derived, that are also stated in the legend.

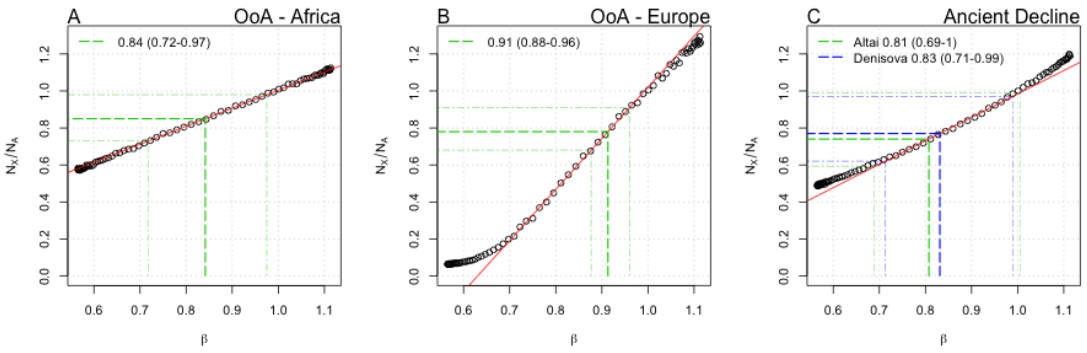
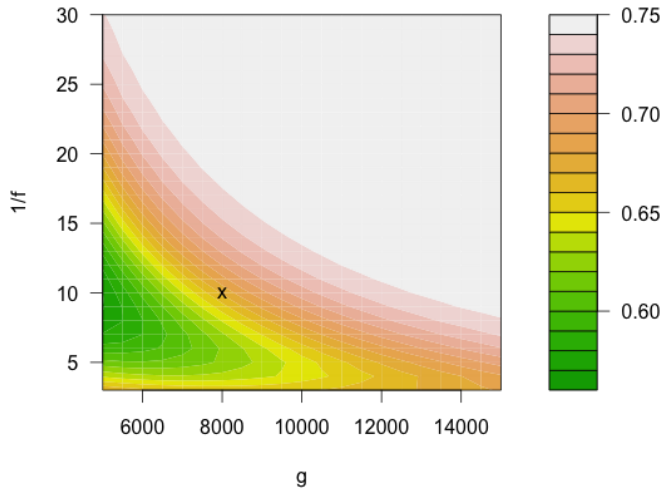


Figure S11.3: We use equation (4) from Pool & Nielsen (2007) to estimate the Nx/Na ratio (colors) if the population changed by 1/f g generations ago and N_0=1,000. The x marks the parameter combination that was used for the simulations in Figure S11.2 C.



References

1 Prufer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**, 527-531, doi:10.1038/nature11128 (2012).

2 Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

3 Haubold, B., Pfaffelhuber, P. & Lynch, M. mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Molecular ecology* **19 Suppl 1**, 277-284, doi:10.1111/j.1365-294X.2009.04482.x (2010).

4 Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529-533, doi:10.1038/nature09687 (2011).

5 Pool, J. E. & Nielsen, R. Population size changes reshape genomic patterns of diversity. *Evolution; international journal of organic evolution* **61**, 3001-3006, doi:10.1111/j.1558-5646.2007.00238.x (2007).

6 Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**, 2064-2065, doi:10.1093/bioinformatics/btq322 (2010).

7 Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics* **5**, e1000695, doi:10.1371/journal.pgen.1000695 (2009).

8 Hammer, M. F. *et al.* The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nature genetics* **42**, 830-831, doi:10.1038/ng.651 (2010).

Supplementary Information 12

Population size changes and split times

Heng Li, Swapan Mallick and David Reich*

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

(i) Findings

- The Altai Neandertal was inbred and its ancestral population size was low.
- We estimate that Neandertals and Denisovans split 381-473 kya (assuming $\mu = 0.5\times10^{-9}$/bp/year).
- We estimate that archaic and modern humans split 550-765 kya (assuming $\mu = 0.5\times10^{-9}$/bp/year).
- We estimate that the Neandertal/Denisovan split time was 58-86% of that from modern humans.

(ii) Reconstruction of Neandertal population history

We used the Pairwise Sequential Markovian Coalescent (PSMC)[1] method to infer changes in the Altai Neandertal ancestral population size over time. We first described this approach in SOM 14 of ref. 2.

Briefly, we used the mappings of the Altai Neandertal reads to the human reference genome sequence *hg19*, and called SNPs using SAMtools, which is the SNP calling algorithm for which the PSMC software was optimized[3,4]. We then used the PSMC to infer the distribution of coalescent times between the two copies of the genome each individual carries (one from their mother and one from their father) across all of chromosomes 1-22. The reciprocal of the coalescent probability at a given time depth can be interpreted as the effective population size at that time, assuming that the ancestral population was unstructured. An inferred large population size can also reflect ancient substructure; that is, descent from multiple ancestral populations that only mixed at a later time.

The SAMtools estimate of heterozygosity for the Altai Neandertal (without filtering out the tracts of inbreeding discussed below and in SI 11), is even lower than the estimate for the Denisova finger bone, consistent with the findings from two other methods for estimating heterozygosity (SI 9).

Table S12.1: Relative estimates of heterozygosity from three methods

Altai/other ratio	*GATK calls*	*mlrho*	*SAMtools*
Altai/Denisova	82%	89%	95%
Altai/Karitiana	31%	30%	34%
Altai/San	16%	16%	19%

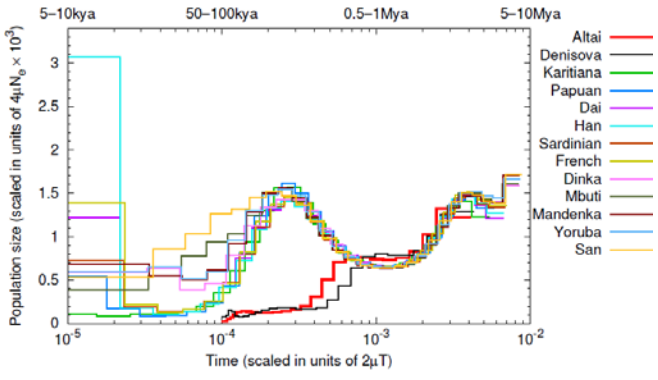
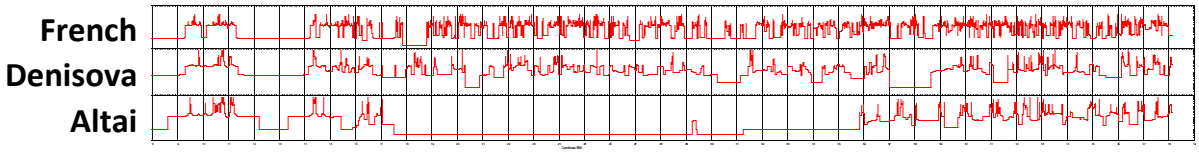


Figure S12.1: Population size changes over time inferred from the PSMC. *We estimate that Neandertals and Denisovans had persistently small ancestral population sizes in the past 150-300 kya, and population sizes similar to modern humans 400-800 kya suggesting a common origin at that time. We compare these archaic samples to 11 deep sequences from present-day humans.*

Figure 6 and Figure S12.1 shows the PSMC output on the Altai Neandertal data, compared with Denisova and 11 diverse present-day human genome sequences[2]. Both Altai and Denisova are inferred to have had small inferred population sizes for the last 150,000-300,000 years of their histories. The population size estimates are inferred to become similar to present-day humans at time depths of >400,000-800,000 years ago. These findings are consistent with Altai and Denisova and present-day humans descending from common ancestral populations at these time depths.

We also examined the posterior decoding of the PSMC, which shows the inferred distribution of the time since the most recent common ancestor (TMRCA) at each position of the genome. Figure S12.2 shows a plot across chromosome 21, with the first row corresponding to the French "Panel A" modern human, the second round corresponding to the Denisova finger bone, and the third to the Altai Neandertal. The plot shows a large region of recent coalescence in the Altai Neandertal spanning 19 Mb on this chromosome (17-36 Mb in *hg19* coordinates), which likely reflects the fact that the individual's mother and father had a shared ancestor in the last few generations. Similar patterns are observed on other chromosomes, with the longest stretch of homozygosity seen on chromosome 14 (>40 Mb). Overall, about 30% of the Altai Neandertal genome is inferred to have a TMRCA in the most recent bin. In SI 11, we show that the pattern of these homozygous chunks in the Altai Neandertal is consistent with an inbreeding coefficient of about what would be expected from a mating of half siblings. In contrast, the Denisovan individual does not look particularly inbred, although she does have ~1-5 Mb chunks of homozygosity in some places in her genome.

Figure S12.2: PSMC inference of the time since the most recent common ancestor of the two chromosomes of a single individual for French (top), Denisova (middle) and Altai (bottom). *Results are on chromosome 21 and are shown on a vertical log scale. The Altai Neandertal is homozygous from 19-36 Mb, showing that she was inbred: her parents were close relatives.*



(iii) Population split dates

It is important to know when the Altai and Denisova ancestral populations separated from each other, and when their common ancestral population separated from the lineage leading to modern humans. We obtained population split date estimates using three complementary methods.

Split date estimate #1: Extension of the PSMC

We extended the PSMC to estimate population split dates. The PSMC in its original formulation works by estimating the distribution of the time since the most recent common genetic ancestor of the two haploid genomes carried by a single individual. If two effectively haploid genomes can be extracted from two different populations, the PSMC statistical machinery makes it easy to put them together—pretending that they are from the same individual—and to infer the distribution of their time since the common ancestor. For this purpose we use a modified version of the PSMC model that introduces an extra parameter (a sudden split time). The method makes the simplifying assumption that the population separation was total so that no coalescent events occurred more recently than the split. As in Figure S12.1, the inferred split time is scaled in units of $2\mu T$, the pairwise sequence divergence between two sequences diverged T generations ago. Simulations have shown that this procedure gives meaningful population divergence time estimates for a range of models of history [5].

To obtain phased haplotypes from present-day humans, we used the experimentally phased haploid genomes of sub-Saharan Africans (San and Mandenka), generated as described in SI 4. For Altai and Denisova, we cannot obtain experimentally phased genomes. However, we can take advantage of the recent history of small sizes in both populations to study those parts of these two individuals' genomes where their two chromosomes coalesced more recently than their split from other

populations. For these segments of the genome, inferences about population relationships do not depend on which allele we pick. For the archaic-modern population split date estimates, we used segments that coalesced more recently than an inferred sequenced divergence of 5×10^{-4} per base pair (bp) corresponding to 94% of the Altai genome and 97% of the Denisova genome. For the more recent Altai-Denisova divergence, we imposed a stronger filter to restrict to parts of the genome where we could be more confident that the two chromosomes of the sequenced individual coalesced more recently than the separation of the two populations: $<2\times10^{-4}$/bp sequence divergence. This filter retained 89% of the Altai genome and 75% of the Denisova genome.

Table S12.2: Population split times inferred from the PSMC

Population split	Gb used	Diver-gence / bp × 10^{-4} *	Date as % of human-chimp (uncorrected for branch shortening)	Date correcting for branch shortening†		
				% human-chimp†	Kya assume Div_{HC}=6,500	Kya assume Div_{HC}=13,000
San-Mandenka	2.01	0.86	0.66%	0.66%	43	86
Altai-Denisova	1.62	2.62	2.02%	2.93%	190	381
Altai-San	1.90	4.94	3.80%	4.31%	280	560
Altai-Mandenka	1.90	4.87	3.75%	4.26%	277	553
Denisova-San	1.94	5.36	4.12%	4.53%	294	589
Denisova-Mandenka	1.94	5.26	4.05%	4.45%	289	579

* The uncorrected date range is based on assuming that human-chimp genetic divergence is 6.5-13 Mya, and multiplying by the divergence/bp divided by an assumed human-chimp divergence of 0.0130 in this subset of the genome (this yields exactly the same date range as assuming a mutation rate of $0.5\text{-}1.0\times10^{-9}$/bp/year). We note that the empirical human-chimp divergence in the subset of the genome we use for the PSMC analysis is 0.0130 for all the analyses in this table except for the Altai-Denisova comparison where it is 0.0135 (which is so similar to 0.0130 that we ignore the difference).

† The adjusted date range is based on correcting for the fact that the samples are ancient, with the degree of branch shortening measured in SI 6b. The branch shortening only affects one side of the tree, so for Altai the correction is 0.51% = 1.02% / 2 of human-chimp divergence and for Denisova it is 0.405% = 0.81% / 2 of human-chimp divergence.

The resulting population split time estimates are presented in Table S12.2. It is important to recognize that our split time estimates correspond to the average branch lengths between the two sequences. These are underestimates if one or both of the samples are of ancient origin. To correct for this branch shortening, we use the estimates of 1.02% of human-chimpanzee divergence for Altai and 0.81% of human-chimpanzee divergence for Denisova reported in SI 6b, and then divide by two to reflect the fact that each branch shortening only affects one half of the tree (for the Altai-Denisova comparison branch shortening affects both sides of the tree so the correction is 0.915% = 0.51% + 0.405%).

We highlight 3 findings:

(1) Altai and Denisova are estimated to have similar split times as Africans, consistent with our previous report that they are an approximate clade[6] ($4.87\text{-}4.94\times10^{-4}$/bp for Altai compared to $5.26\text{-}5.36\times10^{-4}$/bp for Denisova). Taking the union over the ranges and correcting for branch shortening, this translates to 277-294 kya assuming a mutation rate of 1.0×10^{-9}/bp/year and 553-589 kya assuming a mutation rate of 0.5×10^{-9}/bp/year. The slightly older divergence estimate of Africans to Denisova than to Altai may reflect two known ways in which our sudden population split time assumption is incorrect. First, we know that Denisova harbors a small proportion of ancestry from a highly diverged archaic population (SI 16a), and this could make the Denisova estimates a bit older. Second, we know that some African samples harbor a very small amount of Neandertal ancestry due to gene flow from West Eurasians in the last ten thousand years (SI 13). However, the fact that the split time estimates are similar whether we use San and Mandenka to represent Africans, or Altai or Denisova to represent archaic individuals, suggests that these violations of the sudden split time assumption may not be causing large biases.

(2) The estimated split of Altai and Denisova is 2.62×10^{-4}, corresponding to 190 kya assuming a mutation rate of 1.0×10^{-9}/bp/year and 381 kya assuming a mutation rate of 0.5×10^{-9}/bp/year (after correcting for branch shortening). This corresponds to 65-69% of the archaic/modern split date.

(3) The estimated population split time for San and Mandenka is 15-16% of the archaic/modern split date. This is 43 kya assuming a mutation rate of 1.0×10^{-9}/bp/year and 86 kya assuming a mutation rate of 0.5×10^{-9}/bp/year. This is substantially more recent than has been inferred based on other studies that also compared San to West African populations; for example ~130,000 years in refs. 7 and 8. We caution that the PSMC has poor resolution for inferring recent demographic because its inference of recent population size is poor, while its inference for more ancient events is better[1]. Thus, it is possible that our recent dates are biased while our more ancient dates are more accurate. Another reason why the split date estimate here may be biased low is that the San are known to have a minimum of 4% ancestry from populations that are more closely related to Mandenka (West African, East African and West Eurasian) than the majority ancestry in the San[9]. Some of the genomic segments in the San that are inherited from these ancestral populations are thus likely to be coalescing with genomic segments in Mandenka more recently than the main San / West African population separation, which could cause our method to infer a too-recent split.

Split date estimate #2: Probability of being derived in population *A* at a heterozygous site in pop. *B*

In the papers describing the Neandertal draft genome[7] and the Denisova high-coverage genome[2], we estimated population split times based on discovering single nucleotide polymorphisms (SNPs) as heterozygous in a single individual from population *B*, and then measuring the rate *F(A|B)* at which the derived allele not seen in apes is seen in a randomly chosen genome from another population *A*.

If a polymorphic site is identified as being heterozygous in an individual from population *B*, the probability *F(A|B)* of the derived allele being observed in a random chromosome from another population *A* is a function of the split time. For a short time divergence the probability is at its maximum: close to 1/3 with the exact value depending on the demographic history. For a more ancient population split, *F(A|B)* decreases, reflecting the fact that some mutations may have arisen since the split of *A* and *B*. We carried out computer simulations using the *ms* software[10] to infer the monotonically decreasing function that relates the derived allele probability *F(A|B)* to time (Figure S12.3). For this purpose, we simulated exactly the models of population size change over time inferred by the PSMC (Figure 6 and Figure S12.1).

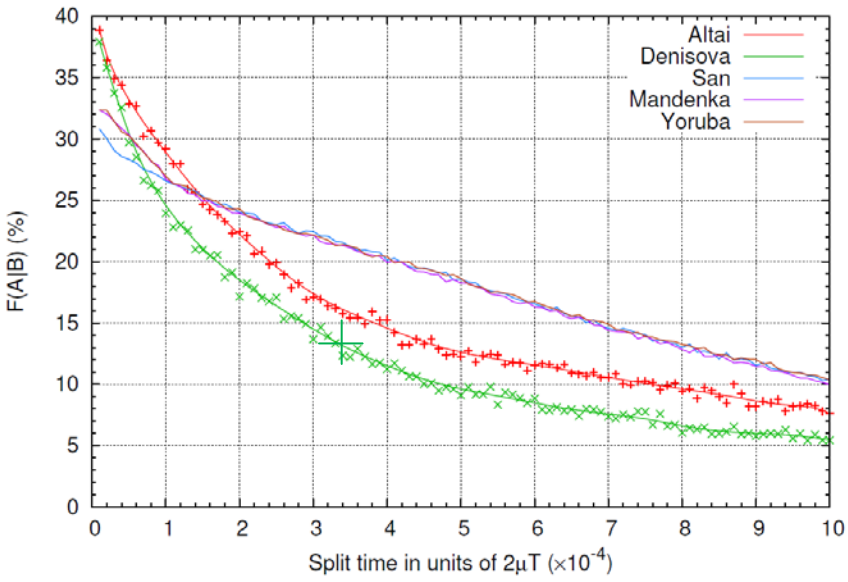


Figure S12.3: Calibration curves for translating F(A|B) to split time. *F(A|B) is obtained from simulations of population B's history, using the PSMC model fit. F(A|B) is the probability that at a heterozygous site in an individual from population B, the derived allele will be observed at a randomly sampled chromosome from population A. The green crosshairs marks F(Altai|Denisova)=13.0%, which translates to an inferred sequence divergence of d(Altai|Denisova)=0.00034 for genomic segments that coalesce at the population split time depth. After dividing by human-chimpanzee divergence of 0.013, adding 0.0081 to correct for branch shortening on the Denisova lineage, and multiplying by 6.5-13 Mya for human-chimpanzee divergence, we obtain the divergence time estimate of 223-445 kya shown in Table S12.3.*

We infer the inferred split time as *d(A|B)* in units of per-base-pair sequence divergence expected for two sequences separated at that time depth. This allows us to convert to absolute time units for a given mutation rate per year. We can convert this quantity to time using the equation $T(A|B)=d(A|B)/2\mu$.

An important feature of this method is that the history of population *A* since the split has no impact on the validity of the split time inference, even if population *A*'s history is complicated involving population size changes or substructure. The reason is that we are randomly sampling only a single chromosome from population *A*. This must trace its ancestry back to the split from population *B* without coalescing with either of the two *B* chromosomes we used to ascertain the SNP, and thus the probability of it carrying the derived allele is the same as the probability in its ancestor (after the *A*-*B* population split), which was of unaffected by population *A*'s history. In two previous papers[2,7], we chose *B* to be Yoruba Nigerians, and found that multiple models for Yoruba history gave calibration curves similar to the PSMC. Thus, it seemed reasonable to only fit PSMC-based reconstructions here.

A second advantage of this date estimation procedure—new to the analysis in this paper compared with the similar analyses we reported previously[2,6]—is that we repeat the analysis two times for each pair of populations: once in a way that is only affected by the demographic history of *B (F(A|B))*, and once in a way that is only affected by the demographic history of *A (F(B|A))*. If consistent results are obtained for computations on both sides of the tree, our confidence in the results increases.

A complication in measuring *F(A|B)* is that recurrent mutation can cause miscalling of the ancestral allele. We previously addressed this by restricting to transversions (where the mutation rate and thus the probability of recurrent mutation is lower), by restricting to sites where there is data for at least two primates (out of chimp, gorilla, orangutan and macaque) and the primate allelic states agree, and by making a statistical correction for undetected recurrent mutations on the human-specific linage since separation from chimpanzee[7].

Here, we follow a similar procedure. We only analyze transversion polymorphisms at sites where we have coverage from both chimpanzee and macaque. We then compute the statistic in two ways: once using chimpanzee only to determine the ancestral allele ($F_{chimp\text{-}only}(C|D)$), and once using both chimpanzee and macaque and restricting to sites where they agree ($F_{chimp+macaque}(C|D)$). If the rate of recurrent mutations in $F_{chimp\text{-}only}(C|D)$ is *P*, then we expect that the rate in $F_{chimp+macaque}(C|D)$ will be about *P*/2 (the latter measurement screens out the half of recurrent mutations that occurred on the chimpanzee side of the tree but cannot screen out the mutations on the human side of the tree). Thus, we can write $F_{chimp\text{-}only}(C|D) = (F(C|D))\times(1\text{-}P) + (1\text{-}F(C|D))\times(P)$ and similarly $F_{chimp+macaque}(C|D) = (F(C|D))\times(1\text{-}P/2) + (1\text{-}F(C|D))\times(P/2)$. Substituting *P*, and after some algebra, we obtain $F(C|D) = 2F_{chimp+macaque}(C|D)\text{-}F_{chimp\text{-}only}(C|D)$, which we use to generate the numbers in Table S12.3.

Table S12.3: Population split times inferred from the probability of an allele being derived

	SNP discovery in Population B						*SNP discovery in Population A*					
A – B	***F(A|B)***	***d(A|B)* (×10^{-4})**	**% HC (uncorr-ected)**	**% HC (corr-ected)**	***Kya* Div_{HC}= 6,500**	***Kya* Div_{HC}= 13,000**	***F(B|A)***	***d(B|A)* (×10^{-4})**	**% HC (uncorr-ected)**	**% HC (corr-ected)**	***Kya* Div_{HC} 6,500**	***Kya* Div_{HC} 13,000**
San - Yoruba	25.8%	1.3	1.00%	1.00%	65	130	26.0%	1.2	0.92%	0.92%	60	120
Altai - Den	13.0%	3.4	2.62%	3.43%	223	445	16.0%	3.4	2.62%	3.64%	236	473
Yoruba - Altai	12.8%	5.0	3.85%	4.87%	316	633	17.3%	5.5	4.23%	4.23%	275	550
San - Altai	12.8%	5.0	3.85%	4.87%	316	633	17.1%	5.8	4.46%	4.46%	290	580
Mandenka - Alt	12.8%	5.0	3.85%	4.87%	317	633	17.4%	5.6	4.31%	4.31%	280	560
Yoruba - Den.	7.8%	6.6	5.08%	5.89%	383	765	16.9%	5.8	4.46%	4.46%	290	580
San - Denisova	8.0%	6.6	5.08%	5.89%	383	765	16.7%	5.9	4.54%	4.54%	295	590
Mandenka - Den	7.8%	6.6	5.08%	5.89%	383	765	16.9%	5.8	4.46%	4.46%	290	580

Note: We convert a divergence per base pair to a date by dividing *d(A|B)* or *d(B|A)* by an assumed human-chimp genetic divergence per base pair of 0.0130. To correct for branch shortening, we add 1.02% to the branches when the discovery individual is Altai and 0.81% when it is Denisova (SI 6b). We then multiply by 6.5-13 Mya for human-chimp divergence.

Table S12.3 shows the results. We highlight four observations:

(1) We find that *d(San|Yoruba)* is approximately the same as *d(Yoruba|San)*. Quoting the range over these two estimates, the inferred population split time is 17-22% of the archaic/modern split. This corresponds to 60-65 kya assuming a mutation rate of 1.0×10^{-9}/bp/year and 120-130 kya assuming a mutation rate of 0.5×10^{-9}/bp/year.

(2) The Archaic-African split times are estimated to be similar on either side of the tree and using either archaic sample. Assuming a mutation rate of 1.0×10^{-9}/bp/year, the estimates range from 275-383 kya. Assuming a mutation rate of 0.5×10^{-9}/bp/year, the estimates range from 550-765 kya.

(3) The Altai-Denisova split times are estimated to be similar on both sides of the tree. Assuming a mutation rate of 1.0×10^{-9}/bp/year, the estimates range from 223-236 kya. Assuming a mutation rate of 0.5×10^{-9}/bp/year, the estimates range from 445-473 kya (Table S12.3). These dates are only modestly different from the 190 kya and 381 kya point estimates from the PSMC (Table S12.2).

(4) The fact that our date for the Denisova-Altai split falls after the split dates of Africans and archaic samples (range of 58-86%) supports the notion that Denisova and Altai are (approximately) a clade relative to modern humans as previously reported[6]. Focusing on discovery as SNPs in present-day humans, we expect F(*Altai|Yoruba)=F(Altai|San)=F(Altai|Mandenka)*, and indeed we observe a range of 17.1-17.4%. Similarly, *F(Denisova|Yoruba)=F(Denisova|San)=F(Denisova|Mandenka)* has a range of 16.7-16.9%. The rate of derived alleles in Altai is higher than Denisova in all three African comparisons at 0.4-0.6%, consistent with a proportion of the genome of Denisovans derived from an unknown archaic population that diverged from the modern human lineage earlier than the split from Neandertals (SI 16a,b).

(iv) Summary

In this note, we have estimated that modern and archaic humans split between 275-765 kya (union of all point estimates). If we assume that the mutation rate is 0.5×10^{-9}/bp/year the range is 550-765 kya.

We also estimated the population split time for Altai and Denisova. Quoting the range over the point estimates in Table S12.2 and Table S12.3, and ignoring the effects of gene flow after the main split (which appears to have a small effect as the estimates for Denisova and Altai are similar; SI 16a), we estimate that Altai-Denisova split occurred at a time depth corresponding to 58-86% of the archaic-African split (Table S12.4). In absolute terms, this corresponds to an Altai-Denisova population divergence of 190-473 kya, or 381-473 kya if we assume a mutation rate of 0.5×10^{-9}/bp/year.

Most of the uncertainty in the date estimates reported in this note is due to lack of confidence in the true value of the mutation rate (at present uncertain by a factor of two). In Table S12.2 and Table S12.3, we therefore also provide estimates of time splits as a fraction of human-chimp for all quantities. These numbers can be converted into more accurate date estimates when the true value of the human mutation rate becomes known with more certainty.

Table S12.4 – Summary of population split times inferred from two different methods

Human-chimp div. time assumption:	PSMC method			Probability allele is derived			Union of both methods		
	% HC	6,500 kya	13,000 kya	% HC	6,500 kya	13,000 kya	% HC	6,500 kya	13,000 kya
San - West African	0.66%	43	86	0.92-1.00%	60-65	120-130	0.66-1.00%	43-65	86-130
Altai-Denisova	2.93%	190	381	3.43-3.64%	223-236	445-473	2.93-3.64%	190-236	381-473
Archaic - African	4.26-4.53%	277-294	553-589	4.23-5.89%	275-383	550-765	4.23-5.89%	275-383	550-765
$\frac{San-West\ African}{Archaic-African}$		15-16%			17-22%			15-22%	
$\frac{Altai-Denisova}{Archaic-African}$		65-69%			58-86%			58-86%	

Note: This table is based on the union of point estimates that appear in Table S12.2 and Table S12.3.

References

[1] Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475, 493-6.

[2] Meyer M., et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222-6.

[3] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and samtools. Bioinformatics 25, 2078-9.

[4] Li H (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987-93.

[5] Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubí C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andrés AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T (2013) Great ape genetic diversity and population history. Nature 499, 471-5.

[6] Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, 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 JJ, Kelso J, Slatkin M, Pääbo S (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468, 1053-60.

[7] Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PL, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S (2010) A draft sequence of the Neandertal genome. Science 328, 710-22.

[8] Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. Nat Genet. 43, 1031-4.

[9] Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D (2013) Ancient west Eurasian ancestry in southern and eastern Africa. arXiv:1307.8014.

[10] Hudson RR (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics 18, 337-8.

Supplementary Information 13

Population relationships inferred from phased haplotypes in present-day humans

Sriram Sankararaman, Swapan Mallick, Priya Moorjani, Joseph Pickrell and David Reich*

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

(i) Findings

- We can localize Neandertal- and Denisovan-introgressed segments in present-day humans.
- The Neandertals who introgressed into non-Africans shared ancestry with Altai 77-114 kya (assuming a mutation rate of $\mu = 0.5\times10^{-9}$/bp/year).
- The Denisovans who introgressed into Oceanian populations shared ancestry with Siberian Denisovans 276-403 kya (assuming a mutation rate of $\mu = 0.5\times10^{-9}$/bp/year).
- We detect Denisovan ancestry in eastern non-Africans at a low level (0.19-0.24% of their genomes). There is significantly more Denisovan ancestry in eastern non-Africans than in Europeans.
- We detect likely West Eurasian gene flow into the ancestors of Yoruba West Africans within the last ten thousand years, which indirectly contributed a small amount of Neandertal ancestry to Yoruba.

(ii) Phased haplotypes from present-day humans

We experimentally phased 13 genomes from present-day humans: 10 of the 11 Panel A individuals (all the genomes from ref. 1 except the Dinka individual), and 3 of the 14 Panel B genomes (2 Australians and 1 Mixe Native American). For each of these genomes, we sequenced pools of fosmids, following the method of Kitzman et al[2] (SI 4). We then combined the fosmid data with whole genome shotgun sequencing data that we also had on the same samples, resulting in phased contigs with median sizes (N50) of 222 kb for the most poorly phased sample (Yoruba) to 839 kb for the most accurately phased sample (Australian1) (SI 4). We restricted analyses of the phased data to sites passing the stronger of the two sets of filters described in SI 5b (Map35_100%), and only used genotypes that agreed between the SAMtools calls used by our phasing algorithm (SI 4)[3], and the calls from the Genome Analysis Toolkit (SI 3)[4]. We co-analyzed the data with Altai, Denisova and chimp.

(iii) Using the phased genomes to study archaic introgression into non-Africans

In the Neandertal draft genome paper, we used phased haplotypes from the human genome reference sequence as the basis of one of the three lines of evidence for Neandertal gene flow into non-Africans presented in that study[5]. Here we extend this analysis using the 13 phased genomes. A particular strength of our new analysis compared with the one we previously reported is that we now have high quality genome sequence from archaic humans. This allows us to measure divergence on both sides of the tree for any comparison of archaic to present-day humans, which allows us to obtain unbiased estimates of genetic divergence time.

We restricted our analysis to segments of the genomes that were inferred to be part of the same phased haplotype, and divided these segments into non-overlapping windows of 0.01 centimorgans (cM; assessed based on the Oxford linkage disequilibrium-based genetic map[6]), corresponding to regions expected to be undisrupted by recombination on each lineage for the last 1/0.0001 = 10,000 generations (290,000 years assuming 29 years per generation[7]). We divided each chromosome into non-overlapping windows of size at most 0.01 cM by moving across the chromosome in a p-arm to q-arm direction and breaking windows either when we encountered a phased contig that had not previously been encountered in the window, or when we reached the maximum span of 0.01cM.

We consecutively applied the following further filters, retaining windows where:
- The number of nucleotides passing filters and with coverage in chimpanzee was at least ≥25,000 bp.
- The fraction of these nucleotides with coverage in macaque was at least 60%.
- The window is in the central 95% of the genomic distribution of chimp-macaque divergence per bp.
- The window is in the central 95% of the genome distribution of test haplotype-chimpanzee divergence divided by the chimpanzee-macaque divergence.

In every window of the genome where we have alignment of an archaic haplotype (A), a test haplotype from a present-day human (*T*), and the phased other haplotype from the same present-day human (*O*), we define the following quantities:

n^j_A = number of derived mutations only seen in *A*
n^j_T = number of derived mutations only seen in *T*
n^j_O = number of derived mutations only seen in *O*
n^j_{AT} = number of derived mutations seen in both *A* and *T* (but not *O*)
n^j_{AO} = number of derived mutations seen in both *A* and *O* (but not *T*)
n^j_{TO} = number of derived mutations seen in both *T* and *O* (but not *A*)
n^j_{HC} = number of human-chimpanzee divergent sites
n^j_{CM} = number of chimpanzee-macaque divergent sites
$norm_{HC}$ = $\frac{n^j_{HC}/n^j_{CM}}{Div_{HC}/Div_{CM}}$ human-chimp divergence in window divided by chimp-macaque divergence, divided by the genome-wide ratio of this quantity

We also define genome-wide sums of human-chimpanzee and chimpanzee-macaque divergence:

Div_{HC} = number of human-chimpanzee divergent sites summed over the genome
Div_{CM} = number of chimpanzee-macaque divergent sites summed over the genome

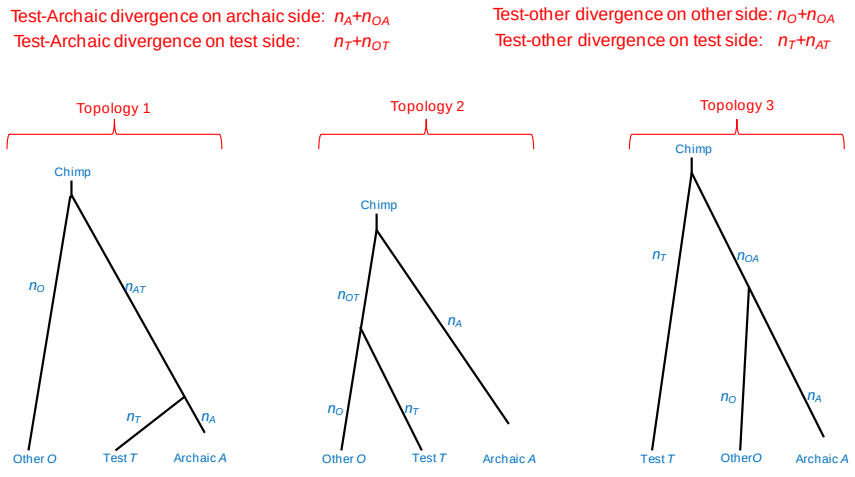


Figure S13.1: Pairwise divergence computations. *Consider an archaic haplotype ("A"), a phased test haplotype from a present-day human ("T") and the phased other haplotype from the same human ("O"). There are three possible topologies, and in all three the formulae at the top provide valid estimates of pairwise divergence.*

We rank-ordered all windows for each phased genome based on divergence to an archaic sample (Altai or Denisova) computed only using mutations on the archaic side of the tree (mutations on the *A+OA* lineages in Figure S13.1). We combined data from both haploid genomes in each individual and then binned the data into 500 equal sized rank-ordered bins. For each bin, we then capitalized on divergent sites that were not used in the rank-ordering to obtain unbiased estimates of two quantities:

(Quantity 1) *Test haplotype – archaic divergence time as a fraction of human-chimp divergence time.* To estimate the divergence between the test haplotype (*T*) and an archaic sample (*A*) in a way that is not biased by the $n^j_A+n^j_{OA}$ lineage mutations used to ascertain and rank order them, we used the divergence on the test haplotype side of the tree per bin measured as a fraction of human-chimp

divergence: $2(n^j_T+n^j_{OT})/n^j_{HC}$ (Figure S13.1) We then multiply by a normalization term equal to the local ratio of human-chimpanzee to chimpanzee-macaque divergence to the genome-wide estimate of this ratio, $(n^j_{HC}/n^j_{CM})/(Div^j_{HC}/Div^j_{CM})$, to correct for variation in the local mutation rate and variation in the time since the most recent common ancestor of humans and chimpanzees across the genome. This gives a number that we can convert into years if we make an assumption about the genome-wide average human-chimpanzee genetic divergence time.

(Quantity 2) *Test-other haplotype divergence (TMRCA of the two haplotypes from the same phased individual) as a fraction of the genome average of this quantity*. To obtain an unbiased estimate of the divergence between the test haplotype from a present-day human (*T*) and the other phased haplotype from the same present-day human (which is proportional to heterozygosity in that window), we restrict to mutations that occurred on the test haplotype side of the tree ($n^j_T+n^j_{AT}$ lineages in Figure S13.1). We ignore mutations on the other haplotype side of the tree as they include n^j_{OA} sites that were used in rank-ordering so they would produce a bias (Figure S13.1). We normalize by local human-macaque divergence in the bin and then divide by the genome average of this ratio to obtain an estimate of the local TMRCA as a fraction of the genome average.

Figure S13.2 plots (Quantity 1) against (Quantity 2), restricting to segments where the two haplotypes from the archaic genome have an inferred time since the most recent common ancestor from the Pairwise Sequential Markovian Coalescent (PSMC) of <0.0002/bp. This restriction means that we are restricting to regions where the coalescence of the two haplotypes in the archaic individual post-dates Denisova-Altai population divergence so that the archaic individual is effectively haploid (SI 12). (We repeated the analysis without this threshold and obtained qualitatively similar results.)

We highlight two theoretical predictions.

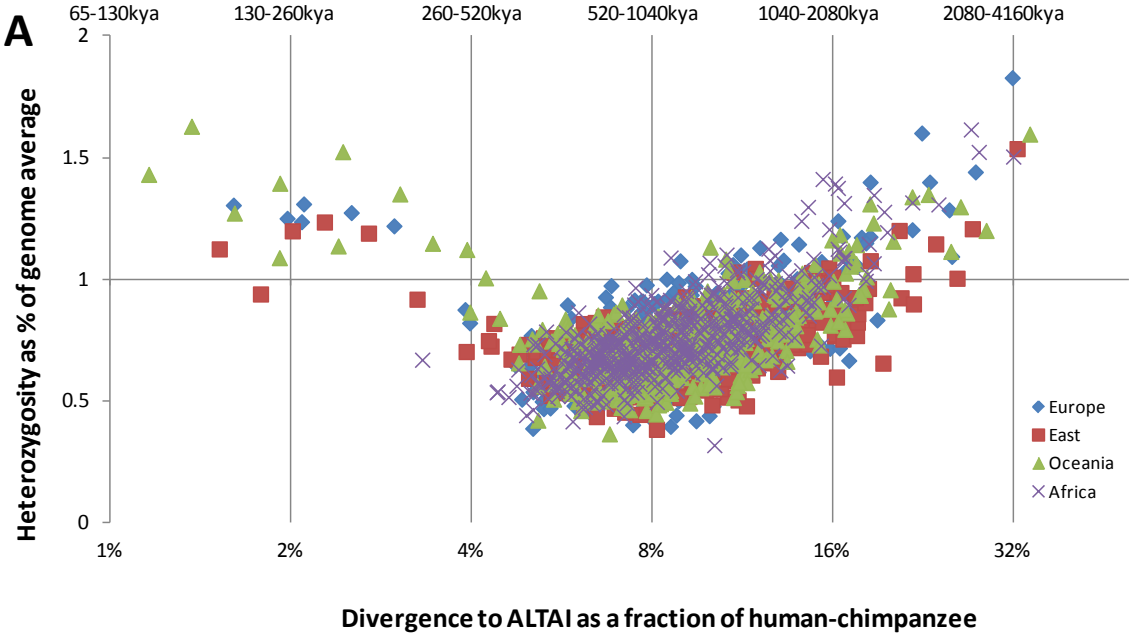
- In the absence of gene flow, present day human haplotypes with low divergence to archaic humans should have REDUCED divergence to other modern human haplotypes. This is because when we are restricting to locations in the test phased genome with low divergence to archaic humans, we are restricting to locations where we already know that two haplotypes have a recent coalescent time. To the extent that other haplotypes are correlated with them, they will have low coalescences too. Thus, we should see a monotonically increasing pattern as in sub-Saharan Africans (Figure S13.2).

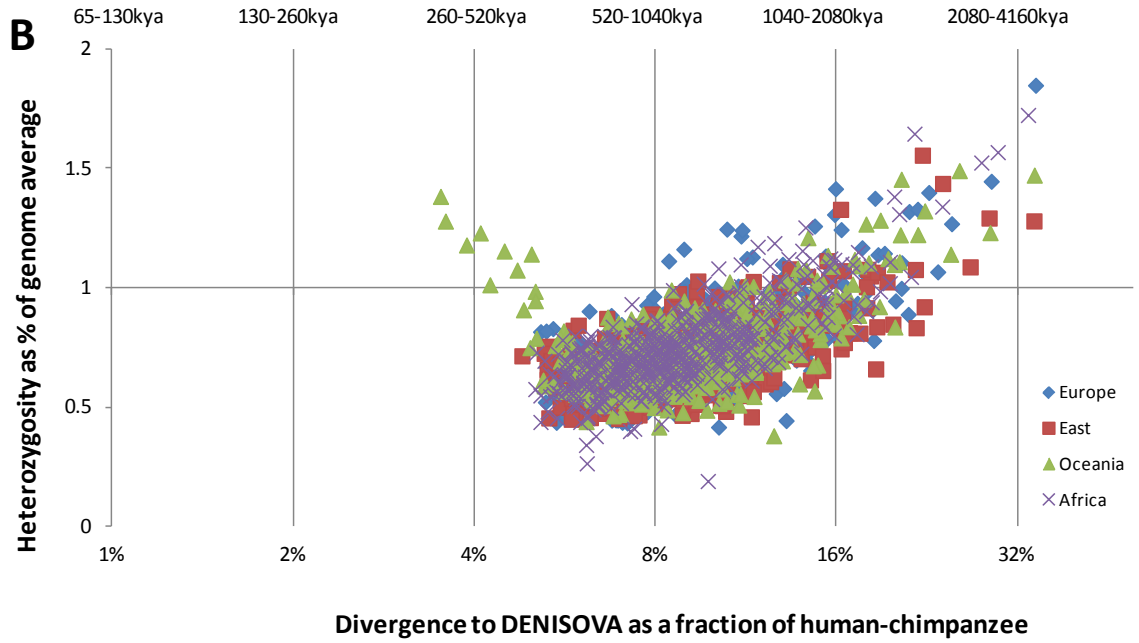
- In contrast, if the haplotype is archaic it should have INCREASED divergence to modern human haplotypes. Thus, if we plot test-archaic haplotype divergence against test-other, we expect to see a non-monotonic pattern, with low divergence to Neandertal predicting high divergence to most present-day human haplotypes. We observe this inflated left tail in non-Africans (Figure S13.2).

Figure S13.2A shows the signal of Neandertal gene flow into non-Africans that emerges by plotting the normalized $n^j_T+n^j_{OT}$ rate (test-archaic divergence) against the normalized $n^j_T+n^j_{AT}$ rate (test-other divergence). Pooling the two phased African genomes with the least evidence of back-to-Africa gene flow (Yoruba and Mbuti), we observe monotonically increasing curves, yielding no evidence of Neandertal-related flow into Africans. In non-African phased genomes, the curves jump at the far left, an unambiguous signal of Neandertal-related gene flow into these populations.

The *x*-axis in Figure S13.2A has the particularly important feature that it provides unbiased estimates of the divergence time in each of the 500 bins. This is because we only used data from the archaic genome side of the tree for rank-ordering the windows, and reserved the divergence on the present-day human side for unbiased time estimation. We thus use the points at the far left of Figure S13.2A (ascertained as locations where the test-archaic divergence time measured on the archaic side of the tree is lowest) to provide valid upper bounds on population divergence time (assuming no continued gene flow after the main population separation). For each point, the sampling error is negligible: around 3% of the point estimates because each bin contains so many sites. Thus, we can ignore statistical error as a contributor to uncertainty in the estimate of the mean divergence time between the phased genome and the archaic sample for segments included in the bin.

Figure S13.2: Some phased haplotypes in non-Africans have low divergence to archaic humans and high divergence to other present-day human haplotypes, revealing archaic ancestry. *We rank-order windows based on divergence on the archaic side of the tree, and divide into 500 bins. (A) Low divergence between a present-day non-African and Altai is associated with a high divergence between that haplotype and the other carried by the same individual (measured on the other side of the tree to avoid ascertainment bias). This is expected if the non-African haplotype derives from Neandertals. We see no signal replacing non-Africans with Africans. (B) Low divergence between an Oceanian population and Denisova is associated with high divergence between the individual's two haplotypes.*





The lowest bin in Figure S13.2 corresponds to 3.32% of genome average human-chimpanzee divergence for Altai divergence to Africa (pool of Yoruba + Mbuti), 1.61% for Altai divergence to Europeans (Sardinian + French), 1.53% for Altai divergence to Eastern non-Africans (Karitiana + Han + Dai + Mixe), and 1.16% for Altai divergence to Oceanians (Papuan + Australian1 + Australian2). This corresponds to upper bounds on the time of final divergence between Altai and these population

pools of 216 kya, 105 kya, 198 kya and 151 kya respectively assuming that human-chimp genetic divergence occurred 6.5 Mya (mutation rate of μ=1.0×10^{-9}/bp/generation). It corresponds to upper bounds on the time of final divergence between Altai and these population pools of 432 kya, 209 kya, 199 kya and 151 kya respectively assuming that human-chimp genetic divergence occurred 13 Mya (mutation rate of μ=0.5×10^{-9}/bp/generation). Table S13.1 presents results for the lowest 500th divergence bin for each of the 13 phased genomes separately.

Figure S13.2B reports the same analysis but now comparing to Denisova. The scatterplot is monotonically increasing for both Africans and Europeans. In Oceanian populations (Papuans and Australians), in contrast, we see a non-monotonic pattern: the lowest bins of test-Denisovan divergence have high heterozygosity in Oceanians. For the lowest bin, divergence is 3.52% of human-chimp, corresponding to 229 kya assuming 6.5 Mya for human-chimp divergence, and 458 kya assuming 13 Mya for human-chimp divergence. This is 2-3-times the upper bounds for divergence of Altai to the introgressing Neandertal, consistent with the introgressing Neandertal material being more closely related to Altai than the introgressing Denisovan material to the Siberian Denisovan.

Table S13.1: Africans share more segments of low divergence with Altai than with Denisova

	Span of windows passing filters in Altai & Denis. (Mb)	Altai		Denisova		Z for Altai sharing more genome than Denisova with divergence:	
		Divergence in 500th of genome of lowest div. as % of human-chimp	Het. in lowest 500th / mean	Divergence in 500th of genome of lowest div. as % of human-chimp	Het. in lowest 500th / mean	<1.5%	<4.0%
Yoruba	290	2.96% ± 0.22%	0.7	5.41% ± 0.32%	0.6	1.2	3.2
Mbuti	268	4.07% ± 0.29%	0.5	4.51% ± 0.31%	0.5	0.7	3.8
San	331	3.62% ± 0.23%	0.6	5.38% ± 0.30%	0.5	2.4	3.5
Mandenka	319	3.12% ± 0.22%	0.8	5.81% ± 0.32%	0.7	1.6	1.9
All Africa*	n/a	3.32% ± 0.17%	0.7	5.16% ± 0.23%	0.5	1.3	4.4
Sardinian	326	1.34% ± 0.14%	1.0	5.38% ± 0.30%	0.5	12.2	7.2
French	347	1.90% ± 0.16%	1.6	5.09% ± 0.29%	0.8	7.9	5.3
All Europe	n/a	1.61% ± 0.11%	1.3	5.28% ± 0.21%	0.6	11.3	7.8
Karitiana	363	1.21% ± 0.12%	1.2	4.80% ± 0.27%	0.8	7.8	4.5
Han	329	1.73% ± 0.16%	1.0	4.91% ± 0.28%	0.7	3.6	4.6
Dai	285	1.72% ± 0.17%	1.3	4.79% ± 0.31%	0.7	8.8	4.9
Mixe	390	1.51% ± 0.13%	1.0	5.59% ± 0.28%	0.8	15.8	6.5
All East	n/a	1.53% ± 0.07%	1.1	4.81% ± 0.14%	0.7	9.8	7.2
Papuan	356	1.24% ± 0.13%	1.5	3.47% ± 0.23%	1.7	9.5	2.6
Australian1	399	1.09% ± 0.11%	1.4	3.49% ± 0.22%	1.1	10.8	3.4
Australian2	389	1.29% ± 0.12%	1.4	3.37% ± 0.22%	1.3	13.1	3.5
All Oceanian	n/a	1.16% ± 0.07%	1.4	3.52% ± 0.13%	1.4	15.0	4.0

Note: We analyze 0.01cM bins with ≥25,000 screened bases where the archaic individual has an inferred upper bound on its TMRCA likely to be less than Denisova-Altai population divergence. Results are for the 1/500th of the genome of lowest divergence on the archaic side of the tree. Standard errors are from a binomial distribution. To test for differences in low-diverged bins in Altai vs. Denisova in the final two columns, we only use data from the present-day human side of the tree (to avoid complications due to different degrees of branch shortening), and we use a Block Jackknife over 20 equally sized blocks to compute a standard error.
* "All Africa" is a pool of Yoruba and Mbuti. We leave out San and Mandenka because of evidence of relatively large introgression of non-African ancestry into these individuals' ancestors (compared with Yoruba and Mbuti)[8].

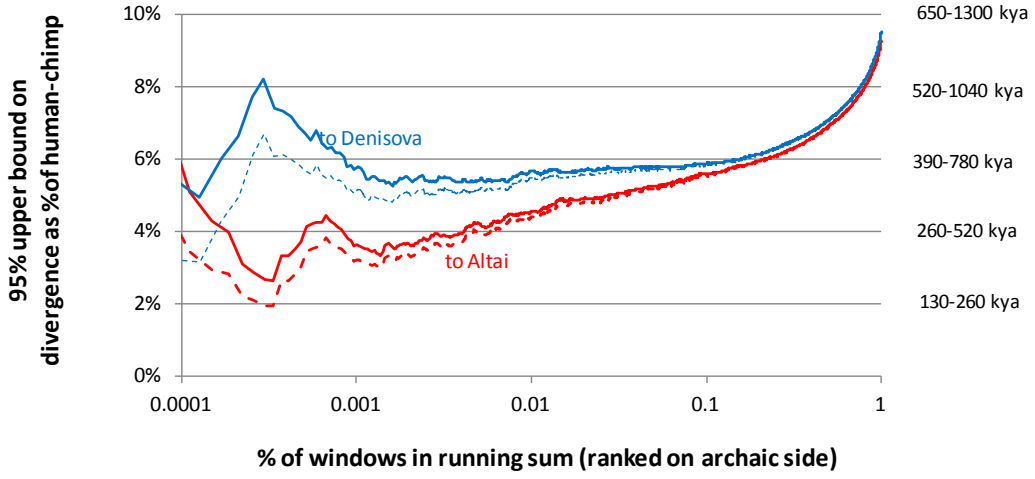
(iv) We detect some Neandertal ancestry in Yoruba, likely reflecting back-to-Africa migration
A notable pattern in Table S13.1 and Figure S13.1 is that the bin of lowest African-Altai divergence has a lower divergence as a fraction of human-chimpanzee (3.32%) than the bin of lowest African-Denisova divergence (5.16%) (Table S13.1). This is also apparent in the *x*-axis position of the lowest bin for the African data series comparing the two panels of Figure S13.2.

To test if the apparent excess of segments of low divergence to Africans in Altai compared with Denisova is statistically significant, we rank-ordered all windows in the genome that passed our filters based on the divergence computed on the present-day human side of the tree (we do not use the archaic side of the tree for this analysis as Altai and Denisova have different degrees of branch-

shortening, which in theory could produce artifactual signals). We then computed the fraction of the genome less than specified thresholds of divergence and obtained a standard error from a Block Jackknife breaking the genome into 20 equally chunks. The last column of Table S13.1 shows that there is a significant enrichment in the fraction of African genomes that have a divergence of <4% to Altai compared with the fraction with divergence of <4% to Denisova (Z=4.4 when we use a pool of Yoruba and Mbuti to represent Africa). The fact that these results are statistically significant in a Block Jackknife indicates that our signal is not driven by a few odd loci; it is present genome-wide.

Figure S13.3 compares African-Altai and African-Denisova divergence on the same plot, with the *x*-axis reporting results for different fractions of the genome. Focusing on the lines showing the 95% confident upper bounds on archaic-African divergence, we observe that the minimum of these lines corresponds to 2.62% of human-chimpanzee genetic divergence for Altai, compared with 4.93% for Denisova. (We note that these numbers are different from those in Table S13.1: Table S13.1 reports the 500th smallest bin, whereas these numbers report the minimum at any threshold, and furthermore report an upper bound taking into account statistical error.) These results suggest some shared ancestry between Neandertals and Africans in the last 341,000 years = 2.62% × 13 Mya.

Figure S13.3: Divergence of present-day Africans to archaic genomes. *We rank-order windows using the test-archaic divergence computed on the archaic side of the tree and measure divergence on the present-day African (Yoruba+Mbuti) side (not used for rank-ordering so that it provides an unbiased estimate of divergence). The x-axis corresponds to the percentage of the ranked windows used to compute the divergence. The dashed lines are point estimates and the solid lines are 95% confident upper bounds on the true divergence.*



What history could explain the signal of more African haplotypes being related within the last 4% of human-chimpanzee divergence to Altai than to Denisova?

We ruled out the possibility that these results are an artifact of not having phased genomes for Altai and Denisova. In particular, we were concerned that if the probability with which the two Altai haplotypes coalesce more recently than the split from Africans is higher than that for the two Denisova haplotypes, there could be more opportunity in Altai than Denisova for the two haplotypes to recently coalesce with an African haplotype, biasing our statistics. However, our analyses are restricted to segments where Altai and/or Denisova have an inferred heterozygosity from the PSMC (SI 12) <0.0002/bp; less than the date of Altai-Denisova divergence. Moreover, when we remove the PSMC threshold we obtain similar results; thus, our signal is not sensitive to this issue.

We hypothesize that these results could be explained by the presence of Neandertal genetic material in sub-Saharan Africans that owes its origin to back-to-Africa admixture, and that occurs at a sufficiently low-level that we only find a signal in the 500th of the genome of lowest Altai-Denisova divergence in Figure S13.2. If the signal was sufficiently weak that there were a substantial number of false-positives in the 500th of the genome of lowest divergence to Altai (segments that are not in fact

Neandertal-derived), the signal in the 500th of the genome of lowest divergence could be diluted enough to be consistent with several observations that on their surface seem like they might not be compatible with Neandertal ancestry in Yoruba:

(a) Our upper bound on the divergence time between Yoruba and Mbuti segments and the Altai Neandertal of 2.62% is substantially higher than the point estimates of 1.09-1.90% in the bin of 500th lowest divergence of non-African segments from Altai (Table S13.1). While this seems nominally inconsistent with Neandertal ancestry, it could be explained if the signal was diluted.

(b) We observe very few windows where Altai forms a clade with an African phased haplotype. In particular, Table S13.4 (below) shows that the rate at which this occurs in the Yoruba and Mbuti is <1/10000, two orders of magnitude lower than the rate seen in non-Africans. Thus, if our signal is due to back-to-Africa migration, it would have to be due to gene flow that explains on the order of a percent of the ancestry of these populations, which is consistent with our hypothesis.

(c) We only see a modest increase in Yoruba and Mbuti heterozygosity in the bin of lowest divergence to Altai (Figure S13.2 and Table S13.1), corresponding to 10-20% of the genome average. This contrasts with what we observed in non-Africans, where we see a large rise in the *y*-axis of Figure S13.2 by about 100% of the genome- average. However, if the proportion of Neandertal material was two orders of magnitude lower in Africans than in non-Africans, the signal even in the 500th of the genome of lowest divergence could in theory be diluted enough that the heterozygosity in the lowest bin would only rise modestly, as we observe.

We also found two confirmatory lines of evidence suggesting that the patterns we observe are indeed due to back-to-Africa gene flow, which indirectly brought some Neandertal ancestry into Africans.

Evidence of Neandertal ancestry in the Yoruba from the joint allele frequency distribution

The first line of confirmatory evidence comes from SI 16a, where we examine the joint allele frequency spectrum between archaic humans and 80 YRI West African alleles. We find that at mutations with a low derived allele frequency in YRI (<10%), the derived allele matches Altai at a higher rate than Denisova (right panels of Figure S16.1A and Figure S16.1B). The most extreme pattern is observed at derived alleles that occur only once in 80 YRI alleles sampled. Define the number of sites where Altai but not Denisova carries the derived allele as $nd10_1$, and the number of sites where Denisova but not Altai carries the derived allele as $nd01_1$. In this case, we find an excess of 9.7% = $(nd10_1\text{-}nd01_1)/(nd10_1+nd01_1)$ (Figure S16.1A, right). This is what would be expected from gene flow into the ancestors of the YRI from a population carrying Neandertal-related ancestry that occurred recently enough that this ancestry did not drift much in frequency. This signal is stronger in the Yoruba than in the Mbuti or Dinka, as in the left panels of Figure S16.1A and Figure S16.B which pool data from 10 chromosomes from Yoruba, Mbuti and Dinka, we see no rise in the African matching rate to Altai for derived alleles that only occur once. This is not an artifact of a low number of alleles sampled as when we downsample the YRI in the right panels to 10 we still see a rise in the YRI matching rate to Altai for derived alleles that only occur once.

Evidence of Neandertal ancestry in Yoruba that arrived indirectly through back-to-Africa gene flow

The second line of confirmatory evidence comes from analyzing weighted linkage disequilibrium (LD) in 113 unrelated YRI individuals genotyped at 606,071 single nucleotide polymorphisms (SNPs) (using a merged dataset from refs. 9 and 10). Specifically, we used the *ALDER* software to test for the presence of admixture linkage disequilibrium (LD) in the YRI. We would only expect to observe a correlation between LD in the YRI and the allele frequency difference between San Bushmen and different non-African groups (in total 45 groups) if the LD owes its origin to mixture between populations related (even distantly) to the reference population8.

Applying this analysis to our data, we find a significant correlation that is strongest when we use West Eurasians as the non-African reference population, suggesting a history of West Eurasian related gene flow into the ancestors of Yoruba (Table S13.2). Figure S13.4 shows the decay that we observe when we treat YRI as an admixture of West Africans and Europeans. The decay of LD is highly statistically

significant (Z = 5.3 for the significance of detection of admixture LD), and has a genetic distance scale of its decay corresponding to 332 ± 63 generations or 9618 ± 1825 years assuming 29 years per generation7. The *ALDER* method is also able to use the amplitude of the LD decay (the value at small genetic distances) to infer a minimum estimate of the proportion of gene flow. We estimate a lower bound of 2.7 ± 0.9% of lineages in the YRI tracing their origins to West Eurasia (±1 standard error), consistent with the other lines of evidence reported above that suggest that the rate of Neandertal ancestry in the Yoruba may be on the order of 1% of the rate in non-Africans.

In conclusion, the Yoruba harbor a small amount of Neandertal ancestry, which they likely inherited indirectly through gene flow from a Neandertal-admixed modern human population. The Mbuti may also have some Neandertal ancestry based on the analysis of Table S13.1, albeit probably at a lower level than the Yoruba. These results mean that we have not identified any sub-Saharan African sample that we are confident has no evidence of back-to-Africa migration. Our best candidate at present is the Dinka but it is possible that with a phased genome or large sample sizes we would detect evidence of non-African ancestry in this population as well. We note that in many analyses in this manuscript, we use Yoruba and Mbuti as reference modern humans assumed not to have any Neandertal ancestry. While we have shown here that this assumption is not accurate, the proportion of Neandertal ancestry is very small (~1% of ~2%), and we would not expect it to have an effect on many of the inferences in this study. In particular, it would not explain our key signal of Africans sharing more derived alleles with Altai than with Denisova, especially at sites with high African derived frequency (SI 16a).

Table S13.2: Test for mixture in YRI, using 45 populations as surrogates for the minor mixing group

World region	Populations tested (*X*)	% of pops with signals	List of pops that pass ALDER 1-ref test of admixture
Europe	French, Basque, Sardinian, Italian, Orcadian, Tuscan, Adygei, Russian	75%	French, Sardinian, Italian, Orcadian, Tuscan, Adygei
Central/South Asia	Brahui, Balochi, Hazara, Makrani, Sindhi, Pathan, Kalash, Burusho	38%	Hazara, Kalash, Burusho
Middle East	Bedouin, Druze, Palestinian	33%	Druze
East Asia	Han, Han-Nchina, Tujia, Yi, Miao, Daur, Mongola, Hezhen, Xibo, Uygur, Oroqen, Dai, Lahu, She, Tu, Yakut, Japanese, Cambodian, Naxi	0%	-
Oceania	Papuan, Melanesian	0%	-
Americas	Pima, Maya, Colombian, Karitiana, Surui	0%	-

Note: We ran *ALDER* with YRI as the target and San and each non-African group *X* in turn as the reference population. We report the number of populations that pass the ALDER test for admixture (P<0.05).

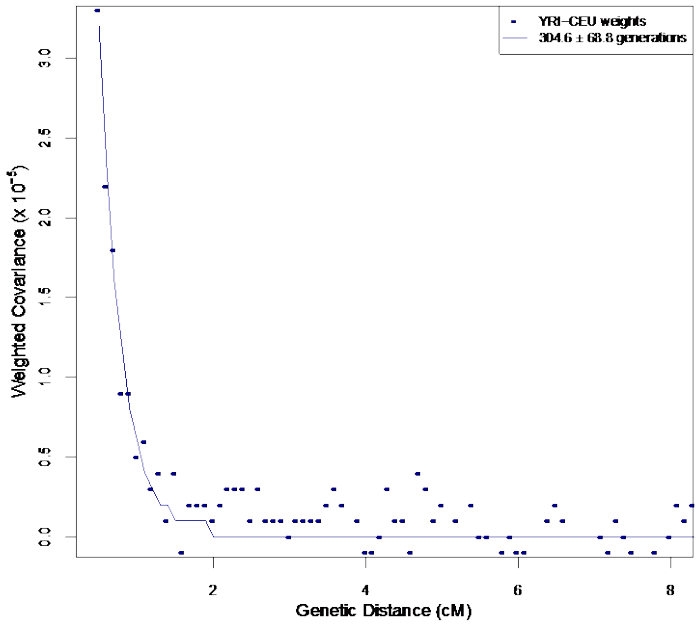


Figure S13.4 Weighted LD curve for West Africans using West Africans and European Americans as the putative admixing populations. *We analyze 113 YRI individuals genotyped at 606,071 SNPs using weights based on the allele frequency difference between YRI Nigerians and CEU European Americans. We observe a significant decay of weighted LD. We plot the admixture LD curve with the fit starting at the ALDER-computed LD correlation threshold (0.5cM).*

(v) Identifying segments of archaic ancestry in present-day human phased genomes

The previous analyses suggest that there is enough information in the phased genomes to support local ancestry inference; that is, to pull out segments of the genomes of present-day non-Africans that are highly likely to be derived from Neandertals. In this section we describe a Hidden Markov Model (HMM) that scans through the genome, identifying such segments in a principled way.

The HMM that we present here is specifically designed to support computation of *D*-statistics and studies of population history. Because the HMM only uses data from a single phased genome from a test individual, as well as from outgroups that we believe have not been affected by introgression from populations related to Neandertals (sub-Saharan Africans or primates), the only non-African data is from the test haplotype itself. This means that the method is expected to be equally efficient at identifying introgressed haplotypes in Europeans and East Asians. Specifically, each haplotype is guaranteed to trace its ancestry back to the split from the outgroup samples without coalescing with them, so that differences in demographic history across populations (however large) should not affect coalescence rates among the analyzed samples and hence should not affect the HMM's sensitivity.

It is important to point out that this feature of our method—its equal sensitivity in theory to archaic introgression in all populations—contrasts with other Neandertal local ancestry inference methods by Wall et al.[11] and Sankararaman et al.[12] that detect introgressed segments by using data from multiple samples from a population. Using multiple haplotypes increases power but has the potential drawback that demographic history differences across populations may affect the sensitivity of the method so that it may not be valid to compare the proportion of present-day genomes inferred to be of archaic ancestry to infer which populations have most ancestry. For example, stronger genetic drift in the history of a present-day human population—as is known to have occurred in East Asia compared with Europe since they separated[13]—may in theory cause true Neandertal haplotypes to stand out more clearly against East Asian haplotype structure than European haplotype structure, and in principle could lead to the detecting of more Neandertal haplotypes in East Asians than in Europeans even if the true proportion of Neandertal ancestry is in fact the same. Our method has lower sensitivity than the other local ancestry inference methods, but it should be equally low in all populations so that a test of the relative proportion of the genome called as introgressed should still be meaningful.

Labeling each site in a test haplotype as consistent or not with being a clade with an archaic individual

In each of the phased genomes from present-day humans, we examined sites with information from chimpanzee (to determine the ancestral allele), the test haplotype, an archaic sample used to fish out introgressed haplotypes (a single archaic sample or a pool of archaic samples), and a pool of outgroup samples which always included sub-Saharan Africans and sometimes included additional archaic samples. We represented sub-Saharan Africans by using data from 107 YRI individuals from the 1000 Genomes Project[14] sequenced to an average coverage of 7.1× (http://www.1000genomes.org/), and only analyzing sites where there were at least 80 YRI each of whom had coverage from at least 3 reads with map quality of MAPQ≥37 and base quality ≥30. For each site, we sampled a single allele to represent each YRI individual, obtaining a higher confidence call by taking the majority allele (almost always supported by at least 2 reads since we had at least 3 reads for each individual we analyzed). In addition to restricting to sites where we had at least 80 YRI allele calls identified in this way, we further restricted to sites where we had data available from at least 3 of 4 deeply sequenced sub-Saharan Africans we had not phased (Dinka_A, Dinka_B, Mbuti_B and Yoruba_B). Thus, at all sites we had coverage from at least 86 = 80+3×2 sub-Saharan African chromosomes. We next defined:

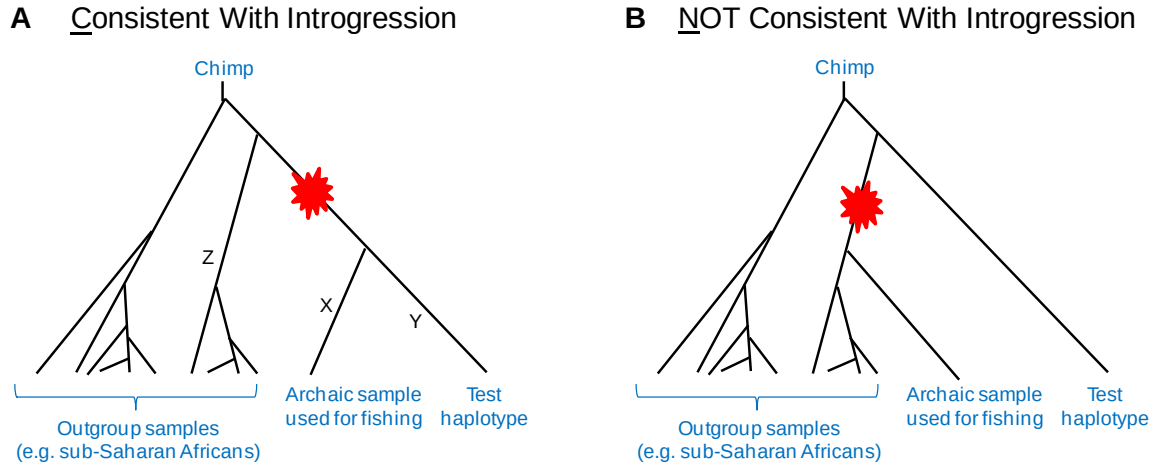
f_T = Frequency of the derived allele in the test haplotype (f_T=0 or f_T=1)
f_A = Frequency of the derived allele in the archaic genomes ($0 \le f_A \le 1$)
f_Y = Frequency of the derived allele in sub-Saharan Africans ($0 \le f_Y \le 1$)[14].

We restricted our analysis to two classes of sites that are unambiguously informative (in the absence of recurrent mutation) about whether a test haplotype is more closely related to one of the archaic samples used to fish than to any of the outgroup samples (Figure S13.5):

"Consistent" f_T-f_Y=*1* and $|f_A$-$f_T|$<*1*.
Sub-Saharan Africans are all ancestral while the test haplotype is derived, and at least one of the archaic allele matches the derived allele in the test haplotype. In the absence of recurrent mutation or sequencing error, this shows that the test haplotype is more closely related to an archaic haplotype than to any African.

"Not consistent" *0*<f_Y<*1* and $|f_A$-$f_T|$=*1*.
The test haplotype is definitely more closely related to some sub-Saharan African haplotypes than it is to any of the archaic haplotypes.

Figure S13.5: Schematic depiction of the 2 classes of mutations we use for local ancestry inference *(A) "Consistent" mutation showing that the test haplotype is more closely related to an archaic haplotype used for fishing than any outgroup haplotype. (B) "Not Consistent" mutation showing the test haplotype is more closely related to some outgroup haplotypes than to any test archaic haplotype.*



We highlight three points:

(1) Our definitions of "Consistent" and "Not consistent" mutations are meaningful even if we use more than one haplotype for fishing out introgressed haplotypes. If we had a phased archaic haplotype to use as our "bait", our two definitions would be (f_T-f_Y=*1* and $|f_A$-$f_T|$=0) for "Consistent" and unchanged for "Inconsistent". The inequality in the definition allows a site to be defined as "Consistent" if any sample in the pool of archaic samples used for fishing is the closest match to the test haplotype.

(2) The great majority of mutations do not fall in the "Consistent" or the "Not consistent" category. While some of these provide information about introgression (and are used in the accompanying paper by Sankararaman et al. to increase power[12]), they are ignored here to avoid bias that could arise from co-analyzing multiple samples from a population.

(3) The identification of a region of the genome where an archaic haplotype is most closely related to the test haplotype does not prove introgression. Such a pattern could alternatively arise by incomplete lineage sorting (i.e. the test haplotype might not coalesce with any outgroup haplotype all the way back to the common ancestral population of modern humans and the archaic samples used for fishing). In this case, we expect the genetic span of shared haplotypes to be small (because they have been broken up by many hundreds of thousands of years of recombination). We reduce the detection of such false-positive introgressed segments by programming a Hidden Markov Model (HMM) that searches for the much longer segments of shared ancestry (0.0005 Morgan switch rate) expected from true introgression.

Hidden Markov Model to infer the probability of Neandertal introgression at each point in the genome
We initially binned each phased haploid genome (we obtained two phased haplotypes from each diploid genome giving 26 haploid genomes in total) into non-overlapping windows of 0.00005

Morgans, chosen to be a tenth of the typical genetic span of archaic segments today (0.0005 = 1/2000 Morgans, the length expected from admixture around 2,000 generations ago[15]).

Within each bin of size 0.00005 Morgans, we counted the total number of "Consistent" and "Not consistent" sites, and defined the bin to be informative about ancestry if $n_C>0$, $n_N=0$ (supporting the test haplotype being in a clade with archaic haplotypes), or $n_C=0$, $n_N>0$ (contradicting such a history). If $n_C=0$, $n_N=0$ or if $n_C>0$, $n_N>0$, we treated the bin as uninformative.
We wrote a simple Hidden Markov Model (HMM) to process the windows of "all C" or "all N" sites to infer the hidden state (archaic or not). The HMM has three parameters:

s = Ancestry switch rate, which we set in practice to be 0.0005 Morgans
p = Prior probability for archaic ancestry at any locus, which we set to be 0.01.
u= Probability of archaic ancestry conditional on all SNPs in the window being of state "C", which we require for the sake of reducing our parameter space to also be the probability of modern human ancestry conditional on all SNPs in a window being of state "N". We set this to 0.99.

The HMM produces a posterior decoding at each window *i* (γ_i), which if the model was correct would be the probability of introgression. To optimize the HMM, we varied *s*, *p* and *u*, identifying combinations that maximized the likelihood of the data as well as the separation between the non-African (excluding Oceanian) and African phased genomes. Thus, we used our African data to empirically calibrate the HMM.

In practice we found that *s*=0.0005, *p*=0.01, *u*=0.99 produces excellent separation between Africans and non-Africans. This set of parameters is also intuitively sensible, since a switch rate of *s*=0.05=1/2000 corresponds to previous estimates of an admixture date of around 2,000 generations ago[15], *p*=0.01 corresponds to a Neandertal admixture proportion of 1% which is conservatively at the low end of the range of what has previously been reported, and *u*=0.99 allows for some probability of a test haplotype not being introgressed even if locally it is a clade with an archaic haplotype due to incomplete lineage sorting. This setting means that a locus is only called as confidently introgressed if at least two windows in a row are "C", thus discriminating against haplotypes that owe their origin to incomplete lineage sorting and that we do not wish to call as Neandertal introgressed. Figure S13.6 shows an example of the local ancestry inference along chromosome 3 comparing a real non-African (Sardinian) and African (Mbuti), empirically documenting the excess of segments in non-Africans.

Figure S13.6: HMM results on chromosome 2, comparing Sardinian and Mbuti. *We run the HMM using s=0.0005, p=0.01, u=0.99, and use Altai to pull out introgressed haplotypes. The y-axis shows the expected number of archaic alleles summing up results from the two phased haplotypes. We see many more inferred Neandertal introgressed segments in Sardinians than Mbuti (qualitatively similar plots are observed in other sub-Saharan Africans: Mandenka, Yoruba and San).*

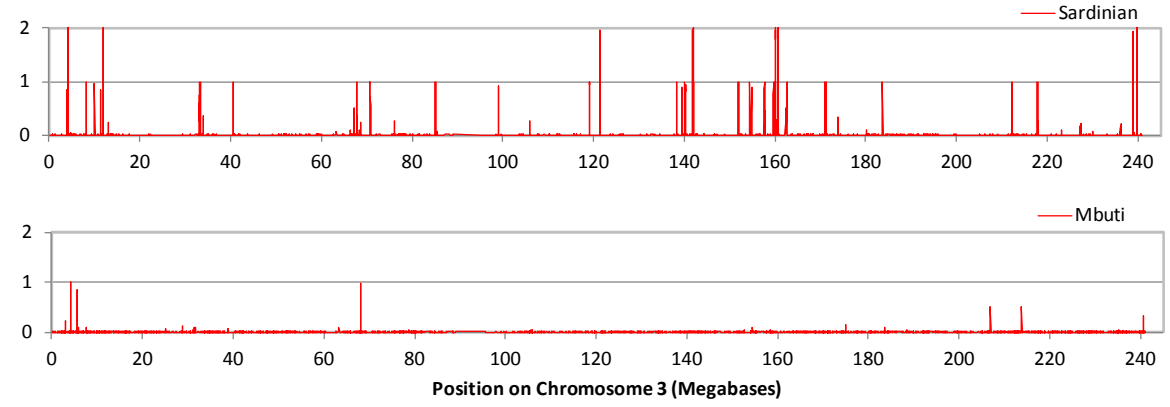


Table S13.3 shows a histogram of γ_i for all 13 samples, using Altai to fish out introgressed segments and assessing ancestry based on the proportion of 0.0005 Morgan windows in each sample that have specified ranges of γ_i. We have several observations:

- There is an almost 100-fold excess of sites with elevated γ_i values in non-Africans as compared with sub-Saharan Africans, consistent with the sparsity of peaks in Yoruba in Figure S13.6. Pooling over the sub-Saharan Africans, we infer that the proportion of the genome with γ_i>0.9 is 0.01% averaging over the 4 sub-Saharan Africans and 0.76% averaging over the 9 non-Africans.

- For all four sub-Saharan Africans, the distribution of the genome into bins of different γ_i is similar. Given the extremes differences in the histories of these African individuals, this suggests that the excess of γ_i>0.9 in non-Africans compared to Africans is genuinely reflecting Neandertal ancestry.

The HMM we use to infer γ_i is not an accurate approximation of history or the recombination process, so we cannot interpret it as the literal probability of archaic ancestry at any position in the genome. However, we can empirically recalibrate γ_i into a probability of reflecting true Neandertal ancestry by measuring the excess rate of values in a particular γ_i range compared with sub-African samples used as a baseline (assumed to have no Neandertal ancestry). For our African baseline, we use two samples: Mbuti and Yoruba. The reason for this is that analysis of admixture linkage disequilibrium has shown that the San and Mandenka inherit more of their ancestry than the Yoruba or Mbuti from West Eurasia due to gene flow events in the last few thousand years[16,17]. In section (iv) of this note, we show that Yoruba and Mbuti probably also have some West Eurasian ancestry, but that it is less.

Table S13.3: Posterior decoding histogram using Altai to fish ancestral haplotypes (parts per 10,000)

	Sub-Saharan African				European		Eastern non-African				Oceanian			
γ_i	Mbuti	Yoruba	San	Mandenka	French	Sardinian	Dai	Mixe	Han	Karitiana	Papuan	Australian1	Australian2	% of sites in non-Oceanian non-Africans likely due to introgression*
0-0.01	9984	9997	9948	9991	9820	9817	9803	9779	9773	9774	9654	9653	9677	0%
0.01-0.05	7	2	22	3	25	29	31	32	38	33	53	52	50	87%
0.05-0.1	3	1	10	2	16	18	20	20	24	21	32	32	29	90%
0.1-0.25	3	1	9	2	26	29	31	33	36	34	51	48	44	94%
0.25-0.5	2	0	5	1	27	26	31	32	35	33	47	45	41	97%
0.5-0.75	0	0	2	1	17	16	19	21	21	23	33	32	29	99%
0.75-0.9	0	0	2	0	13	13	14	15	15	15	25	24	21	99%
0.9-0.95	0	0	1	0	6	5	7	7	6	7	10	11	11	99%
0.95-0.99	0	0	1	1	14	13	13	14	14	15	25	25	23	99%
0.99-1	0	0	1	1	37	33	33	48	38	45	71	79	73	100%

Note: Entries are in parts per 10,000, averaging over all 0.00005 Morgan windows in each sample. We run the HMM with *s=0.05, p=0.01, u=0.99* and use Altai to fish out archaic haplotypes. We pool data from the two haplotypes of each person to increase precision. * Referring to $G_{j,k}$ as the entries in the table (j is the row and k the column), we define $G_{j,Africa} = (G_{j,Mbuti} + G_{j,Yoruba})/2$ and $G_{j,Non\text{-}African} = (G_{j,French} + G_{j,Sardinian} + G_{j,Dai} + G_{j,Mixe} + G_{j,Han} + G_{j,Karitiana})/6$. To estimate the percent of sites in a specified rang of γ_i that reflect introgression, we use the excess beyond the rate in Africans; that is, we quote the ratio $(G_{j,Non\text{-}African} - G_{j,African})/G_{j,Non\text{-}African}$. For γ_i<0.01 we clip the estimates to be no less than 0%.

The last column of Table S13.3 shows the fraction of the genome in non-Africans that we infer reflects archaic admixture comparing to Africans as a baseline (this analysis leaves out populations from Oceania who have an extra complication due to their substantial Denisova admixture). We infer that 96% of sites with γ_i>0.01 and nearly 100% of sites with γ_i>0.9 in non-Oceanian non-Africans are reflecting a history of archaic introgression. We caution that enrichment is not the same as introgression: some sites with γ_i values in this range in non-African haploid genomes are not in fact introgressed, but only near introgressed segments. Thus in what follows, we focus on sites with γ_i>0.9, corresponding to a nominal >90% probability of being introgressed according to our HMM.

We next varied the archaic samples used for fishing out archaic ancestry, as well as the set of samples used as outgroups. Table S13.4 compares the results of five experiments in which we screened for introgressed fragments using Altai, Denisova, or both as baits, and varied the outgroup panel. All these experiments provide extraordinary enrichment in non-Oceanian non-Africans compared with Africans, suggesting that the HMM is genuinely detecting archaic ancestry. However, the HMMs vary in the proportion of the genome that they call confidently introgressed (0.02% to 0.65% averaging across the six samples). Even for the most sensitive of the HMMs, based on pooling Altai and Denisova as the archaic baits, the proportion called as confidently introgressed is below the proportion of Neandertal ancestry in non-Oceanian non-Africans that we estimate in Table S14.8 of SI 14 (1.72 ± 0.12% in Europeans and 1.89 ± 0.13% in eastern non-Africans). This reflects the fact that the sensitivity of our HMMs to true Neandertal-derived segments is far from perfect, and poorer than for more sophisticated methods like the one reported in the accompanying paper by Sankararaman et al.[12]

An important feature of Table S13.4 is that when we use Denisova as bait for fishing out archaic material in the HMM rather than Altai, we identify an excess of segments in Oceanian populations compared with the other phased non-Africans (average of 0.93% of the genome versus 0.11%). The excess becomes even shaper when we include Altai in the panel of outgroups so that we are effectively filtering out Neandertal introgressed fragments from the segments inferred by the HMM (average of 0.55% of the genome versus 0.02%). Thus, with our HMM we have not only a method for calling Neandertal introgressed fragments, but also a method for calling potential Denisovan ones.

Table S13.4: Genome introgressed (γ_i>0.9) using different baits and outgroups (parts per 100,000)

Archaic sample used for fishing	Outgroup panel	African				European		Eastern non-African				Oceanian			% of sites in non-Africans due to introgression*	% of Oceanian sites from non-Neandertal introgression*
		Mbuti	Yoruba	San	Mandenka	French	Sardinian	Dai	Mixe	Han	Karitiana	Papuan	Australian1	Australian2		
Altai	Africans	5	2	27	19	563	511	525	684	590	677	1058	1144	1077	100%	52%
Altai	Afr.+Den.	3	0	17	14	437	386	419	554	468	563	734	819	778	100%	47%
Den.	Africans	2	0	16	3	91	89	112	134	97	122	914	964	919	100%	90%
Den.	Afr.+Altai	0	0	6	0	7	3	32	23	23	27	505	593	546	100%	100%
Both	Africans	7	3	42	19	601	545	607	756	652	749	1872	2026	1948	100%	71%

Note: Entries are in parts per 100,000. We use *s=0.05, p=0.01, u=0.99*, and pool data from the two haplotypes. We call a segment as introgressed if it is assigned γ_i>0.9 by the HMM.
* The non-African enrichment over Africans is computed as in Table S13.3. The Oceanian enrichment over non-Africans is computed as $(G_{j,Oceanian} - G_{j,Europe})/G_{j,Oceanian}$.

(vi) Confirmation of signal of less Neandertal ancestry in Europe than in Eastern non-Africans

Table S13.4 shows that when we use Altai as the bait to pull out introgressed segments and Africans+Denisova as outgroups (to screen out both modern human and Denisova ancestry and highlight the Neandertal material), the proportion of the genome called as confidently introgressed in Europe is French=0.44% and Sardinian=0.39%, which is only modestly lower than in eastern non-Africans where it is Dai=0.42%, Mixe=0.55%, Han=0.47% and Karitiana=0.56%. If we compute the ratio of inferred archaic ancestry in the pool of 2 Europeans to that in the pool of 4 Eastern non-Africans and compute a standard error on this quantity using a Block Jackknife (50 equal-sized bocks), we obtain a 95% confidence interval of 71-93%. (The 95% confidence interval for the HMM using Altai as bait and just Africans as the outgroup is 75-99%.)

Table S13.5 compares estimates of the relative proportions of European to eastern non-African archaic ancestry from the analyses of this study (SI 13 and SI 14) to other estimates of this ratio. For the local ancestry inference based methods, the ratio of the Neandertal proportion in Europe to that in East Asia reported here (95% CI of 71-93%) is not quite as low as was inferred by Wall et al.[11] who reported a point estimate of 67%. It overlaps with the local ancestry inference range obtained by

Sankararaman et al. (76-88% over all 15 pairwise population comparisons[12]). For the methods based on genome-wide *S*-statistics or *f*-statistics, the ratio of the Neandertal proportion in Europe to that in East Asia reported in this study (95% CI of 82-110%; SI 14) is consistent with that reported in the analysis of the high coverage Denisovan genome1 (95% CI of 64-88%; Table S26), and potentially consistent with the point estimate of 71% computed based on the statistics reported in Wall et al[11] if we recognize that there was unreported statistical uncertainty in that estimate.

Table S13.5: Archaic ancestry estimates are slightly lower in Europeans than in eastern non-Africans

Method	Type of method	Ratio in Europe to East	Reference
% of genome called as introgressed	Local ancestry inference	71-93% *	This study (SI 13)
% of genome called as introgressed	Local ancestry inference	76-88% †	Sankararaman et al.[12]
% of genome called as introgressed	Local ancestry inference	67%	Wall et al.[11]
Ratio of non-enhanced S-statistics	Genome-wide statistics	79-112% *	This study (SI 14)
Ratio of enhanced S-statistics	Genome-wide statistics	64-88% *	Meyer et al.1 (Table
% of genome called as introgressed	Genome-wide statistics	71%	Wall et al.[11]

Note: No confidence interval for the Wall et al. study is reported so we only give a point estimate.
* 95% CI from a Block Jackknife. For the SI 13 analysis, we use the ratio of the % of the genome called as introgressed in the 2 Europeans to the 4 eastern non-Africans using the HMM with Altai as bait and Africans+Denisova outgroups.
† Range of the Europe/East ratio observed over all pairwise population comparisons in the 1000 Genomes Data.

(vii) Evidence for Denisovan introgression into eastern non-Africans

We found that when we run our HMM with Denisova as the archaic bait and present-day Africans+Altai as the outgroup panel, the rate of European or African phased genomes called as archaic is less than 1 part in 10,000 (Table S13.4). However, the same method identifies some of the genome as confidently introgressed in all four eastern non-African samples (Dai 0.032%, Han 0.023%, Mixe 0.023%, Karitiana 0.027%). The proportion of the genome called as Denisovan introgressed by this HMM is extremely small: 4.8 ± 0.9% of that called as Denisovan in the 3 Oceanian genomes. This is too small to be confidently detectable by F_4 *Ratio Estimation* and *D*-statistics which have standard errors of a few tenths of a percent, which perhaps explains why we did not detect this signal previously1. If this is a real signal, we are detecting it here because the local ancestry inference method provides more sensitivity than genome-wide statistics.

Table S13.6: Tests for whether segments called as introgressed are closer to Altai or Denisova

D(Den,Alt;X,Y)	*Bait:* / *Outgroup:*	Altai Africans	Denisova Africans	Both Africans	Altai Afr.+Den.	Denisova Afr.+Altai
X=Europe Y=Chimp	n_{BABA}	348	318	786	270	214
	n_{ABBA}	15,403	946	16,011	14,255	8
	D	-0.956	-0.497	-0.906	-0.963	0.930
	Std. Err.	0.005	0.104	0.011	0.005	0.022
X=East Y=Chimp	n_{BABA}	762	1,706	2,685	517	1,412
	n_{ABBA}	27,780	1,639	29,238	25,901	48
	D	-0.947	0.020	-0.832	-0.961	0.934
	Std. Err.	0.006	0.090	0.019	0.005	0.019
X=Oceania Y=Chimp	n_{BABA}	2,121	28,401	31,583	587	25,285
	n_{ABBA}	33,205	3,227	35,605	31,065	556
	D	-0.880	0.796	-0.06	-0.963	0.957
	Std. Err.	0.013	0.020	0.061	0.004	0.003

To further investigate this signal, we computed *D*-statistics restricting to subsets of the genome where we had confidently called archaic ancestry (γ_i>0.9), and repeating the analysis for each of the five HMMs described in the previous section. For this analysis, we divided our phased haplotypes into

three population samples: 6 Oceanian haplotypes ("Oceanian"), 4 European haplotypes ("Eur"), and 8 eastern non-African non-Oceanian haplotypes ("East"). We reported the derived allele frequency in the population sample at each position where there was at least one archaic haplotype called. When only one haplotype was called as archaic, the derived allele frequency was always 0 or 1; when more than one haplotype was called, the frequency could be intermediate.

Table S13.6 shows *D*-statistics of the form *D(Denisova, Altai; Introgressed, Chimp)*, which can evaluate whether the non-African haplotypes called introgressed by our HMM share more derived alleles with Denisova or with Altai. On average only 1.9% of the genome is included in the *D*-statistic computations (because such a small fraction of the genome is called as introgressed). The average amount of genome used in each cell is 52 Mb, and all cells are based on at least 2 Mb except for the analysis that fishes for archaic segments in Europeans using Denisova as bait and Africans+Altai as outgroups (this cell only reflects 297 kb of data, and we hypothesize that the segments that the HMM captures in this cell contain many false-positives: segments of Neandertal or modern human genomes that simply match Denisova due to incomplete lineage sorting). We compute standard errors with a Block Jackknife, using 50 contiguous blocks each with about an equal number of screened bases.

- *When Altai is used as the bait to enrich for Neandertal archaic segments, the fragments called archaic are similar in their ancestry characteristics in Europeans and eastern non-Africans*

  The first column in Table 13.5 uses Altai only as bait and Africans as outgroups. When we test the rate at which haplotypes identified by this HMM share derived alleles with Denisova vs. Altai, we observe statistically similar *D*-statistics for Europeans and eastern non-Africans: D_{Eur} = -0.956 ± 0.005 (one standard error) and D_{East} = -0.947 ± 0.006. This is consistent with the hypothesis that the archaic material pulled out in both population is of the same (Neandertal) origin. In contrast, the *D*-statistic is significantly smaller for Oceanians ($D_{Oceanian}$ = -0.880 ± 0.013), consistent with their having additional archaic material related to Denisovans1 that is (inefficiently) pulled out by the HMM and that has a different historical relationship to the Altai and Denisova genomes.

  The fourth column in Table S13.6 again reports results using Altai only as the bait, but now using both Africans and Denisova as outgroups. Thus, the material pulled out by this HMM is enriched for exclusively Neandertal ancestry. Consistent with the Denisovan material being screened out, we observe consistent *D*-statistics for the material called as introgressed in all including Oceanians: D_{Eur} = -0.963 ± 0.005, D_{East} = -0.961 ± 0.005 and $D_{Oceanian}$ = -0.963 ± 0.004.

- *When Denisova is used as the bait to enrich for Denisovan archaic segments, Denisova is distinctly closer to segments called as introgressed in eastern non-Africans than in Europeans.*

  The second column in Table S13.6 uses Denisova as bait and Africans only as outgroups. For Oceanians, the segments identified by this HMM more closely match Denisova than Altai ($D_{Oceanian}$ = 0.796 ± 0.020), consistent with previous reports of Denisovan related genetic material in Oceania[1,18]. For Europeans, we see the opposite signal consistent with all the material being of Neandertal origin, even when we enrich for any Denisova-related material that might be present by using Denisova as a bait (D_{Eur} = -0.497 ± 0.104 in the direction of closer matching to Altai). What is striking is that when we repeat the computation in eastern non-Africans, D_{East} = 0.020 ± 0.090, reflecting closer matching to Denisovans. This is consistent with some fraction of the archaic material in eastern non-Africans being of non-Neandertal origin.

  The final column of the table again uses Denisova as bait, and uses both Altai and Africans as outgroups to nearly completely eliminate Neandertal segments. In segments of the genome identified by this analysis, Europeans, eastern non-Africans, and Oceanians have similar *D*-statistics, but this is not unexpected (even if there are no real Denisovan segments in Europeans) since this HMM is specifically is pulling out segments that are phylogenetically closer to Denisova than Altai, biasing the statistics to be very skewed toward Denisova as we observe.

To test formally whether different populations harbor more archaic ancestry than others, we computed the difference between the proportion of the genome called as archaic for all pairwise combinations of non-Africans (Table S13.7). Four features of Table S13.7 are notable:

Table S13.7: Z-scores for differences in the proportion of the genome called as introgressed

			Bait:	Altai	Denisova	Den+Altai	Altai	Denisova
	Pop.1	**Pop.2**	*Out:*	**African**	**African**	**African**	**Afr+Den**	**Afr+Alt**
Europe (within)	French	Sardinian		1.3	0.1	1.3	1.6	1.5
East (within)	Dai	Han		-1.6	1.0	-1.0	-1.3	0.8
	Dai	Mixe		-3.1	-1.0	-2.7	-3.1	0.7
	Dai	Karitiana		-3.0	-0.5	-2.7	-3.3	0.4
	Han	Mixe		-2.1	-2.4	-2.2	-2.1	0.0
	Han	Karitiana		-1.6	-1.7	-1.7	-1.9	-0.3
	Mixe	Karitiana		0.1	0.6	0.1	-0.2	-0.4
Oceania (within)	Papuan	Australian1		-1.3	-1.0	-1.8	-1.4	-2.4
	Papuan	Australian2		-0.3	-0.1	-0.8	-0.9	-1.0
	Australian1	Australian2		1.0	0.7	0.8	0.7	1.0
Europe-East	French	Dai		0.7	-1.1	-0.1	0.5	-2.7
	French	Han		-0.5	-0.4	-0.9	-0.7	-2.0
	French	Mixe		-2.3	-2.5	-2.7	-2.8	-2.0
	French	Karitiana		-2.3	-1.8	-2.9	-3.2	-2.0
	Sardinian	Dai		-0.2	-1.1	-1.0	-0.7	-8.5
	Sardinian	Han		-1.6	-0.4	-2.0	-2.0	-7.1
	Sardinian	Mixe		-3.1	-2.2	-3.5	-3.6	-6.3
	Sardinian	Karitiana		-3.1	-1.7	-3.7	-4.1	-6.0
	Europe	Dai		0.2	-1.2	-0.6	-0.1	-4.7
	Europe	Han		-1.1	-0.4	-1.6	-1.5	-3.7
	Europe	Mixe		-3.0	-2.5	-3.5	-3.6	-3.5
	Europe	Karitiana		-3.1	-1.9	-3.7	-4.1	-3.3
	French	East		-1.3	-1.7	-2.0	-1.9	-2.4
	Sardinian	East		-2.3	-1.6	-2.9	-3.0	-7.9
	Europe	East		-2.1	-1.9	-2.9	-2.9	-4.3
Europe-Oceania	French	Papuan		-6.3	-30.6	-13.1	-4.9	-59.0
	French	Australian1		-6.8	-30.0	-14.4	-5.7	-68.5
	French	Australian2		-7.0	-32.3	-15.1	-5.8	-65.2
	French	Oceanian		-7.1	-31.5	-14.7	-5.9	-64.4
	Sardinian	Papuan		-8.1	-28.0	-15.5	-6.8	-177.2
	Sardinian	Australian1		-8.6	-28.8	-16.7	-7.4	-197.2
	Sardinian	Australian2		-8.4	-30.3	-16.5	-7.4	-186.6
	Sardinian	Oceanian		-9.0	-29.5	-16.9	-7.9	-187.7
	Europe	Oceanian		-7.6	-31.9	-15.3	-6.2	-102.3
	Europe	Papuan		-8.1	-31.8	-16.6	-6.9	-116.8
	Europe	Australian1		-8.3	-34.6	-17.1	-7.1	-112.7
	Europe	Australian2		-8.7	-33.4	-17.0	-7.4	-111.1
East-Oceania	Dai	Papuan		-7.2	-28.9	-14.3	-5.3	-27.8
	Dai	Australian1		-7.4	-28.1	-14.5	-5.9	-30.4
	Dai	Australian2		-8.4	-27.4	-17.0	-6.3	-27.7
	Dai	Oceanian		-8.1	-28.9	-15.7	-6.2	-28.9
	Han	Papuan		-7.3	-36.5	-14.2	-5.0	-46.7
	Han	Australian1		-7.4	-33.3	-14.1	-5.6	-56.1
	Han	Australian2		-7.9	-36.5	-16.4	-6.0	-49.3
	Han	Oceanian		-8.3	-36.3	-15.5	-6.1	-51.6
	Mixe	Papuan		-4.9	-22.5	-11.5	-3.0	-38.0
	Mixe	Australian1		-4.9	-21.7	-11.6	-3.4	-44.1
	Mixe	Australian2		-5.4	-22.6	-13.6	-3.6	-39.3
	Mixe	Oceanian		-5.4	-22.7	-12.7	-3.6	-40.9
	Karitiana	Papuan		-4.9	-25.4	-12.3	-2.6	-23.9
	Karitiana	Australian1		-5.1	-23.5	-12.0	-3.2	-27.8
	Karitiana	Australian2		-5.3	-24.0	-12.9	-3.2	-25.5
	Karitiana	Oceanian		-5.5	-24.8	-12.9	-3.2	-25.9
	East	Papuan		-7.0	-35.9	-15.3	-4.6	-52.7
	East	Australian1		-6.9	-31.7	-14.8	-4.9	-57.7
	East	Australian2		-8.0	-34.1	-18.1	-5.5	-51.0
	East	Oceanian		-8.0	-35.1	-16.9	-5.6	-55.1

Note: We compute a Z-score for the difference between the proportion of the genome called introgressed in Pop1 - Pop2. Standard errors are from a Block Jackknife. Values ≥4 standard errors from zero are highlighted.

(a) For the HMM with Denisova as bait and Africans+Altai as outgroups, the analysis confirms a significant excess of Denisovan-related ancestry in Oceania compared with mainland Eurasia (|Z|=111 standard errors for Europe-Oceania and |Z|=55 standard errors for East-Oceania).

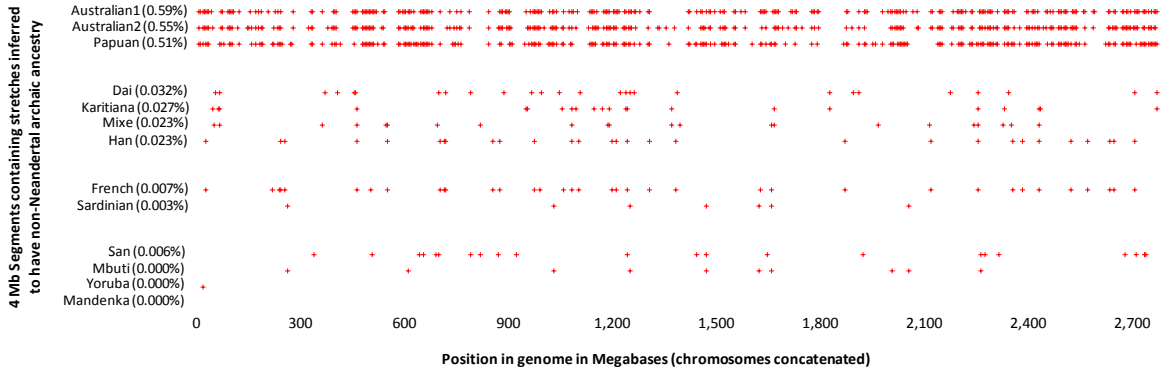
(b) For the HMM with Denisova as bait and Africans+Altai as outgroups, the analysis also confirms a significant excess of Denisovan-related ancestry in the pool of 4 eastern non-Africans compared with the pool of 2 Europeans (Z=|4.3| standard errors). A potential pitfall in this analysis is that the West Eurasian gene flow into the Yoruba population (of at most a couple of percent as documented in section (iv) above) biases our outgroup panel to be genetically closer to Europeans than to eastern non-Africans, which in theory could cause an artifactual finding of more archaic ancestry by our method in one group than the other. To test whether this could be influencing our results,

we reran our HMM now adding a Han Chinese individual (from the B Panel) to the outgroup panel used for the HMM, so that the outgroup panel had approximately 2% of its ancestry being Han Chinese, as large or larger than the estimated West Eurasian contribution to Yoruba. The results of this analysis confirm a significant excess of Denisovan-related ancestry in the pool of 4 eastern non-Africans compared with the pool of 2 Europeans (Z=|3.4| standard errors). Figure S13.7 presents a plot across the genome showing the patterns of Denisovan introgression, visually confirming the excess in Eastern non-Africans compared with Europeans.

(c) For the HMM with Altai as bait and Africans+Denisova as outgroups, the analysis provides weak confirmation of a significant excess of Neandertal-related ancestry in the pool of 4 Eastern non-Africans compared with the pool of 2 Europeans (Z=|2.9| standard errors).

(d) For the HMM with Altai as bait and Africans+Denisova as outgroups, the analysis suggests a significant excess of Neandertal-related ancestry in the 3 Oceanian samples compared with the 4 Eastern non-African samples (Z=|5.6| standard errors). One possibility is that this pattern is due to genuinely more Neandertal ancestry in Oceanian populations, which is an important signal if confirmed. However, we cannot currently rule out the possibility that some truly Denisovan ancestry in Oceanians is being misclassified as Neandertal by the local ancestry inference engine, (even with the inclusion of Denisova as outgroup) because the introgressing Denisovan population is not so closely related to the Siberian one and hence far from a perfect outgroup. Furthermore, as shown in SI 15, some of the Denisovan genome is of Neandertal ancestry, which might further compromise its usefulness for screening out true Denisova segments.

Figure S13.7: Plot of Denisovan-related archaic segments for the 13 phased genomes. *We divide the genome into 4 Mb segments, run the HMM with Denisova as bait and Africans+Altai as outgroups, and plot a point for each segment containing any locus in which either of the inferred haplotypes has γ_i>0.9. The results show an excess of archaic ancestry in the order Oceania > East > Europe. The actual proportion of inferred ancestry is shown in parentheses beside each sample name.*



Under the assumption that the segments of the genome called as introgressed by the HMM using Denisova as bait and Africans+Altai as outgroups are of Denisova origin, we estimated the ratio of Denisovan ancestry in each non-Oceanian non-African population as a fraction of that in Oceanians. Table S13.8 gives these results along with standard errors from a Block Jackknife. It is important to recognize that false positives in the local ancestry inference (sections of the genome erroneously called as introgressed) will have a larger proportionate impact on populations with small proportions of Denisovan ancestry than in populations with larger proportions like Oceanians. If we assume that Europeans represent a baseline with truly 0% Denisovan admixture, then the proportion of Denisovan ancestry in eastern non-Africans is 3.8% = 4.8% - 1.0% of that in Oceanian populations. An alternative possibility is that the segments in Europeans that we call as Denisovan include some true fragments of Denisovan ancestry. A plausible scenario by which this could occur is that these fragments owe their origin to gene flow from northeast Asia into Europe that is known to affect the French to a greater extent than Sardinians[19] (this gene flow might explain why we detect more than 2-fold more Denisovan fragments in the French than in Sardinians (Table S13.4) although the excess is not statistically significant; Z=1.3 (Table S13.7)). The proportion of this ancient northeast Asian ancestry in the French has recently been estimated[20] to be in the range 18-39%, and hence could have

brought a substantial amount of Denisovan ancestry into this group, consistent with the fact that we in practice observe 0.007% inferred Denisovan segments in the French compared with about four times that in the eastern non-Africans (0.023-0.032%) (Figure S13.7).

Integrating over all these scenarios, we can assume that the true proportion of Denisovan ancestry in Eastern non-Africans as a fraction of that in Papuans is 3.8-4.8%. Multiplying by a previously published point estimate of 5.1% Denisovan ancestry in Oceanian populations[1,18], this corresponds to about 0.19-0.24% of eastern non-African genomes deriving from Denisovans.

Table S13.8: Amount of genome called as Denisovan-related as a fraction of that in Oceania

	Denisova ancestry estimate as a fraction of Oceania	**Standard error**
French	1.4%	0.8%
Sardinian	0.6%	0.3%
Europe (Sardinian+French)	1.0%	0.4%
Dai	5.8%	1.7%
Han	4.2%	1.0%
Mixe	4.2%	1.2%
Karitiana	4.9%	1.9%
East (Dai+Han+Mixe+Karitiana)	4.8%	0.9%

Note: Fraction of genome called as introgressed using Denisova as bait and Africans+Altai as outgroups.

We note that this is not the first study to show evidence for Denisovan introgression in eastern non-Africans: Skoglund and Jakobsson also suggested more Denisova genetic material in eastern non-Africans than in Europeans using genome-wide *D*-statistics[21]. The Skoglund and Jakobsson result contrasted with two previous studies that we published that tested for such a signal, and failed to find any evidence of Denisovan ancestry in mainland Eurasia1[,18]. We do not understand why Skoglund and Jakobsson's study had the power to detect a signal while ours did not given that we believe that the statistics we and they previously used had similar power. Nevertheless, the fact that they made an inference that we are validating in this new study suggests that they did detect a true signal. The Skoglund and Jakobsson paper also suggested more Denisovan ancestry in southeast Asians like the Dai than in northeast Asians and Native Americans like Han, Karitiana and Mixe[21]. When we test for this using our local introgression analysis, our point estimate of Denisovan ancestry is higher in the Dai than in the other three samples, although the excess is not significant given our current resolution: Z=1.2 for Dai-Han, Z=0.9 for Dai-Mixe, and Z=0.6 for Dai-Karitiana; Table S13.7).

(viii) Upper bound on population split times exploiting the archaic segments

Our HMM identifies segments of archaic ancestry using mutations that are ancestral to both the test haplotype and the most closely related archaic haplotypes used as bait (Figure 13.5A). Mutations that occurred since the split of these two haplotypes (on lineages *X* and *Y* in Figure 13.5A), or on the outgroup side of the tree (e.g. lineage *Z*), were not used in ascertainment, and thus can provide unbiased estimates of the time elapsed along these lineages.

The introgressing Neandertal and Altai populations split <176 kya assuming $\mu=0.5\times10^{-9}$/bp/year

We began by examining the segments of the genome identified as likely to be of Neandertal ancestry ($\gamma>0.9$) in 9 non-African samples (French, Sardinian, Karitiana, Mixe, Han, Dai, Papuan, Australian1 and Australian2) based on the HMM with Altai as bait and the sub-Saharan African pool as outgroup (we did not use Denisova to screen out Denisovan segments because if we are capturing Denisovan segments by our method they will not interfere with our estimate of the minimum population split date for Neandertals and introgressing Neandertals which is the main goal of the present section). We took the ~0.5% of the genome confidently identified as archaic in each of the 9 samples (Table S13.4), and then combined these haplotypes and filtered and ranked them as follows:

(a) We scanned through the autosomes in the p-arm to the q-arm direction. We started each analysis window at a position where $\gamma>0.9$, and then ended the window either because we encountered a site with $\gamma\leq0.9$, or because the distance from the start to the end of the window exceeded >0.01 cM (to restrict to regions that are unlikely to be disrupted by recombination over the last few hundred thousand years[15]). We then started the next window.

(b) We further restricted analysis to windows where at least 25,000 base pairs passed filters, and where at least 60% of these sites had alignment to macaque.

This procedure identified 1,784 haplotypes of at least 25,000 bp (mean of 44,078 bp), which altogether covered 78.6 Mb. These haplotypes are mostly independent, in the sense that 76% of them are completely non-overlapping with each other.

We sorted the haplotypes inferred to be Neandertal-introgressed by the haplotype-Neandertal divergence per base pair on the Altai side of the tree (lineage *X* in Figure S13.5A). To obtain an unbiased estimate of the coalescence time of these lineages for the sum of all haplotypes below a specified fraction of haplotypes on the list, we used the sum of all divergent sites from the test haplotype side of the tree, which we carefully did not use for ascertaining the regions so that these data can be used to provide an unbiased estimate of the population split time. We expressed this quantity as a fraction of local human-chimpanzee divergence over the same set of screened bases. We then further multiplied by the ratio of (local test haplotype-chimp genetic divergence)/(local chimp-macaque divergence) divided by the genome-wide ratio of this quantity $(n^j_{HC}/n^j_{CM})/(Div^j_{HC}/Div^j_{CM})$, to adjust for local variation in mutation rate (this is the same procedure used to correct for local mutation rate variation in section (iii) above). For estimates in years, we multiplied by 6.5-13 Mya for human-chimpanzee divergence, covering the range of uncertainty in the current literature[22].

Figure S13.8: Divergence between archaic material in present-day humans and ancient genomes. *We compute the divergence of Altai to the Neandertal segments in 9 non-Africans using Altai as bait and Africans as outgroups (blue lines). We also compute the divergence of Denisova to the Denisovan segments fished in Oceania using Denisova as bait and Altai+Africans as outgroups (red lines). The dip for the Denisova curve at the far left is due to windows with low divergence to Denisova on both sides of the tree which appear to be Neandertal segments falsely classified as Denisovan, perhaps because the Denisova genome we use for fishing has some Neandertal introgression (SI 15). We can filter out many of these likely cases of false classification by removing the 1% of the genome of lowest divergence between the inferred archaic haplotypes in Oceanians and Denisova measured on the Denisova side of the tree (purple lines). All points on these lines are genetic divergences, so they are upper bounds on population divergence. Conversions to dates in years are shown on the y-axis.*

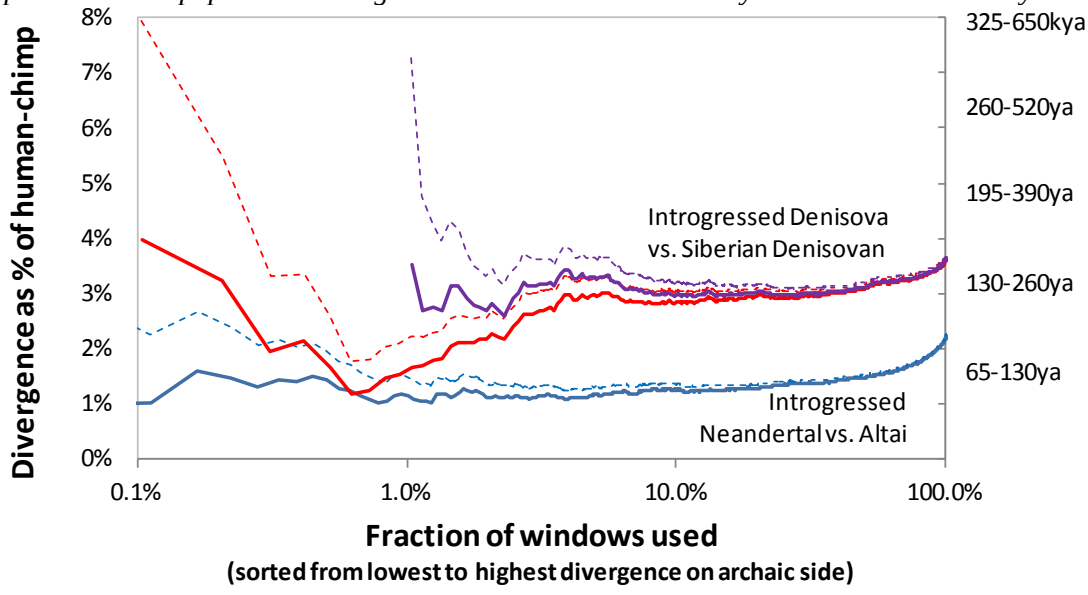


Figure S13.8 shows the results. We compute a 95% confident upper bound (dashed line) at each point on the curve. At the far left, the statistical error is large because there is little data and thus the upper bound is not a strong constraint. At the far right, we are including segments of the genome that coalesce substantially earlier than the population divergence time so there is not a strong constraint even though we have large amounts of data. Focusing on the data from the bottom 20% of windows which is enough to drive down the standard error to a low value, we conclude that the test haplotype – Altai divergence is less than 1.35% of human-chimpanzee genetic divergence, corresponding to an upper bound on the population split of <88 kya assuming a mutation rate of μ=1.0×10^{-9}/bp/year, and <176 kya assumingμ=0.5×10^{-9}/bp/year. This post-dates the time that Neandertals appeared in the non-African paleontological record and also post-dates the time that Neandertal and anatomically modern human forms had begun differentiating morphologically by several hundred thousand years[23]. Thus, the archaic material we see in non-Africans is from Neandertals and not due to ancient sub-structure predating the split of Neandertals and Africans.

The introgressing Denisovan and Siberian Denisovans split <394 kya assuming μ=0.5×10^{-9}/bp/year

We analyzed Denisovan segments identified in the Oceanian phased genomes (Papuan, Australian1, and Australian2) by using Denisova as bait and Africans+Altai as outgroups. As shown in Table S13.4, this procedure screens out the great majority of Neandertal introgressed segments (European phased genomes have essentially no Neandertal introgressed segments when this method is applied), and identifies 0.55% of the Oceanian genomes as confidently introgressed from proto-Denisovans. Restricting to haplotypes of no more than 0.01cM and covering at least 25,000 screened bases, we identified 962 putative Denisovan-introgressed haplotypes with a mean size of 43,823 bp and altogether covering 42 Mb. As above, we rank ordered these segments by divergence on the archaic side of the tree, and then plotted the cumulative divergence on the present-day human side of the tree.

Figure S13.8 shows the results. We first observe what is likely to be an artifact, which is that the divergence between the inferred Denisovan segments in Oceanian populations and the Siberian Denisovan genome is very low in the 1% of windows of lowest divergence measured on the Denisova side of the tree. In theory this observation could be due to genuine close relatedness of Denisovan introgressed segments to the Denisovan ancestry in the Siberian Denisovan individual. However, we believe that a more likely explanation is that these segments correspond to segments of the Siberian Denisovan genome that are actually of Neandertal ancestry (SI 15), so that our HMM identifies Neandertal rather than Denisovan segments in Oceanian genomes and thus infers divergence dates that are much younger for these segments (corresponding to the date of divergence of Altai to the introgressing Neandertal discussed in the previous section). While such errors of misclassification are rare as shown in Table S13.4, their effect is greatly enhanced by our analysis which focus on the extreme tail of low divergence to the Denisova genome, and so they could plausibly have an effect. To be conservative, we therefore filtered out the 1% of the genome of lowest divergence measured on the archaic side of the tree. (We note that applying the same filtering to the previous analysis would not substantially change the date estimate for the population split of the introgressing Neandertal and Altai.) This still provides a valid upper bound on population divergence between the Siberian Denisovan and the introgressing Denisova material since we are computing this using data from the present-day human side of the tree not used for ascertainment.

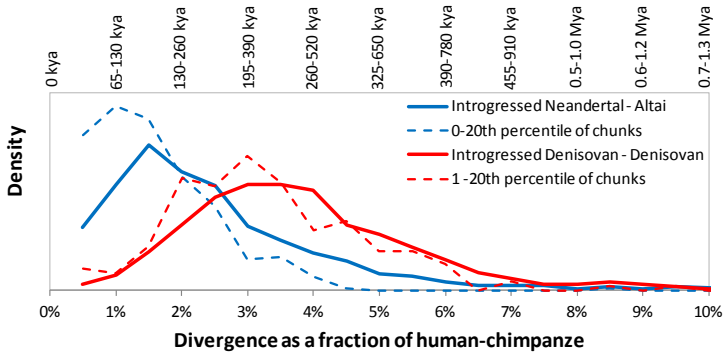


Figure S13.9: Histogram of divergence of archaic introgressed segments to other genomes. *We measure on the archaic side of the tree. We analyze all segments, and the fifth of lowest divergence (excluding the 1% of lowest divergence for Denisova because of misclassification in this subset of the data).*

Filtering out the 1% of windows of lowest divergence measured on the archaic side of the tree, and then focusing on windows between the 1st and 20th percentiles, we find that the upper bound on the divergence between the proto-Denisova material in the Denisova genome and the proto-Denisova material in Oceanian genomes reaches a minimum of 3.18% of human-chimpanzee divergence; that is, about two times as old as the upper bound for the divergence of Altai and the introgressing Neandertal material. This corresponds to <197 kya assuming a mutation rate of μ=1.0×10^{-9}/bp/year, and <394 kya assumingμ=0.5×10^{-9}/bp/year. We conclude that the Siberian Denisovan is consistent with being more distantly related to the Denisova-related genetic material in Oceanians than the Altai Neandertal is to the Neandertal-related genetic material in all non-Africans.

Unbiased estimates for population divergence times: 77-114 kya for Altai & introgressing Neandertals and 276-403 kya for the Siberian Denisovan & introgressing Denisovans assuming μ=0.5×10^{-9}/bp/yr

The method above provides a valid upper bound on population divergence time, as the introgressed segments are all guaranteed to have a genetic divergence time that is older than the population split time. However, we would ideally like to obtain an unbiased estimate of population divergence time.

To obtain unbiased estimates of the population divergence time, we analyzed the histograms of the genetic divergences of the introgressed archaic segments to various other genomes (Figure S13.9). Two of the histograms correspond to the same data used in Figure S13.8 (introgressed Neandertal divergence to Altai, and introgressed Denisova divergence to the Siberian Denisovan). The third histogram corresponds to the 20% of windows of inferred Neandertal ancestry in 9 non-Africans that are likely to have the lowest divergence time between Altai and the Neandertal introgressed haplotypes as none of them have any divergent sites measured on the archaic side of the tree. The fourth histogram corresponds to the windows of inferred Denisovan ancestry in 3 Oceanian individuals that are in the 1st to 20th percentiles of divergence between Denisovan and the Denisovan introgressed haplotypes as measured on the archaic side of the tree (we exclude the segments below the 1st percentile of divergence because of the possibility that they might be Neandertal segments falsely classified as Denisovan as discussed in detail in the previous section).

We assume that the histograms reflect processes in which the introgressed segments are drawn from a true distribution with a fixed start time (T = the population divergence we want to estimate), and a tail of older divergence. To be concrete, this might be an offset exponential of the form that would be expected for a simple population split at time T, and a constant diploid population size N ancestral to the split. However, the observed histogram does not visually look like such a function, which we attribute to the limited size of the segments we are analyzing, which results in stochasticity. The observed histogram is a convolution of the true distribution of the genetic divergence time $f(t)$, and the stochastic sampling process corresponding to how many mutations actually occurred at each locus. We therefore fit a model in which the true distribution of coalescence times has the following form:

$$f(t) = 0 \quad \text{if } t < T$$
$$f(t) = e^{-(t-T)/2N} \quad \text{if } t \geq T \qquad \text{(S13.1)}$$

If there were not an ascertainment bias in the way that our segments were identified, the value of the effective population size N would be historically interesting. However, the way that our introgressed segments were ascertained means that they do not coalesce too much more anciently than the population split, and thus we are likely to underestimate the true value of N by our method. Thus, N is a nuisance parameter that we are not concerned with as our goal is to estimate T.

We also need to model stochasticity in our data due to the limited number of mutations observed per window: Given a locus with divergence time t (generated by $f(t)$), the observed number of mutations on the measured side of the tree is expected to be Poisson distributed with parameter $e_{HC}t$:

$$d = \text{Poisson}(e_{HC}t) \qquad \text{(S13.2)}$$

In this equation, t is the genetic divergence expressed as a fraction of genome-wide average human-chimpanzee divergence. In addition, e_{HC} is the number of divergent sites that would be expected for a pair of lineages that coalesced at the date of genome-wide average human-chimpanzee divergence. We infer this quantity in practice by multiplying the number of screened nucleotides at the locus by the genome-wide average rate of human-chimpanzee divergence, and making a correction for locus-specific variation in local mutation rate and human-chimp divergence by dividing local human-chimp divided by chimp-macaque divergence per base pair by the genome-wide value of this quantity.

The Appendix of this note describes in practice how we deconvoluted the distributions computationally. This procedure produces a standard error, which we use to infer a 95% confidence interval as a fraction of human-chimpanzee divergence as shown in Table S13.9. We can then convert this to absolute dates by making an assumption about human-chimp average genetic divergence time.

We applied this procedure to our data. For Altai divergence to 9 non-African phased genomes, we report results based on the lowest 20 percent of windows in divergence measured on the archaic side of the tree. For the Siberian Denisovan divergence to 3 Oceanian phased genomes, we report results based on the 1st through the 20th percentiles of divergence measured on the archaic side of the tree, because of the evidence for misclassification of truly Neandertal segments as Denisovan segments in the lowest percentile (see above) Our population divergence time estimates have the following notable features that are all consistent with the upper bounds on population divergences reported above:

- The unbiased estimate for the population divergence of Altai and the introgressing archaic material in non-Africans is 38-57 assuming a mutation rate of $\mu=1.0\times10^{-9}$/bp/year, and 77-114 kya assuming a mutation rate of $\mu=0.5\times10^{-9}$/bp/year. These results indicate that the introgressing population was definitely Neandertal, as it is genetically related to the Altai Neandertal within the time frame in which late Neandertals lived.

- The unbiased estimate of the population divergence of the Siberian Denisovan and the introgressing archaic material in Papuans is 2.12-3.10%, corresponding to 138-202 kya assuming a mutation rate of $\mu=1.0\times10^{-9}$/bp/year, and 276-403 kya assuming a mutation rate of $\mu=1.0\times10^{-9}$/bp/year. This suggests that the Denisovan material in Papuans is not particularly closely related to that in Siberian Denisovans. By comparison, the Altai-Denisova divergence is estimated to be only modestly larger, at 2.86-3.41% as described in Table 1 of the main text and Table S12.2 and Table S12.3 of SI 12.

Table S13.9: Unbiased estimates of pop. divergence between introgressing and sequenced archaics

Comparison	Data	Seg-ments	% of HC*	†Kya assume Div_{HC}=6500	†Kya assume Div_{HC}=13000
Altai – introgressed segments, 9 non-Africans	0th to 20th percentile divergence measured on archaic side of tree	356	0.59-0.88%	38-57	77-114
Denisova – introgressed segments, 3 Oceanians	1st to 20th percentile divergence measured on archaic side of tree	183	2.12-3.10%	138-202	276-403

* The 95% confidence interval is obtained from the point estimate ± 1.96 times the standard error using the method in the Appendix.
† The range of absolute times is obtained by multiplying the estimates of human-chimp divergence by 6.5-13 Mya.

References

[1] Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV,

Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222-6.
[2] Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE, Shendure J (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol. 29, 59-63.
[3] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.
[4] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43, 491-8.
[5] Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PL, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S (2010) A draft sequence of the Neandertal genome. Science 328, 710-22.
[6] Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310, 321-4.
[7] Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol. 128, 415-23.
[8] Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B (2013) Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193, 1233-54.
[9] The International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations (2010) Nature 467, 52-58.
[10] Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100-4.
[11] Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, Gignoux C, Woerner A, Hammer MF, Slatkin M (2013) Higher levels of Neanderthal ancestry in East Asians than in Europeans. Genetics 194, 199-209.
[12] Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, Patterson N, Reich D (2013) The landscape of Neandertal ancestry in present-day humans. Nature in submission.
[13] Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet. 39, 1251-5.
[14] 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65
[15] Sankararaman S, Patterson N, Li H, Pääbo S, Reich D (2012) The date of interbreeding between Neandertals and modern humans. PLoS Genet. 8, e1002947.
[16] Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B (2013) Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193, 1233-54.
[17] Pickrell JK, Patterson N, Loh PR, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D (2013) Ancient west Eurasian ancestry in southern and eastern Africa. arXiv:1307.8014.
[18] Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AM, Ko YC, Jinam TA, Phipps ME, Saitou N, Wollstein A, Kayser M, Pääbo S, Stoneking M (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet. 89, 516-28.
[19] Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D (2012) Ancient admixture in human history. Genetics 192, 1065-1093.

[20] Lipson M, Loh PR, Levin A, Reich D, Patterson N, Berger B (2013) Efficient moment-based inference of admixture parameters and sources of gene flow. Mol Biol Evol 30, 1788-1802.
[21] Skoglund P, Jakobsson M (2011) Archaic human ancestry in East Asia. Proc Natl Acad Sci USA 108, 18301-6
[22] Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 13, 745-53.
[23] Hublin JJ (2009) Out of Africa: Modern human origins special feature: The origin of Neandertals. Proc Natl Acad Sci USA 106, 16022-7.

Appendix S13.1 Statistical model for inferring split times of populations from haplotype data

We are given two haplotypes that align to the same locus – one modern human, the second archaic. The allelic state at both haplotypes as well as the chimpanzee genome sequence are observed at L_i bases. Given the allelic states of the three sequences, we determine n_i mutations occurred on the modern human lineage since the most recent common ancestor of the two haplotypes (under an infinite sites model of mutation). We also observe d_i differences between human and chimpanzee sequences at this locus. The mutation rate at this locus is μ_i. We ignore intra-locus recombination. S denotes the mean human-chimp genetic divergence .

The parameter of interest are the split times of the two populations $\tilde{t}_0$ and the effective population size of the ancestor $\tilde{N}_0$.

$$\begin{aligned} n_i|T,t_0,N_0 &= Poi\left(\mu_i L_i(\tilde{t}_0+2\tilde{N}_0 T)\right) \\ T &= Exp(1) \\ d_i &= 2\mu_i L_i S \end{aligned} \tag{1}$$

We rewrite this as

$$\begin{aligned} n_i|T,t_0,N_0 &= Poi\left(\frac{d_i}{2}\left(\frac{\tilde{t}_0}{S}+2\frac{\tilde{N}_0}{S}T\right)\right) \\ &= Poi\left(\frac{d_i}{2}(t_0+2N_0T)\right) \end{aligned}$$

Here t_0 is the split time as a fraction of Human-chimp divergence and N_0 is the effective population size also rescaled by human-chimp divergence.

We can write the likelihood of t_0, N_0 as

$$\begin{aligned} Pr(n_i|t_0,N_0) &= \int_0^\infty dT\exp(-T)Poi(n_i|\frac{d_i}{2}(t_0+2N_0T)) \\ &= \int_0^\infty dT\exp(-T)\exp\left(-\frac{d_i}{2}(t_0+2N_0T)\right)\frac{\left(\frac{d_i}{2}(t_0+2N_0T)\right)^{n_i}}{n_i!} \\ &= \frac{1}{n_i!}\frac{1}{N_0 d_i}\int_{\frac{d_i}{2}t_0}^\infty dz\exp\left(-\frac{1}{2N_0}\left(\frac{2z}{d_i}-t_0\right)\right)\exp(-z)z^{n_i} \\ &= \frac{1}{n_i!}\frac{\exp(\frac{t_0}{2N_0})}{N_0}\frac{{d_i}^{n_i}}{\left(d_i+\frac{1}{N_0}\right)^{n_i+1}}\int_{\frac{t_0}{2}\left(d_i+\frac{1}{N_0}\right)}^\infty dy\exp(-y)y^{n_i} \end{aligned} \tag{2}$$

The log likelihood of (t_0, N_0) given n loci can be written as (ignoring constants)

$$\mathcal{L}(t_0,N_0) = \sum_{i=1}^n \frac{t_0}{2N_0} - \log(N_0) - (n_i+1)\log\left(d_i+\frac{1}{N_0}\right) + \log\Gamma\left(\frac{t_0}{2}\left(d_i+\frac{1}{N_0}\right);n_i+1,1\right) \tag{3}$$

where Γ denotes the truncated gamma function. We computed maximum likelihood estimates, $(\hat{t}_0, \hat{N}_0)$, using the optim function in R. We estimated 95% confidence interval on t_0 using the profile likelihood $\mathcal{L}_{t_0}(t_0) = \max_{N_0}\mathcal{L}(t_0,N_0)$ as $[t_l, t_u]$ where

$$\begin{aligned} f(t) &= \mathcal{L}_{t_0}(t) - \mathcal{L}(\hat{t}_0,\hat{N}_0) + \frac{1}{2}{\chi^2}_{0.95} \\ t_l &= argmin_t \mathbf{1}\{f(t)\geq 0\} \\ t_u &= argmax_t \mathbf{1}\{f(t)\geq 0\} \end{aligned}$$

Here ${\chi^2}_{0.95}$ is the 0.95 quantile of the Chi-squared distribution with one degree of freedom.

Supplementary Information 14

Neandertal population relationships and mixture proportions

Nick Patterson*, Swapan Mallick and David Reich*

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

(i) Findings

- The introgressing Neandertals were genetically closer to Mezmaiskaya than to Altai or to Vindija.
- The proportion of Neandertal ancestry is 1.48-1.96% in Europeans and 1.64-2.14% in eastern non-Africans (95% confidence intervals).

(ii) Inferring population relationships

Here we co-analyze the high coverage Neandertal genome (Altai) with the low coverage Mezmaiskaya and Vindija genomes as well as multiple present-day humans (the 25 individuals from Panels A and B). The goal is to infer population relationships.

Our main tools are *D*-statistics, used now in multiple papers[1,2,3]. Our convention in this note is that if *A, B, C, D* are 4 populations, then:

$$D(A,B;C,D) = \frac{n_{BABA} - n_{ABBA}}{n_{BABA} + n_{ABBA}} \quad \text{(S14.1)}$$

where n_{BABA} is a count of alleles agreeing in populations A, C and also in B, D (but different in A, B) and n_{ABBA} is a count of alleles agreeing in populations A, D and also in B, C (but different in A, B). With this sign convention, a positive *D* is an indication of gene flow between the corresponding population pairs *(A, C)* or *(B, D)*.

Standard errors on statistics in this note are computed using a Weighted Block Jackknife[4,5] with a block size of 5 million base pairs (5 Mb) unless otherwise stated.

To explore whether artifacts in the low coverage Neandertal data might be affecting our inferences about population relationships, we co-analyzed the following subset of samples. (We chose all the present-day humans to be from the same panel (Panel A) so that our analyses would not be complicated by differences in sequencing and data processing protocols between Panels A and B.)

Altai	(This study)
Mezmaiskaya	(This study)
Denisova	(ref. 2)
$Dinka_A$	(ref. 2)
$French_A$	(ref. 2)
$Sardinian_A$	(ref. 2)
Han_A	(ref. 2)
San_A	(ref. 2)
Chimpanzee	(*panTro2*)

For this robustness analysis, we restricted to alignments of the hominin reads to the chimpanzee reference genome (*panTro2*) as it is equally distant to all hominin populations. We picked a random allele from the GATK genotype calls for all samples except Mezmaiskaya. For Mezmaiskaya we chose a base at random from a high quality read, as in ref. 1. We restricted analysis to sites in the genome where we have coverage from at least one high quality read from Mezmaiskaya, where all the remaining samples pass the stronger version of the filters described in SI 5 (Map35_100%), and where exactly two alleles are observed across samples.

We were concerned that in spite of the chemical treatment of Altai, Mezmaiskaya and Denisova to remove uracils due to ancient DNA damage, residual ancient DNA errors might compromise analyses. We obtain the following counts, displaying the alleles in the sample order given above.

```
TCCCCCCCT 223
CTCCCCCCT 279
CCCCCCCCT 267206
```

For example, the count of 279 is for *Mezmaiskaya* and *Chimpanzee* having a 'T' allele and all other samples having a 'C'. Errors contributing to the first 2 patterns are overwhelmingly likely to arise from C → T changes at sites where the truth is a fixed human-chimpanzee difference (CCCCCCCCT). The counts confirm that deamination typical of ancient DNA occurs at a very low rate in our dataset (reflecting chemical treatment to remove uracils and our bioinformatic procedure of setting the last couple of bases of the reads to low quality; SI 2). Thus, we can use transitions as well as transversions in comparisons of Mezmaiskaya, Altai, Denisova and present-day humans. However, as discussed below, in what follows we restrict our *D*-statistic computations to transversions only so that we can co-analyze the other samples with the Vindija Neandertal.

Alignments to the human and chimpanzee genomes give largely concordant results

We were concerned that *D*-statistics might differ depending on whether we aligned to the human reference genome *hg19* or to *Chimpanzee* (*panTro2*). We especially wanted to find cases where the sign of *D* depended on the reference sequence as this could lead to errors in reconstructing population relationships. We find differences between the *hg19* and *panTro2* alignments when *Chimpanzee* is also used as a population in the *D*-statistic. For instance, we find a Z-score of 8.0 for the statistic

$$D(Mandenka_A, Yoruba_A; Dai_A, Chimpanzee) \quad \text{(S14.2)}$$

The implied phylogeny here is implausible, but the Z-score is just -0.01 when aligning to *hg19*. We suspect that different mapping error rates in our sequenced samples result in different correlation patterns to *Chimpanzee*. Since the *hg19* genome assembly is of much higher quality than the *panTro2* genome assembly, these errors are reduced for *hg19*.

Ignoring *D*-statistics explicitly involving *Chimpanzee*, the correlation between *D*-statistics using the *hg19* and *panTro2* alignments is 0.998. We remark that statistics using *panTro2* alignment are systematically smaller (regression coefficient = 0.95) probably reflecting greater alignment noise. In the remainder of this note we focus on statistics from the *hg19* alignment.

For analysis of Vindija data, we need to restrict to transversion polymorphisms

The data from the three Vindija bones (Vi25.16, Vi25.25 and Vi25.26)[1] are different from the data from the other archaic samples in multiple ways:

(a) A substantially shorter average read length.
(b) An older library preparation method.
(c) Use of an earlier version of the Illumina sequencing technology.
(d) No chemical treatment to remove uracils.
(e) The samples were subjected to a restriction enzyme pre-treatment to remove sequences rich in CG dinucleotides (to reduce the proportion of bacterial sequences).

These differences are likely to result in different sequencing and mapping error processes, which could bias inference of population relationships.

The largest difference between the data from the 3 Vindija samples and the data from the other ancient samples is that the Vindija samples were not treated to remove uracils, which results in a high rate of deamination. Confirming this, when we replace Mezmaiskaya with Vindija (picking a random high quality base from one of the 3 samples as in ref. 1) we obtain the following counts:

```
TCCCCCCCT 988
CTCCCCCCT 23791
CCCCCCCCT 1298682
```

The gross imbalance makes it in practice impossible to use transitions for studying the relationship of Vindija to other samples, and hence the analyses that follow all restrict to transversions.

(iii) Denisova is genetically closer to Altai than to the other Neandertals we sequenced

We first examined *D*-statistics relating Denisova to the Neandertals. This analysis, which restricts to transversion polymorphisms, shows that Denisova shares significantly more derived alleles with Altai than with the other Neandertals we sequenced. The finding is robust to whether the non-Altai Neandertal we analyze is Mezmaiskaya (first block of the table) or Vindija (second block). This suggests a history of gene flow between Altai-related Neandertals and Denisovans, an inference we confirm in SI 15 with a second line of evidence based on haplotypes.

Table S14.1: Denisova is more closely related to Altai than to the low coverage Neandertals

	hg19 mapping			*panTro2 mapping*		
	***D*-stat**	**Std. Err.**	***Z*-score**	***D*-stat**	**Std. Err.**	***Z*-score**
D(Altai, Mezmaiskaya; Denisova, Chimp)	0.132	0.022	5.9	0.151	0.018	8.5
D(Altai, Mezmaiskaya; Denisova, San_A)	0.192	0.022	8.7	0.181	0.019	9.7
D(Altai, Mezmaiskaya; Denisova, $Dinka_A$)	0.193	0.023	8.5	0.193	0.018	10.4
D(Altai, Vindija; Denisova, Chimp)	0.079	0.014	5.6	0.177	0.010	17.1
D(Altai, Vindija; Denisova, San_A)	0.107	0.015	7.1	0.150	0.012	13.0
D(Altai, Vindija; Denisova, $Dinka_A$)	0.104	0.015	6.8	0.155	0.012	13.1
D(Mezmaiskaya, Vindija; Denisova, Altai)	0.032	0.026	1.2	0.017	0.026	0.6

(iv) Introgressing Neandertals were genetically closer to either Mezmaiskaya or to Vindija

We studied *D*-statistics where one pair of samples are Neandertals and the other present-day humans:

$$D(Neandertal_1, Neandertal_2; X, Y) \quad \text{(S14.3)}$$

We further restricted our analysis to statistics where the second pair of samples consists of present-day humans prepared in the same way (e.g. two from Panel A). Thus, any differences between these samples are expected to be uncorrelated to differences among Neandertals under the assumption of no gene flow from relatives of Neandertals since the ancestors of *X* and *Y* separated.

We first observe from Table S14.2 that:

$$D(Neandertal_1, Neandertal_2; X=San_A, Y=Dinka_A) \sim 0 \quad \text{(S14.4)}$$

The fact that $D(Neandertal_1, Neandertal_2; San_A, Dinka_A)$ is indistinguishable from 0, regardless of the Neandertal pair we analyze, is consistent with present-day Africans being about equally closely related to the diverse Neandertals we have sequenced.

Table S14.2: The introgressing Neandertal is closest to Mezmaiskaya

		D(Altai, Mez; X, Y)		*D(Altai, Vindija; X, Y)*		*D(Mez, Vindija; X, Y)*	
X	***Y***	***D***	***Z*-score**	***D***	***Z*-score**	***D***	***Z*-score**
San_A	$Dinka_A$	0.015	0.6	-0.016	-1.0	-0.045	-1.1
$French_A$	$Dinka_A$	-0.164	-5.8	-0.070	-4.3	0.071	1.8
$Sardinian_A$	$Dinka_A$	-0.111	-4.0	-0.043	-2.5	0.072	1.7
Han_A	$Dinka_A$	-0.114	-4.2	-0.074	-4.5	0.130	3.2
$French_A$	San_A	-0.159	-6.4	-0.048	-3.0	0.148	3.9
$Sardinian_A$	San_A	-0.139	-5.4	-0.033	-2.1	0.098	2.7
Han_A	San_A	-0.123	-5.0	-0.057	-3.6	0.131	3.3

We next examined statistics involving *X* = non-African, which allow us to study the relationship between the introgressing Neandertals and those we sequenced. For this analysis, we again restrict to transversion polymorphisms. The results are reported in Table S14.2.

A summary of these results is that:

$$\begin{array}{ll} D(Altai, Mezmaiskaya; X=Eurasian, Y=African) << 0 & (Z = -4.0 \text{ to } -6.4) \\ D(Altai, Vindija; X=Eurasian, Y=African) << 0 & (Z = -2.1 \text{ to } -4.5) \\ D(Mezmaiskaya, Vindija; X=Eurasian, Y=African) > 0 & (Z = 1.7 \text{ to } 3.9) \end{array} \quad \text{(S14.5)}$$

These signals are consistent in suggesting that the introgressing Neandertals were genetically closer to Mezmaiskaya than to the other two sequenced Neandertals, and closer to Vindija than to Altai. If we have to choose a single tree to represent the relationship among the three sequenced Neandertals and the introgressing Neandertals, we would therefore use the one shown in Figure S14.1

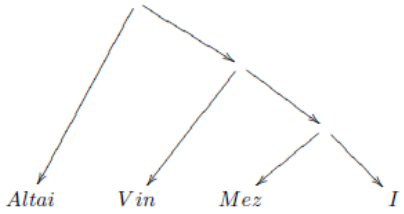


Figure S14.1: The best fitting tree relating the sequenced and introgressing Neandertals. *The introgressing Neandertals are labeled "I" and are shown as closest to Mezmaiskaya. We caution, however, that other D-statistics suggest that this is not be a perfect fit to the data, suggesting the possibility of gene flow in Neandertal history.*

While Figure S14.1 is the best fit to the data if one forces a single tree to the data, other aspects of the data suggest that the true relationship among these Neandertals was more complex, involving additional gene flows. In SI 15, Table S15.2, we show statistics of the form *D(Mezmaiskaya, Vindija; Denisova, Dinka or Yoruba or Mbuti or Chimp).* If Mezmaiskaya and Vindija were consistent with being a perfect clade, we would expect these statistics to be consistent with 0. In fact, however, they are all significantly negative (*Z* = -3.6 to -5.6), suggesting additional complexity in Neandertal history that at present with the low coverage Mezmaiskaya and Vindija genomes we do not fully understand.

(v) Contamination cannot explain the inferred Neandertal population relationships

We were concerned that our finding that the introgressing Neandertals are more closely related to some sequenced Neandertals (especially Mezmaiskaya) than to others might not reflect the historical relationships of the Neandertals, but instead differences in contamination by present-day humans.

Our key evidence in section (iv) that the introgressing Neandertal *I* is genetically closer to Mezmaiskaya than to Altai is that *D(Altai, Mezmaiskaya; X, Y)* << 0 where *X* is any non-African and *Y* is a sub-Saharan African (Table S14.2). However, suppose as an alternative explanation that Mezmaiskaya is contaminated by present-day human DNA and that in truth the uncontaminated Mezmaiskaya sample and Altai form a clade with respect to the introgressing Neandertals. This could contribute to an artifactual inference of non-Africans being closer to Mezmaiskaya than to Altai.

If our observations are explained by contamination, the contamination is most likely to come from a non-African, as Mezmaiskaya was excavated in Russia and the ancient DNA laboratory work was carried out in Germany. We can estimate the proportion of contamination by an f_4 ratio statistic[3] where *X* is any Eurasian sample and *Y* is any sub-Saharan African sample (Equation S14.6). Intuitively, this statistic measures how far the skew from zero is with Mezmaiskaya, compared to what is seen in an individual of all non-African ancestry $(French_A)$:

$$\hat{\beta} = \frac{f_4(Altai,\ Mezmaiskaya;\ X,\ Y)}{f_4(Altai,\ French_A;\ X,\ Y)} \quad \text{(S14.6)}$$

Here we assume that the Altai data are effectively uncontaminated, which seems reasonable based on our estimate of low contamination rates per read in SI 5, and the expectation that any effects of contamination will be greatly reduced by consensus genotype calling on the high coverage data.

We estimated $\hat{\beta}_{Mezmaiskaya}$ using X = $Karitiana_A$, Han_A, or $Sardinia_A$, and Y = San_A (Table S14.3). The 95% confidence intervals for $\hat{\beta}_{Mezmaiskaya}$ are 2.04-4.66% (for X=$Karitiana_A$), 2.01-4.91% (for X=Han_A), and 3.04-5.44% (for X=$Sardinia_A$). These are outside the bounds of direct contamination estimates (0.49-0.65% from mtDNA, 0.28-79% from the Y chromosome, and 0-1.12% from the nuclear genome; SI 5a). Thus, contamination in Mezmaiskaya cannot explain our findings.

Table S14.3: Mezmaiskaya's relatedness to the introgressing Neandertal is beyond what can be explained by the empirically estimated levels of contamination in Mezmaiskaya of <1.12%.

Population X	$\hat{\beta}$ Point estimate	$\hat{\sigma}$ Std. Err.	$\hat{\beta} - 1.96\hat{\sigma}$ to $\hat{\beta} + 1.96\hat{\sigma}$ 95% confidence interval
$Karitiana_A$	3.35%	0.67%	2.04 - 4.66%
Han_A	3.46%	0.74%	2.01 - 4.91%
$Sardinia_A$	4.24%	0.61%	3.04 - 5.44%
Union of 3 estimates	3.35 - 4.24%	n/a	2.04 - 5.44%

Note: We compute the f_4 ratio of equation S14.6 with X equal to the specified sample and Y = San_A.

(vi) Present-day non-Africans carry a little less than 2% Neandertal ancestry

Past papers[1] have shown that Neandertals are genetically closer to non-Africans than to Africans, and that Denisovans are closer to Oceanian populations (a term we use here to refer to aboriginal people from New Guinea, Australia and the Philippines) than to mainland Asians[10]. We replicate these results (Table S14.4; Table S14.5).

Table S14.4: Neandertals closer to non-Africans than to Africans: D(Non-Afr, Afr; Altai, Chimp)

Non-African	African	Z-score
$French_A$	San_A	8.1
$French_B$	San_B	8.0
$French_A$	$Dinka_A$	9.2
$French_B$	$Dinka_B$	9.2
Han_A	San_A	10.6
Han_B	San_B	9.9
Han_A	$Dinka_A$	11.4
Han_B	$Dinka_B$	10.4

Table S14.5: Denisovans closer to Oceanians than Eurasians: D(Eurasia, Oceania; Den, Altai)

Eurasian	Oceanian	Z-score
$French_A$	$Papuan_A$	-5.5
$French_B$	$Papuan_B$	-5.0
$French_B$	$Australian_{B1}$	-5.3
$French_B$	$Australian_{B2}$	-6.0
Han_A	$Papuan_A$	-7.3
Han_B	$Papuan_B$	-6.9
Han_B	$Australian_{B1}$	-7.6
Han_B	$Australian_{B2}$	-8.0

Note: The two present-day humans in these *D*-statistics are always drawn from the same sequencing panel (both Panel A or both Panel B) to minimize the potential for sequencing artifacts to produce false positive skews from zero.

In addition to the two signals above, several studies have also found evidence for more archaic genetic material in eastern non-Africans (East Asians and Native Americans) than in Europeans[2,6,7]. One of the main lines of evidence for differences in the archaic ancestry in eastern non-Africans and in Europeans was "enhanced *D*-statistics" (SOM 11 of ref. 2). These statistics are based on restricting to sites where a pool of sub-Saharan African samples that we assume all carry the ancestral allele. Requiring that these sub-Saharan Africans always carry the ancestral allele enriches for mutations that arose as new mutations in an archaic lineage. Here we revisit the "enhanced *D*-statistics" for each of 3 possible pairs of outgroups (*X*, *Y*) ((Altai, Chimpanzee), (Denisova, Chimpanzee), or (Altai, Denisova)) and each of three possible pairs of non-African population. Specifically, we pool samples from "Europe" (French + Sardinian), "East" (Dai + Han + Karitiana + Mixe), and "Oceanian" (Papuan + 2 Australians), and report the statistic separately for the 2 panels of present-day humans (Mixe and Australians are only available for Panel B). We use 3 enhancement strategies:

"All" Basic *D*-statistic without enhancement.

"E12" We require all deeply sub-Saharan African alleles from the 2 Mbuti, 2 Yoruba and 2 Dinka to be ancestral (we require coverage from at least 5 of these 6 individuals)

"E119" We impose the same requirement as E12, but also examine data from 107 YRI individuals from the 1000 Genomes Project, sampling the majority allele from all individuals with at least 3 reads covering a site, restricting to sites with a coverage of at least 80 YRI, and requiring that all the sampled YRI alleles match chimpanzee.

Table S14.6 confirms previous reports of significantly more archaic ancestry in Oceanian populations than in other non-Africans[2,8] (the *Z*-scores range from-3.3 and -27.0 standard errors from zero). The evidence that the Oceanian archaic genetic material has Denisovan affinity derives from the fact that the skews are strongest and the statistics most significant for statistics of the form *D(Non-African, Oceanian; Denisova, Chimpanzee).* Moreover, *D(Non-African, Oceanian; Altai, Denisova)* is positive, indicating that the extra archaic material in Oceanians is closer to Denisova than to Altai.

Table S14.6: **Enhanced statistics of form $D_{enhanced}$(*Non-African$_1$, Non-African$_2$; Outgroup$_1$, Outgroup$_2$*)**

D-stat	Enhance	Panel	X=Altai, Y=Chimpanzee				X=Denisova, Y=Chimpanzee				X=Altai, Y=Denisova			
			n_{BABA}	n_{ABBA}	D	Z	n_{BABA}	n_{ABBA}	D	Z	n_{BABA}	n_{ABBA}	D	Z
D(Eur,East;X,Y)	All	A	180,560	183,969	-0.009	-1.7	167,939	170,728	-0.008	-2.2	243,284	244,527	-0.003	-0.4
		B	179,624	183,180	-0.010	-1.8	167,314	169,878	-0.008	-2	241,135	243,121	-0.004	-0.7
	E12	A	17,612	20,563	-0.077	-3.3	9,445	10,663	-0.061	-3.7	29,968	33,432	-0.055	-2.4
		B	18,036	20,339	-0.060	-2.5	9,497	10,885	-0.068	-4.1	31,065	32,896	-0.029	-1.3
	E119	A	9,649	12,502	-0.129	-3.9	2,891	4,007	-0.162	-5	16,367	19,841	-0.096	-2.9
		B	9,975	12,192	-0.100	-3.2	2,992	4,023	-0.147	-4.3	16,980	19,353	-0.065	-2.2
D(Eur, Oceania;X,Y)	All	A	172,064	183,205	-0.031	-4.3	155,626	181,673	-0.077	-10.1	259,506	229,696	0.061	5.5
		B	185,435	197,807	-0.032	-4.4	167,528	196,349	-0.079	-11.7	280,259	247,359	0.062	6.4
	E12	A	16,555	25,399	-0.211	-6.3	8,713	23,324	-0.456	-18	45,191	33,658	0.146	3.9
		B	18,134	27,268	-0.201	-7	9,429	25,283	-0.457	-18.5	49,841	36,402	0.156	4.9
	E119	A	9,059	16,412	-0.289	-6.5	2,693	14,407	-0.685	-27	29,814	21,090	0.171	3.3
		B	10,068	17,543	-0.271	-7.3	2,996	15,730	-0.680	-25.9	33,188	22,669	0.188	4.4
D(East,Oceania;X,Y)	All	A	167,265	175,101	-0.023	-3.3	150,494	174,177	-0.073	-9.7	251,737	220,042	0.067	6.3
		B	179,981	188,880	-0.024	-4.2	161,686	188,088	-0.076	-11	272,460	237,452	0.069	7.9
	E12	A	18,356	24,456	-0.143	-4.9	9,254	22,834	-0.423	-15.7	47,971	33,014	0.185	5.5
		B	19,467	26,328	-0.150	-5.9	10,204	24,743	-0.416	-15.1	51,336	35,979	0.176	6.4
	E119	A	11,121	15,853	-0.175	-4.5	3,534	14,297	-0.604	-19.8	33,070	21,008	0.223	5
		B	11,747	17,030	-0.184	-5.5	3,853	15,614	-0.604	-19.6	35,689	22,732	0.222	6

Notes: The analyses in this table include both transitions and transversions. We only compute comparisons of present-day humans from the same panel (both Panel A or both B, indicated in the third column) to avoid artifacts due to systematic experimental differences between these samples. Standard errors are from a Block Jackknife with 100 equally sized blocks.

The enhanced *D*-statistics reported in Table S14.6 are also consistent with our previous findings based on such statistics of significantly more archaic material in Eastern non-Africans than in Europeans[2,6]. In particular, both *D(Europe, East; Altai, Chimp)* and *D(Europe, East; Denisova, Chimp)* show some enhanced *D*-statistics that are significantly below zero (the *Z*-scores range from -1.8 to -5.0). To provide additional insight about this signal, we examined *D*-statistics of the form *D(Europe, East; Altai, Denisova),* which are negative in the direction of a closer proximity of the introgressing material in eastern non-Africans to Altai than to Denisova (Table S14.6). While the skew is not highly significant (maximum of 2.8 standard errors below zero), the direction is opposite to the positive skew seen for *D(Non-African; Oceanian, Altai, Denisova),* indicating that the extra archaic material in eastern non-Africans may have a more Neandertal-like ancestry than that in Oceania (indeed, this was our interpretation in Note 11 of the SOM of ref. 2). Thus, we cannot simply explain the extra archaic material in eastern non-Africans compared with in Europeans by the small amount of Denisova-like genetic material that we detect in eastern non-Africans in SI 13 through local ancestry analysis.

It is tempting to hypothesize that the evidence for extra archaic genetic material in eastern non-Africans than in Europeans, which is not completely explained by the excess Denisovan ancestry in eastern non-Africans that we document in SI 13, reflects more than one Neandertal introgression into the ancestors of non-Africans. However, an alternative explanation for these patterns is gene flow between West Eurasians and sub-Saharan African populations, after the main out of Africa migration[2]. If the gene flow were into Europeans, it would dilute the proportion of Neandertal ancestry in Europeans, potentially contributing to our observation that Europeans have less Neandertal relatedness than East Asians. More generally, gene flow in either direction would also complicate the interpretation of enhanced *D*-statistics, as a significantly negative $D_{enhanced}$*(Europe, East; Altai, Africa)* might not reflect differences in Neandertal ancestry between Europeans and eastern non-Africans, but rather the effects of the gene flow in making Europeans and Africans look more similar. Indeed some small African admixture into Sardinia and other southern European populations has been detected[9], and in SI 13, we report evidence of some West Eurasian gene flow into sub-Saharan African populations like the YRI which is the main population we use for enhancement. Because of these considerations, we are cautious about using enhanced *D*-statistics to compute mixture proportions (as in ref. 2); we are concerned that the enhancement strategy might amplify the biases due to these subtle gene flows. Instead, in what follows, we estimate Neandertal mixture without enhancement.

Estimates of Neandertal mixture proportions in present-day non-Africans

We next estimate the proportion of Neandertal ancestry in various non-Oceanian non-African populations by taking advantage of the new data reported in this study.

For our ancestry estimation we use the fact that we have sequences from multiple Neandertals. We further assume that their relationships to each other and to other hominins are as shown in Figure S14.1 and Figure S14.2, so that Vindija, Mezmaiskaya and the introgressing Neandertals are a clade with respect to Altai and the Denisova finger bone. We note that Figure S14.2 is an unrooted tree, so it is a correct description of the population relationships even in light of the Altai-related gene flow into Denisova that we document in SI 15.

To estimate mixture proportions, we apply F_4 ratio estimation, described elsewhere[2,3,10]. *F_4 ratio estimation* is based on a ratio of f_4 statistics where f_4*(A, B; C, D)* is an unbiased estimate of the mean of allele frequencies a_i, b_i, c_i and d_i in populations *A*, *B*, *C* and *D* respectively, averaging over SNPs:

$$f_4(A,B;C,D) = \frac{1}{n}\sum_{i=1}^{n}(a_i - b_i)(c_i - d_i) \qquad \text{(S14.7)}$$

We have previously referred to f_4 ratios as *S*-statistics[2], and another way to think of f_4 statistics is as the numerators of *D*-statistics.

Consider the population relationships depicted in Figure S14.2. Here *R* is the ancestral population that existed at the time of the split of archaic and modern humans; *F* is an unadmixed Eurasian population prior to Neandertal admixture but after the split from present-day Africans; *X* is a present-day non-African population; and α is the proportion of Neandertal admixture. We note that this figure does not capture several real complexities in the population relationships: for example, Altai-related Neandertal gene flow into the Siberian Denisovan (SI 15), or differences between all Neandertals and the Siberian Denisovans in their proportion of ancestry from an unknown archaic population (SI 16). However, as argued below, these histories in no way compromise our admixture estimates.

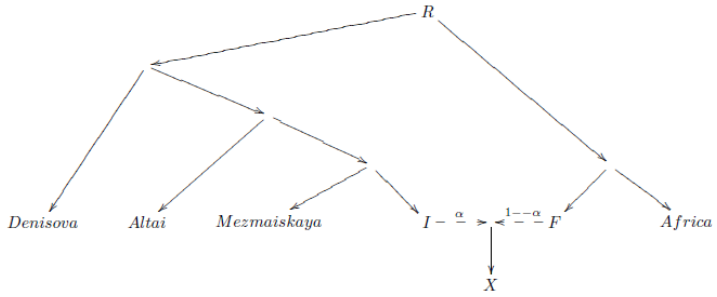


Figure S14.2: Model assumed for mixture estimation. *This unrooted tree is accurate even in the presence of Altai-related gene flow into Denisova. R is the root, I the introgressing Neandertal, F the ancestral population of non-Africans, and X the admixed population.*

Based on this figure, we apply the mixture estimation methodology that is described in detail in ref. 3. As in ref. 3, in what follows we used "f_4" to represent statistics (functions of data) and F_4 to represent expected values that depend on population history and the ascertainment of polymorphic sites.

We can write symbolically

$$X = \alpha I + (1-\alpha)F \quad \text{(S14.8)}$$

meaning that a modern Eurasian population X is admixed between I and F where I is a population of Neandertals, and F is a modern human population that is a sister group to present-day sub-Saharan Africans and that does not harbor Neandertal ancestry. We now obtain the following expressions by partitioning the ancestry of X into its two components:

$$F_4(Den, Alt; Africa, X) = \alpha F_4(Den, Alt; Africa, I) + (1-\alpha)F_4(Den, Alt; Africa, F) \quad \text{(S14.9)}$$

Because *Africa* and F are a clade with respect to Altai and Denisova (Table S14.2), their allele frequency differences are expected to be uncorrelated to those between Denisova and Altai. Thus:

$$F_4(Den, Alt; Africa, F) = 0 \quad \text{(S14.10)}$$

Because I and *Mezmaiskaya* are a clade, their allele frequency differences are expected to have the same correlation pattern when compared to Denisova, Altai and Africans, and thus:

$$F_4(Den, Alt; Africa, I) = F_4(Den, Alt; Africa, Mezmaiskaya) \quad \text{(S14.11)}$$

This gives us the key equation

$$F_4(Den, Alt; Africa, X) = \alpha F_4(Den, Alt; Africa, Mezmaiskaya) \quad \text{(S14.12)}$$

which we can rewrite as an estimator for the mixture proportion α:

$$\hat{\alpha} = \frac{f_4(Den,\ Alt; Africa,\ X)}{f_4(Den,\ Alt; Africa,\ Mezmaiskaya)} \quad \text{(S14.13)}$$

This argument remains valid given

(a) Gene flow from Altai into Denisova (SI 15) because it does not change the unrooted topology.
(b) Different proportions of an unknown archaic lineage in the Siberian Denisovan and all Neandertals (SI 16a,b) because it multiplies the numerator and denominator of Equation S14.13 by the same factor and thus its effect cancels.
(c) Replacement of Mezmaiskaya by Vindija because it gives the same topology (Figure S14.1).
(d) Arbitrary ascertainment of polymorphisms as long as it is restricted to sub-Saharan Africans.

Table S14.7: **Neandertal ancestry estimate** $\hat{\alpha} = \frac{f_4(Denisova,\ Altai; Africa,\ X)}{f_4(Denisova,\ Altai; Africa,\ Other\ Neandertal)}$

	Other Neandertal = Mezmaiskaya				Other Neandertal = Vindija			
	Panel A		Panel B		Panel A		Panel B	
	$\hat{\alpha}$	Std. Err.	$\hat{\alpha}$	Std. Err.	$\hat{\alpha}$	Std. Err.	$\hat{\alpha}$	Std. Err.
French	0.020	0.003	0.019	0.003	0.016	0.002	0.017	0.002
Sardinian	0.019	0.002	0.017	0.003	0.018	0.002	0.018	0.002
Han	0.022	0.003	0.018	0.003	0.023	0.002	0.019	0.002
Dai	0.019	0.003	0.016	0.003	0.019	0.002	0.016	0.002
Karitiana	0.020	0.003	0.019	0.003	0.018	0.002	0.019	0.002
Mixe	-	-	0.018	0.003	-	-	0.017	0.002

Table S14.7 shows estimates of Neandertal mixture proportions for populations from Eurasia and for Native Americans (we do not include Oceanians here because of the complication of Denisovan admixture into these groups[2,10]). We use a pool of Dinka, Mbuti and Yoruba samples to represent "Africa". There is good concordance in the estimates whether we use Vindija or Mezmaiskaya. This finding is important as in the analyses above we found that the tree shown in Figure S14.1 is not a perfect fit to the data, in the sense that a formal test for whether Mezmaiskaya and Vindija area clade with respect to Denisova (as suggested by Figure S14.1) fails (Table S15.2 of SI 15). The fact that our ancestry estimates are robust to whether we use Mezmaiskaya or Vindija in the analysis suggests that errors in our model of history are not likely to be compromising our ancestry inference.

To further increase the accuracy of our ancestry estimates, we next followed the procedure of ref. 2, assuming that the introgression fraction is the same across Europe, and also the same across eastern non-Africans. We can then pool frequencies for 4 Europeans (2 French and 2 Sardinian) and for 7 eastern non-Africans (2 Han, 2 Dai, 2 Karitiana and 1 Mixe). We also pool data for the two Neandertals (Mezmaiskaya and Vindija). Table S14.8 shows the results, which correspond to 95% confidence intervals of 1.48-1.96% for Europeans and 1.64-2.14% for eastern non-Africans.

We conclude that if the population relationships assumed in Figure S14.2 are correct, then Neandertal introgression has contributed a bit less than 2% of the ancestry of present-day Eurasians and Native Americans. The standard errors here are smaller than we have previously reported, reflecting our better data and our ability to use the inferred population relationships among multiple Neandertals in *F_4 Ratio Estimation*. We note that our previous Neandertal ancestry estimates are statistically consistent with those that we report here to within a couple of standard errors. In particular, we estimated 1.7 ± 0.2% and 1.1 ± 0.8% (SOM 18 of ref. 1, Table S58); 2.5 ± 0.6% (SI 8 of ref. 2, Table S8.4), and 1.0 ± 0.3% for Europeans and 1.7± 0.4% for eastern non-Africans (Table S28 of ref. 2).

A notable feature of Table S14.8 is that the evidence for more extensive introgression in eastern non-Africans than in Europeans[2,6] is weak. A previous estimate based on enhanced *D*-statistics suggested that the proportion of Neandertal ancestry in Europeans was 64-88% of that in eastern non-Africans (95% confidence intervals) (Note S11 of ref. 2), and Wall and colleagues reported a point estimate of 67% in ref. 6. Here, the 95% confidence interval is 79-112%. While this is consistent with the upper end of one of the previously reported ranges, it is also consistent with no effect and represents a substantial weakening of the previously reported signal. We discuss this issue further in SI 13, where we systematically compare all estimates and conclude that the proportion of Neandertal ancestry is indeed probably smaller in Europeans than in eastern non-Africans, but that the effect is slight.

Table S14.8: **Estimates of Neandertal ancestry pooled across multiple samples per population**

Quantity	**Statistic**	**Est.**	**Std. Err.**	**95% CI**
Neandertal ancestry in Europeans	$\frac{f_4(Denisova,\ Altai;\ Africa,\ 4\ Europeans)}{f_4(Denisova,\ Altai;\ Africa,\ Vindija+Mez)}$	1.72%	0.12%	1.48-1.96%
Neandertal ancestry in Eastern non-Africans	$\frac{f_4(Denisova,\ Altai;\ Africa,\ 7\ Eastern)}{f_4(Denisova,\ Altai;\ Africa,\ Vindija+Mez)}$	1.89%	0.13%	1.64-2.14%
Neandertal in Europeans as a fraction of Eastern	$\frac{f_4(Denisova,\ Altai;\ Africa,\ 4\ Europeans)}{f_4(Denisova,\ Altai;\ Africa,\ 7\ Eastern)}$	96%	9%	79-112%

References

[1] Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PL, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S (2010) A draft sequence of the Neandertal genome. Science 328, 710-22.

[2] Meyer M, Kircher M, Gansauge MT, Li H, Racimo F et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222-6.

[3] Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D (2012) Ancient admixture in human history. Genetics, 192, 1065–1093.

[4] Busing FMTA, Meijer E, van der Leeden R (1999) Delete-*m* jackknife for unequal *m*. Statistics and Computing, 9, 3-8.

[5] Künsch HR (1989) The jackknife and the bootstrap for general stationary observations. Ann. Statistics 17, 1217-1241.

[6] Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, Gignoux C, Woerner A, Hammer MF, Slatkin M (2013) Higher levels of Neanderthal ancestry in East Asians than in Europeans. Genetics 194, 199-209.

[7] Skoglund P, Jakobsson M (2011) Archaic human ancestry in East Asia. Proc Natl Acad Sci USA 108, 18301-6.

[8] Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, 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 JJ, Kelso J, Slatkin M, Pääbo S (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468, 1053-60

[9] Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D (2011) The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet. 7, e1001373.

[10] Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AM, Ko YC, Jinam TA, Phipps ME, Saitou N, Wollstein A, Kayser M, Pääbo S, Stoneking M (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Gen 89, 516-28.

Supplementary Information 15

Gene flow from Neandertals into Denisovans

Nick Patterson, Swapan Mallick, Janet Kelso and David Reich*

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

(i) Overview

- Denisova shares more derived alleles with Altai than with the other Neandertals we sequenced
- Haplotype analysis provides an independent line of evidence for Neandertal gene flow into Denisova
- The gene flow was recent and contributed at least 0.5% of the Siberian Denisovan's ancestry.
- A clear signal of Neandertal introgression into Altai occurs at the HLA locus. HLA introgressed alleles in modern humans have recently been found to have risen in frequency[6], suggesting that they have been selected after they were introgressed from Neandertal and Denisova. Similarly, the Neandertal HLA allele may have been selected for in Denisovans, although further data from Denisovans are required to test this hypothesis.

(ii) Denisova shares more derived alleles with Altai that with the other Neandertals we sequenced

For this note, we use the same data filtering as in SI 14. For the deeply sequenced genomes, we restrict to sites passing the stronger of the two sets of filters in SI 5 (Map35_100%), perform all analyses on genotypes extracted from the VCF files, and further restrict to sites with genotype quality of GQ≥45. For the two Neandertals with light sequencing coverage (Mezmaiskaya and Vindija), we use randomly sampled high quality bases from the BAM files (SI 14). For Vindija, we restrict to transversions (no A/G or C/T polymorphisms). All analyses use reads mapped to *hg19*, but mappings to *pantro2* are occasionally used for verification.

Table S15.1: Sub-Saharan Africans are a clade relative to all archaic genomes sequenced to date

H_1	H_2	H_3	H_4	$D(H_1, H_2; H_3, H_4)$	Z-score
Denisova	Altai	$Dinka_A$	$Yoruba_A$	0.016	2.2
Denisova	Altai	$Dinka_A$	$Mbuti_A$	0.014	1.9
Denisova	Altai	$Yoruba_A$	$Mbuti_A$	-0.001	-0.1
Denisova	Mezmaiskaya	$Dinka_A$	$Yoruba_A$	0.008	0.6
Denisova	Mezmaiskaya	$Dinka_A$	$Mbuti_A$	-0.008	-0.6
Denisova	Mezmaiskaya	$Yoruba_A$	$Mbuti_A$	-0.018	-1.3
Denisova	Vindija	$Dinka_A$	$Yoruba_A$	0.012	1.4
Denisova	Vindija	$Dinka_A$	$Mbuti_A$	0.006	0.7
Denisova	Vindija	$Yoruba_A$	$Mbuti_A$	-0.006	-0.8
Altai	Mezmaiskaya	$Dinka_A$	$Yoruba_A$	0.049	1.7
Altai	Mezmaiskaya	$Dinka_A$	$Mbuti_A$	-0.041	-1.4
Altai	Mezmaiskaya	$Yoruba_A$	$Mbuti_A$	-0.091	-3.2
Altai	Vindija	$Dinka_A$	$Yoruba_A$	0.014	0.9
Altai	Vindija	$Dinka_A$	$Mbuti_A$	-0.026	-1.5
Altai	Vindija	$Yoruba_A$	$Mbuti_A$	-0.039	-2.6
Mezmaiskaya	Vindija	$Dinka_A$	$Yoruba_A$	-0.084	-2.0
Mezmaiskaya	Vindija	$Dinka_A$	$Mbuti_A$	-0.038	-0.9
Mezmaiskaya	Vindija	$Yoruba_A$	$Mbuti_A$	0.031	0.8

Note: These analyses restrict to transversions. All comparisons to present-day humans are to Panel A individuals.

We began by computing *D*-statistics for all possible pairs of archaic humans, testing the relative rate at which they share derived alleles with present-day sub-Saharan Africans. We restrict this analysis to

Mbuti, Dinka, and Yoruba, as other studies have suggested the possibility of West Eurasian related ancestry (at low levels) in both San and the Mandenka due to gene flow in the last few thousand years[1,2] (in fact it is likely that West Eurasian gene flow occurred into some of the other populations as well as shown in SI 16a but the proportion may be less). Table S15.1 shows that present-day sub-Saharan Africans are about equally closely related to all pairs of archaic populations to which we compare them. The most extreme statistic is Z = -3.2 for $D(Altai, Mezmaiskaya; Yoruba_A, Mbuti_A)$, which corresponds to P=0.02 after applying a two-sided test and correcting for 18 tested hypothesis using Bonferroni correction. This weakly significant signal potentially reflects a history of Neandertal-related gene flow into the Yoruba to a greater extent than the Mbuti or Dinka as is also suggested by the analyses reported in SI 16. However, as we also discuss in SI 16, this may be a special feature of Yoruba history that is not relevant to all Africans. We proceed with the assumption that to a first approximation, sub-Saharan Africans form a clade relative to the archaic genomes.

Table S15.2 shows that Denisova shares more derived alleles with Altai than with the other two Neandertals we sequenced. (Some of these *D*-statistics are also reported in SI 14, where we further show that the signal is consistent when we analyze reads mapped to the human reference genome *hg19* or the chimpanzee reference genome *panTro2*.) This suggests that the pattern is likely to reflect gene flow between Altai-related Neandertals and Denisovans. In the haplotype-based analysis below, we show that the direction of flow was at least in part from Neandertals into Denisovans.

Table S15.2: Denisovans are closer to Altai than to other Neandertals

H_1	H_2	H_3	H_4	$D(H_1, H_2; H_3, H_4)$	Z-score
Altai	Mezmaiskaya	Denisova	Chimpanzee	0.132	5.9
Altai	Mezmaiskaya	Denisova	$Dinka_A$	0.193	8.5
Altai	Mezmaiskaya	Denisova	$Yoruba_A$	0.225	10.2
Altai	Mezmaiskaya	Denisova	$Mbuti_A$	0.182	7.8
Altai	Vindija	Denisova	Chimpanzee	0.079	5.6
Altai	Vindija	Denisova	$Dinka_A$	0.104	6.8
Altai	Vindija	Denisova	$Yoruba_A$	0.112	7.3
Altai	Vindija	Denisova	$Mbuti_A$	0.089	5.9
Mezmaiskaya	Vindija	Denisova	Chimpanzee	-0.113	-3.6
Mezmaiskaya	Vindija	Denisova	$Dinka_A$	-0.139	-4.2
Mezmaiskaya	Vindija	Denisova	$Yoruba_A$	-0.179	-5.6
Mezmaiskaya	Vindija	Denisova	$Mbuti_A$	-0.159	-5.0

Note: These analyses restrict to transversions. All comparisons to present-day humans are to Panel A individuals.

(iii) Haplotype analysis documents Neandertal → Denisova flow

We used a window-based analysis to provide an independent line of evidence for gene flow between Neandertals and Denisovans, and specifically from Neandertals *into* Denisovans. This approach builds on ideas developed in the Neandertal draft genome paper, where a similar approach was used to document gene flow from Neandertals into non-African ancestors[3].

The key idea is diagrammed in Figure S15.1, and requires comparing diploid data from one archaic sample, to haploid data from the other. Consider Denisova diploid and Altai haploid data:

(a) In the absence of gene flow (Figure S15.1A) , there are two predictions:
- The genetic divergence time of the Altai haplotype to the closer of the two Denisova haplotypes at locus j is guaranteed to be at least as old as the population separation, $T^j_{alt\text{-}min} \geq \tau$.
- The average time to the common ancestor of the two Denisova haplotypes T^j_{den} is expected to have a weakly positive or no correlation to Denisova-Altai genetic divergence, as most coalescence events in Denisova occur since the divergence from the Altai ancestors (SI 12)

(b) In the case of Neandertal→Denisova gene flow (Figure S15.1B), the genetic divergence time of the Altai haplotype to the closer of the two Denisova alleles can be less than the population divergence ($T^j_{alt\text{-}min}$ can be $< \tau$). Thus, at loci where there is Neandertal introgression the genetic

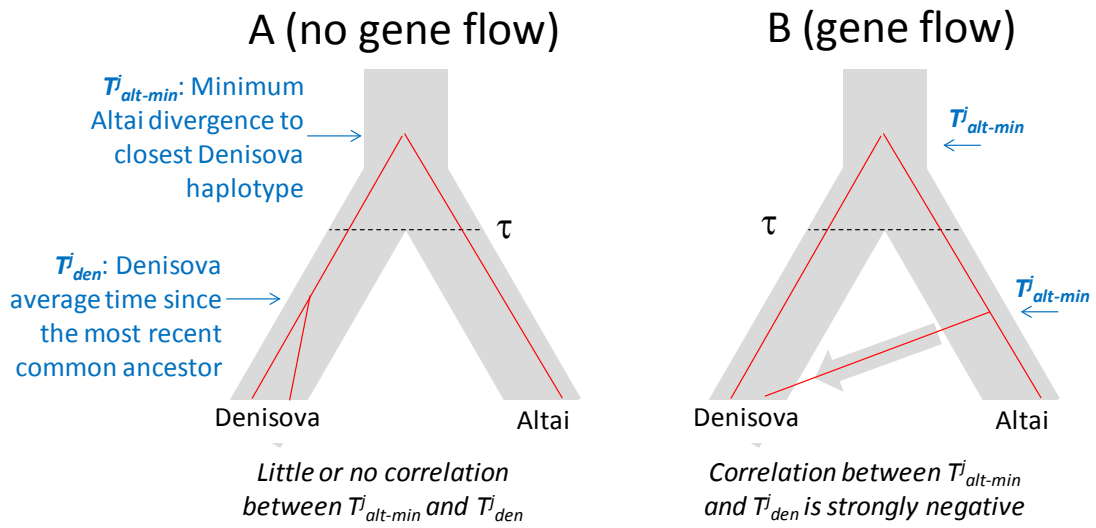
divergence may be reduced relative to the genome average and sometimes very small indeed. At these loci, Denisova is carrying one haplotype of Neandertal and one haplotype of Denisova origin, and hence there is expected to be a negative correlation between $T^i_{alt\text{-}min}$ and T^i_{den}.

Thus, we have a qualitatively different prediction in the case of no gene flow and gene flow.

Similarly, when Altai is diploid and Denisova haploid, the absence of gene flow predicts a weakly positive correlation of $T^i_{den\text{-}min}$ and T^i_{alt}, and Denisova→Neandertal a negative correlation.

A challenge in applying this test of gene flow is obtaining effectively haploid archaic data. We addressed this by restricting to subsets of each archaic genome where the inferred time since the most recent common ancestor (from the Pairwise Sequential Markovian Coalescence (PSMC); SI 12) is likely to be more recent than the main Denisova-Neandertal population divergence, which in SI 12 we estimate corresponds to loci with an expected per base pair divergence of less than around 0.0002/bp. Thus, we restrict to subsets of the genome where the TMRCA in one of the two archaic individuals being analyzed is estimated by the PSMC to be less than this threshold, which for the 50 kb screened windows below corresponds to 86% of windows in Altai and 76% of windows in Denisova. Under the null hypothesis of no gene flow, this means the two haplotypes that the archaic individual carries will form a clade relative to the haplotypes of the other archaic individual. Their two lineages converge to a single ancestral lineage more recently than the ancestral population is reached so that it does not matter which of the two alleles we choose and they can be treated as a single haplotype.

Figure S15.1: A haplotype-based test for gene flow between Neandertals and Denisovans. *We examine divergence of an Altai haplotype to the closer of the two Denisova haplotypes $T^i_{alt\text{-}min}$, measured on the Altai side of the tree. (If the two Denisova haplotypes are a clade as in the left panel, the divergence will be the same.) (A) In the absence of gene flow, this minimum divergence is uncorrelated or weakly correlated to Denisova average time since the most recent common ancestor T^i_{den}. (B) For Neandertal→Denisova gene flow, a strong negative correlation is expected*



To implement this approach, at each site *i* in the genome, we recorded:

$tmrca^i_{alt}$ =	the estimated time since the most recent common ancestor of the two Altai chromosomes as a fraction of the genome average from the PSMC (SI 12)
$tmrca^i_{den}$ =	the estimated time since the most recent common ancestor of the two Denisova chromosomes as a fraction of the genome average from the PSMC (SI 12)
p^i_{alt} =	derived allele frequency in Altai (0, 0.5 or 1)
p^i_{den} =	derived allele frequency in Denisova (0, 0.5 or 1)
p^i_{afr} =	derived allele frequency in sub-Saharan Africans (0 to 1)
p^i_{HC} =	probability that the site differs between chimpanzee and an African haplotype (0 to 1)

p^i_{CM} = probability that the site differs between chimpanzee and macaque (0 or 1); this is set to missing for sites for which we do not have macaque data.
h^i_{alt} = indicator variable for Altai heterozygote: 1 if p^i_{alt}=0.5 and 0 otherwise
h^i_{den} = indicator variable for Denisova heterozygote: 1 if p^i_{den}=0.5 and 0 otherwise
h^i_{het} = probability that the site is heterozygous in a pool of 10-12 African alleles (0 to 1)
$p^i_{alt\text{-}min}$ = derived allele frequency in Altai at sites where p^i_{den}=0. When Altai is effectively haploid ($tmrca^i_{alt}$ < 0.0002/bp) this is proportional to the divergence to the closer of the two Denisova haplotypes, since when we require there to be no Denisova derived alleles, we count all the mutations in Altai since diverging from the closer haplotype.
$p^i_{den\text{-}min}$ = derived allele frequency in Denisova at sites where p^i_{alt}=0. When Denisova is haploid, this is proportional to the divergence time to the closer of the two Altai haplotypes.

We wish to infer quantities proportional to divergence time. Since we normalize divergent sites counts by human-chimpanzee divergence in the same windows, we need to correct for variation in mutation rate across the genome as well as variation in human-chimpanzee genetic divergence. To obtain our normalizing quantities we screened all *G* bases of the genome that not only pass the filters described above but that also have data from macaque, and computed the number of African-chimpanzee and chimpanzee-macaque divergent sites:

$Div_{HC} = \sum_i^G p^i_{HC}$ = total number of human-chimpanzee divergent sites
$Div_{CM} = \sum_i^G p^i_{CM}$ = total number of chimpanzee-macaque divergent sites

We also wished to estimate the amount of branch-shortening genome-wide on the Altai and Denisova lineages relative to present-day humans as a fraction of human-chimpanzee genome-wide average genetic divergence, so that our estimates of divergence time could be interpreted in terms of years since the present. Thus, we computed the following genome-wide estimate of branch shortening. We note that these branch shortening estimates in practice differ slightly from those in Figure S16b, due to the different filters we used. For this analysis, we thought it best to correct for the degree of branch shortening that we empirically observe when applying the same set of filters.

$S_{alt} = \left(\sum_i^G \left[p^i_{afr} - p^i_{alt}\right]\right)/Div_{HC}$
= branch shortening on the Altai side of the tree as a fraction of human-chimp
$S_{den} = \left(\sum_i^G \left[p^i_{afr} - p^i_{den}\right]\right)/Div_{HC}$
= branch shortening on the Denisova side of the tree as a fraction of human-chimp

For computing local estimates of divergence, we divide the genome into 0.01cM non-overlapping windows using recombination rates from the Oxford linkage disequilibrium map[4], restricting to windows where the number n_j of nucleotides passing our filters was at least 50,000 bp of which at least 60% had a macaque call (we required data from a large number of nucleotides so as to be able to make an accurate divergence computation). This requirement filtered out most of the genome: the span of genome available for analysis fell from 1,426 Mb passing the basic filters in SI 5 to 168 Mb. For each window, we then computed the following statistics:

$T^j_{alt} = \sum_i^{n_j} tmrca^i_{alt}/n_j$ = average Altai TMRCA in the n_j bases in the window
$T^j_{den} = \sum_i^{n_j} tmrca^i_{den}/n_j$ = average Denisova TMRCA in window
$N^j_{HC1} = \sum_i^{n_j} p^i_{HC}$ = number of human-chimp divergent sites in window
$N^j_{HC2} = \sum_i^{n_j(macaque\ presnt)} p^i_{HC}$ = number of human-chimp divergent sites in window restricting to sites where we have a macaque call
$N^j_{CM} = \sum_i^{n_j(macaque\ present)} p^i_{CM}$ = number of chimp-macaque divergent sites in window
$norm_{HC} = \frac{N^j_{HC2}/N^j_{CM}}{Div_{HC}/Div_{CM}}$ = human-chimp divergence time in window divided by genome average correcting for mutation rate variation

$D^j_{alt\text{-}min} = 2\left[S_{alt} + (norm_{HC})\frac{\sum_i^{n_j} p^i_{alt.min}}{N^j_{HC1}}\right]$ = Altai divergence time to closest Denisova haplotype

$D^j_{den\text{-}min} = 2\left[S_{den} + (norm_{HC})\frac{\sum_i^{n_j} p^i_{den.min}}{N^j_{HC1}}\right]$ = Denisova divergence time to closest Altai haplotype

$D^j_{alt\text{-}ave} = 2\left[S_{alt} + (norm_{HC})\frac{\sum_i^{n_j} p^i_{alt}(1-p^i_{afr})}{N^j_{HC1}}\right]$ = Altai div. time to average Denisova haplotype

$D^j_{den\text{-}ave} = 2\left[S_{den} + (norm_{HC})\frac{\sum_i^{n_j} p^i_{den}(1-p^i_{afr})}{N^j_{HC1}}\right]$ = Denisova div. time to average Altai haplotype

$D^j_{alt\text{-}max} = 2D^j_{alt\text{-}ave} - D^j_{alt\text{-}min}$ = Altai divergence to more distant Denisova haplotype

$D^j_{den\text{-}max} = 2D^j_{den\text{-}ave} - D^j_{den\text{-}min}$ = Denisova divergence to more distant Altai haplotype

$D^j_{alt\text{-}afr} = 2(norm_{HC})\frac{\sum_i^{n_j} p^i_{alt}(1-p^i_{afr})}{N^j_{HC1}}$ = Altai divergence to average African on African side

$D^j_{den\text{-}afr} = 2(norm_{HC})\frac{\sum_i^{n_j} p^i_{afr}(1-p^i_{den})}{N^j_{HC1}}$ = Denisova divergence to average African on African side

$T^j_{alt} = (norm_{HC})\frac{\sum_i^{n_j} h^i_{alt}}{N^j_{HC1}}$ = Altai average TMRCA

$T^j_{den} = (norm_{HC})\frac{\sum_i^{n_j} h^i_{den}}{N^j_{HC1}}$ = Denisova average TMRCA

The left panels of Figure S15.2 plot $D^j_{alt\text{-}min}$ against T^j_{den} for windows where Altai is effectively haploid (T^j_{alt}<0.0002/bp). At windows of the genome where Altai divergence to the closest Denisova haplotype is low (the fiftieth of the genome of lowest $D^j_{alt\text{-}min}$), the average TMRCA of the two Denisova chromosomes is elevated compared with the genome-wide average, which we see both in the scatterplot (Panel A) and binned view of the data (Panel B).

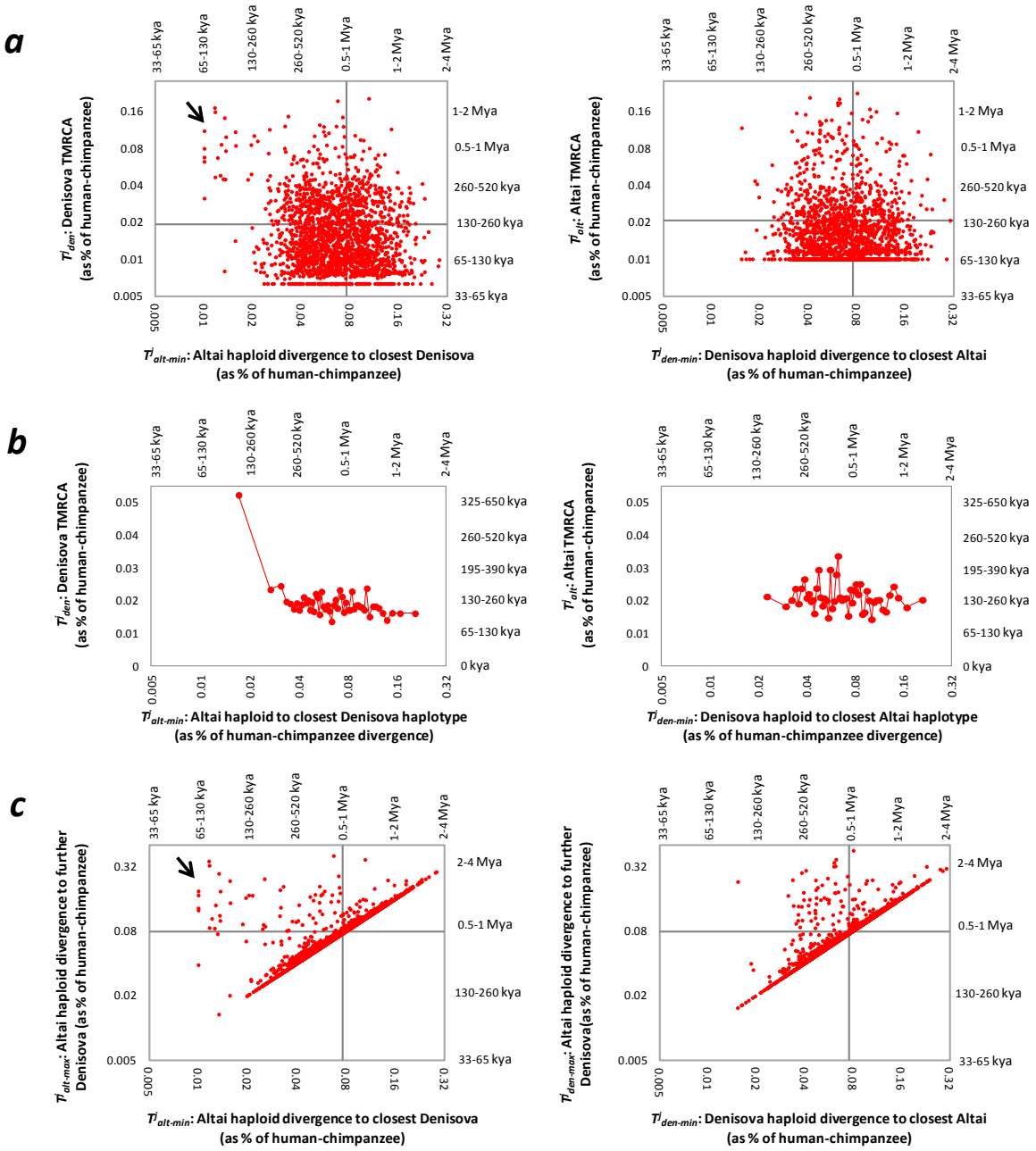
The right panels of Figure S15.2 test for the alternative history of Denisova gene flow into Neandertal, plotting $D^j_{den\text{-}min}$ against T^j_{alt} for windows where Denisova is effectively haploid (T^j_{den}<0.0002/bp). There is no evidence of windows with low divergence of Denisova to the closest Altai haplotype that also have elevated Altai heterozygosity, either in the scatterplot or binned data.

Another way to see that the only clear haplotype-based signal of gene flow is from Neandertal into Denisova (and not in the reverse direction) is that there are 17 windows where $D^j_{alt\text{-}min}$<0.015 compared with 0 windows where $D^j_{den\text{-}min}$<0.015. This divergence as a fraction of human-chimp corresponds to 100-200 kya, likely less that the main Neandertal-Denisova divergence (SI 12).

(iv) Neandertal-into-Denisova introgression was recent and occurred in a low proportion
Figure S15.2 also reveals two further features of the Neandertal-into-Denisova introgression.

(a) The Neandertal introgression occurred relatively recently in the history of the Siberian Denisovan
Strikingly, when we plot the divergence to the closer Denisova haplotype against the divergence to the more distant Denisova haplotype ($D^j_{alt\text{-}min}$ vs. $D^j_{alt\text{-}max}$), we observe that loci where the Altai Neandertal is closer to one of the Denisova haplotypes are almost never particularly close to the other (the other haplotype has a divergence to Altai that is typical or even slightly higher than the genome average). Given the very high level of genetic drift that we have documented occurred in the last hundreds of thousands of years of Denisovan history in the PSMC analyses of ref. 5 and SI 12, we would expect that if the Neandertal-into-Denisova introgression was more than a few tens of thousands of years old, it would often be the case that the two Denisova haplotypes would share a common ancestor more recently than the separation from the Altai haplotype to which we are comparing them, and thus we would tend to see a very low $D^j_{alt\text{-}max}$ when we see a very low $D^j_{alt\text{-}min}$. The fact that we see no evidence for this suggests that the gene flow is recent, post-dating almost all the genetic drift that occurred in the history of the Denisova individual.

Figure S15.2: Haplotype-based evidence of Neandertal→Denisova gene flow. *This analysis is based on 0.01cM windows with a minimum of 50 kb of screened bases where one archaic sample is effectively haploid (inferred TMRCA from the PSMC < 0.0002/bp). All divergences are corrected for branch shortening and are expressed as a fraction of human-chimpanzee divergence. (a) In windows of the genome with low Altai divergence to the closest Denisova haplotype we also see elevated Denisova heterozygosity (arrow) as expected from introgression (left panel). We see no similar evidence of Denisova introgression into Altai (right panel). (b) The signal is also evident in an analysis where we group the data into 50 bins ranked by the x-axis coordinate. (c) At loci of low Altai divergence to the closest Denisova haplotype, the other Denisova haplotype almost never shows sign of Altai introgression, as its divergence to Altai is typical of the genome average (horizontal line).*



(b) Neandertal introgression contributed at least 0.5% of the Denisovan genome

We have not been able to derive a point estimate of the proportion of Neandertal ancestry in the Denisovan genome, because a robust estimate would require being able to reconstruct a model of relationships among the sequenced Neandertals that is a good statistical fit to the data and we have not

been able to do this (SI 14). However, we were able to use our haplotype analysis to obtain a rough lower bound. To do this, we leverage the fact that from Figure S15.2B (left panel), it is clear that the signal of increased heterozygosity in Denisova is highly enriched in the bin of the genome of lowest haploid Altai divergence to the closest Denisova haplotype. We can use this observation to estimate the fraction of the windows in the lowest 50th of the genome that are likely introgressed, which in turn provides a minimum ancestry estimate.

In detail, in the 50th of the genome with the lowest divergence of Altai to the closest Denisova haplotype, the average time since the most recent common ancestor of the two Denisova haplotypes is 5.2% of the human-chimpanzee divergence. This value is about half way between the genome-wide average divergence of two Denisova haplotypes of 1.95% (Figure S15.2B left panel), and the genome-wide average divergence of Altai and Denisova haplotypes of 8.04%. Thus, we can conservatively estimate that at least half of the windows in this 1/50th of the genome reflect Neandertal introgression. We then further multiply by a factor of ½ to reflect the fact that at these loci there is usually one Denisovan and one Neandertal origin haplotype (there has been very little genetic drift in Denisova since the Neandertal introgression; see above). Thus, our lower bound is 0.5% = (1/2)×(1/50)×(1/2).

It is tempting to view 0.5% as a point estimate for the proportion of Neandertal ancestry in the Denisovan genome. However, we caution that this is in fact a conservative lower bound corresponding to the amount of the genome we were able to directly measure as having Neandertal ancestry. One reason it is a lower bound is that it is likely that the lowest fiftieth of the genome with respect to Neandertal divergence to the closest Denisovan haplotype does not in fact contain all the truly Neandertal introgressed segments. The second reason is that the windows of the genome we are analyzing may harbor recombinant haplotypes that combine truly Neandertal introgressed segments with Denisovan segments, which would raise the average divergence (diluting our signal) making these loci too difficult to detect. While we are analyzing windows that are supposed to be 0.01cM in size to avoid the effects of recombination, we are also requiring the windows to span a large physical distance (at least 50 kb), which means that if the recombination map has shifted between archaic and modern humans recombination may very plausibly have occurred within the regions we are analyzing.

(v) Specific loci that are introgressed from Neandertals into Denisovans

To obtain further insight into the architecture of regions of Altai introgression into Denisova, we plotted $D^{j}_{alt\text{-}min}$ across the genome for non-overlapping windows that are of a minimum size of 200 kb of data passing our basic filters (that is, we merge consecutive 0.01cM windows until they span at least 200 kb of covered nucleotides). For this analysis, we do not restrict to loci where Altai is effectively haploid (we no longer filter based on the average T^{j}_{alt} in the window), because if Altai is effectively diploid in the region, this will overestimate the divergence to the closest Denisova haplotype, which causes a search for loci of low divergence to be more conservative.

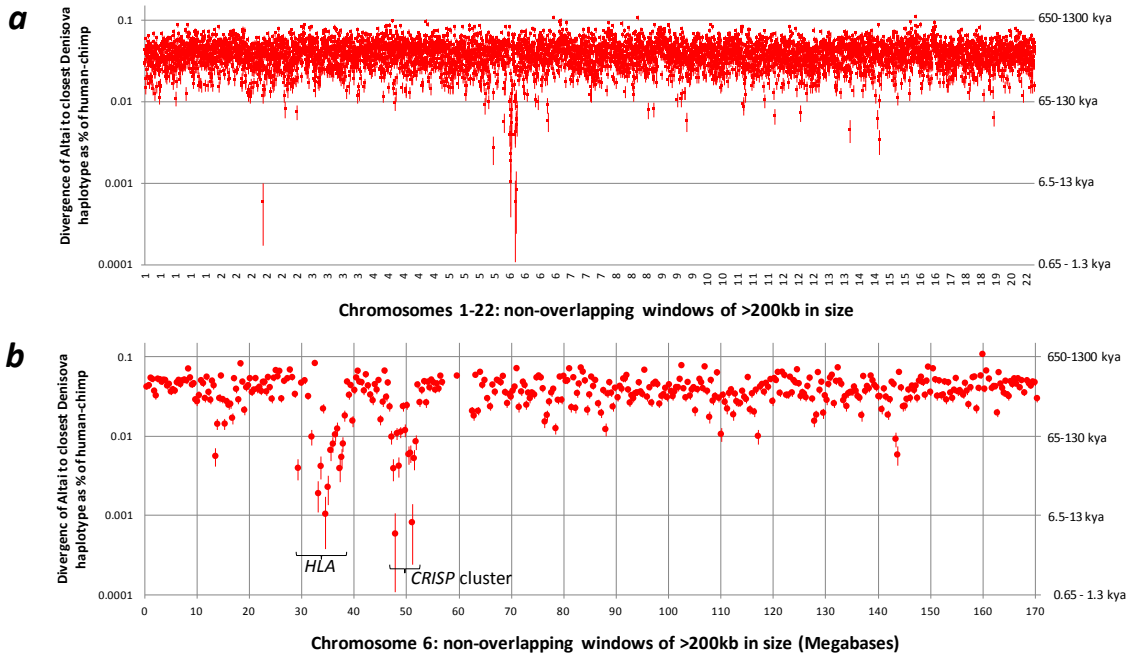
Figure S15.3 shows the results. There are multiple loci in the genome where $D^{j}_{alt\text{-}den}$ is extraordinarily reduced compared with the genome average: sometimes to <1% of human-chimpanzee divergence over multiple consecutive windows which corresponds to <130,000 years before the date when the Altai individual lived assuming (conservatively) that the mutation rate is at the slow end of the range that has been reported in the literature: μ=0.5×10^{-9}/bp/year. We note that we are not correcting for branch-shortening in this analysis, and so this is not the date in the past. The lower panel shows a blow-up of chromosome 6, which contains the two largest loci:

(a) The first locus spans 47.5-51.3 Mb in *hg19*, and includes 22 genes including the *CRISP* cluster. *CRISP1* and *CRISP2* are known to have a role in sperm maturation and egg fertilization, while CRISP3 is known to have a role innate immunity. Here, 12 of 14 windows have divergence between Altai and the closest Denisova haplotype of <130,000 before the date that Altai Neandertal lived assuming a mutation rate of μ=0.5×10^{-9}/bp/year, and one window has an estimated divergence of 8,000 and a 95% confident upper bound of <18,000 years before the date that the Altai Neandertal lived.

(b) The second locus includes *HLA*. This locus spans 29.5-37.5 Mb in *hg19*, and includes 220 genes including MHC class I and class II genes with a central role in immunity. Here, 13 of 18 windows have divergence between Altai and the closest Denisova haplotype of <130,000 years before the date that Altai lived assuming a mutation rate of μ=0.5×10^{-9}/bp/year, and one window has an estimated divergence of 14,000 years and a 95% confident upper bound of <27,000 years before Altai lived. The Neandertal introgressed allele may have evolved under selection in Denisova and may have contributed to the variation of immunological traits in Denisovans, similar to the evidence of selection for Neandertal and Denisova derived haplotypes in modern humans[6]. However, sequence from more Denisova individuals is needed to substantiate this hypothesis.

We were concerned that our strong signal at *HLA* might be an artifact of balancing selection, which is known to occur at the *HLA* locus[6-9]. However, we ruled this out as a likely explanation, since for each window in the genome, we express the divergence as a ratio of local divergence per base pair divided by local human-chimpanzee divergence, times a normalization factor $norm_{HC}$ = $\left(N^j_{HC2}/N^j_{CM}\right)/(Div_{HC}/Div_{CM})$ that uses macaque data to correct for variation in mutation rates across the genome and variation in human-chimpanzee genetic divergence time. For this normalization to work, the chimpanzee-macaque genetic divergence time locally needs to be the same (or at least similar to) what it is genome-wide. Encouragingly, when we measure chimpanzee-macaque genetic divergence per base pair across the >200 kb windows, we find that the fluctuation is at most in the range of a few tens of percent: it ranges over 93-116% of the genome average rate in the *CRISP* cluster, and over 83-121% of the genome average in the *HLA* cluster. These ranges are not atypical for >200 kb segments in the genome, and thus in fact may mostly entirely variation in mutation rate rather than real variation in genetic divergence. In any case, fluctuations of a few tens of a percent are not sufficient to explain the observation that the genetic divergence between Altai and Denisova haplotype in the *HLA* region is one to two orders of magnitude lower than the genome average.

Figure S15.3: Scan of Altai divergence to the closest Denisova haplotype. *The scan is comprised of windows of a minimum of 200kb of nucleotides passing our filters, formed by joining windows of 0.01cM in size. (a) The scan reveals multiple loci of low divergence between Neandertal and the closest Denisova haplotype, with upper bounds of divergence (bars show ±1 standard error) far below the estimate of several hundred thousand years for the main population divergence inferred in SI 12. (b) The largest loci of low divergence are on chromosome 6, and include HLA.*



(vi) Conclusion

In this note, we have documented a history of Neandertal gene flow into the Siberian Denisovan genome, which we conservatively estimate contributed at least 0.5% of the Siberian Denisovan's ancestry. We have further shown that this gene flow occurred relatively recently in the history of the Siberian Denisovan, as her ancestors experienced very little genetic drift since the gene flow. We have finally created a map of specific locations in the genome that are likely to be introgressed, which includes phenotypically important loci such as HLA.

References

[1] Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B (2013) Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193, 1233-54.

[2] Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking N, Pakendorf B, Reich D (2013) Ancient west Eurasian ancestry in southern and eastern Africa. arXiv:1307.8014.

[3] Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PL, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S (2010) A draft sequence of the Neandertal genome. Science 328, 710-22.

[4] Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310, 321-4.

[5] Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222-6.

[6] Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G, Donnelly P, McVean G, Przeworski M (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 339, 1578-82.

[8] Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335, 167-70.

[9] Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci USA, 1989, 958-62.

Supplementary Information 16a

Denisova has some ancestry from an archaic population not related to Neandertals

Sriram Sankararaman, Nick Patterson, Swapan Mallick and David Reich*

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

(i) Overview

- Sub-Saharan Africans shared more derived alleles with Neandertals than with Denisovans, a signal that grows stronger for alleles that occur at 100% frequency in Africans. This can be parsimoniously explained by gene flow from an unknown archaic population into Denisovans.

- We estimate that the unknown archaic population contributed 2.7-5.8% of the ancestry of Denisovans, and diverged from modern humans and Neandertals 0.90-1.40 million years ago (assuming a mutation rate of 0.5×10^{-9}/bp/year).

(ii) Dataset

All the analyses in this note are based on alignments of reads to the human reference genome *hg19*. For the deeply sequenced genomes, we applied the stronger of the two sets of filters described in SI 5 (Map35_100%), and further required that genotype quality scores were GQ≥45. For analyses that require knowledge of the ancestral allele, we restrict to sites with coverage from at least 2 great apes (chimp and at least one of gorilla and orangutan) and require that the great ape alleles agree. We also restrict to transversion polymorphisms which have a lower probability of recurrent mutation.

For some analyses, we capitalized on the fact that the Altai individual is inbred, and that both Altai and Denisova have low diversity. Thus, for substantial fractions of the genome, the two chromosomes from these two archaic individuals coalescence more recently than their split from other samples to which we are comparing them. At these segments, we can randomly pick one of the two haplotypes at each nucleotide and our results will not be affected by which one we choose. To infer where the segments of recent coalescence are, we used the Pairwise Sequential Markovian Coalescent (PSMC) (SI 12), and restrict analyses to subsets of the genome of these two individuals where the genetic divergence time of the two haplotypes measured in per-base-pair units is inferred to be <0.0002/bp. This is around the time of the population divergence time of the Altai and Denisova genomes (SI 12).

For some purposes, we wished to analyze data from a large number of sub-Saharan Africans. We downloaded read data from 107 unrelated YRI Nigerians from the 1000 Genomes Project project website (http://www.1000genomes.org/), specifically analyzing individuals NA19160, NA18865, NA18864, NA18867, NA18861, NA18868, NA18908, NA18909, NA18907, NA19118, NA19119, NA19113, NA19116, NA19117, NA19114, NA19175, NA19171, NA19172, NA18873, NA18870, NA18871, NA18876, NA18877, NA18874, NA18878, NA18879, NA18486, NA18489, NA18488, NA19206, NA19207, NA19204, NA19200, NA19201, NA19209, NA19149, NA19141, NA19143, NA19144, NA19147, NA19146, NA18881, NA18498, NA18499, NA19121, NA19213, NA19099, NA19098, NA19096, NA19095, NA19093, NA19092, NA18520, NA18522, NA18523, NA19152, NA19153, NA19257, NA19256, NA19159, NA19239, NA19238, NA19236, NA19235, NA18934, NA18933, NA19248, NA19247, NA19185, NA19184, NA19225, NA19189, NA19222, NA19223, NA18924, NA18923, NA19130, NA19131, NA18517, NA18516, NA18511, NA18510, NA19137, NA19138, NA18519, NA18507, NA19190, NA19197, NA19210, NA19198, NA19214, NA18508, NA18858, NA18910, NA18912, NA18915, NA18917, NA18916, NA18853, NA18856, NA18502, NA18501, NA19108, NA18504, NA18505, NA19107 and NA19102. The data that we downloaded consists of an average of 7.1× coverage from 100 base pair paired-end reads. Because this coverage is too low to support diploid calls, we sampled just one of the two alleles for sites and individuals where

we could make this "haploid" call with high reliability. To increase the reliability of the call, we restricted to reads with map quality score of MAPQ≥37 and base quality≥30. To further increase reliability, we only called alleles for individuals with at least 3 reads passing these filters, so that we could call the allele based on which one appeared in the majority and have at least 2 reads supporting the call (we took the most common allele out of A, C, G or T, and broke ties randomly).

(iii) Present-day Africans share more derived alleles with Neandertals than with Denisovans
We computed the divergence per base pair to Altai and to Denisova of 6 sub-Saharan Africans in Panels A and B (2 Mbuti, 2 Yoruba and 2 Dinka). To perform this computation, we examined all nucleotides passing the filters described in the previous section, and that furthermore had coverage from at least 5 of the 6 African samples, Altai, Denisova, and chimpanzee. We then computed two times the number of African-specific divergent sites, divided by the number of human-chimpanzee divergent sites, averaged over all of the African chromosomes analyzed. This calculation only uses data on the African side of the tree, and is thus unaffected by missing evolution on the archaic lineage. We compute a standard error using a Block Jackknife with 100 contiguous equally sized blocks.

Table S16a.1 reports the results, which show that sub-Saharan Africans are significantly more diverged on a per-base-pair basis from Denisova than from Altai (similar results are obtained with a different set of filters in SI 6). This is not an artifact of the different ages of the Altai and Denisova samples—or different mutation rates in Altai and Denisova—since we are computing the divergence on the present-day African side of the tree where there is no branch shortening and where the mutation rate is guaranteed to be the same for comparisons to both samples. The significance of the difference between these two estimates is in truth greater than might be expected from the standard errors in Table S16a.1, as the standard errors are correlated for the Altai and Denisova computations.

Table S16a.1: Africans are less diverged genetically from Altai than from Denisova

	Altai		*Denisova*	
African sample	***Divergence***	***Std. Err.***	***Divergence***	***Std. Err.***
$Dinka_A$	11.49%	0.05%	11.72%	0.06%
$Yoruba_A$	11.47%	0.05%	11.71%	0.05%
$Mbuti_A$	11.74%	0.05%	11.98%	0.06%
$Dinka_B$	11.34%	0.05%	11.57%	0.05%
$Yoruba_B$	11.32%	0.05%	11.56%	0.05%
$Mbuti_B$	11.33%	0.05%	11.56%	0.05%

Note: The analyses in this table are based on both transitions and transversions. We report divergence per base pair on the African side of the tree divided by African-chimp divergence, and multiply by two to correct for using only one side of the tree. Standard errors are from a Block Jackknife with 100 blocks.

To compute the significance of the difference between the genetic divergence of Africans to Altai and Denisova, we use *D*-statistics, which were first described in ref. 1 and which we define precisely as in SI 14. Specifically, we compute *D(Altai, Denisova; Africa, Chimpanzee)* to evaluate if there is evidence for Africans sharing more derived alleles with one archaic group or the other. (In this note we do not analyze the statistic *D(Altai, Denisova; Non-African, Chimpanzee*) because it is confounded by the history of Neandertal gene flow into modern humans.) We analyzed transversion polymorphisms where we could determine an ancestral allele based on data from at least two apes.

Table S16a.2 reports the results, which show that Africans share significantly more derived alleles with Altai than with Denisova. We represented Africans for this analysis in two ways. First, we represented Africans by a pool of 6 deeply sequenced individuals (2 Dinka, 2 Yoruba, and 2 Mbuti), restricting to sites with coverage from at least 5 individuals (thus we had between 10-12 chromosomes at each analyzed site). Second, we represented sub-Saharan Africans by a pool of 107 YRI individuals sequenced to an average coverage of 7.1×, in each of whom we sampled a single chromosome and restricted to locations with coverage from at least 80 individuals (thus, we had a coverage of 80-107

chromosomes per site). For both datasets, we observe that Africans share 7.0% more derived alleles with Altai than Denisova (significant at Z=11.6 and Z=13.0 standard errors from zero) (Table S16a.2). We also performed a secondary analysis in which we restricted the computation to sites where all African alleles analyzed were derived. Here, Africans share 13.4% and 15.9% more derived alleles with Altai than with Denisova for the deep sequences and YRI individuals respectively.

Table S16a.2: Africans shared more derived alleles with Altai than they do with Denisova

No. of chromosomes	*Source of sequencing data*	D_{basic}		D_{fixed}	
		D-statistic	*Z-score*	*D-statistic*	*Z-score*
10-12	Panel A and B deep sequences	0.070	11.6	0.134	10.0
80-107	YRI from 1000 Genomes	0.070	13.0	0.159	11.8

Note: The analyses in this table restrict to transversion polymorphisms. We restrict to sites where we have coverage from at least 10 sub-Saharan African chromosomes (drawn from $Mbuti_A$, $Mbuti_B$, $Dinka_A$, $Dinka_B$, $Yoruba_A$ and $Yoruba_B$) and at least 80 light-coverage YRI sequences. For D_{basic} we weight the site by the probability of Africans carrying the derived allele, and for D_{fixed}, we restrict to sites where all African samples are fixed for the derived allele.

The signal grows stronger at sites with high African derived allele frequencies

Under the null hypothesis that Altai and Denisova are a clade with respect to Africa, the expected value of *D(Altai, Denisova; Africa, Chimp)* is 0, regardless of an allele's frequency in Africans. Thus, testing for consistency of a frequency-stratified *D*-statistic with 0 is a valid way to test for gene flow. Moreover, if there is a dependence of the *D*-statistic on the African allele frequency, it may be informative about the history that led to the signal of Africans being closer to Altai than to Denisova.

We define our frequency-stratified statistics as follows. Let $f_{i,n}$ and $f_{i,d}$ be the derived allele frequencies at site *i* in Altai and Denisova respectively (0, 0.5 or 1). We then compute frequency-stratified *D*- and *S*- statistics, which we call D_j and S_j, restricting to sites where sub-Saharans carry *j* derived alleles. Let $nd10_j$ be the expected number of sites where Neandertal is derived and Denisova is ancestral, and $nd01_j$ be the expected number of sites with the opposite pattern (this is the expectation for what we would get if we randomly sampled a single allele to represent each archaic sample).

$$nd10_j = \sum_i f_{i,n}(1 - f_{i,d}) \quad \text{(S16a.1)}$$

$$nd01_j = \sum_i (1 - f_{i,n}) f_{i,d} \quad \text{(S16a.2)}$$

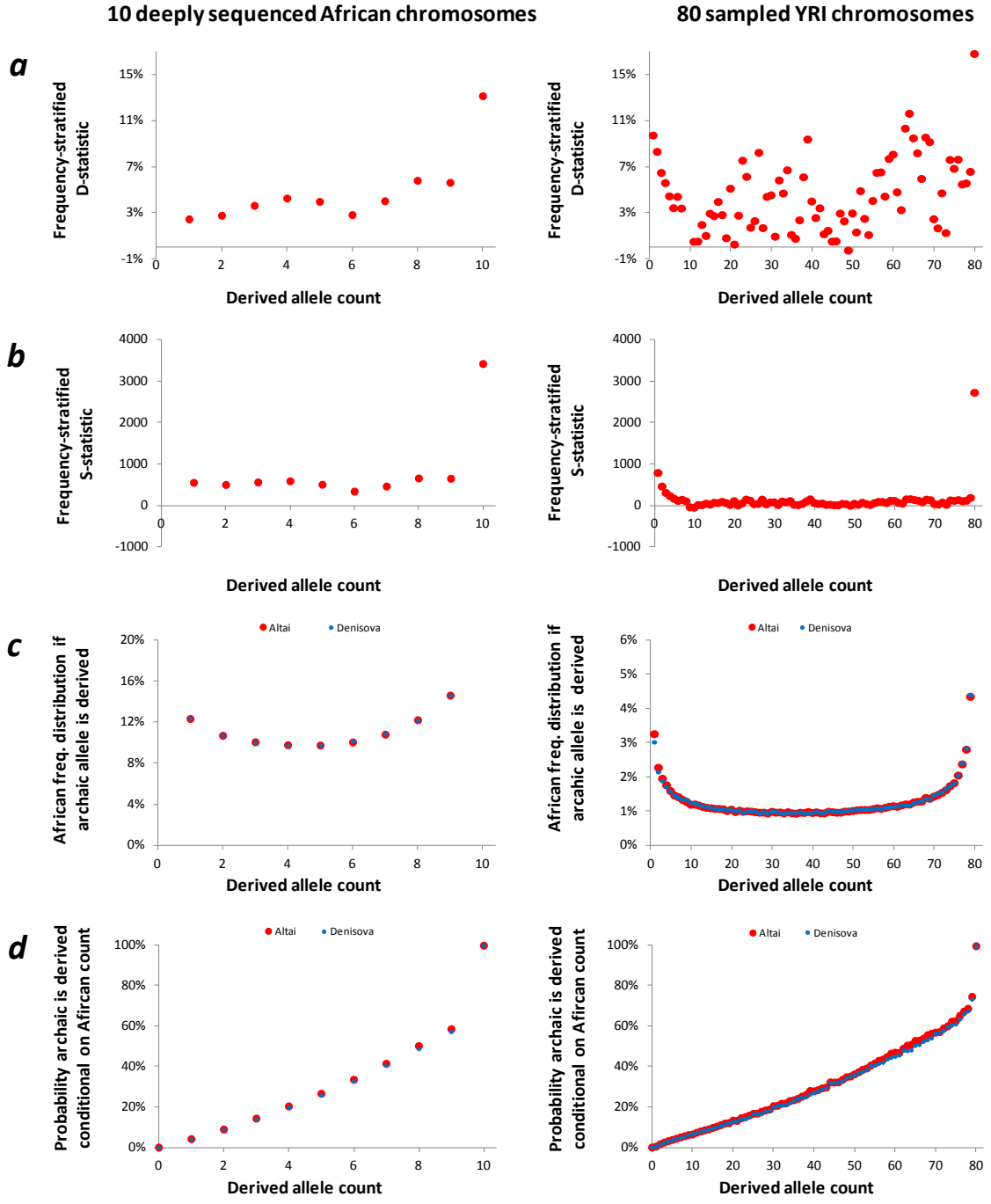
$$D_j = \frac{nd10_j - nd01_j}{nd10_j + nd01_j} \quad \text{(S16a.3)}$$

$$S_j = nd10_j - nd01_j \quad \text{(S16a.4)}$$

Figure S16a.1 shows the dependence of D_j and S_j on derived allele count for 10 deeply sequenced sub-African chromosomes (randomly down-sampled from a maximum of 12 using a hypergeometric distribution), as well as 80 YRI chromosomes (randomly down-sampled from a maximum of 107).

The most striking feature of Figure S16a.1—and the focus of the modeling analyses that follow—is that both D_j and S_j show a discontinuous jump at sites where Africans all carry the derived allele. Defining the statistic where all Africans are derived as D_{fixed}, we find that for the 10-12 deep sequences, $D_{basic} = 0.070 \pm 0.006$ increases to $D_{fixed} = 0.134 \pm 0.013$, and for the 80-107 YRI alleles, $D_{basic} = 0.070 \pm 0.005$ increases to $D_{fixed} = 0.159 \pm 0.013$ (Table S16a.2). In fact, about half of the deviation of the *S*-statistic from 0 (the absolute excess of *nd10-nd01* sites at positions where a randomly sampled African chromosome is derived) is contributed by sites where all sampled Africans carry the derived allele (56% for the deep sequences, and 47% for the light-coverage YRI samples).

Figure S16a.1: Summary statistics as function of derived allele frequency. *We restrict to sites with ≥10 of a maximum of 12 diverse African chromosomes or ≥80 of 107 YRI chromosomes, and to transversion polymorphisms where we determine the ancestral allele based on agreement of two apes. We show the number of derived alleles from randomly down-sampling of 10 (left) or 80 (right) chromosomes. (a) D_j: frequency stratified D-statistics. (b) S_j: frequency stratified S-statistics. (c) African derived frequency for sites that are polymorphic in Africans and where a randomly selected archaic allele is derived. (d) Probability that the archaic is derived condition on African frequency.*



A second notable feature of these analyses is that the D_j and S_j decrease for low YRI derived allele frequencies (derived allele counts of 1-9), before increasing (derived allele counts 10-80) (right panels of Figure S16a.1). A decreasing value of D_j and S_j for low derived allele frequencies is what would be expected due to gene flow from a population related to Neandertals into the ancestors of the YRI. Such a history would inject Neandertal-related mutations into the modern human lineage at a frequency no larger than the mixture proportion, thus producing effects at the low derived allele

frequency end of the spectrum. If this gene flow occurred recently enough that there has not been much genetic drift in the YRI since mixing into the sub-structured population, then the signal could be concentrated at alleles of <10% derived frequency as we in fact empirically observe. In SI 13, we show that this and several other features of genetic data in YRI can parsimoniously be explained by a very small amount of West Eurasian gene flow into the YRI in the last approximately ten thousand years. While this is an interesting pattern, we do not focus on it in this note as such a history cannot explain the main pattern driving the asymmetry signal (the great strengthening of the signal at sites where all sub-Saharan Africans carry the derived allele). In passing, we note that the effect of the likely West Eurasian gene flow into sub-Saharan Africans appears to be stronger in the YRI than in a pool of Yoruba, Dinka or Mbuti. When we down-sample the YRI data to 10 chromosomes we continue to see a signal at low derived African frequencies (not shown), even though we do not see a signal for the Mbuti-Dinka-Yoruba pool of 10 chromosomes (right side of Figure S16a.1).

The significant *D*-statistics are not data artifacts

(a) Robustness to variability in read coverage

In the paper on the draft sequence of the Denisova genome[2], we found that read coverage for Denisova and for Vindija Neandertal were strongly correlated to the *D*-statistic measuring whether present-day Africans are more closely related to one archaic group or the other (SI 19 of ref. 2). The issue was sufficiently concerning that the paper refrained from claiming that Africans are more closely related to Neandertals than to Denisovans, even though the *D*-statistic suggested exactly such an effect. (Waddell and colleagues also explored this pattern, and showed that it could be explained if Denisovans harbor unknown archaic ancestry[3,4].)

To explore the effect of read coverage on deep coverage data, for each nucleotide *i* in the genome passing our filters we computed a normalized coverage statistic by taking the difference between the observed coverage x^i_{Altai} and $x^i_{Denisova}$ at site *i* and the mean coverage (μ_{Altai}=52.49, $\mu_{Denisova}$=30.73), and dividing by the standard deviation in read coverage (σ_{Altai}=11.58, $\sigma_{Denisova}$=7.01). This produced a coverage statistic $Cov^i = (x^i_{archaic} - \mu_{archaic})/\sigma_{archaic}$ that was approximately normally distributed so that we could bin the nucleotides in the genome based on deciles of a normal distribution.

Figure S16a.2: The D-statistic is robust to stratification by read coverage. *The left plot shows the effect of Altai and Denisova coverage on D(Altai, Denisova; Africa, Chimp), while the right plot shows the effect of difference in coverage. The error bars give one standard error, while the black and gray lines give the genome-wide mean and standard error. While the D-statistic is clearly affected by coverage, when we require the coverage of Altai and Denisova to be similar (that is, we require coverage difference within the 20-80th percentile so we are likely screening out regions where one of the samples is mismapping), we obtain results consistent with the genome average.*

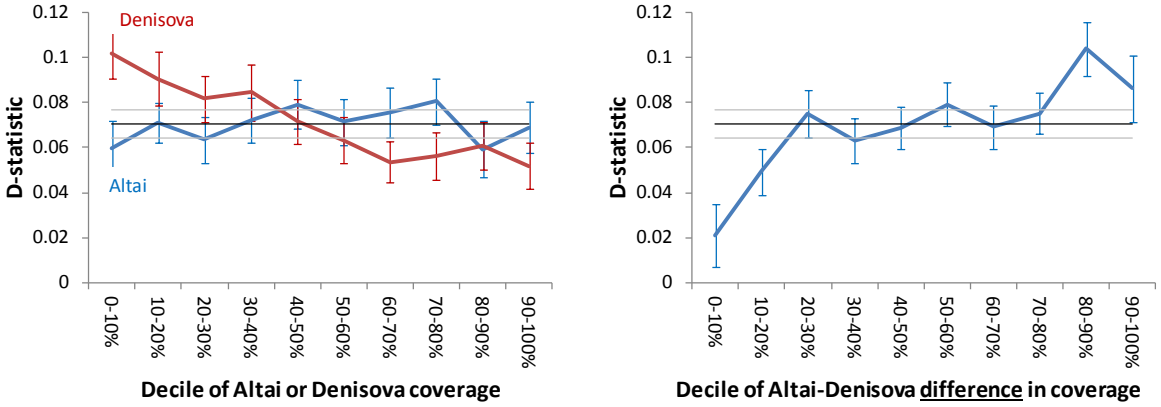


Figure S16a.2 plots *D(Altai, Denisova; Africa, Chimp)* as a function of the decile in coverage. There is no clear effect of Altai coverage, but the *D*-statistic decreases with higher Denisova coverage. We also computed a normalized statistic corresponding to the difference in Altai and Denisova coverage

(based on the applying the normalizing transformation to Cov^{i}_{Altai}-$Cov^{i}_{Denisova}$), and found that the *D*-statistic increases with increasing Altai-Denisova coverage difference.

A possible explanation for the fact that the Altai-Denisova coverage difference is correlated to the *D*-statistic is that at sites that are heterozygous in one of the archaic individuals, the reads with the non-reference allele tend to fail to map, causing not only low coverage, but also an excess rate of matching to the reference sequence. Thus, for example, when Denisova coverage is low, it sometimes shows an excess matching to chimpanzee compared with what should be the case, increasing the value of the *D*-statistic compared with the genome-wide value, and contributing to the coverage effect we observe.

Crucially, however, our observation that Africans share more derived alleles with Altai than with Denisova cannot be explained by mapping error. While Figure S16a.2 shows that coverage differences between Altai and Denisova influence the *D*-statistics—implying that mapping artifacts have some influence—it cannot explain our asymmetry signal. First, we observe a significant *D*-statistic in every coverage bin in Figure S16a.2 indicating that we observe the signal both for sites where Altai reads are more likely to be mis-mapped, and for sites where Denisova reads are more likely to be mis-mapped. Second, when we restrict to sites where coverage is similar between Altai and Denisova so that mismapping errors are likely to be reduced (20-80th percentiles in Figure S16a.2), the *D*-statistics match the genome average.

We note in passing that the effect of coverage documented in Figure S16a.2 is in the opposite direction to that observed in the Denisova draft genome paper[2], where low coverage made the tested sample seem less archaic. There, the coverage was so low that there was no power to study loci of unusually low coverage (the modal coverage for both archaic samples was 1×). Similarly, the loci of high coverage in that paper are likely filtered out in our new study because they are in the >97.5% tail of coverage. We are dealing with a different, more subtle set of artifacts here.

(b) Transition-transversion ratio

If sequencing or mapping error is artifactually causing a deviation of *D(Altai Denisova, Africa,Chimp)* from zero, then it would also be expected to produce a skew in the transition : transversion ratio (ts:tv) away from the expectation for genuine divergent sites. This is a good test because the expected ts:tv ratio for real mutations and errors is very different: 2-3 for real polymorphisms, and <1 for random error. Concretely, we hypothesized that if the excess of *nd10* over *nd01* sites is due to increased errors in one archaic population over the other, there would be a different ts:tv ratio for *nd10* and *nd01* sites.

Table S16a.2 show the empirical ts:tv ratio in the deeply sequenced sub-Saharan Africans, randomly down-sampled to what would be expected for studying 10 chromosomes using a hypergeometric distribution. We report results by African derived allele count. We find no statistically significant differences in ts:tv between *nd10* and *nd01* in any category. It is notable that ts:tv increases for high African derived allele counts for both *nd10* and *nd01* (from around 2.05 to around 2.25). The cause is unclear to us, but since the pattern is seen for both *nd10* and *nd01*, it cannot be contributing to our signal of Africans sharing more derived alleles with Altai than with Denisova.

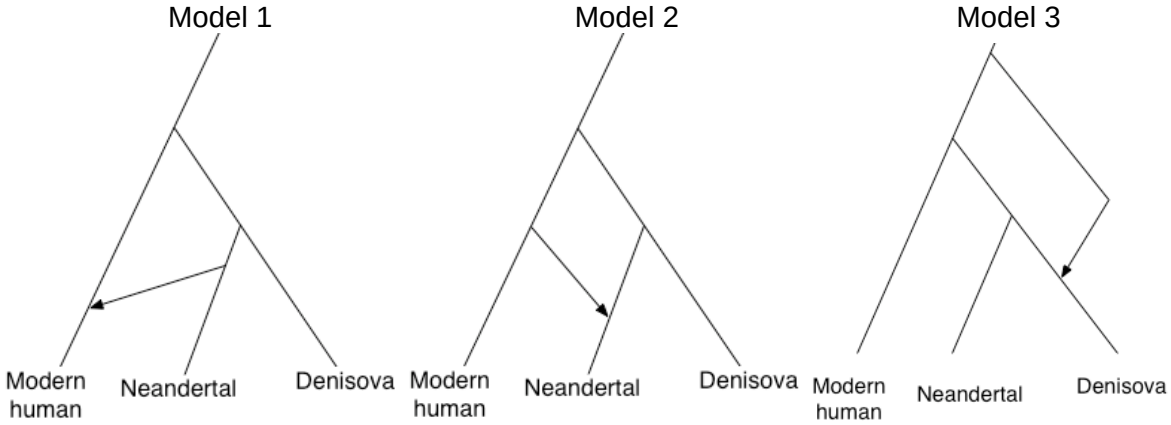
Table S16a.2: Transition:transversion ratio reveals no evidence for an artifactual D-statistic

African	*D-statistic*		*transition : transversion ratio*		
derived count	*transitions*	*transversions*	*nd10*	*nd01*	*P-value for diff.*
1	0.021	0.024	2.07	2.05	0.65
2	0.023	0.028	2.03	2.01	0.63
3	0.035	0.036	2.00	2.00	0.89
4	0.043	0.043	2.01	1.99	0.93
5	0.035	0.040	2.02	2.00	0.64
6	0.036	0.028	1.99	2.05	0.49
7	0.052	0.040	1.99	2.03	0.31
8	0.056	0.058	2.03	2.04	0.85
9	0.060	0.056	2.05	2.04	0.76
10	0.121	0.131	2.27	2.23	0.16

(iv) Unknown archaic gene flow into Denisova predicts the pattern observed in real data
We considered three demographic models that could potentially explain the evidence of Africans being more closely related to Altai than to Denisova (Figure S16a.3):

(Model 1) Gene flow from the Neandertal lineage into the early modern human (EMH) lineage
(Model 2) Gene flow from the EMH lineage into the Neandertal lineage
(Model 3) Gene flow into the Denisova lineage from a lineage unrelated to Neandertals or EMH

Figure S16a.3: Three demographic models consistent with D(Altai, Denisova; Africa, Chimp)<<0.



We note that we are not considering ancient structure models here, whereby the ancestral population of Neandertals, Denisovans, and modern humans was not fully homogenized when the modern human lineage separated from the common ancestral lineage of Neandertals and Denisovans. In such a scenario, the proto-Neandertals would have been genetically closer to the proto-Denisovans at the time of final separation from the modern human lineage, and retained this extra proximity at the time of final separation from modern humans. It is worth pointing out, however, that this ancient structure scenario is similar in some ways to Model 3, as both Model 3 and ancient structure both specify that Denisovans have a component of ancestry that is relatively more diverged from modern humans than do Neandertals. The fact that our data are in fact best fit by Model 3 (see below) could thus in principle be explained either by the type of pulse admixture event we specify in Model 3, or a continuous gene flow scenario among diverged lineages as expected from ancient structure.

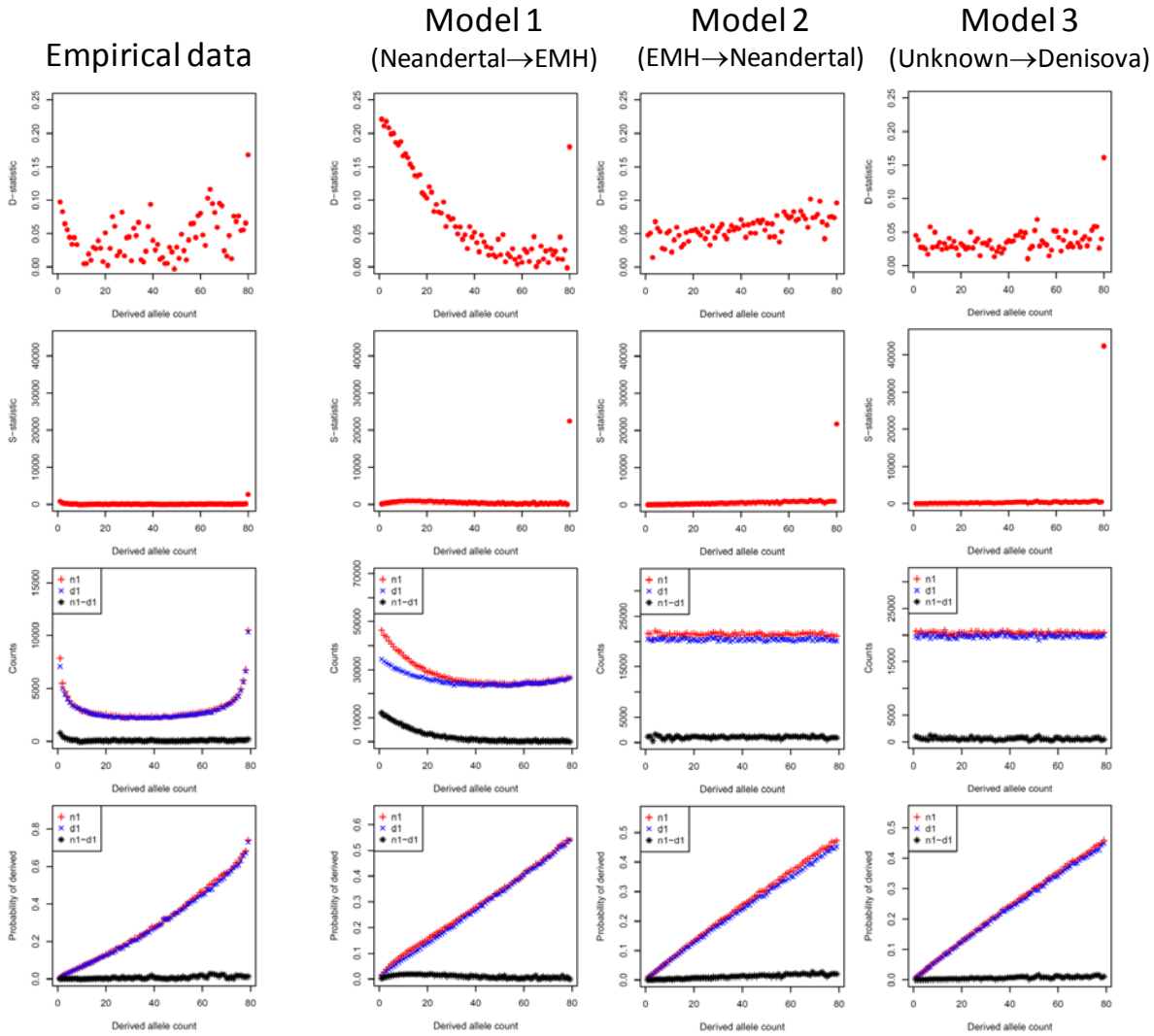
In the remainder of this section, we present multiple lines of argument that allow us to reject Models 1 and 2 as sufficient to explain the key features of the data, and to show in contrast that Model 3 does match important features. The approach here is not to match fully parameterized models to the data, as in the Approximate Bayesian Computation (ABC) analysis of SI 16b, but rather to find qualitative patterns than distinguish the predictions of the three models.

(a) Simulation show that Model 1 predicts qualitative features of the data that are not observed
We ran coalescent simulations to assess the qualitative patterns expected under a concrete example of each of Models 1, 2 and 3.

For each simulation series, we used the *ms* software[5] to generate 3 Gb of simulated sequence data (300,000 loci of 10 kb each), specifying free recombination between loci, an intra-locus recombination rate of 1.3×10^{-8} per bp per generation, and assuming a mutation rate of 2.5×10^{-8} per bp per generation. (This mutation rate is at the high end of recent estimates[6], but an overestimate of the mutation rate should not matter for the inferences reported in this section because it is not expected to affect the allelic configuration at polymorphic sites which is what we analyze here.) We fixed the population split time of the ancestors of modern humans and the common ancestor of Neandertal and Denisova at 12,000 generations, the population split time of the ancestors of Neandertal and Denisova at 8,000 generations, and the time of gene flow at 4,000 generations.

While the three simulation series are all the same with respect to the parameters specified above, the direction of gene flow and the mixture proportion α are different in each of the three models. We set the mixture proportions in each of the three simulated models to approximately match the observed statistic *D(Altai, Denisova; Africa, Chimpanzee)* = 0.07. For Model 1, we simulated gene flow from Neandertal to EMH with a Neandertal admixture proportion of α =0.20. For Model 2, we simulated gene flow from EMH to Neandertal with an EMH contribution of α =0.10. For Model 3, we simulated admixture into Denisova from an unknown archaic population with a proportion of α =0.05 (the unknown archaic population in this model split from the common ancestor of Neandertals, Denisovans and modern humans 30,000 generations ago). We note that our simulations incorrectly treat the ancient samples as sampled from the present day. However, this is not a problem for our analysis as the statistics we analyze are not affected by the branch shortening documented in SI 6b.

Figure S16a.4: A simple simulation series is a better qualitative match to Model 3 than to either Model 1 or 2. *The top row is the D_j spectrum for data and simulations, the second row is the S_j spectrum, the third row is the spectrum conditional on a single derived allele in an archaic sample (Neandertal red and Denisova blue, and the fourth row is the probability of YRI being derived conditional on the specified archaic sample being derived. In this particular set of simulations, no attempt has been made to quantitatively match all the aspects of real data.*



We compared data and simulation results for both the D_i and S_i statistics. We observe that:

- Model 1 (Neandertal gene flow into modern human ancestors) is qualitatively inconsistent with the data in that D_j and S_j decrease with the derived allele count, opposite to the increase we observe.

Intuitively, the reason why Neandertal gene flow into modern human ancestors causes a decrease in D_j and S_j as a function of African frequency is that it introduces Neandertal-related alleles at low frequency (lower that the admixture proportion) into the modern human ancestral population. Since African populations have been large for the last couple of hundred thousand years (SI 12), the Neandertal-derived alleles do not drift much, and remain at the low frequency end of the spectrum causing the asymmetry signal to be concentrated there as we observe.

We note that in the real YRI data, we do see a decrease in D_j and S_j for the very low end of the derived allele frequency spectrum (derived allele counts of <10) (Figure S16a.1 right). This is consistent with Model 1 making a minor contribution to the patterns in real data even if it cannot explain the main pattern. As we show in SI 13, a parsimonious explanation for these patterns is a small amount of West Eurasian gene flow into YRI ancestors in the last ten thousand years, which indirectly introduced a very small amount of Neandertal ancestry explaining our signal.

- In both Models 2 and 3, the D_j- and S_j-statistics are positively correlated with derived allele frequency, matching the main pattern in our data. In the next sections, however, we show that other aspects of the data are more consistent with Model 3 than with Model 2. The key qualitative pattern that distinguishes Model 2 and Model 3—and is already evident in the simulations of Figure S16a.4—is that in Model 3 there is a discontinuous jump in the D_j statistic for alleles that are fixed derived in Africans, while for Model 2 this jump is much smaller. Similarly, the jump in the S_j statistic is larger for Model 3 than for Model 2.

We note that our rejection of Model 1 here is only based on the specific set of parameters used in the simulation of Figure S16a.4 and the intuition that Model 1 should produce a downward sloping curve as we observe in the specific simulation series we report in SI 16a.4 and that we fail to observe in real data. However, in SI 16b we also report an Approximate Bayesian Computation analysis in which we report simulations of Models 1, 2 and 3 over a larger range of parameters than is shown in Figure S16a.4. These simulations also reject Model 1 as sufficient to explain the main asymmetry signal.

(b) More thorough exploration of parameter space using a robust statistic for distinguishing models

A limitation of the simulations reported in Figure S16a.4 is that they explore a small part of parameter space and that their parameters are not fitted to real data. In practice, the parameter space of the models we needed to fit is very large and we found that it is challenging to fully explore this parameter space with simulations. (Nevertheless, we report such a study in the Approximate Bayesian Computation analysis of SI 16b where we make some simplifying assumptions about demographic history to make the simulations more tractable.)

In this section we use an alternative approach for distinguishing between Model 2 and Model 3, which involves focusing on a statistic that is highly informative about the degree of drift on the early modern human lineage since the time of gene flow, and that is not very sensitive to quantities that are difficult for us to estimate like the proportion of gene flow or population sizes at different times in history (Appendix S16a.1). The robust statistic that we analyze is $S_{fixed}/S_{polymorphic}$: the ratio of the S_j-statistic (Equation S16a.4) at sites that are fixed for the derived allele in YRI to the sum of sites that are polymorphic in YRI weighted by the African derived frequency. Empirically this is 0.885.

To assess whether Model 2 could explain the large increase in these statistics at fixed derived vs. polymorphic sites, we simulated sequence data in the same way as above. The amount of data and the recombination and mutation rate assumptions are also the same as in the simulations above. Specifically, for these simulations, we hold the time of gene flow to be 4,000 generations ago for both Models 2 and 3, assume a constant population size of 10,000 in all lineages, and vary the gene flow proportion to match the observed *D*-statistics. While there are many parameters that we could vary (for example, population size changes and mixture proportions), the analysis of Appendix S16a.1 suggests that they are not expected to have much effect on the $S_{fixed}/S_{polymorphic}$ ratio and thus we can study this ratio in a more limited part of parameter space to understand its behavior.

Table S16a.4 shows that as we increase the simulated split time of modern and archaic humans (T_{NA}) from 9,000 to 24,000 generations ago, we observe an increase in $S_{fixed}/S_{polymorphic,}$ as expected if a larger proportion of introgressing mutations are ones that have become fixed in their frequency. We compute a Z-score for the match of the simulated $S_{fixed}/S_{polymorphic}$ to the observed ratio by obtaining a standard error on the observed quantity using a Block Jackknife with 30 equally sized blocks. The empirical ratio is only consistent with the data (D=0.07 and $S_{fixed}/S_{polymorphic}$=0.885) for T_{NA}>=21,000, corresponding to genetic drifts of $\tau_2 = t/2N_e$ >=1.05 on the modern human lineage since the split from the archaic population (Table S16a.3).

Table S16a.3: Simulations of Model 2 cannot match the observed S_F/S_P ratio

T_{NA}	*D-statistics (observed = 0.070)*	$S_{fixed}/S_{polymorphic}$ *(observed = 0.885)*	*Z-score of* $S_{fixed}/S_{polymorphic}$	*Genetic drift ($t/2N_e$) (observed =0.79)*
9,000	0.065	0.414	18.1	0.45
12,000	0.072	0.540	11.4	0.60
15,000	0.074	0.637	6.9	0.75
18,000	0.069	0.716	4.8	0.90
21,000	0.067	0.808	1.2	1.05
24,000	0.067	1.096	-2.6	1.20

Genetic drift on the YRI lineage since the split from Neandertals of τ_2>=1.05, however, is too large to be consistent with the amount of genetic drift that we know has occurred on the modern human lineage since the split from Neandertals. Assuming as a null hypothesis that Model 2 is correct, one set of simplifying assumptions allows us to infer that the drift since T_{NA} was τ_2=0.79 (Appendix S16a.2). In a more general analysis (making fewer assumptions about size constancy on the YRI lineage) and assuming as a null hypothesis that Model, we can bound the amount of genetic drift on this lineage as being $0.479 < \tau_2 < 0.618$ (Equation S16a.27, below). In either case, the estimated genetic drift from the modeling is far less than the amount required in simulations to match our data.

Intuitively, what allows us to reject Model 2 is the degree of excess matching of African derived alleles to Altai compared with matching to Denisova at sites *fixed* for the African derived allele.

Under Model 2, the gene flow is coming from an early modern human population that split from the Neandertal lineage after the main Neandertal-modern human split. From other aspects of the genetic data, we KNOW how much genetic drift occurred in this period. Furthermore, we know that this amount of genetic drift would not be sufficient to generate the observed degree of fixation for the derived allele at the sites that are driving our signal. Thus, we can conclude the gene flow must have occurred from a more anciently diverged archaic population that drifted more.

Under Model 3, in contrast we can easily generate the degree of discontinuity between S_{fixed} and $S_{polymorphic}$ that we observe, simply by moving the split of the unknown archaic lineage far enough back in time to allow more fixation to have occurred. It is important to recognize that this inference is robust to the details of the population size changes that occurred in the ancestral YRI; all that matters to our statistic is the degree of genetic drift that occurred (Appendix S16a.1).

(c) We can reject Model 2 and fit Model 3 without fully specifying demographic history

We next considered more general versions of Models 2 and 3, and studied the predictions that these models make about the joint frequency spectrum for YRI, Altai and Denisova and whether these patterns are consistent with the data. All that these models specify is the topology of population relationships, and that populations were constant in size and freely mixing prior to the separation of modern and archaic humans. They do not specify population split times or admixture times, or the details of population size changes either on the archaic lineages since divergence from modern humans, or on the modern human lineage since divergence from the archaic populations. Thus, the inferences that we make in this section are robust to uncertainty about these quantities.

We begin by analyzing Model 2, using the specific topology and notation specified in Figure S16a.5. The version of Model 2 that we discuss here has 4 parameters that affect the statistics we are measuring. There are 3 genetic drift terms τ_1, τ_2, and τ_3, and an admixing fraction α. In what follows we will use *D* for Denisova, *N* for Altai and *Y* for YRI. We then write $\beta = e^{-\tau_3}$, where β is the probability that 2 random alleles from Denisova and the archaic ancestry in Neandertals N_1 have not coalesced by the time that they join the ancestral population of Neandertals and Denisovans *B*. Our strategy is to derive analytical formulae for various statistics of the joint distribution of derived alleles in (Denisova, Altai, YRI). We define 4 functions of the 4 model parameters, each of which we can readily estimate as shown below and in Appendix S16a.3. This gives us four equations and four unknowns, with some additional constraints. We show in what follows that the system of equations has no feasible solution for Model 2, given the constraints.

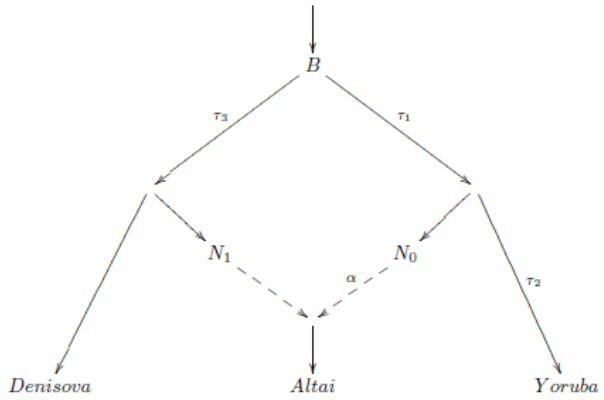


Figure S16a.5: Parameters of Model 2. *The archaic and modern lineages descend from a common ancestral population B. After the split, Neandertals receive a proportion α of ancestry from early modern humans. We do not specify absolute split times or the details of population size changes on the various lineages. The statistics we compute are fully determined by the topology of population relationships, the mixture proportion α, and the drifts τ_1, τ_2, and τ_3, if we assume size constancy in the modern human lineage prior to the split from the archaic samples.*

We always take the derived allele as 1, the ancestral allele as 0, and assume constant population sizes both ancestral to *B* and on the *B*→YRI lineage (we do not need to specify the details of population size changes in the lineages that are specific to the archaic samples). We define the following events:

H	ascertaining two YRI chromosomes and discovering they are heterozygous
N=1	randomly sampling a single Altai allele and discovering that it is derived
D=1	randomly sampling a single Denisova allele and discovering that it is derived
ND10	randomly sampling a single Altai allele and discovering that it is derived and randomly sampling a single Denisova allele and discovering that it is ancestral
ND01	randomly sampling a single Altai allele and discovering that it is ancestral and randomly sampling a single Denisova allele and discovering that it is derived
Y1	randomly sampling a single YRI allele and discovering that it is derived

We first define a normalized difference spectrum conditional on YRI being heterozygous:

$$X_1 = \frac{E(N=1|H) - E(D=1|H)}{E(D=1|H)} \quad \text{(S16a.6)}$$

As described in Appendix S16a.3, the value of X_1 under our model is

$$X_1 = \alpha(e^{\tau_1} - 1) \quad \text{(S16a.7)}$$

We next define

$$X_2 = \frac{E(ND10|H) - E(ND01|H)}{2E(ND10|Y1) - 2E(ND01|Y1)} \quad \text{(S16a.8)}$$

As described in Appendix S16a.3, the expected value of X_2 under our model is

$$X_2 = \frac{e^{-\tau_2}(1-e^{-\tau_1})}{3\tau_1} \quad \text{(S16a.9)}$$

Two features of X_2 are notable and make it particularly valuable for distinguishing Model 2 and 3.

First, the expected value of X_2 is independent of the admixing fraction α, similar to the $S_{fixed}/S_{derived}$ statistic discussed in the previous section. The independence from α means that our inference about the model fit is robust to uncertainty about the true value of the mixing proportion.

Second, the denominator of X_2 is sensitive to alleles in YRI that are fixed derived (we are applying the conditioning *Y1:* a single randomly chosen allele from YRI is derived). This is important, as inspection of Figure S16a.3 shows it is unlikely to be possible to distinguish Model 2 and Model 3 using statistics like X_1 that require YRI to be polymorphic. Specifically, as shown in Figure S16a.3 (third row right), the polymorphic YRI spectrum conditional on *D=1* or *N=1* is essentially independent of which archaic sample is examined. In contrast, we know that there is a large difference between Altai and Denisova at sites where YRI is fixed derived: $S_{fixed}>>0$ (Figure S16a.1).

For the last 2 equations, let *y* be the derived allele frequency in YRI ($0 < y < 1$). Consider the conditional probability

$$Q = P(N = 1|D = 1, y) \quad \text{(S16a.10)}$$

We can show that *Q* is linear in *y* under Model 2 (Appendix S16a.3). Write $Q = X_3 y + X_4$. Then

$$X_3 = \beta(1-\alpha)e^{-2(\tau_1+\tau_2)} + \alpha e^{-2\tau_2} \quad \text{(S16a.11)}$$

$$X_4 = \frac{[(1-\beta)+\beta(1-e^{-2(\tau_1+\tau_2)})](1-\alpha)+\alpha(1-e^{-2\tau_2})}{2} \quad \text{(S16a.12)}$$

Figure S16a.6 shows a plot of the empirical value of *P(N=1|D=1,k)* against *k*/80, where *k* is the number of derived YRI alleles (from *n*=80, using the data as in the right of Figure S16a.1).

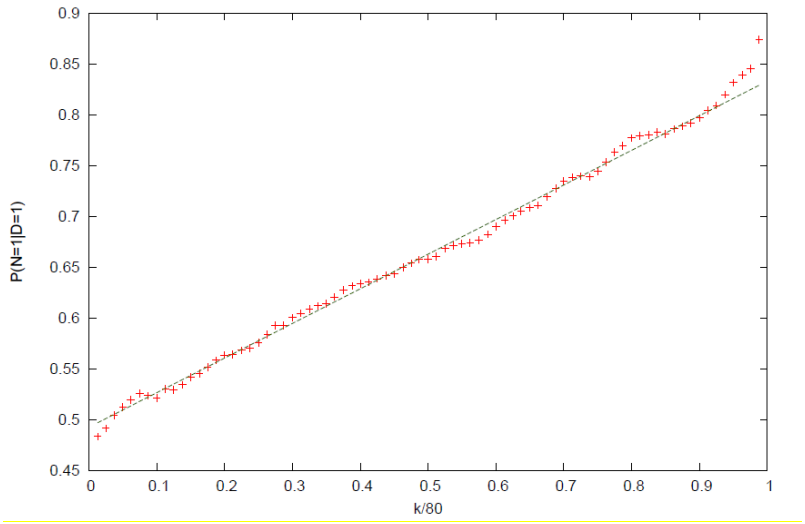


Figure S16a.6: Empirical plot of P(N=1|D=1, k) for 80 YRI chromosomes. *The curve is linear except for deviations at the low and high ends (this deviation may reflect complexity in YRI history that occurred after the archaic split, resulting in the U-shaped distribution in the third row of Figure S16a.1, and which we make no attempt to fit here). We fit the line for 10<k<70.*

The empirical curve is linear except for deviations at the low and high ends that may reflect complexity in YRI history that are not related to the relationship to archaic samples, and that we make no attempt to fit here. We therefore fit the proportion of the spectrum that is linear, consisting of derived allele counts of 10<*k*<70 (Figure S16a.6). We obtain the regression line:

$$P(N = 1|D = 1, k) = u\left(\frac{k}{80}\right) + v \quad \text{(S16a.13)}$$

where

$$\{u, v\} = \{0.3532, 0.4921\} \quad \text{(S16a.14)}$$

As described in Appendix S16a.3, we can show that X_3, X_4 are related to u, v by:

$$(n+2)u = X_3(n+2) + X_4 \quad \text{(S16a.15)}$$
$$v = X_4 n \quad \text{(S16a.16)}$$

where n=80 is the total number of alleles. Thus, estimates of u, v give estimates of X_3, X_4.

Empirically, we obtain the following estimates R_i for X_i:

$$\{R_1, R_2, R_3, R_4\} = \{0.02599, 0.1540, 0.3346, 0.5048\} \quad \text{(S16a.17)}$$

We have constraints. First, τ_1, τ_2 and τ_3 are all non-negative (thus $\beta = e^{-\tau_3}$ must be between 0 and 1). Second, the admixture proportion α must lie between 0 and 1. With these constraints, we now show that there is no feasible solution to the equation $\{R_1,R_2,R_3,R_4\} = \{X_1,X_2,X_3,X_4\}$.

To see the contradiction visually, fix α. From the Equation S16a.7, τ_1 is determined by α, and from Equation S16a.9, τ_2 is determined by τ_1. But Equation S16a.11a is linear in β, so we can solve for β. A plot of β as a function of α shows that $\beta > 2$, which is impossible (Figure S16a.7).

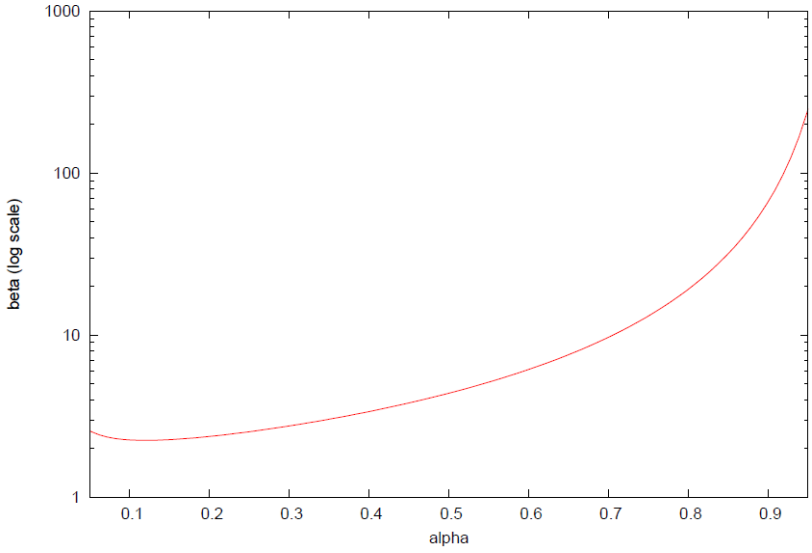


Figure S16a.7: Plot of β vs. α shows that Model 2 is not feasible. *We solve the system of four equations (Equation S16a.6). We find that for all values of the proportion α of early modern human related ancestry in Neandertals ($0 \leq \alpha \leq 1$), the inferred probability of coalescence on the common archaic lineage $\beta = e^{-\tau_3}$ is at least 2, which is greater than the maximum possible of 1.*

We also give a proof that avoids examining the plot, as follows.

First, the left hand of Equation S16a.11 is monotonically increasing in β. Thus, if we set $\beta = 1$:

$$G = (1-\alpha)e^{-2(\tau_1+\tau_2)} + \alpha e^{-2\tau_2} \geq R_3 \quad \text{(S16a.18)}$$

It is clear that $e^{-2\tau_2} \geq G \geq R_3$, and therefore

$$\tau_2 \leq 0.548 = \frac{-\log R_3}{2} \quad \text{(S16a.19)}$$

Crucially, from Equation S16a.7 and Equation S16a.17 together it is clear that τ_2 is a monotonically decreasing function of α.

Further, we can rewrite Equation S16a.9 as

$$e^{-\tau_2} = 3R_2 f(\tau_1) \quad \text{(S16a.20)}$$

where

$$f(\tau) = \frac{\tau}{1-e^{-\tau}} \quad \text{(S16a.21)}$$

We will show that τ_2 is a monotonically decreasing function of τ_1. To do this, it is enough to show that *f(t)* is monotonically increasing. To assess this, we differentiate *f(t)* and check its sign:

$$f'(\tau) = \frac{1}{1-e^{-\tau}} - \frac{\tau e^{-\tau}}{(1-e^{-\tau})^2} = \frac{1-(1+\tau)e^{-\tau}}{(1-e^{-\tau})^2} \quad \text{(S16a.22)}$$

For $\tau > 0$, $1 + \tau < e^{\tau}$, and so $f'(\tau) > 0$. This shows that τ_2 is a monotonically decreasing function of τ_1 and hence a monotonically increasing function of α. It follows that an upper bound for τ_2 yields a lower bound for τ_1. Substituting into Equation S16a.9 and Equation S16a.17 we obtain

$$\tau_1 \geq 0.477 \quad \text{(S16a.23)}$$
$$\alpha \leq 0.044 \quad \text{(S16a.24)}$$

But then we get from Equation S16a.11, using the crudest bounds corresponding to $\alpha e^{-2\tau_2} \leq \alpha$ and $\beta(1-\alpha) \leq 1$ and substituting in the value from Equation S16a.24:

$$R_3 - \alpha e^{-2\tau_2} = \beta(1-\alpha)e^{-2(\tau_1+\tau_2)}$$
$$\Rightarrow R_3 - 0.044 \leq e^{-2(\tau_1+\tau_2)} \quad \text{(S16a.25)}$$

Taking logarithms, and substituting R_3=0.3346, allows us to infer that $\tau_1 \leq (\tau_1+\tau_2) \leq 0.618$.

Now, substituting $\tau_1 \leq 0.618$ into Equation S16a.9 and Equation S16a.17, we obtain:

$$\tau_2 \geq 0.479 \quad \text{(S16a.26)}$$

Taken together, we now have

$$\tau_1 \geq 0.477, \tau_2 \geq 0.479, \text{ and } (\tau_1 + \tau_2) \leq 0.618 \quad \text{(S16a.27)}$$

a contradiction.

Informally, the observed values of X_1 and X_2 imply that $(\tau_1 + \tau_2)$ is large. Specifically, X_1 and X_2 show that a large amount of genetic drift must have occurred on the YRI lineage since the separation from the archaic populations to produce the observed similarity of the Altai and Denisova conditioned polymorphic spectra (X_1), and the observed increase in matching to Altai at fixed derived sites (X_2). This implies that the introgressing material we are analyzing must be from a population that is highly genetically drifted relative to YRI. At the same time, the observed value of X_3 (related to the slope of the regression line in Figure S16a.6), shows that the amount of genetic drift in this period is limited, thus, giving an upper bound on $(\tau_1 + \tau_2)$ which contradicts the lower bound. We emphasize that this inference is robust to any population size changes on the YRI lineage since the separation from archaic populations; it does not require any assumptions about population sizes having been constant in this period as all that matters is the drift τ_2.

We caution that we have not ruled out the possibility that there could be more general versions of Model 2 not specified in Figure S16a.7—for example, substructure in the ancestral population, or

deviation from size constancy prior to the separation of modern and archaic populations—that could explain our data. However, we do not think that in practice such a model could work. First, there is an extreme mismatch of our observed statistics to those expected from theory; the effect is not subtle. Second, the observed concentration of the signal at sites that are fixed derived in YRI can only be explained by a high probability of coalescence of YRI lineages since separation from the introgressing population. Many aspects of genetic data reported here and in other studies suggest that genetic drift on the YRI lineage since separation from Neandertals is not so high.

We next fit a version of Model 3 to the data (Figure S16a.8), using the same strategy as we did to try to fit Model 2. The version of Model 3 that we fit here again has four parameters: three 3 drift terms τ_1, τ_2, and τ_3, and the unknown archaic admixing fraction α (we note that some of these terms are different from those in the previous section due to the different topology). As before, we take the derived allele as 1, the ancestral allele as 0, and assume constant and equal population sizes both ancestral to *A* and on the *A→YRI* lineage.

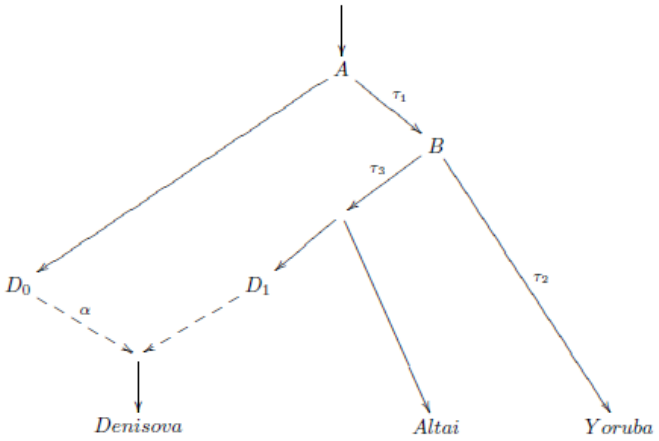


Figure S16a.8: Parameters of Model 3. *An ancient (unknown archaic) population split at A. Denisova received a proportion α of gene flow from this population.*

We define $\beta = e^{-\tau_3}$ as before. In Model 3, as opposed to Model 2, Neandertal has a simple phylogenetic relationship to the other samples. This makes it technically convenient to use statistics where we avoid Denisova appearing in the denominator. We reuse the variable names but caution that they are different in the Model 2 and Model 3 analyses.

We define X_1 and X_2 as follows.

$$X_1 = \frac{E(N=1|H)-E(D=1|H)}{E(N=1|H)} \quad \text{(S16a.28)}$$

$$X_2 = \frac{E(ND10|H)-E(ND01|H)}{2E(ND10|Y1)-2E(ND01|Y1)} \quad \text{(S16a.29)}$$

Their values in terms of our parameters are derived in Appendix S16a.3 and given by:

$$X_1 = \alpha(1 - e^{-\tau_1}) \quad \text{(S16a.30)}$$

$$X_2 = \frac{e^{-\tau_2}(1-e^{-\tau_1})}{3\tau_1} \quad \text{(S16a.31)}$$

We need two more equations, which we defined similarly as in Model 2.

We first define

$$Q = P(D = 1|N = 1, y) \quad \text{(S16a.32)}$$

Q is linear in y (Appendix S16a.3). Write $Q = X_3 y + X_4$. Then

$$X_3 = (\alpha e^{-\tau_1} + (1-\alpha)\beta)e^{-2\tau_2} \quad \text{(S16a.33)}$$

$$X_4 = (1-\alpha)(1-\beta) + (\alpha e^{-\tau_1} + (1-\alpha)\beta)\frac{1-e^{-2\tau_2}}{2} \quad \text{(S16a.34)}$$

We obtain the following estimates:

$$\{R_1, R_2, R_3, R_4\} = \{0.02525, 0.1535, 0.3137, 0.4980\} \quad \text{(S16a.35)}$$

and we solve the equations $X_i = R_i$, i={1, 2, 3, 4} numerically. There is a unique feasible solution after applying the further transformation $\tau_3 = -\ln(\beta)$. Here we quote results with Block Jackknife standard errors (dividing the genome into 100 contiguous equally sized blocks).

$$\alpha = 0.0424 \pm 0.0077 \quad \text{(S16a.36)}$$
$$\tau_1 = 0.906 \pm 0.183 \quad \text{(S16a.37)}$$
$$\tau_2 = 0.356 \pm 0.014 \quad \text{(S16a.38)}$$
$$\tau_3 = 0.431 \pm 0.009 \quad \text{(S16a.39)}$$
$$\tau_1 + \tau_2 = 1.262 \pm 0.185 \quad \text{(S16a.40)}$$

We note that the estimate of $\alpha = 0.0424 \pm 0.0077$ implies that a 95% confidence interval for the proportion of unknown archaic ancestry in Denisova under our model is 2.7-5.8%.

We next took these inferences and converted them into estimates of absolute time. To do this, we use the inference from the PSMC analysis of SI 12. The standard PSMC output infers population size change over time. The PSMC also infers the probability of coalescence of two chromosomes as a function of time $\gamma = 1 - e^{-\tau}$, which translates to drift on the lineage of the two sampled chromosomes using the transformation $\tau = -\ln(1-\gamma)$.

Figure S16a.9 presents this plot for the two Yoruba chromosomes analyzed in SI 12, along with the estimated genetic drift on the YRI lineage since the separation from the common ancestral population with Altai and Denisova of τ_2 = 0.329-0.383 (95% confidence interval), and the estimated genetic drift on the YRI lineage since the separation from the unknown archaic lineage of $\tau_1 + \tau_2$ = 0.90-1.63 (95% confidence interval). The figure provides a calibration curve that allows us to translate genetic drifts to absolute split times (Table S16a.4).

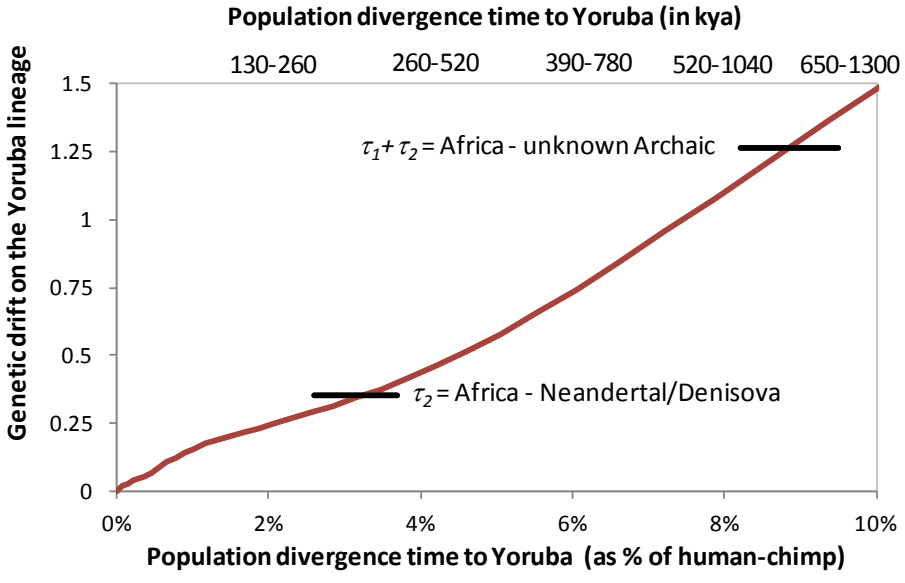


Figure S16a.9: Calibration curve from PSMC that we use to translate genetic drift on the YRI lineage to absolute divergence time from YRI. *The estimates of genetic drift τ_2 between the archaic populations (Neandertal and Denisova) and the unknown archaic populations ($\tau_1 + \tau_2$) are shown.*

We highlight two observations that emerge from this analysis.

The estimated genetic drift on the YRI lineage since the separation from the common ancestral population of Altai and Denisova has a 95% confidence interval corresponding to 3.00-3.50% of human-chimp divergence, or 391-461 kya assuming 13 Mya for human-chimp divergence. This is below the estimates of 4.26-4.53% in Table S12.2 of SI 12, as well as the 4.23-5.89% in Table S12.3. However, the high end of the range is only modestly below the low end of the range reported in SI 12, and so we think these numbers are not implausible.

The estimated genetic drift on the YRI lineage since population divergence from the unknown archaic population has a 95% confidence interval of 6.9-10.8% of human-chimpanzee divergence, or 900-1,401 kya assuming a full range of uncertainty for human-chimpanzee divergence.

Table S16a.4: Inferences of parameters of history under Model 3 based on this section's analysis

Para-meter	***Description***	***Estimate ± std. err.***	***95% conf. interval***	***Div. as % of human-chimp (95% conf. int.)***	**Date if Div_{HC} = 6,500 kya *(95% conf. int.)***	**Date if Div_{HC} = 13,000 kya *(95% conf. int.)***
τ_2	Drift in YRI since separation from Neand. & Denisova	0.356 ± 0.014	0.329 - 0.384	3.00 - 3.50%	195 - 231 kya	391 - 461 kya
$\tau_1 + \tau_2$	Drift in YRI since separation from unknown archaic	1.262 ± 0.185	0.900 - 1.625	6.90 - 10.8%	450 - 700 kya	900 - 1401 kya
α	Unknown archaic mixture proportion	4.24 ± 0.77%	2.73 - 5.75%	n/a	n/a	n/a

We conclude by noting that while the confidence intervals that we report in Table S16a.4 are substantial, there are additional sources of uncertainty in the parameter estimates that are not taken into account by the confidence intervals. In particular, while we have attempted to make relatively few assumptions about demographic history, all models make simplifications, and any errors in our modeling here could in principle result in biased estimates of the true values of historical parameters.

(vi) Summary

This note documents evidence of unknown archaic gene flow into the ancestors of Denisova, which must have occurred to a greater extent than unknown archaic gene flow into the ancestors of Altai.

The key evidence for an asymmetry in the relationship of Africans to Altai and Denisova is that Africans share more derived alleles with Altai than with Denisova, a signal that becomes stronger with increasing African derived allele frequency and is especially strong for alleles that are fixed derived in Africans. We carried out several analyses reported in section (iii) of this note that suggest that this is reflecting real history, and is not an artifact of sequencing or mapping error.

We also report a series of analyses that show that the signals we observe cannot be entirely explained by gene flow between modern human and Neandertal ancestors after the split from Denisova. Intuitively, the reason why we can reject this signal is that the amount of genetic drift that has occurred on the African (Yoruba) lineage since the split from Neandertals—which we can measure well from other aspects of the data—is far too little to explain the degree of fixation in Africans of sites that are driving the signal. This allows us to conclude the gene flow is likely to be coming from an unknown population that diverged prior to the separation of Neandertals and Africans.

References

[1] Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PL, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S (2010) A draft sequence of the Neandertal genome. Science 328, 710-22

[2] Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222-6.

[3] Waddell PK, Ramos J, TanX (2011) Homo denisova, correspondence spectral analysis, finite sites reticulate hierarchical coalescent models and the Ron Jeremy hypothesis. arxiv:1112.6424.

[4] Waddell PJ, Tan X (2012) New g%AIC, g%AICc, g%BIC, and power divergence Fit Statistics Expose Mating between Modern Humans, Neanderthals and other Archaics. arXiv:1212.6820.

[5] Hudson RR (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics 18, 337-8.

[6] Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 13, 745-53.

Appendix S16a.1 The expectation of the ratio of the S-statistic at fixed vs polymorphic sites under Model 2 does not depend on the gene flow proportion or the demographic history in the archaic populations.

Supplementary Note SI16a shows that all African populations share more derived alleles with Neandertals than Denisovans and considered three demographic models that are consistent with this observation (Figure S16a.3). We reject model 1 (gene flow from Neandertals to Early Modern Humans (EMH)) on the grounds that the slope of the D-statistic as a function of derived allele frequency under this model is opposite to the observation (see SI16b for an ABC analysis that explores a larger range of parameters). To distinguish between models 2 and 3, we need to explore a large number of parameters in each model. To deal with this challenge, we identify a statistic, the ratio of the S-statistic at fixed vs polymorphic sites, that is robust to several of the unknown parameters. In this note, we derive the expected values of the S-statistic and its ratio at fixed and polymorphic sites under demographic model 2 *i.e.*, gene flow from EMH into Neandertal and show that the expected value of this statistic does not depend on the gene flow proportion or on the details of the demographic history of Neandertals and Denisovans. This result motivates us to set up a simple simulation scheme in which we vary the split time of modern humans and the ancestors of Neandertal and Denisovans to test if Model 2 can fit the data.

Details

Consider the demography shown in Figure S16a.14 that relates a west African population (y), Altai Neandertal (n) and Denisova (d). The west African population that we analyze is the Yoruba from Nigeria, denoted YRI. Each node is labeled with the derived allele frequency at the node *e.g.*, y denotes the derived allele frequency in west Africans. All edges are labeled with the drifts except the edge representing the ancestral population which is labeled with the effective population size. We would like to compute the expectation of the S-statistic S(Altai, Denisova; YRI, Chimp) at fixed and polymorphic sites in YRI under this model. The S-statistic is $S = nd10 - nd01$ where nd10 is the count of sites at which Altai carries the derived while Denisova carries the ancestral allele and nd01 is the opposite. We derive the expected value of the S-statistic at polymorphic sites and at fixed sites under model 2.

We consider two ascertainments

1. $\{\mathcal{P}\}$: sites that are polymorphic in YRI *i.e.*, $0 < y < 1$
2. $\{\mathcal{F}\}$: sites that are fixed for the derived allele in YRI *i.e.*, $y = 1$.

In these discussions, we only consider ascertainments on the population allele frequency but the extension to sample frequencies is straightforward.

We are interested in computing $\mathbb{E}\left[S\{\mathcal{F}\}\right], \mathbb{E}\left[S\{\mathcal{P}\}\right]$ as well as in the ratio $\mathbb{E}\left[\frac{S\{\mathcal{F}\}}{S\{\mathcal{P}\}}\right] \approx \frac{\mathbb{E}[S\{\mathcal{F}\}]}{\mathbb{E}[S\{\mathcal{P}\}]}$.

To compute $\mathbb{E}\left[S\{\mathcal{P}\}\right]$, we write

$$\{\mathcal{P}\} \quad = \quad \cup_{i=1}^{3}\{\mathcal{P}_i\}$$

where $\{\mathcal{P}_1\} = \{0 < x_1, x_2, y < 1\}$ are sites that are polymorphic in the ancestral population of modern humans and Neandertals and remain polymorphic all the way down to y, $\{\mathcal{P}_2\} = \{x_1 = 0, 0 < x_2, y < 1\}$ are sites where a new mutation arose on the edge connecting x_1 and x_2 and that is polymorphic at x_2 and y, and $\{\mathcal{P}_3\} = \{x_1 = 0, x_2 = 0, 0 < y < 1\}$ are sites where a mutation arose after x_2 and that are segregating in y.

We can then write

$$\begin{aligned}\mathbb{E}\left[S\{\mathcal{P}\}\right] &= \sum_{i=1}^{3}\Pr(\{\mathcal{P}_i\})\mathbb{E}\left[S|\{\mathcal{P}_i\}\right] \\ &= \sum_{i=1}^{3}\Pr(\{\mathcal{P}_i\})\left[\mathbb{E}\left[nd10|\{\mathcal{P}_i\}\right]-\mathbb{E}\left[nd01|\{\mathcal{P}_i\}\right]\right] \qquad (1)\end{aligned}$$

The following lemma will be useful:

Lemma 1.

$$\mathbb{E}\left[nd10|\{\mathcal{C}\}\right]=(1-\alpha)\mathbb{E}\left[x_0(1-x_0)|\{\mathcal{C}\}\right]+\alpha\mathbb{E}\left[x_2(1-x_0)|\{\mathcal{C}\}\right]$$
$$\mathbb{E}\left[nd01|\{\mathcal{C}\}\right]=(1-\alpha)\mathbb{E}\left[x_0(1-x_0)|\{\mathcal{C}\}\right]+\alpha\mathbb{E}\left[(1-x_2)x_0|\{\mathcal{C}\}\right]$$

where $\{\mathcal{C}\}$ *is an ascertainment in* y.

Proof.

$$\begin{aligned}\mathbb{E}\left[nd10|\{\mathcal{C}\}\right] &= \mathbb{E}\left[n(1-d)|\{\mathcal{C}\}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[n(1-d)|z,x_0\right]|\{\mathcal{C}\}\right], \text{Tower property} \\ &= \mathbb{E}\left[\mathbb{E}\left[n|z\right]\mathbb{E}\left[(1-d)|x_0\right]|\{\mathcal{C}\}\right], \text{Conditional independence of } n,d,\mathcal{C} \text{ given } x_0 \\ &= \mathbb{E}\left[z(1-x_0)|\{\mathcal{C}\}\right], \text{Martingale property} \\ &= \mathbb{E}\left[\left((1-\alpha)z_0+\alpha z_2\right)(1-x_0)|\{\mathcal{C}\}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left((1-\alpha)z_0+\alpha z_2\right)(1-x_0)|x_0,x_2\right]|\{\mathcal{C}\}\right], \text{Tower property} \\ &= \mathbb{E}\left[(1-x_0)\mathbb{E}\left[\left((1-\alpha)z_0+\alpha z_2\right)|x_0,x_2\right]|\{\mathcal{C}\}\right] \\ &= \mathbb{E}\left[(1-x_0)\left(\mathbb{E}\left[((1-\alpha)z_0|x_0,x_2\right]+\mathbb{E}\left[\alpha z_2|x_0,x_2\right]\right)|\{\mathcal{C}\}\right] \\ &= \mathbb{E}\left[(1-x_0)\left((1-\alpha)\mathbb{E}\left[z_0|x_0\right]+\alpha\mathbb{E}\left[z_2|x_2\right]\right)|\{\mathcal{C}\}\right] \\ &= \mathbb{E}\left[(1-x_0)\left((1-\alpha)x_0+\alpha x_2\right)|\{\mathcal{C}\}\right], \text{Martingale property}\end{aligned}$$

The result for $\mathbb{E}\left[nd01|\{\mathcal{C}\}\right]$ follows by symmetry. □

Each term in the sum in Equation 1 evaluates to:

1. $\mathbb{E}\left[nd10|\{\mathcal{P}_3\}\right]=\mathbb{E}\left[nd01|\{\mathcal{P}_3\}\right]=0$ since these sites will be fixed ancestral in n and d.
2. Application of Lemma 1 gives

$$\begin{aligned}\mathbb{E}\left[S|\{\mathcal{P}_1\}\right] &= \alpha\mathbb{E}\left[(x_2-x_0)|\{\mathcal{P}_1\}\right] \\ &= \alpha\mathbb{E}\left[(x_2-x_1)|\{\mathcal{P}_1\}\right], \text{ using the Martingale property } \mathbb{E}\left[x_0|x_1\right]=x_1 \\ &= \alpha\frac{1}{\Pr(\{\mathcal{P}_1\})}\int\int\int dx_1dx_2dyf_0(x_1)K(x_2;x_1,\tau_{12})K(y;x_2,\tau_{2y})\left[x_2-x_1\right]\end{aligned}$$

3. Similarly

$$\mathbb{E}\left[S|\{\mathcal{P}_2\}\right] = \alpha\mathbb{E}\left[x_2|\{\mathcal{P}_2\}\right]$$

We can then write

$$\mathbb{E}\left[S\{\mathcal{P}\}\right] = \alpha\mathbb{E}\left[(x_2-x_1)\{\mathcal{P}\}\right] \qquad (2)$$

where according to whether the allele frequencies (x_1,x_2) satisfy $\mathcal{P}_i$, one of the above derived relations holds.

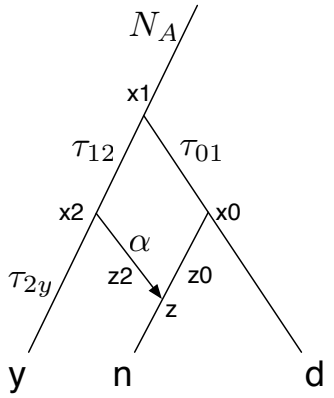


Figure S16a.14: Model of gene flow from Early Modern Humans (EMH) (denoted y) into Neandertals (denoted n) after the split of the ancestors of Neandertals and Denisovans. This is one of the models that is consistent with the greater sharing of derived alleles between Africans and Neandertals relative to Africans and Denisovans (See S16.3). Labels on the edges denote the drift on the lineage except the upper most edge where N_A denotes the ancestral population size. α denotes the admixture proportion.

In words, $\mathbb{E}[S\{\mathcal{P}\}]$ is a linear function of the admixture fraction α (increasing with α) and depends on the average divergence of the two populations, *i.e.*, x_1 and x_2 that contributed genes to n. This average divergence has contributions from polymorphic sites that segregated in the ancestral population ("old" polymorphisms) as well as mutations that arose between x_1 and x_2 ("new" polymorphisms). Thus, we expect that $\mathbb{E}[S\{\mathcal{P}\}]$ depends both on the frequency spectrum at x_1 and on the demographic history between x_1 and y (the contribution of old polymorphisms can be summarized by the drifts τ_{12} and τ_{2y} while the contribution of new polymorphisms will depend on the entire history of population sizes). Importantly, details of the demography relating n and d do not affect $\mathbb{E}[S\{\mathcal{P}\}]$ except through the split times when modern humans and the archaics diverged *i.e.*, the time of node x_1 and when n received gene flow from EMH *i.e.*, the time of node x_2.

To compute $\mathbb{E}[S\{\mathcal{F}\}]$, we write

$$\begin{aligned}
\mathbb{E}[S\{\mathcal{F}\}] &= \sum_{i=1}^{5} \Pr(\{\mathcal{F}_i\})\mathbb{E}[S|\{\mathcal{F}_i\}] \\
\{\mathcal{F}\} &= \cup_{i=1}^{5}\{\mathcal{F}_i\} \\
\{\mathcal{F}_1\} &= \{y=1, 0<x_1,x_2<1\} \\
\{\mathcal{F}_2\} &= \{y=1, x_2=1, 0<x_1<1\} \\
\{\mathcal{F}_3\} &= \{y=1, 0<x_2<1, x_1=0\} \\
\{\mathcal{F}_4\} &= \{y=x_2=1, x_1=0\} \\
\{\mathcal{F}_5\} &= \{y=1, x_1=x_2=0\}
\end{aligned}$$

Applying Lemma 1

$$\begin{aligned}
\mathbb{E}\left[S|\{\mathcal{F}_1\}\right] &= \alpha\mathbb{E}\left[(x_2 - x_1)\,|\{\mathcal{F}_1\}\right] \\
\mathbb{E}\left[S|\{\mathcal{F}_2\}\right] &= \alpha\mathbb{E}\left[(1 - x_1)\,|\{\mathcal{F}_2\}\right] \\
\mathbb{E}\left[S|\{\mathcal{F}_3\}\right] &= \alpha\mathbb{E}\left[x_2|\{\mathcal{F}_3\}\right] \\
\mathbb{E}\left[S|\{\mathcal{F}_4\}\right] &= \alpha \\
\mathbb{E}\left[S|\{\mathcal{F}_5\}\right] &= 0
\end{aligned}$$

As with the case of $\mathbb{E}\left[S\{\mathcal{P}\}\right]$, we see that

$$\mathbb{E}\left[S\{\mathcal{F}\}\right] = \alpha\mathbb{E}\left[(x_2 - x_1)\,\{\mathcal{F}\}\right] \tag{3}$$

This equation has a similar interpretation as the case of polymorphic sites. The details of the demography relating n and d do not affect $\mathbb{E}\left[S\{\mathcal{F}\}\right]$

Finally, from Equations 2 and 3, the ratio of S-statistics at fixed derived sites to polymorphic sites is approximately (approximating the expectation of ratios by the ratio of expectations)

$$\frac{\mathbb{E}\left[S\{\mathcal{F}\}\right]}{\mathbb{E}\left[S\{\mathcal{P}\}\right]} = \frac{\mathbb{E}\left[(x_2 - x_1)\,\{\mathcal{F}\}\right]}{\mathbb{E}\left[(x_2 - x_1)\,\{\mathcal{P}\}\right]} \tag{4}$$

To summarize, the ratio of S-statistics at fixed verus polymorphic sites does not depend on the admixture fraction α in expectation. Further, this ratio does not suffer from the conflation of admixture fraction and demographic history in the D-statistic or the S-statistic. Also, from the earlier remarks, the numerator and denominator do not depend on the history of population sizes ancestral to n and to d and hence, neither does the ratio.

This should make it easy to test if the ratio statistic is consistent with a demographic model. We only need to vary the ancestral population size N_A, the times of nodes x_1 and x_2 and the population size of y (before and after node x_2) and check if the model can match empirical data.

Appendix S16a.2 Estimate of the drift in Africans since the split from the ancestors of Neandertals and Denisovans under Model 2

We estimate the drift in Africans since the split from the ancestors of Neandertals and Denisovans using the frequency spectrum of polymorphic sites in west Africans conditioned on observing a derived allele in Denisovans. This corresponds to estimating $\tau_{1y} = \tau_{12} + \tau_{2y}$ in Figure S16a.14. We assume a simple split between the two populations (with no subsequent gene flow), random mating in the African population and that the African population is in mutation-drift equilibrium at the time of split and has constant population size since the split. Assume that we sample n chromosomes randomly from the African population.

For this model, the frequency spectrum when we ascertain a derived allele in the archaic is [1]

$$AFS_D(i) = \frac{\theta}{n+1}\exp\left(-\tau_{1y}\right), i \in \{1,\ldots,n-1\} \quad (5)$$

Model 2 predicts no gene flow between Denisova and ancestors of Africans. So we can use the frequency spectrum of Africans conditioned on observing a derived allele in Denisova, AFS_D, to estimate τ_{1y} (using Equation 5)

$$\hat{\tau}_{1y} = -\log\left(\frac{(n+1)\overline{afs}_d}{\hat{\theta}}\right) \quad (6)$$

Here we use $\hat{\theta} = \hat{\theta}_\pi$, the estimator of population-scaled mutation rate based on heterozygosity, and $\overline{afs}_d = \frac{\sum_{i=d_{min}}^{d_{max}} afs_d(i)}{d_{max}-d_{min}+1}$. A caveat is that the observed AFS is not flat and ,hence, violates the assumptions of the theory – so we exclude the counts in the first and last nine bins of the spectrum ($d_{min} = 10, d_{max} = 62$). Further work needs to be done to learn models that explain the entire spectrum in Africans.

References

[1] Hua Chen, Richard E. Green, Svante Pbo, and Montgomery Slatkin. The joint allele-frequency spectrum in closely related species. *Genetics*, 177(1):387–398, September 2007.

Appendix S16a.3 Mathematical details of the diffusion theory analysis we use to reject Model 2 and estimate parameters of history under Model 3

We give details of the derivation of our formulae under Model 3 for the expected values of the 4 statistics that we use. The derivation under model 2 is similar. We first give some easy general results on the Wright-Fisher diffusion. All follow from Kimura's theorem on the transition function of the diffusion [2], or work by Ewens [1]. We suppose:

$$U \xrightarrow{\tau} V$$

and u, v are frequencies of the derived allele in U, V, respectively. We suppose a constant population size X, on the whole lineage of V, both ancestral to U and on the path from U to V. Thus mutations arise at a constant rate, and have an initial frequency

$$\epsilon = \frac{1}{2X}$$

We will use the diffusion approximation, ignoring terms that are $O(\epsilon^2)$. We write the *Kimura function* $K(u, v; \tau)$ to be the transition probability $P(v|u, \tau)$ for $0 < u, v < 1$.

We have

1. A result of Ewens. For $\epsilon < u < 1$, the allelic spectral density of u is

$$P(u) = \int_0^\infty K(\epsilon, u; \psi)\ d\psi = \frac{2\epsilon}{u}$$

2. Let

$$I = \int_0^1 f(u) K(u, v; \tau)\ du$$

Suppose $\int_0^1 f^2(u) u(1-u)\ du$ is finite. We can write $f(u) = \sum_{i=0}^\infty c_i J_i(u)$ where J_i are Jacobi polynomials (here in fact Gegenbauer polynomials). Full details can be found in [3], which uses the same notation as this note. Then

$$I = \sum_{i=0}^\infty c_i J_i(v) e^{-\lambda(i)\tau}$$

where

$$\lambda(i) = \frac{(i+1)(i+2)}{2}$$

We give 2 special cases

(a)

$$\int_0^1 K(u, v; \tau)\ du = e^{-\tau}$$

(b)

$$\int_0^1 uK(u, v; \tau)\ du = (v - \frac{1}{2})e^{-3\tau} + \frac{1}{2}e^{-\tau}$$

3. We next give some expectations of functions of v, for the special cases we use.

$$\begin{aligned} E(v|u,\tau) &= u \\ E(v(1-v)|u,\tau) &= u(1-u)e^{-\tau} \\ E(v^2|u,\tau) &= u - u(1-u)e^{-\tau} \end{aligned}$$

Note that in the above equations we do *not* require that V is polymorphic so we include the case that v has fixed,

In what follows, H is the event that we ascertain a het in Yoruba, Y is the event that a random Yoruba allele is derived, and y is the derived allele frequency in Yoruba, for $0 < y < 1$.

For X_1 we have

$$\begin{aligned} P(D=1,y) &= 2\epsilon(\alpha e^{-(\tau_1+\tau_2)} + (1-\alpha)e^{-\tau_2}) \\ P(N=1,y) &= 2\epsilon e^{-\tau_2} \end{aligned}$$

from which our formulae for the expected value follows.
For X_2, We must evaluate

$$\frac{E(ND10,H) - E(ND01,H)}{2\left(E(ND10,Y) - E(ND01,Y)\right)}$$

write

$$Q = \frac{E(ND10,H) - E(ND01,H)}{2}$$

and

$$R = E(ND10,Y) - E(ND01,Y)$$

so that $X_2 = Q/R$. We note that both Q, R are linear in α, so we can assume that $\alpha = 1$. Next note that

$$E(ND10,H) - E(ND01,H) = E((N=1,H) - E(D=1,H))$$

$$\begin{aligned} E(N=1,H) &= 2\epsilon e^{-\tau_2}\int_0^1 2y(1-y)\,dy = e^{-\tau_2}\frac{2\epsilon}{3} \\ E(D=1,H) &= 2\epsilon e^{-(\tau_1+\tau_2)}\int_0^1 2y(1-y)\,dy = e^{-(\tau_1+\tau_2)}\frac{2\epsilon}{3} \end{aligned}$$

It follows that

$$Q = \frac{\epsilon e^{-\tau_2}(1-e^{-\tau_1})}{3}$$

For R, there are 3 terms to evaluate

1. A_1: A is polymorphic, $D=0, N=1, Y=1$.
2. A_2 : A mutation arises on the lineage $A \to B$. Furthermore $N=1, Y=1$.
3. A_3 : A is polymorphic, $D=1, N=0, Y=1$.

We find that:

$$\begin{aligned} A_1 &= 2\epsilon\int_0^1 \frac{(1-a)E(b^2|a,\tau_1)}{a}\,da \\ A_2 &= \int_0^{\tau_1} E(b^2|\epsilon,\psi)\,d\psi \\ A_3 &= 2\epsilon\int_0^1 \frac{aE(b(1-b)|a,\tau_1)}{a}\,da \end{aligned}$$

We obtain

$$A_1 - A_3 = 2\epsilon \int_0^1 (1-a)(1-e^{-\tau_1})\ da = \epsilon(1-e^{-\tau_1})$$

Next

$$A_2 = \int_0^{\tau_1} \epsilon - \epsilon(1-\epsilon)e^{-\psi}\ d\psi = \epsilon\tau_1 - \epsilon(1-e^{-\tau_1}) + O(\epsilon^2)$$

Therefore we get the simple expression

$$R = \epsilon\tau_1$$

Now we get

$$X_2 = Q/R = \frac{e^{-\tau_2}(1-e^{-\tau_1})}{3\tau_1}$$

For X_3, X_4 we must consider

$$S(y) = P(D = 1|N = 1, y)$$

We find

$$S(y) = (1-\alpha)(1-\beta) + \frac{(1-\alpha)\beta B_1 + \alpha B_2}{e^{-\tau_2}}$$

where

$$\begin{aligned} B_1 &= \int_0^1 bK(b, y; \tau_2)\ db \\ B_2 &= \int_0^1 \int_0^1 K(a, b, \tau_1) bK(b, y; \tau_2)\ da\ db \end{aligned}$$

Now

$$B_1 = (y - \frac{1}{2})e^{-3\tau_2} + \frac{1}{2}e^{-\tau_2}$$

For B_2,

$$\int_0^1 K(a, b, \tau_1)\ da = e^{-\tau_1}$$

Thus

$$\begin{aligned} B_2 &= e^{-\tau_1} \int_0^1 bK(b, y; \tau_2)\ db \\ &= e^{-\tau_1} B_1 \end{aligned}$$

This shows that

$$S(y) = (1-\alpha)(1-\beta) + ((1-\alpha)\beta + \alpha e^{-\tau_1}) \left((y - \frac{1}{2})e^{-2\tau_2} + \frac{1}{2} \right) \quad (1)$$

Our equation for slope and intercept now follow.

Finally we derive the relationship of the regression on k/n (k derived alleles from n) to a linear model on the underlying allele probability y. Let $D[k]$ be the empirical probability $P(D = 1|N = 1, k)$ where k is is the number of derived Yoruba alleles (from $n = 80$). We fit

$$D[k] = u(k/n) + v$$

We use the formula (easily derived from the beta integral)

$$\frac{\int_0^1 y^{k+1}(1-y)^{n-k}\ dy}{\int_0^1 y^k(1-y)^{n-k}\ dy} = \frac{k+1}{n+2}$$

We suppose that $E(D[k]|y) = \alpha y + \beta$. Then it follows that

$$\begin{aligned} E(D[k]|k) &= \frac{\int_0^1 (\alpha y + \beta) y^k (1-y)^{n-k}\, dy}{\int_0^1 y^k (1-y)^{n-k}\, dy} \\ &= \alpha \frac{k+1}{n+2} + \beta \end{aligned}$$

Hence

$$\alpha + \beta \frac{k+1}{n+2} = u \frac{k}{n} + v$$

and we get:

$$\alpha = u \frac{n+2}{n} \tag{2}$$

$$\beta = v \frac{n+2}{n} - \alpha \tag{3}$$

This yields the required relationship.

There is a Bayesian interpretation of this result. Note that the polymorphic spectrum of Yoruba, conditional on $N = 1$ is uniform. Thus on observing $N = 1$ and k derived alleles from n ($1 \le k \le n-1$), the posterior mean of y is $\frac{k+1}{n+2}$. If the regression had been carried out with $\frac{k+1}{n+2}$ as the independent variable, instead of the more obvious $\frac{k}{n}$ no adjustment to the estimates of slope and intercept would have been needed.

References

[1] W. J. Ewens. The pseudo-transient distribution and its uses in genetics. *Journal of Applied Probability*, 1(1):pp. 141–156, 1964.

[2] M. Kimura. Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. U.S.A.*, 41(3):144–150, Mar 1955.

[3] N. J. Patterson. How old is the most recent ancestor of two copies of an allele? *Genetics*, 169(2):1093–1104, Feb 2005.

Supplementary Information 16b

Archaic ancestry in Denisova

Fernando Racimo* and Montgomery Slatkin

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

Table of contents

1 – Introduction

We observe an excess of Denisova-ancestral over Neandertal-ancestral alleles at sites where modern humans are derived. This allelic imbalance is especially pronounced at sites that are completely fixed for the derived allele in modern humans. The pattern persists regardless of whether we use only the Europeans, the Yoruba or all 1000 Genomes individuals to determine the derived frequency of the allele (Figure S16b.1).

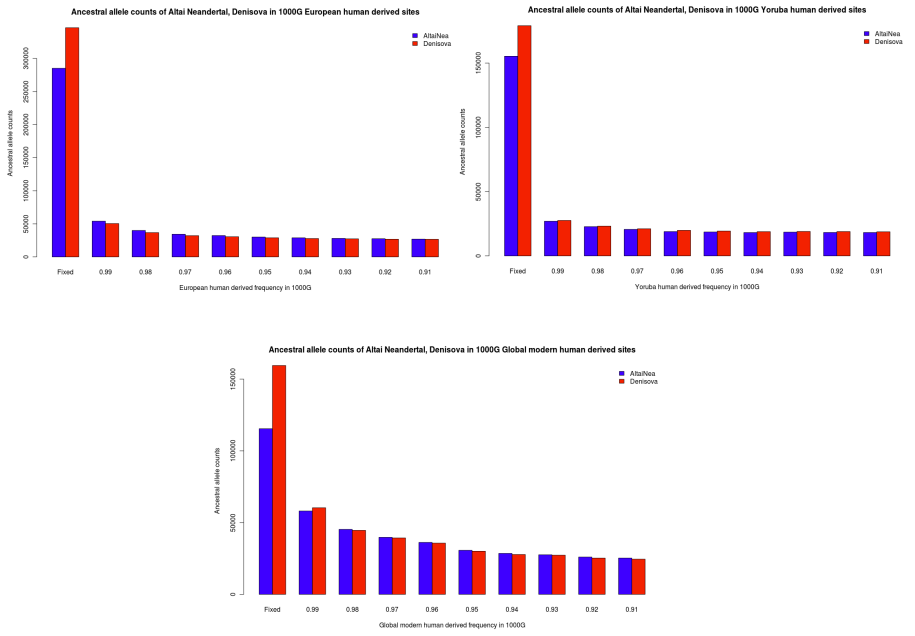


Figure S16b.1. An excess of ancestral alleles in Denisova relative to Altai Nandertal is observed at fixed derived sites in modern humans, regardless of which modern human panel is used: 1000G Europeans, 1000G Yoruba or all 1000G individuals.

We distinguish three models of population history that could explain this pattern (Figure S16b.2):

Model 1: Neandertal admixture into Early Modern Humans
Model 2: Early Modern Human admixture into Neandertals
Model 3: Admixture from a super-archaic population into Denisovans

To distinguish between these models, and to estimate parameters of the best-supported model, we implemented an Approximate Bayesian Computation (ABC) method described below. The observed data were obtained from the Yoruba panel in SI 16a, down-sampled to 72 chromosomes and restricting to transversions only.

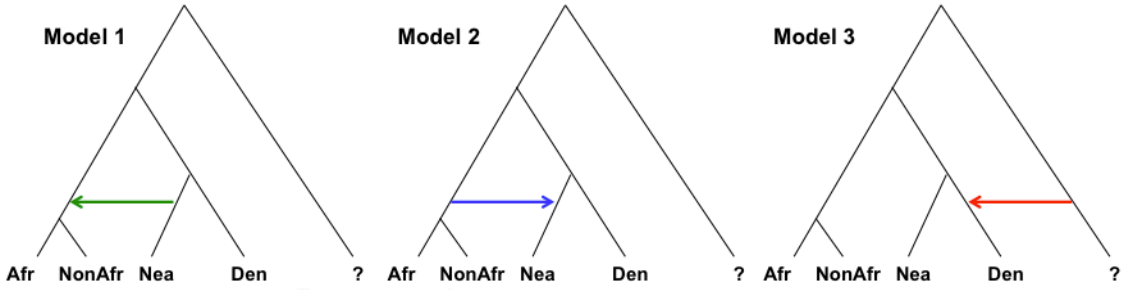


Figure S16b.2. Three different models of archaic admixture that could explain the ancestral excess observed in Denisova.

2 – Methods

We define a "simulated sample" as a set of 5,000 replicate coalescent simulations obtained under the same parameters using ms[1] with theta=10, on which we calculated a set of statistics. For each model, we compared a large number of simulated samples (see below) against the observed data using the following statistics:

- Entire D_j statistic spectrum, where $D_j = \frac{nd10_j - nd01_j}{nd10_j + nd01_j}$ corresponds to a D statistic stratified by the derived frequency *j* of present-day humans in a panel of 72 Yoruba chromosomes, such that:
  - nd01j is the number of sites where j out of the 72 Yoruba chromosomes have the derived allele, Denisovan has the ancestral allele and Altai Neandertal has the derived allele
  - nd10j is the number of sites where j out of the 72 Yoruba chromosomes have the derived allele, Denisovan has the derived allele and Altai Neandertal has the ancestral allele
- Ratio of S statistic computed at fixed derived sites over S statistic computed at all polymorphic sites, weighted by the derived allele frequency in Yoruba: $S_{72} / \sum_{j=1}^{n-1} S_j$ where $S_j = (j/n) * (nd10_j - nd01_j)$.

Note that these statistics are also used in a complementary analysis of archaic admixture into the Denisovan individual in SI16a.

Before comparing models, statistics were linearized via Box-Cox transformation and orthogonal components were extracted from them using partial least squares discriminant analysis (PLSDA) trained on 1,000 simulated samples for each of the pairwise model comparisons[2-4]. The PLSDA is a multivariate discrimination method used to obtain linear components of the data that serve best to classify samples. We used the first 5 components of the PLSDA to compare the simulated samples to the observed data.

For model choice, we obtained 10,000 simulated samples from each model. The 1% of samples closest to the observed data from Table S16.1 were retained and all others were rejected. Model choice was performed via Bayes factors, which is the ratio of the marginal probability of one demographic model over the marginal probability of another demographic model. Marginal probabilities were obtained using a general linear model based on post-sampling regression adjustment[5] via ABCtoolbox[6].

Parameter estimation was then performed from the approximate posterior density of the parameter of interest under the model that was best supported (1% samples retained; Dirac peak width = 0.01). In this case, instead of transforming the summary statistics via PLSDA, we transformed them via partial least squares (PLS) regression (again, calibrated using 1,000 simulated samples) and used the first 5 components to fit the general linear model to the observed data (using 10,000 simulated samples with 1% rejection) via ABCtoolbox.

We fixed some parameters at the values stated below:

- Denisova-Neandertal population split t_{DN} = 8,000 generations ago
- Neandertal-Modern human population split t_{NM} = 9,000 generations OR 12,000 generations OR 16,000 generations ago
- Human-Chimpanzee population split t_{HC} = 200,000 generations ago

For each model, we evaluated both a version with constant population size at 2N = 20,000 and a version with population sizes roughly following the PSMC model from SI12.

We assumed a uniform prior distribution of the following parameters and used ABCtoolbox (Wegmann et al. 2010) to obtain their posterior distribution:

- Archaic admixture time $t_f \sim$ Unif[$t_{AfrNonAfr}$ to t_{DN}]
- Ghost population divergence time $t_S \sim$ Unif[t_{NM} to t_{HC}] (only relevant to model 3)

We also sampled values of the admixture proportion f under two admixture regimes: one in which we limited the exploration of parameter space to admixture proportions that are limited (0 to 10%) under any of the three models, and one in which admixture in any of the three models can reach levels of up to 50%.

"Limited admixture" regime: archaic admixture proportion f ~ Unif [0 to 10%]

"Broad admixture" regime: archaic admixture proportion f ~ Unif [0 to 50%]

3 – Results

Table S16b.1 shows Bayes factors for each pairwise model comparison. Under the "limited admixture" regime (0-10% admixture allowed), model 3 is the one best supported across multiple demographic scenarios, with highly positive Bayes factor in all but one of the scenarios (D, where models 1 and 3 indistinguishable: Bayes factor = 0.98). In addition, the p-value of the fit of the general linear model (GLM) to the observed data is large and not significant for model 3 in all comparisons, while the p-values for the fits for both model 1 and model 2 are in most cases significant and in all cases smaller than the p-value for model 3.

Under the "broad admixture" regime (0-50% admixture allowed), we find slight support for model 1 over model 3 (Bayes factors < 7). However, our power to discriminate between models is much more reduced. In all cases, the two models under comparison are either both good ($p >> 0.1$) or both bad ($p < 0.1$) fits to the GLM, so we are unable to confidently determine whether model 1 or 3 (or both) best explains the data under this admixture regime.

Scenario	Admixture regime	Population sizes	t_{NM} (generations)	Bayes factors			P-values of GLM fitting for each model under each comparison		
				Model 3 / Model 1	Model 3 / Model 2	Model 2 / Model 1	Model 3, Model 1	Model 3, Model 2	Model 2, Model 1
A	**Limited**	Constant	9000	2,694	107	1	0.98, 0	0.98, 0.01	0.04, 0.1
B	**Limited**	Constant	12,000	78,944	21,602	2,906	1, 0	1, 0	0.99, 0.02
C	**Limited**	Constant	16,000	20,320,495	2,893	8,449	1, 0	0.95, 0	1, 0.03
D	**Limited**	PSMC	9,000	0.96	46	0.0067	1, 0.83	1, 0.02	0.03, 1
E	**Limited**	PSMC	12,000	4	2	10	1, 0.78	0.96, 0.21	1, 0.98
F	**Limited**	PSMC	16,000	10	74	4	1, 0.03	1, 0	1, 0.93
G	**Broad**	Constant	9,000	0.15	0.05	2	0.99, 1	1, 1	0.99,0.99
H	**Broad**	Constant	12,000	0.34	0.2	0.15	1, 0.95	1,0.89	0.02, 0.11
I	**Broad**	Constant	16,000	0.38	200	0.00024	0.99, 0.98	0.68, 0	0, 0.03
J	**Broad**	PSMC	9,000	0.49	0.02	4	0.93, 0.38	1, 1	1, 1
K	**Broad**	PSMC	12,000	0.2	0.23	2	0.89, 0.9	0.98,0.76	1,0.99
L	**Broad**	PSMC	16,000	0.16	0.94	0.073	0.98, 1	0, 0.01	0.98, 0.97

Table S16b.1. Bayes factors (ratios of marginal probabilities) for each model comparison and p-values of general linear model (GLM) fitting for each model, under each comparison, for different demographic scenarios. Large Bayes ratios represent strong posterior support for the model in the numerator relative to the model in the denominator.

We also compared the first two Partial Least Squares (PLS) components of the best 1% simulated statistics under each model with the first two PLS components of the observed statistics. For each model, the observed and simulated statistics were transformed using PLS regression trained on simulations generated under that particular model. Figure S16b.3 shows that the observed statistics always fall within the distribution of model 3 simulations, but this does not occur for models 2 and 1.

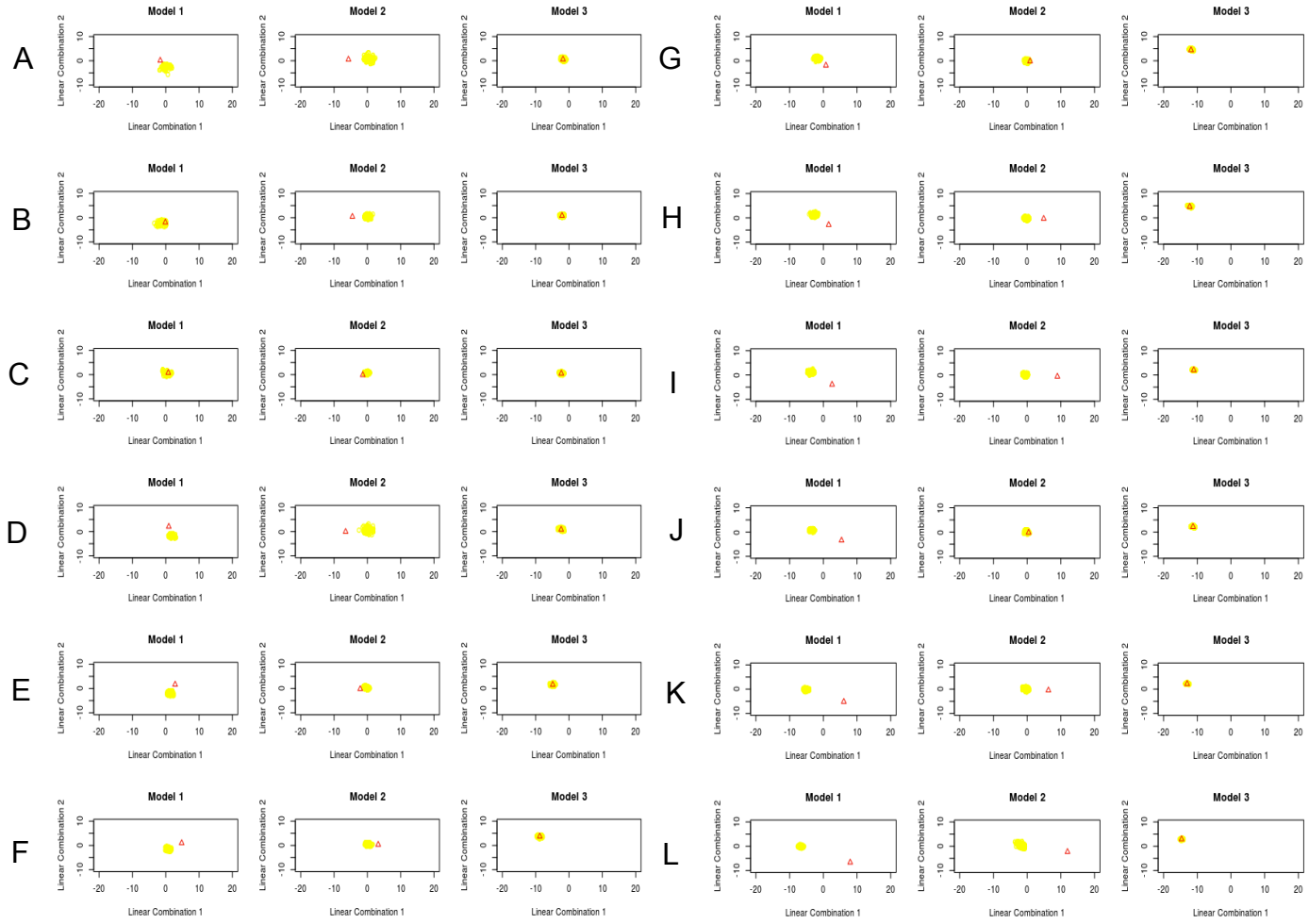


Figure S16b.3. First two PLS components of the best 1% simulated statistics (yellow circles) under each component with the first two PLS components of the observed statistics (red triangle). For each model, the observed and simulated statistics were transformed using PLS regression trained on simulations generated under the same model. Letter labels correspond to scenarios in Table S16b.1.

Assuming model 3 is correct, we attempted to estimate 3 parameters: the ghost admixture proportion f, the ghost admixture time t_f and the ghost population split time t_D (Figure S16b.4).

The posterior distribution for the admixture time t_f is very similar to the prior regardless of the value of t_{NM}, so we conclude that we are unable to estimate the admixture time under this model. However, we also can conclude that the estimates for the two other parameters are fairly insensitive to the time of admixture. We show point estimates and 95% highest posterior density (HPD) intervals under model 3 in Table S16b.2.

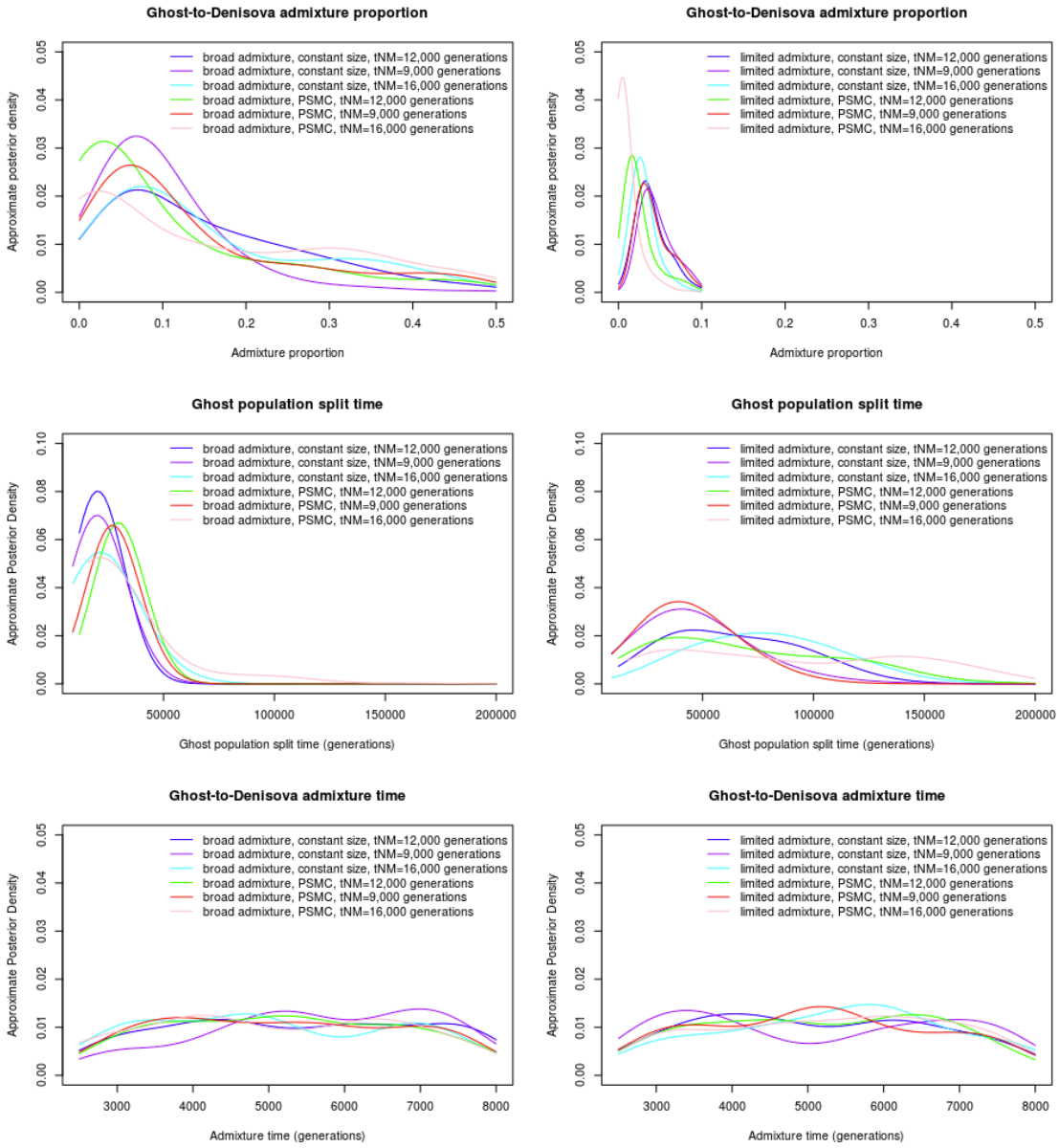


Figure S16b.4. Approximate posterior estimation of 3 parameters for the super-archaic admixture event, assuming model 3 is correct. Left panels correspond to the "broad admixture" regime, while the panels on the right correspond to the "limited admixture" regime. The x-axis boundaries of the curves in each graph correspond to the boundaries of the prior uniform distributions from which the parameter values are sampled for the ABC simulations. t_{NM} = Neandertal-Modern human population split time. PSMC = model following an approximation to the PSMC estimates for population sizes from SI12.

Admixture regime	Population sizes	t_{NM} (generations)	Parameter modes [95% HPD]		
			f	t_D (generations)	t_f (generations)
Limited	Constant	9000	4.04% [1.73% - 8.57%]	37,956 [9,027 – 86,173]	6,055 [2,671 – 7,717]
Limited	Constant	12,000	3.43% [0.72% - 7.76%]	65,167 [15,802 – 120,228]	4,722 [2,726 – 7,608]
Limited	Constant	16,000	3.03% [0.31% - 7.77%]	69,894 [17,874 – 129,346]	5,389 [2,673 – 7,495]
Limited	PSMC	9,000	4.04% [1.62% - 8.98%]	34,077 [9,005 – 63,008]	5,728 [2,727 – 7,717]
Limited	PSMC	12,000	1.72% [0.01% - 6.35%]	40,483 [12,005 – 139,219]	6,500 [2,950 – 7,829]
Limited	PSMC	16,000	0.61% [0.01% - 4.23%]	38,301 [16,005 – 173,951]	6,055 [2,615 – 7,495]
Broad	Constant	9,000	8.08 % [0.01% - 23.22%]	22,523 [9,024 – 37,951]	7,166 [3,115 – 7,995]
Broad	Constant	12,000	8.08% [0.01% - 39.89%]	17,706 [12,015 – 32,891]	4,389 [2,558– 7,609]
Broad	Constant	16,000	7.07% [0% - 36.35%]	23,437 [16,009 – 51,308]	4,556[2,892 – 7,885]
Broad	PSMC	9,000	7.08% [0.02% - 40.9%]	28,301 [12,873 – 43,728]	5,055 [2,726 – 7,718]
Broad	PSMC	12,000	4.04% [0.02% - 35.35%]	27,204 [12,020 – 44,286]	5,222 [2,895 – 7,772]
Broad	PSMC	16,000	2.53% [0.01% - 41.91%]	21,576 [16,004 – 94,050]	6,611 [2,672 – 7,661]

Table S16b.2. Posterior modes and highest posterior density (HPD, in brackets) of parameters for model 3 under different demographic scenarios. f = ghost-to-Denisova admixture proportion. t_D = ghost population split time. t_f = ghost-to-Denisova admixture time.

In Figure S16b.5, we jointly plot the three parameter values for the 1% of the simulated samples that are closest to the observed data under a demographic scenario of PSMC-based population size changes. This graph shows that t_S is inversely proportional to f, while t_f is approximately uniformly distributed across the simulations. Pairwise joint posterior distributions of the three parameters under model 3 are shown in Figure S16b.6.

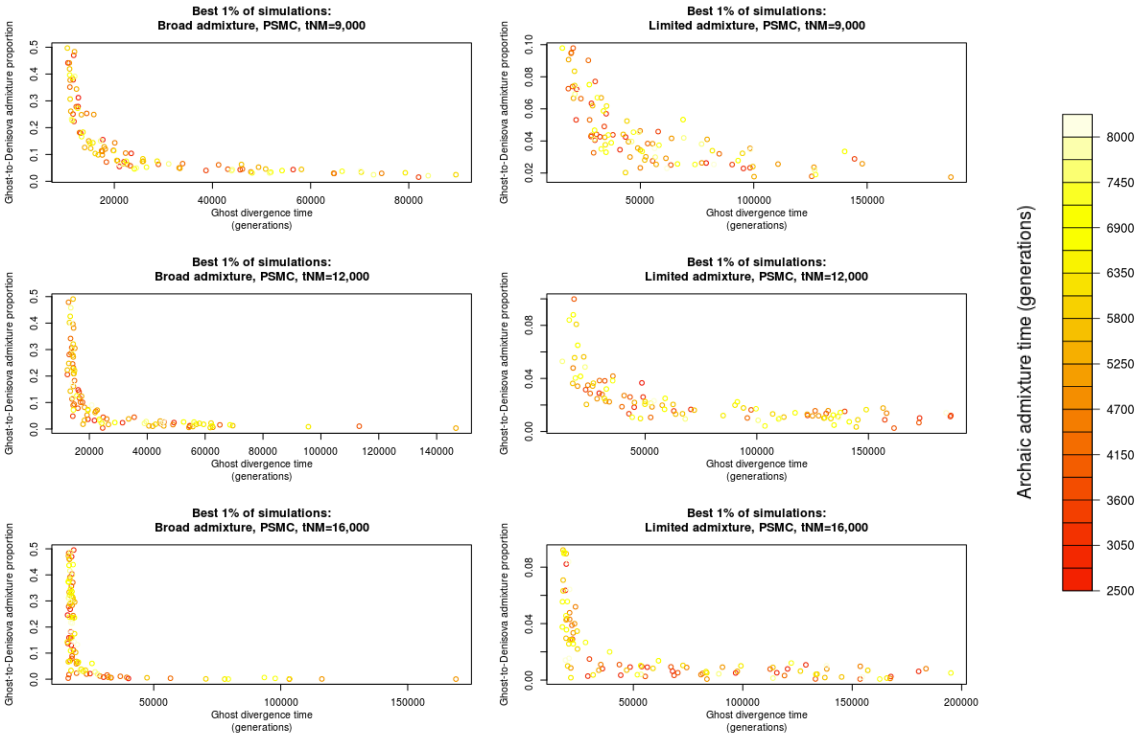


Figure S16b.5. Approximate posterior values of the 3 parameters estimated for the 1% of simulations that are closest to the observed data under model 3 and demographic scenarios with PSMC-based population size changes.

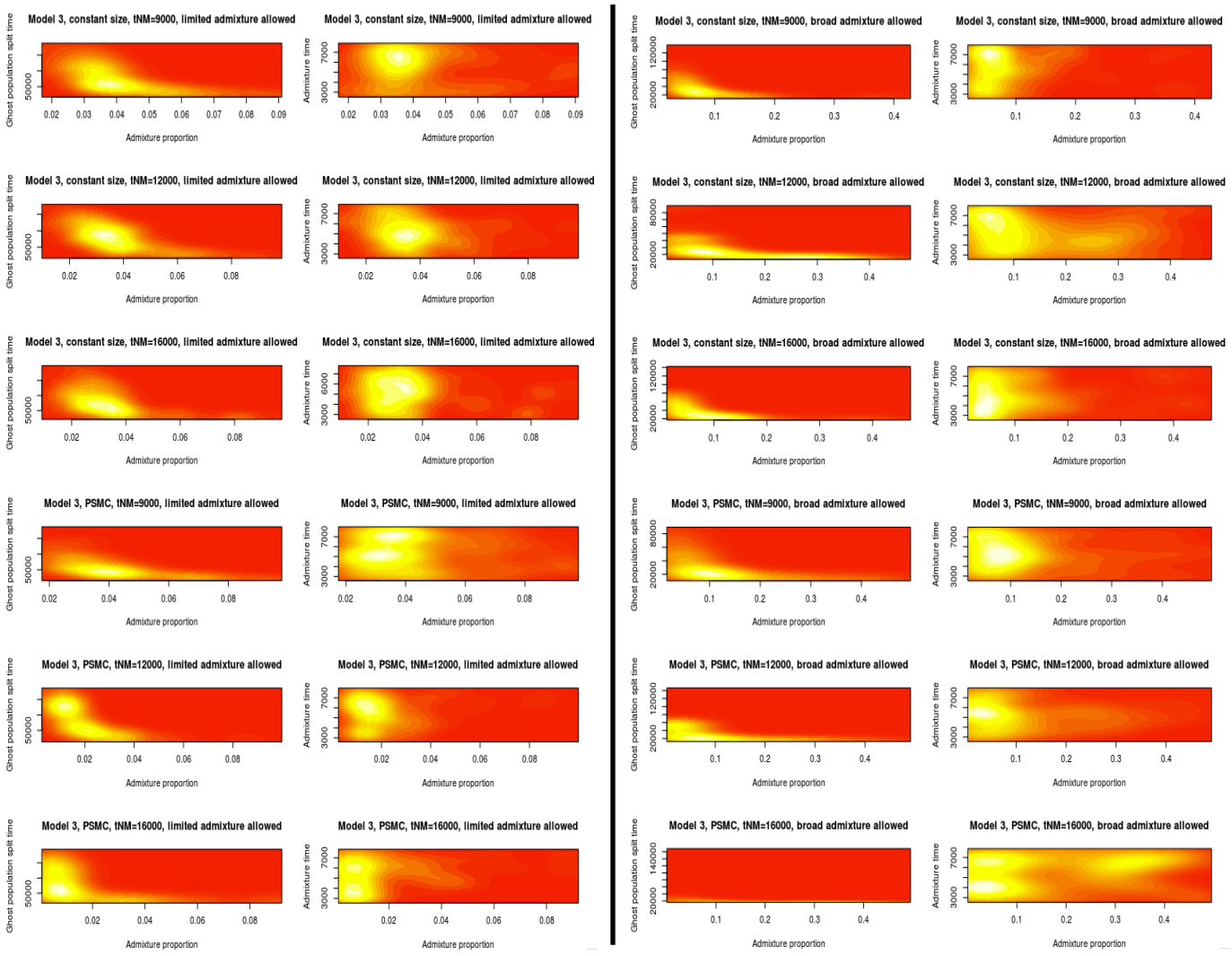


Figure S16b.6. Approximate joint posterior distributions of parameter values for model 3 under different demographic scenarios. Regions of low probability are in red while those of high probability are in yellow.

4 – Two models of ghost admixture

A possibly more parsimonious explanation for the ghost admixture signal is a model where the Denisova individual is the product of recent hybridization between Neandertals and a super-archaic population that is an outgroup to both modern humans and Neandertals. Under this model, the Denisova individual does not belong to a sister clade to Neandertals, but appears as such because of the large proportion of Neandertal ancestry it contains (SI 15). We aimed to distinguish between this model (Figure S16b.7, model 3A) and a model where Denisovans are a sister group to Neandertals but also contain both Neandertal and ghost ancestry due to 2 separate admixture events at different times in their history (Figure S16b.7, model 3B).

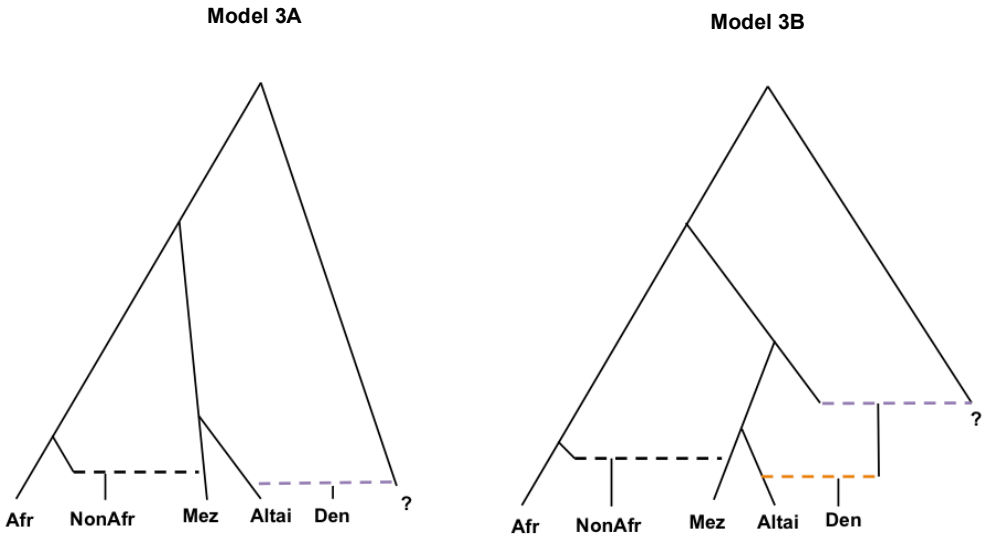


Figure S16b.7. Two models of ghost admixture in Denisova. Model 3A predicts that the Denisovan individual is the product of hybridization between a Neandertal closely related to the Altai Neandertal and a ghost population that is an outgroup to Neandertals and modern humans. Model 3B predicts that the Denisovan individual belongs to a sister clade to Neandertals that received Altai Neandertal and ghost population ancestry in two separate admixture events.

For each simulated sample, we computed the following set of statistics (obtained from SI16, SI9, SI14 and SI15), in addition to the Dj statistic spectrum and the ratio of the S statistic under fixed and polymorphic sites described above, as we thought they could contribute to better discriminate these two models:

- Ratio of Denisova heterozygosity to Altai Neandertal heterozygosity
- D(Altai Neandertal, Denisova, Yoruba, Chimp)
- D(Altai Neandertal, Denisova, French, Chimp)
- D(Altai Neandertal, Mezmaiskaya Neandertal, Denisova, Yoruba)
- Ratio of Denisova-Yoruba pairwise differences to Altai Neandertal-Yoruba pairwise differences

We used a rough approximation of the PSMC model from SI12 for population size changes and assumed that the Neandertal-Modern human population split time t_{NM} = 12,000 generations ago, and, for model 3B, that the Neandertal-Denisova population split time t_{DN} = 9,000 generations ago. We fixed the African-NonAfrican population split $t_{AfrNonAfr}$ = 2,500 generations ago. We also fixed the Neandertal to Non-African admixture proportion at 2% and the time for this admixture event t_X to 2,000 generations ago. The parameters of interest were sampled as follows:

Model 3A:
- Ghost population divergence time $t_S \sim$ Unif[t_{NM} to t_{HC}]
- Ghost-to-Denisova admixture proportion fA ~ Unif [0 to 100%]
- Ghost-to-Denisova admixture time $t_{fA} \sim$ Unif[1,000 generations ago to $t_{AltaiMez}$]
- Mezmaiskaya-Altai Neandertal population split time $t_{AltaiMez} \sim$ Unif[t_X to t_{DN}]

Model 3B:
- Ghost population divergence time $t_S \sim$ Unif[t_{NM} to t_{HC}]
- Ghost-to-Denisova admixture proportion fA ~ Unif [0 to 50%]
- Ghost-to-Denisova admixture time $t_{fA} \sim$ Unif[1,000 generations ago to t_{DN}]

- Neandertal-to-Denisova admixture proportion fB ~ Unif [0 to 50%]
- Neandertal-to-Denisova admixture time t_{fB} ~ Unif[1,000 generations ago to $t_{AltaiMez}$]
- Mezmaiskaya-Altai Neandertal population split time $t_{AltaiMez}$ ~ Unif[t_X to t_{DN}]

All other parameters were fixed as in the section above.

The Bayes factor for support of model 3B over model 3A is equal to 22, which is interpretable as "strong" in favor of 3B, though not high enough to be "decisive"[7]. Further, both models are good fits to the GLM (model 3A p-value: 0.93, model 3B p-value: 0.95). Parameters estimated under each of these two models are shown in Figure S16b.8. The modes and medians of the approximate posterior distributions under each model are shown in Table S16b.3. Because model 3A is presumably more plausible under a model where t_S and t_{NM} are extremely close in time (near-trifurcation), due to the nearly equal values of the Altai-African and Denisova-African divergence times, we also repeated this analysis but sampling from a more restrictive prior for the ghost population divergence time (t_S ~ Unif[t_{NM} to 40,000 generations ago]). Under these conditions, we still see support in favor of model 3B, though not as strong as in the case above (Bayes factor = 3).

Posterior parameter estimates	fA (%)		fB (%)		t_S (generations ago)		$t_{AltaiMez}$ (generations ago)		t_{fA} (generations ago)		t_{fB} (generations ago)	
	Mode	Median	Mode	Median	Mode	Median	Mode	Median	Mode	Median	Mode	Median
Model 3A	0.013	6.81	-	-	27,220	36,286	7,454	7,062	6,736	6,542	-	-
Model 3B	5.56	14.47	28.28	28.73	23,394	40,901	6,364	5,555	2,415	4,215	2,241	2,882

Table S16b.3. Posterior parameter modes and medians under each of the two models. fA = Ghost-to-Denisova admixture, fB = Altai Neandertal-to-Denisova admixture, t_s = Ghost-Human divergence time, t_{fA} = Ghost-to-Denisova admixture time, t_{fB} = Neandertal-to-Denisova admixture time, $t_{AltaiMez}$ = Altai-Mezmaiskaya Neandertal population split time.

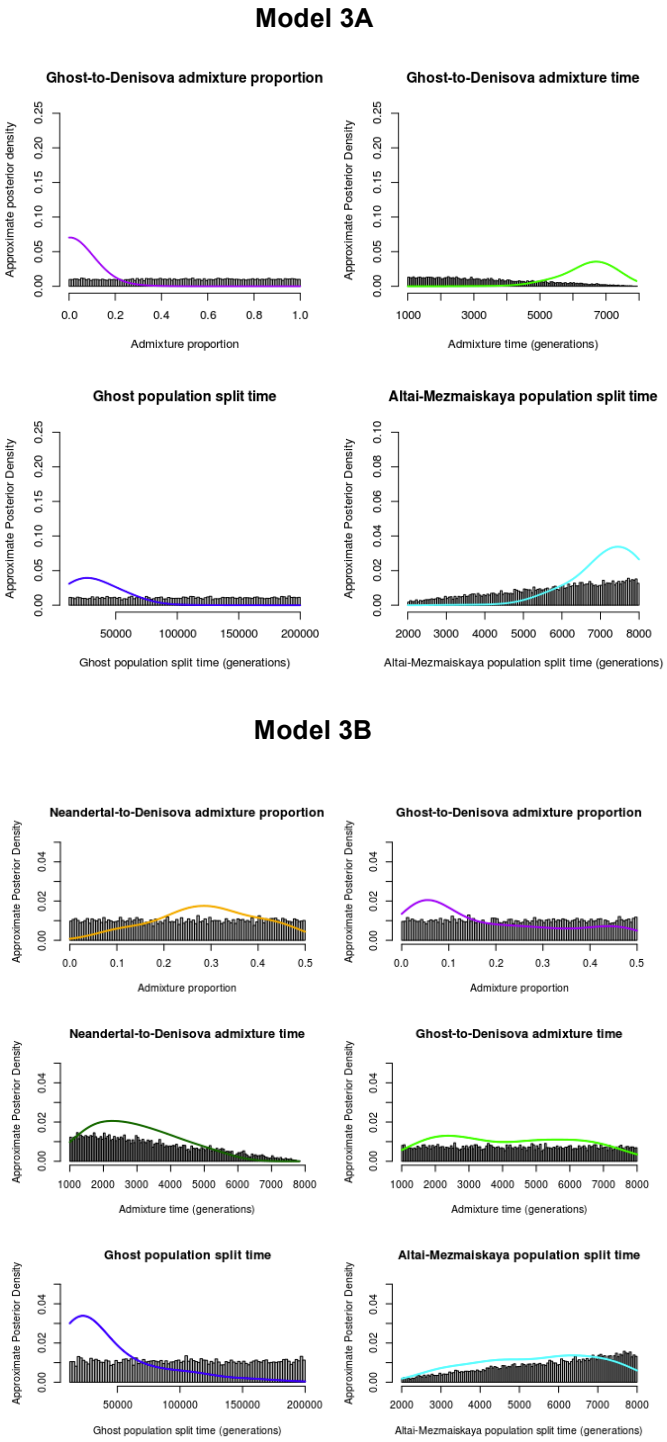


Figure S16b.8. Approximate posterior estimation of sampled parameters under Models 3A and 3B. The curves are the approximate posterior densities, while the bar-plots show the uniformly sampled parameters of all 10,000 simulated samples. The x-axis boundaries in each graph correspond to the boundaries of the prior distributions from which the parameter values are drawn.

5 – Refinement of divergence time distribution using mtDNA coalescence

In Krause et al. (2010)[8], it was observed that the Denisovan mitochondrial genome has a much more ancient coalescence with modern human mitochondrial genomes (1.04 million years ago) than does the Neandertal genome (466,000 years ago). We can further refine our estimate of the archaic divergence time using our ABC distribution for the archaic divergence as a prior, and the probability of the mtDNA sequence divergence given the archaic divergence as a likelihood:

$$P(t_s \mid d) = L(d \mid t_s)P_{ABC}(t_s) = \int P(d \mid \tau)P(\tau \mid t_s)\, d\tau \; P_{ABC}(t_s)$$

where d = observed mtDNA sequence divergence, τ = mitochondrial coalescence time in generations and $P_{ABC}(t_S)$ is the approximate distribution for the archaic divergence time obtained using ABC (Figure S16b.4).

We obtained $P(\tau \mid t_S)$ by assuming a constant population size of 2N = 20,000 and either t_{NM} = 16,000, t_{NM} = 12,000 or t_{NM} = 9,000 generations ago. We assumed an admixture proportion f = 3%, in concordance with the estimated posterior mode above (under a limited admixture regime).

$$P(\tau \mid t_s) = I(\tau < t_s)\left[(1-f)\frac{1}{2N}e^{-\frac{\tau - t_{NM}}{2N}}\right] + I(\tau > t_s)\left[(1-f)e^{-(t_s - t_{NM})}\frac{1}{2N}e^{-\frac{\tau - t_s}{2N}} + f\frac{1}{2N}e^{-\frac{\tau - t_s}{2N}}\right]$$

Here, *I(x)* is the indicator function of the event x. The first term of the sum refers to the event in which the coalescence τ occurs more recently than t_S, in which case there is no admixture and the probability of the coalescence is higher at t_{NM} and decreases into the past. The second term refers to the event in which the coalescence τ occurs more anciently than t_S. In this case, two types of events can happen: either 1) admixture occurs (with probability f) or 2) no admixture occurs (with probability 1-f) and no coalescence occurs between t_{NM} and t_S.

We obtained the number of transversions and transitions separating the Denisovan individual and present-day humans using MEGA[9] (v5.1) on a multiple sequence alignment of the Denisovan individual (FN673705), a Neandertal (NC_011137), the reference human rCRS (NC_012920) and the chimpanzee (NC_001643) mitochondrial genomes aligned via MUSCLE[10] (v3.8.31). We then calculated $P(d \mid \tau)$ by assuming a Kimura 2-parameter model[11] with substitution rate ρ = 1.56e-8 per site per year, 25 years per generation and a transition bias κ = 14.9 (as in ref.[12]).

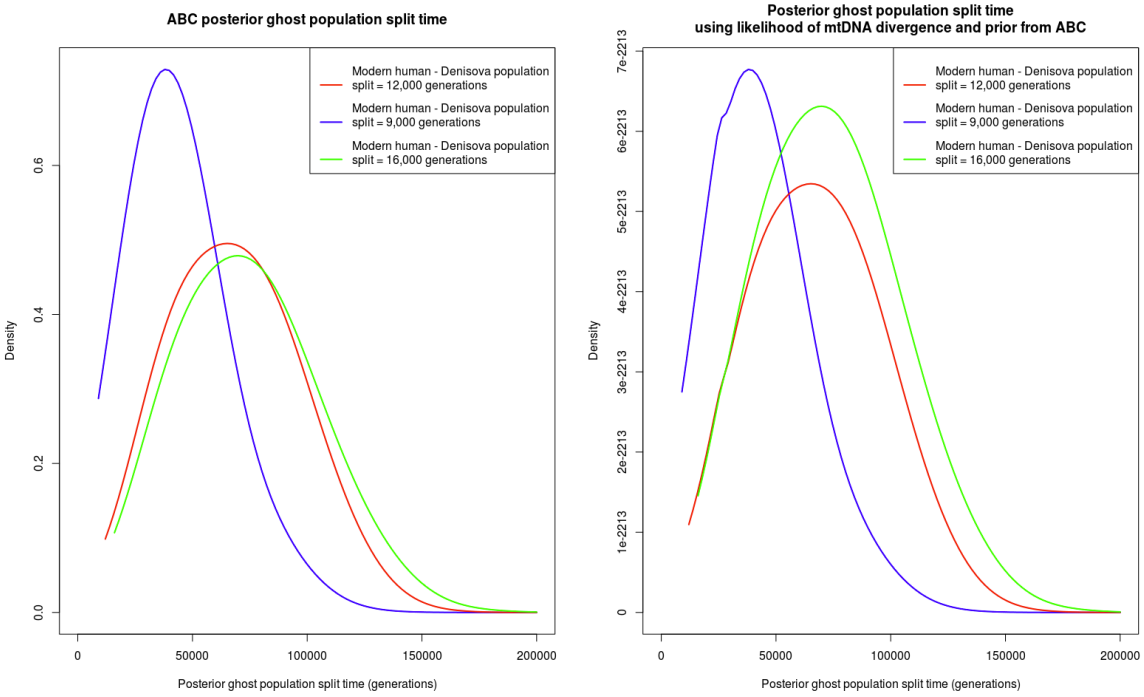


Figure S24b.9. Un-refined (left panel) and refined (right panel) posterior distributions for the divergence time t_S of the ghost lineage that may have admixed with Denisova. The unrefined distributions are the same as in Figure S16b.4 (center-right panel). The refined distributions were obtained by multiplying the likelihood of the mitochondrial divergence given the archaic divergence and the ABC distribution for the archaic divergence from Figure 3 (as a prior) and assuming a constant population size of 2N = 20,000. We computed these distributions assuming t_{NM} = 12,000 generations ago, t_{NM} = 9,000 generations ago or t_{NM} = 16,000 generations ago.

We performed numerical integration in R to obtain the likelihood P(d | τ) and then multiplied it with the approximate archaic divergence time distribution from Figure S16b.4. This gives us new posterior distributions for the archaic divergence time, which we show in Figure S24b.9. The posterior mode of these distributions is a divergence time of 38,385 generations assuming tNM = 9,000 generations ago, 64,445 generations ago assuming t_{NM} = 12,000 generations ago and 69,538 generations ago assuming t_{NM} = 16,000 generations ago.

6 – Conclusion

We find some support for a model where a portion of the Denisovan individual's ancestry comes from an outgroup of Neandertals, Denisovans and modern humans (model 3), assuming admixture was generally low (<10%) among archaic or early modern humans. Assuming model 3 is correct, we find that a model where Neandertals and Denisovans remain sister groups relative to modern humans (model 3B) is better supported than a model where the super-archaic ancestry is due to Denisovans being an outgroup to Neandertals and modern humans (model 3A).

The admixture proportion for the introgression event depends on the demographic history assumed, but is estimated to be small (0.61%-8.08%), even under a broad admixture regime, while the population split time for the introgressing super-archaic lineage is estimated to be close in time to the archaic-modern human split, at 17,706 – 69,894 generations. Due to the fact that the Neandertal and Denisovan individuals were sampled in the past, this split time may be slightly – but not considerably – more ancient. We note that this admixture event could serve to explain the deep coalescence time observed between present-day human mitochondrial genomes and the Denisova mitochondrial genome.

7 – References

1 Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-338 (2002).

2 Le Cao, K. A., Gonzalez, I. & Dejean, S. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* **25**, 2855-2856, doi:10.1093/bioinformatics/btp515 (2009).

3 Peter, B. M., Huerta-Sanchez, E. & Nielsen, R. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS genetics* **8**, e1003011, doi:10.1371/journal.pgen.1003011 (2012).

4 Tenenhaus, M. *La régression PLS: théorie et practique.* . 274 (1998).

5 Leuenberger, C. & Wegmann, D. Bayesian computation and model selection without likelihoods. *Genetics* **184**, 243-252, doi:10.1534/genetics.109.109058 (2010).

6 Wegmann, D., Leuenberger, C., Neuenschwander, S. & Excoffier, L. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC bioinformatics* **11**, 116, doi:10.1186/1471-2105-11-116 (2010).

7 Jeffreys, H. *The Theory of Probability*. 3rd edn, 432 (Oxford, 1961).

8 Krause, J. *et al.* The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464**, 894-897, doi:10.1038/nature08976 (2010).

9 Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution* **28**, 2731-2739, doi:10.1093/molbev/msr121 (2011).

10 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).

11 Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution* **16**, 111-120 (1980).

12 Briggs, A. W. *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**, 318-321, doi:10.1126/science.1174462 (2009).

Supplementary Information 17

Population genetic modelling

Flora Jay* and Montgomery Slatkin

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

Introduction

An excess of similarity between Denisovans and Oceanians relative to other Eurasians was found in Note S14 and previous studies[1,2]. Until now, this excess was attributed to direct gene flow from Denisova or a population closely related to Denisova into Oceanians (Figure 1, Model A).
Notes S16a,b show evidence for archaic gene flow into Denisova. The new model that arises naturally is thus identical to Model A with an additional gene flow from a ghost archaic population to Denisova (Figure 1, Model B). However, another model, where there is no direct gene flow from Denisova to Oceanians, but rather from a ghost archaic into both Denisovans and Oceanians, might also be a candidate (Figure1, Model C). We will show in this note that model C can be discarded.
We then show that the archaic material in Oceanian could be introgressed from a population related to Denisova, and investigate how closely related this population and Denisova have to be depending on the archaic gene flow into Denisova.
In this note, we will use the Papuan individual as a representative of the Oceanian populations.

Method

We derived the D-statistic expectations for all comparisons under different demographic scenarios (see Appendix). For simplicity we assumed constant population size over time in all branches. Although this assumption is not realistic, we can still get insights into the differences between several admixture scenarios. But note that D-statistics do not depend on the effective sizes of the terminal populations[3], so that differences in the size of Neandertals and Denisovans are unlikely to impact our results. We also did not model potential gene flow between Denisovans and Neanderthals (see Note S15 for further details). We used the D-statistics computed in Note S14: D(A;B;C;D)= (nBABA-nABBA) / (nBABA+nABBA), where nBABA is a count of alleles agreeing in population A, C, and also in B, D (but different in A, B).

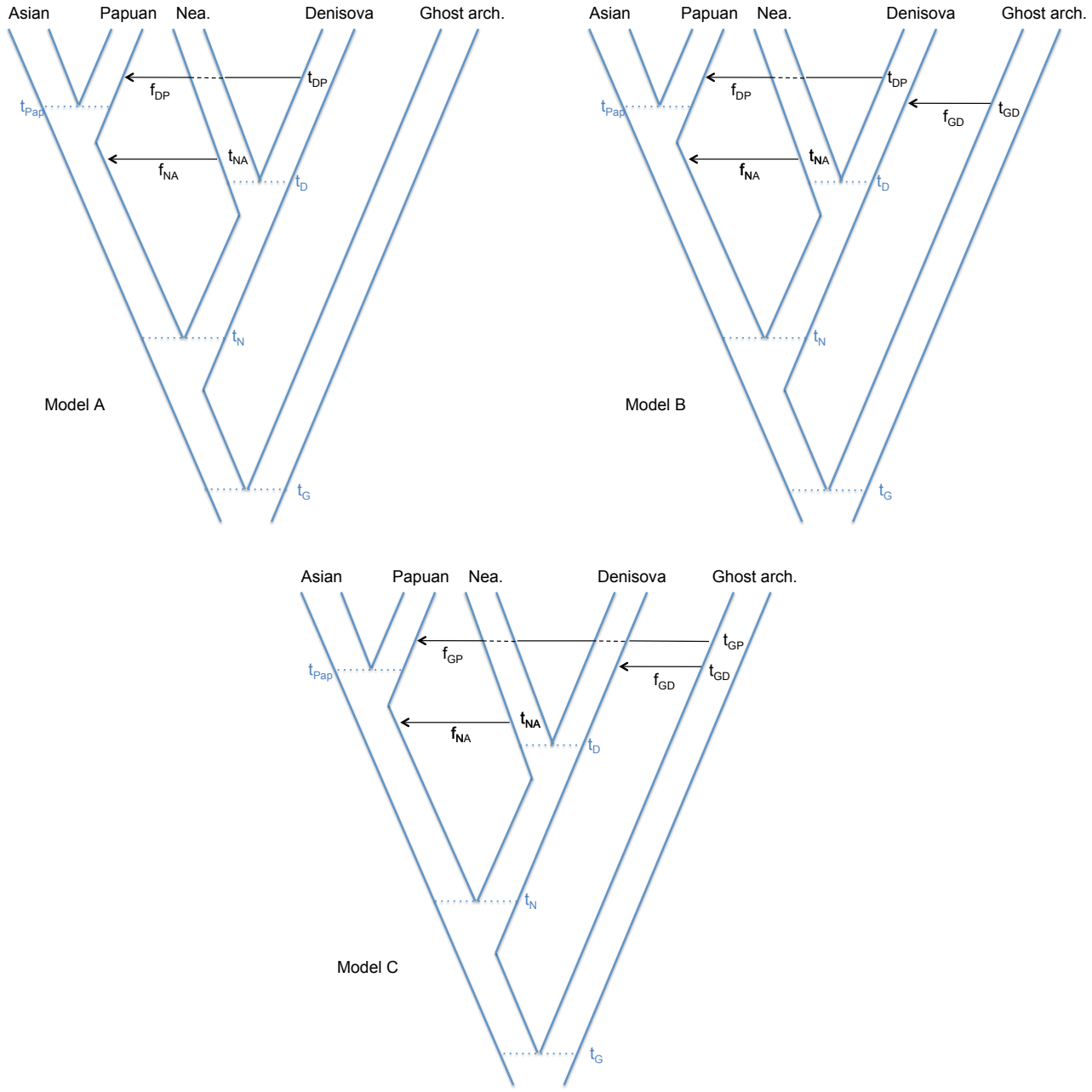


Figure 1. Demographic scenarios. Parameters: split times t_G, t_N, t_D, t_{Pap} ; amounts of gene flow f_{NA}, f_{GD}, f_{GP}, ; times of gene flow: t_{NA} , t_{GD}, t_{GP}. In Model B and C, archaic gene flow to Denisova could happen either before of after gene flow to Papuan. In all models African split from the Asian/Papuan ancestors at a time t_{Afr} comprised in $]t_{Pap}, t_N[$.

Table 1 Parameter space explored for testing for archaic gene flow into Denisova (Model B versus Model A). See also Figure 2. Times are given in generations.

Parameter	Description	Range
t_G	Split time Ghost archaic - others	[20000,60000] ; step=10000
t_N	Split time Modern Humans - Altai	{9000,12000,15000}
t_D	Split time Altai - Denisova	[7000, t_N-1000] ; step=1000
f_{GD}	Amount of gene flow from the ghost archaic to Denisova	{[0.0,0.025],step=0.005, and [0.025,0.3] ; step=0.025 }
t_{GD}	Time of the instantaneous episode of gene flow from the ghost to Denisova	[2000,14000] ; step=2000

Results

Archaic gene flow into Denisova: Model B is more likely than Model A

In agreement with the results in S16a,b, we find that Model A (no archaic gene flow into Denisova) fails to explain the positive value of D(Altai Neandertal, Denisova, African, Chimp) which indicates that Altai is closer to African than Denisova is. The expected D-statistics for Model B are plotted in Figure 2. Our calculations show that, to fit D(Altai, Denisova, African, Chimp), the amount of archaic gene flow into Denisova has to be larger than 0.01 and smaller than 0.2 for a wide range of parameters (Table 1) and population sizes set to 10,000, but see S16a,b for more details and alternative scenarios. This amount represents the fraction of the Denisovan genome that originated in the ghost population.

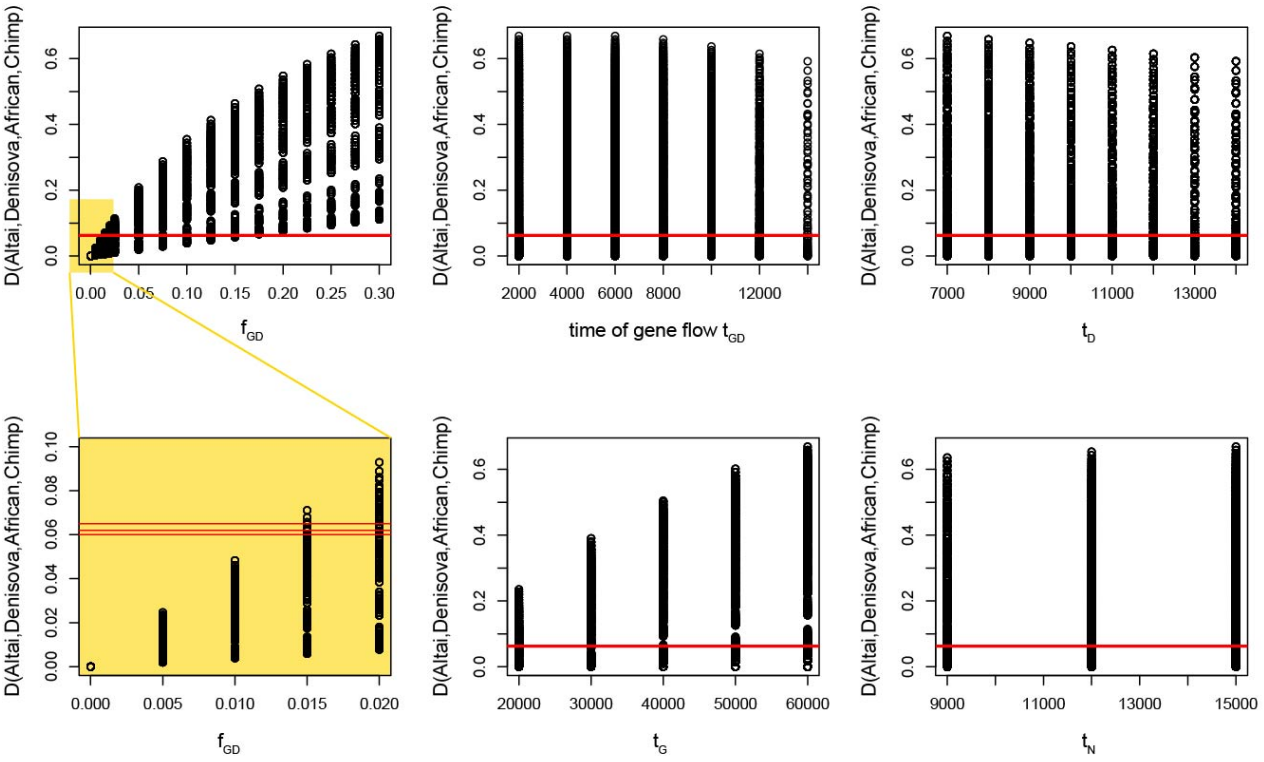


Figure 2. D(Altai, Denisova, African, Chimp) expected for Model B plotted as a function of 5 different parameters. Parameters are described in Table 1 and the population size was set to 10,000. Each dot corresponds to the D value expected for one combination in the parameter space. Red lines indicate the three observed D(Altai, Denisova, African, Chimp) for African=Dinka or Mandenka or Mbuti.

Model C (ghost archaic gene flow into both Denisova and Papuans) can be discarded.

In Model C we test the hypothesis that Papuans have more Denisova material than in any other population (D(Papuan, Asian, Denisova, Chimp)>0) because both populations experienced archaic gene flow; this would create extra time for coalescence when both the Papuan and the Denisova lineage trace their ancestry to the ghost archaic population. However, Model C does not systematically lead to D(Papuan, Asian, Denisova, Chimp)>0: for example if the gene flow from the ghost archaic population is large for Papuans and low for Denisova (e.g. $f_{GP}>>f_{GD}$), D(Papuan, Asian, Denisova, Chimp) is negative. Moreover, under Model C the similarity between Papuan and Neandertal will always be smaller than between Asian and Neandertal, i.e. D(Papuan, Asian, Neandertal, Chimp)<0: if the Papuan lineage traces its ancestry back into the ghost archaic population, it cannot coalesce with the Neandertal lineage before the split of the ghost archaic population from the ancestors of

modern humans, Neandertals and Denisovans (t_G). Thus, the Neandertal lineage is more likely to coalesce with the Asian lineage first.

The expectations of D(Papuan, Asian, Denisova, Chimp) and D(Papuan, Asian, Neandertal, Chimp), for different sets of parameters under model C are plotted in Figure 3 (admixture proportions f_{GP} and f_{GD} ranged from 0 to 0.3, see Table 2 the remaining parameters). The observed D-statistics for different Asian individuals are all positive (red lines), whereas D(Papuan, Asian, Neandertal, Chimp) computed under Model C are all negative (X-axis). Note that Model C with an extra episode of gene flow from Neandertal to Denisova, as found in Note S15, would still produce D(Papuan, Asian, Altai, Chimp). Model C can thus be discarded.

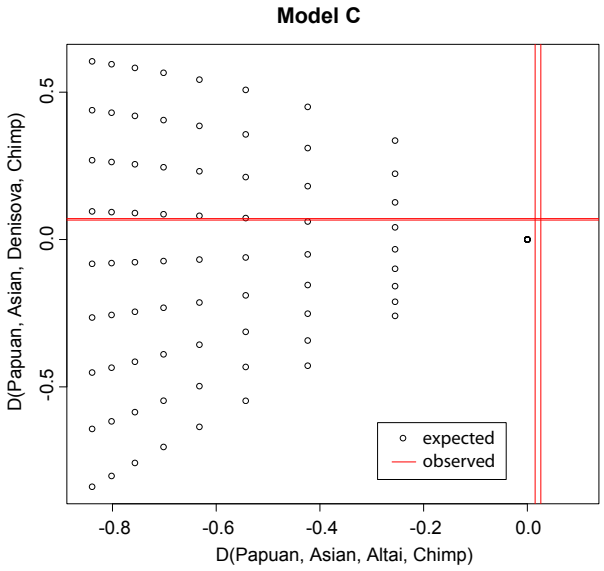


Figure 3. X-axis: D(Papuan, Asian, Neandertal, Chimp) ; Y-axis: D(Papuan, Asian, Denisova, Chimp). The dots indicate the expectations under Model C for different sets of parameters (t_G, t_N, t_D , t_{Pap}, f_{NA}, f_{GD}, f_{GP}, t_{NA}, t_{GD}, t_{GP}). The red lines indicate the observed values where the Asian individual is either from the Han, or the San population. The X-coordinates of the dots are all negatives or null.

Admixture from a close relative of Denisova to Papuans

We investigate a model in which Papuans experienced gene flow from a sister group of Denisova but not Denisova itself. We denote this model 'Model BS'. Note that Model B is actually a specific case of Model BS where the split time between the sister group and Denisova is equal to the time of gene flow to Papuans. We set t_{Pap} and t_{Afr} to 1800 and 2500 generations respectively, and explore the parameter space {N, t_G, t_N, t_D,, t_S, f_{GD}, t_{GD}, f_{NA},, t_{NA}, f_{DP}, t_{DP}, f_{SP}, t_{SP}} described in Table 2. We calibrate these parameters by fitting D(Han, Dinka, Altai, Chimp), D(Han, Dinka, Denisova, Chimp), D(Papuan, Dinka, Altai, Chimp), D(Papuan, Dinka, Denisova, Chimp), D(Papuan, Han, Altai, Chimp), and D(Papuan, Han, Denisova, Chimp). The distance between each observed statistic and the computed one was required to be less than 0.005.

When assuming no archaic admixture in Denisova (f_{GD} =0; Model AS), we observe that the split time between Denisova and its sister group has to be less than 21% of the split time between Altai and Denisova ($t_S < 0.21t_D$). This small percentage indicates that there was little time for Denisova and its sister group to diverge. The more divergent Denisova and its sister group are, the higher the gene flow from the sister group has to be to fit the observed D-statistics. These results are consistent with Supplement 11 of Reich et al. (2010) that showed that this percentage has to be less than ~ 33%.

Table 2 Parameter space explored for fitting Models B and C.

Parameter	Description	Range
N	Population size	{8000, 10000}
t_G	Split Ghost archaic - others	[20000, 80000] ; step=20000
t_N	Split Modern Humans - Altai	{9000, 12000, 15000, 18000}
t_D	Split Altai - Denisova	[5000, t_N-1000] ; step=1000
t_S	Time of split between Denisova and its sister group	[500, t_D] ; step=500
f_{GD}	Gene flow from the ghost archaic to Denisova	[0.015,0.1] ; step=0.005
t_{GD}	Time of gene flow from the ghost to Denisova	[1000, tD-500] ; step=1500
f_{NA}	Gene flow from Neanderthal to Asians and Papuans	[0.01, 0.1] ; step=0.005
t_{NA}	Time of gene flow from Neanderthal to Asians and Papuans	[t_{Pap}, t_{Afr}] ; step=100
f_{DP}	Gene flow from Denisova to Papuans	[0.0, 0.1] ; step=0.005
t_{DP}	Time of gene flow from Denisova to Papuans	[1000, t_{Pap}] ; step=200
f_{SP}	Gene flow from a sister group of Denisova to Papuans	[0.0, 0.1] ; step=0.005
t_{SP}	Time of gene flow from a sister group of Denisova to Papuans	Does not matter, has to be < t_S

When assuming archaic admixture in Denisova (f_{GD} >0), we additionally use D(Altai, Denisova, Dinka, Chimp) as a summary statistic. If the sister group did not experience archaic gene flow (ie $t_{GD} < t_S$; Model BS1), the split time between the sister group and Denisova has to be less than 50% of the split time between Altai and Denisova, and the best fit is obtained for 15%. However, if the ancestors of Denisova and its sister group experienced archaic gene flow (i.e. $t_{GD} > t_S$; Model BS2), the split time could be older (up to 65% of t_D, best fit for 50%). This is because the Papuan and the Denisova lineage would have extra time for coalescence in case they both trace their ancestry in the ghost archaic population. This counterbalances the reduction in time for coalescence in the Denisova population. In summary, the estimation of the split time between Denisova and the relative introgressor based on D-statistic only gets uncertain when the range of plausible models and the number of parameters increase. Note S13 provides an estimation of this time based on the analysis of phased haplotypes.

Models that explain the observed D-statistics

Using Dinka, Han, Papuan, Altai and Denisova, with t_{Pap} and t_{Afr} set to 1800 and 2500 generations respectively, we found that Model BS2, with archaic admixture in the ancestors of Denisova and a sister group, and gene flow from this sister group to Papuans, provides the best fit among all scenarios and parameter sets investigated (Table 3, bottom line; parameter space: Table 2).

Table 3. Parameter sets that provide the best fit for Models B, BS1, and BS2. Top row: B, direct gene flow from Denisova to Papuans (t_S=t_{DP} =t_{SP}). Middle row: BS1, gene flow from Denisova sister group to Papuans, ghost admixture into Denisova only (t_S> t_{GD}). Botton row: BS2, gene flow from Denisova sister group to Papuans, ghost admixture into the ancestors of Denisova and its sister group (t_S<t_{GD}).

Submodel	N_{pop}	t_G	t_N	t_D	t_S	f_{GD}	t_{GD}	f_{NA}	t_{NA}	f_{DP} or f_{SP}	t_{DP} or t_{SP}	Mean error
B	8000	20000	12000	10000	/	0.075	1000	0.045	2500	0.04	1200	0.00235
BS1	10000	20000	12000	10000	1500	0.09	1000	0.06	2000	0.06	<1500	0.00284
BS2	8000	20000	12000	10000	5000	0.075	7000	0.045	2400	0.065	<5000	**0.00205**

Conclusion

This note emphasizes that the extra archaic material in Papuans is not due to gene flow from a same ghost archaic population into both Papuans and Denisovans (Model C discarded), but rather to gene flow from Denisova or a relative of Denisova. The split between this relative introgressor and Denisova could be as recent as the time of gene flow or as ancient as 65% of the Altai/Denisova split time, the best fit being 50% for the parameter space investigated. Note that additionally constraining the split time between Denisova and Altai to be around 50% [30%-70%] of the split time between their ancestors and modern humans (as found in Note S12) leads to similar conclusions. Estimates of the divergence time between the introgressing Denisova material in Papuans and the Siberian Denisovan genome based on phased data are given in Note S13. Estimates of mixture proportions are given in Notes S14 and S16a,b.

References

1 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).
2 Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060, doi:10.1038/nature09710 (2010).
3 Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Molecular biology and evolution* **28**, 2239-2252, doi:10.1093/molbev/msr048 (2011).

APPENDIX FOR SUPPLEMENTARY INFORMATION 17 "POPULATION GENETIC MODELLING"

We derived the D-statistic expectations for all comparisons under different demographic scenarios. For simplicity we assumed constant population size over time in all branches. The scenarios and parameters are described in SM17 Figure 1, Tables 1 and 2. We denote exABBA the excess of ABBA patterns, and exBABA the excess of BABA patterns.

$$D = \frac{ex\text{BABA} - ex\text{ABBA}}{ex\text{ABBA} + ex\text{BABA} + 2 \text{ SHARED}}$$

In this Appendix we only give the expectations of D for Model B and C as these are models that were never considered before (describing potential gene flow from a ghost population to Denisovans and Papuans).

Model B: gene flow from the ghost archaic to Denisovans and from Denisovans to Papuans

We denote f_{GDP} the probability that the Papuan lineage traces back in the ghost population given that it already traces back in the Denisova population.
If $t_{GD} < t_{DP}$, $f_{GDP} = 0$: the Papuan linEage can trace back in the Denisova population but not from the Denisova population to the ghost population because of the order of the gene flow events.
If $t_{GD} >= t_{DP}$, $f_{GDP} = f_{GD}$.

E[Papuan, Asian, Neandertal, Chimp].

$$\begin{aligned}
\text{SHARED} &= (1-f_{NA})f_{DP}(1-f_{GDP})\left(1-\frac{1}{2N}\right)^{t_N-t_D}\frac{2N}{3} \\
&+(1-f_{NA})f_{DP}f_{GDP}\left(1-\frac{1}{2N}\right)^{t_G-t_N}\frac{2N}{3} \\
&+2(1-f_{NA})(1-f_{DP})f_{NA}\left(1-\frac{1}{2N}\right)^{t_N-t_{Pap}}\frac{2N}{3} \\
&+(1-f_{NA})^2(1-f_{DP})\left(1-\frac{1}{2N}\right)^{t_N-t_{Pap}}\frac{2N}{3} \\
&+f_{NA}f_{DP}(1-f_{GDP})\left(1-\frac{1}{2N}\right)^{t_D-t_{NA}}\frac{2N}{3} \\
&+f_{NA}f_{DP}f_{GDP}\left(1-\frac{1}{2N}\right)^{t_G-t_{NA}}\frac{2N}{3} \\
&+f_{NA}^2(1-f_{DP})\left(1-\frac{1}{2N}\right)^{t_{NA}-t_{Pap}}\frac{2N}{3} \\
ex\text{ABBA} &= (1-f_{NA})f_{DP}(1-f_{GDP})(t_N-t_D) \\
&+(1-f_{NA})(1-f_{DP})f_{NA}\left(1-\frac{1}{2N}\right)^{t_{NA}-t_{Pap}}(t_N-t_{NA}) \\
ex\text{BABA} &= (1-f_{NA})f_{DP}f_{GDP}(t_G-t_N)+f_{NA}f_{DP}(1-f_{GDP})(t_D-t_{NA}) \\
&+f_{NA}f_{DP}f_{GDP}(t_G-t_{NA})+f_{NA}(1-f_{DP})(1-f_{NA})\left(1-\frac{1}{2N}\right)^{t_{NA}-t_{Pap}}(t_N-t_{NA})
\end{aligned}$$

E[Papuan, Asian, Denisova, Chimp].

$$
\begin{aligned}
\text{SHARED} \quad = \quad & (1-f_{NA})(1-f_{GP})\{ \\
& f_{DP}(1-f_{GDP})\left(1-\frac{1}{2N}\right)^{t_N-t_{DP}}\frac{2N}{3} \\
& +f_{DP}f_{GDP}\left(1-\frac{1}{2N}\right)^{t_{GD}-t_{DP}+t_G-t_N}\frac{2N}{3} \\
& +(1-f_{DP})f_{NA}\left(1-\frac{1}{2N}\right)^{t_{NA}-t_{Pap}+t_N-t_D}\frac{2N}{3} \\
& +(1-f_{DP})(1-f_{NA})\left(1-\frac{1}{2N}\right)^{t_N-t_{Pap}}\frac{2N}{3} \\
& \}+f_{NA}(1-f_{GP})\{ \\
& f_{DP}(1-f_{GDP})\left(1-\frac{1}{2N}\right)^{t_D-t_{DP}}\frac{2N}{3} \\
& +f_{DP}f_{GDP}\left(1-\frac{1}{2N}\right)^{t_G-t_D}\frac{2N}{3} \\
& +(1-f_{DP})f_{NA}\left(1-\frac{1}{2N}\right)^{t_D-t_{Pap}}\frac{2N}{3} \\
& +(1-f_{DP})(1-f_{NA})\left(1-\frac{1}{2N}\right)^{t_{NA}-t_{Pap}+t_N-t_D}\frac{2N}{3} \\
& \}+(1-f_{NA})f_{GP}\{ \\
& +f_{DP}(1-f_{GDP})\left(1-\frac{1}{2N}\right)^{t_G-t_N}\frac{2N}{3} \\
& +f_{DP}f_{GDP}\left(1-\frac{1}{2N}\right)^{min(0,t_{GD}-t_{DP})+t_G-t_{GD}}\frac{2N}{3} \\
& +(1-f_{DP})f_{NA}\left(1-\frac{1}{2N}\right)^{t_{NA}-t_{Pap}+t_G-t_N}\frac{2N}{3} \\
& +(1-f_{DP})(1-f_{NA})\left(1-\frac{1}{2N}\right)^{t_G-t_{Pap}}\frac{2N}{3} \\
& \}+f_{NA}f_{GP}\{ \\
& f_{DP}(1-f_{GDP})\left(1-\frac{1}{2N}\right)^{t_G-t_D}\frac{2N}{3} \\
& +f_{DP}f_{GDP}\left(1-\frac{1}{2N}\right)^{min(0,t_{GD}-t_{DP})+t_G-t_{GD}}\frac{2N}{3} \\
& +(1-f_{DP})f_{NA}\left(1-\frac{1}{2N}\right)^{t_G-t_{Pap}}\frac{2N}{3} \\
& +(1-f_{DP})(1-f_{NA})\left(1-\frac{1}{2N}\right)^{t_{NA}-t_{Pap}+t_G-t_N}\frac{2N}{3}\}
\end{aligned}
$$

$$
\begin{aligned}
ex\text{ABBA} &= (1-f_{NA})(1-f_{GP})\{ \\
&\quad f_{DP}(1-f_{GDP})(t_N-t_{DP}) \\
&\quad +f_{DP}f_{GDP}(t_{GD}-t_{DP}) \\
&\quad +(1-f_{DP})f_{NA}\left(1-\frac{1}{2N}\right)^{t_{NA}-t_{Pap}}(t_N-t_D) \\
&\quad \} \\
&\quad +f_{NA}(1-f_{GP})f_{DP}(1-f_{GDP})(t_D-t_{DP}) \\
&\quad +f_{GP}f_{DP}(1-f_{GDP})(min(0,t_{GD}-t_{DP})) \\
&\quad +f_{GP}f_{DP}f_{GDP}\left(min(0,t_{GD}-t_{DP})+\left(1-\frac{1}{2N}\right)^{min(0,t_{GD}-t_{DP})}(t_G-t_{GD})\right) \\
ex\text{BABA} &= (1-f_{NA})(1-f_{GP})f_{DP}f_{GDP}\left(1-\frac{1}{2N}\right)^{t_{GD}-t_{DP}}(t_G-t_N) \\
&\quad +f_{NA}(1-f_{GP})\left(f_{DP}f_{GDP}(t_G-t_D)+(1-f_{DP})(1-f_{NA})\left(1-\frac{1}{2N}\right)^{t_{NA}-t_{Pap}}(t_N-t_D)\right)
\end{aligned}
$$

Model C: 2 separate gene flows from the ghost archaic to Papuans and to Denisovans

E[Papuan, Asian, Neandertal, Chimp].

$$
\begin{aligned}
\text{SHARED} &= (1-f_{NA})f_{GP}\left(1-\frac{1}{2N}\right)^{t_G-t_N}\frac{2N}{3} \\
&\quad +(1-f_{NA})(1-f_{GP})f_{NA}\left(1-\frac{1}{2N}\right)^{t_{NA}-t_{Pap}}\left(2*\left(1-\frac{1}{2N}\right)^{t_N-t_{NA}}\frac{2N}{3}+t_N-t_{NA}\right) \\
&\quad +(1-f_{NA})(1-f_{GP})(1-f_{NA})\left(1-\frac{1}{2N}\right)^{t_N-t_{Pap}}\frac{2N}{3} \\
&\quad +f_{NA}f_{GP}\left(1-\frac{1}{2N}\right)^{t_G-t_{NA}}\frac{2N}{3} \\
&\quad +f_{NA}^2(1-f_{GP})\left(1-\frac{1}{2N}\right)^{t_N-t_{Pap}}\frac{2N}{3} \\
ex\text{ABBA} &= \text{SHARED} \\
ex\text{BABA} &= \text{SHARED}+(1-f_{NA})f_{GP}(t_G-t_N)+f_{NA}f_{GP}(t_G-t_{NA})
\end{aligned}
$$

E[Papuan, Asian, Denisova, Chimp].

$$
\begin{aligned}
\text{SHARED} \;=\; & (1-f_{NA})f_{GP}(1-f_{GD})\left(1-\frac{1}{2N}\right)^{t_G-t_N}\frac{2\text{N}}{3} \\
& +(1-f_{NA})f_{GP}f_{GD}\left(1-\frac{1}{2N}\right)^{t_G-max(t_{GP},t_{GD})}\frac{2\text{N}}{3} \\
& +(1-f_{NA})(1-f_{GP})f_{NA}(1-f_{GD})\left(1-\frac{1}{2N}\right)^{t_{NA}-t_{Pap}}\left(\left(1-\frac{1}{2N}\right)^{t_N-t_D}\frac{4\text{N}}{3}+t_N-t_D\right) \\
& +2(1-f_{NA})f_{NA}(1-f_{GP})f_{GD}\left(1-\frac{1}{2N}\right)^{t_{NA}-t_{Pap}+t_G-t_N}\frac{2\text{N}}{3} \\
& +(1-f_{NA})^2(1-f_{GP})(1-f_{GD})\left(1-\frac{1}{2N}\right)^{t_N-t_{Pap}}\frac{2\text{N}}{3} \\
& +(1-f_{NA})^2(1-f_{GP})f_{GD}\left(1-\frac{1}{2N}\right)^{t_G-t_{Pap}}\frac{2\text{N}}{3} \\
& +f_{NA}f_{GP}(1-f_{GD})\left(1-\frac{1}{2N}\right)^{t_G-t_D}\frac{2\text{N}}{3} \\
& +f_{NA}f_{GP}f_{GD}\left(1-\frac{1}{2N}\right)^{t_G-max(t_{GP},t_{GD})}\frac{2\text{N}}{3} \\
& +f_{NA}^2(1-f_{GP})(1-f_{GD})\left(1-\frac{1}{2N}\right)^{t_D-t_{Pap}}\frac{2\text{N}}{3} \\
& +f_{NA}^2(1-f_{GP})f_{GD}\left(1-\frac{1}{2N}\right)^{t_G-t_{Pap}}\frac{2\text{N}}{3} \\
ex\text{ABBA} \;=\; & \text{SHARED}+(1-f_{NA})f_{GP}f_{GD}(t_G-max(t_{GP},t_{GD}))+f_{NA}f_{GP}f_{GD}(t_G-max(t_{GP},t_{GD})) \\
ex\text{BABA} \;=\; & \text{SHARED}+(1-f_{NA})f_{GP}(1-f_{GD})(t_G-t_N)+f_{NA}f_{GP}(1-f_{GD})(t_G-t_D)
\end{aligned}
$$

Supplementary Information 18
Characterization of changes in the modern and archaic human lineages

Fernando Racimo*, Martin Kircher and Janet Kelso

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

Table of Contents

Methods

Data filtering

We retained positions where:

- A GATK call is available in both Denisovan and Altai Neandertal
- Root mean square map quality (MQ) >= 30 in both Denisovan and Altai Neandertal
- Genotype quality (GQ) > 30 in both Denisovan and Altai Neandertal
- The site is within the GC-sensitive coverage cutoffs specified in SI 5b
- The site is within the mapability track specified in SI 5b
- The EPO human-chimpanzee ancestral allele is available and matches at least one other human-ape ancestor (gorilla or orangutan).
- Human and chimpanzee sequences appear only once in the corresponding EPO alignment block
- The site is not flagged as a systematic error identified by GATK using strand bias in 1000G Trio parents ("SysErr"[1]) or by analyzing shared SNPs across humans, chimpanzees and bonobos ("SysErrHCB"[2]).
- The fraction of reads covering the position that have a MQ = 0 is below 10% in both Denisova and Altai Neandertal
- The minor allele frequency is equal to or larger than 25% for heterozygous sites

We flagged positions in CpG and repeat-masked regions, and nearby InDels (+/- 5bp), but did not exclude them from our analysis. We then defined "modern-human-specific sites" as those sites where Denisova or Altai Neandertal have the ancestral (chimpanzee-like) state and where the derived allele

is either fixed or at high frequency (> 90%) in modern humans, using 1000 Genomes Project (1000G) data[3] (20110521 release), as in Supplementary Note 19 of the Denisova high-coverage genome paper[1]. Conversely, we define "archaic-human-specific sites" as those sites where both Altai Neandertal and Denisovan are homozygous for the same derived allele, and the ancestral allele is fixed or at high-frequency (>90%) in modern humans (Figure S18.1).

Annotation and disruption scoring

We used Ensembl's Variant Effect Predictor[4] (VEP 2.5, Ensembl 67 annotation) to annotate SNCs and insertions/deletions (InDels). For predicting the effect of non-synonymous changes we used the HumDiv model in PolyPhen-2[5]. The catalogs of all sites analyzed here are available at: http://bioinf.eva.mpg.de/altai_catalog/. We also ranked all single-nucleotide changes and all small InDels (<12 bp) using a new scoring method that incorporates protein deleteriousness scores, evolutionary conservation scores, and regulatory and expression data from 63 different annotations (including Grantham[6], SIFT[7] and PolyPhen[5] scores, as well as ENCODE[8] data and UCSC genome browser tracks, conservation metrics and gene model information, like GERP[9,10], phyloP[11], phastCons[12], transcription factor binding regions, expression levels and exon-intron boundaries) to determine how disruptive a change may be, using a general linear model trained to score the impact of an observed variant (see ref. [13] for full list of annotations and description of method). This "Combined Annotation Dependent Depletion" score (CADD), or C-score, has been shown to predict pathogenic and causal variants[14]. We are using a pre-publication version of CADD, available for download from http://krishna.gs.washington.edu/download/CADD/v0.5/. By convention, a C-score is positive for changes that are predicted to be disruptive, and negative for changes predicted not to affect function. The scale of this score is arbitrary due to the diversity of the input annotations, so we use a PHRED-like version of the score, ranging from 1 to 99, in all tables where specific change scores are shown. These PHRED-scaled scores are based on the rank that the variant occupies relative to all possible substitution changes in the genome. For example, a change with a PHRED-scaled score of 20 or greater has a C-score in the highest 10^{-2} = 1% of all possible single-nucleotide changes. We do not claim that any of the high-ranking variants actually change function, unless explicitly stated. Nevertheless, we henceforth call changes with arbitrarily highly positive C-scores as "disruptive", as a way to prioritize them for future experimental studies to test their effects.

Results

Single-nucleotide changes

Table S18.1 shows the number of modern-human-specific single nucleotide changes (SNCs) for VEP-predicted effect categories in genic and regulatory regions, as well as ENCODE open chromatin sites. In Figure S18.2, we show the proportion of changes for different genic and regulatory effect categories that occurred after the split from chimpanzees and that are either modern-human-specific, archaic-human-specific, or shared by both lineages. In Figure S18.3 we show the partitioning of archaic genotype states for modern-human-specific changes (where at least one archaic human has at least one ancestral allele), for different effect categories. There is an excess of Denisova-ancestral sites compared to Neandertal-ancestral sites among these changes, which is particularly pronounced when the change is completely fixed for the derived state in present-day humans, a pattern that we address in SI 16a and SI 16b. In Table S18.2 and Figure S18.4, we show the number of archaic-human-specific changes for the same categories.

InDels

We identified InDels that are specific to the modern human or archaic human lineages. In Table S18.3 we show the number of indels for each effect category for the modern human lineage, and in Table S18.4 we show the number for the archaic human lineage.

Binomial ontology term enrichment

We used a binomial enrichment test in FUNC[15] to identify over-represented ontology categories among genes with modern-human or archaic-human specific SNCs, as in the Denisova high-coverage genome paper[1]. Briefly, we compared the number of lineage-specific SNCs that affect genes in a particular ontology category for a test lineage (archaic humans or modern humans) to the number of SNCs that occurred after the human-chimpanzee divergence but before the modern human-archaic human divergence in genes in the same ontology category. We list overrepresented terms ($p < 0.01$, false discovery rate < 0.1) for each effect category in the modern-human-specific or archaic-human-specific lineages in Tables S18.5 and S18.6, respectively.

Excess and depletion of recent SNCs in Segway segments

We tested for a significant excess or depletion of modern-human-specific or archaic-human-specific SNCs relative to all SNCs that occurred since the human-chimpanzee split in particular genomic features defined by the Segway segmentation tracks[16]. The Segway segmentation tracks (http://genome.ucsd.edu/) partition the genome into segments with distinct genic and regulatory features, which were obtained using a dynamic Bayesian model that utilizes ENCODE data from open chromatin assays, Chip-Seq experiments and evidence for histone modifications. The percent of SNCs that fall within each segment type in the modern-human-specific and archaic-human-specific catalogs is shown in Figure S18.5. We also show the percent of SNCs in each segment corresponding to two additional catalogs: the set of all fixed and high-frequency derived SNCs that occurred since the human-chimpanzee split, and the set of all homozygous derived SNCs in archaic humans that occurred since the human-chimpanzee split.

We tested, for each segment type, whether there was an excess or a depletion of modern-human-specific SNCs relative to all modern human SNCs that occurred since the human-chimpanzee split, and also whether there was an excess or a depletion of archaic-human-specific SNCs relative to all archaic human SNCs that occurred since the human-chimpanzee split, using a hypergeometric test. At a Bonferroni-corrected 1% significance level, we observe a significant excess of SNCs in a few segment types (Figure S18.6), including gene start, gene end and "dead" (putatively inactive) zones in the modern-human-specific catalog, and gene end in the archaic-human-specific catalog. There is a significant depletion of SNCs in segments related to transcription factor activity, repression, low zones and histone methylation (H3K9me1) segments in the modern-human-specific catalog, and H3K9me1 segments in the archaic-human-specific catalog. We note that because we have only two archaic humans, we have less power to detect fixation in archaic humans than we do for modern humans. The power of tests for excess or depletions in the two catalogs is therefore not comparable.

Clinically pathogenic variants

We used the NCBI ClinVar Variation Reporter v1.2 to identify potentially pathogenic variants in the modern-human-specific and the archaic-human-specific catalogs. ClinVar is a public archive of reports of human genetic variants and their relationship to particular diseases. We found seven putative pathogenic variants in the modern-human catalog and one in the archaic-human catalog. Predictably, none of them were completely fixed in present-day humans, as the discovery of a pathogenic condition requires the SNP to be present in present-day humans. We list these variants and their corresponding derived allele frequencies in present-day humans in Table S18.7. For all of the modern-human-specific changes, the risk alleles are always ancestral, while in the case of the one archaic-human-specific change, the risk variant is the derived allele in archaic humans (which is at low frequency in present-day humans).

Highly disruptive changes

We used C-scores (described above) to rank all SNCs in each of the archaic and modern human-specific catalogs by how disruptive they are predicted to be. The distributions of the scores for each

catalog are shown in Figure S18.7. We observe an excess of archaic-specific disruptive (positive) C-scores when compared to the modern-human-specific catalog and the catalog of changes that occurred before the modern-archaic divergence. There are 3 possible explanations for this: a) an excess of fixed and high-frequency disruptive alleles in archaic humans due to their low effective population size, b) the fact that we only have two archaic humans to determine whether an allele is fixed or at high-frequency in all archaic humans or c) the fact that C-scores are aggregating annotations some of which include experimental data obtained from present-day humans.

We also observe that SNCs on chromosome X have significantly less disruptive C-scores than autosomal SNCs in the modern-human-specific catalog (Wilcoxon rank-sum test p = 2.7e-5). One reason for this could be that strongly disruptive changes on the X chromosome are more effectively pruned from the population because of the hemizygosity of the X chromosome in males, which makes natural selection more efficient at removing deleterious variants. However, we do not detect a significant difference in C-scores between the X chromosome and the autosomes in the archaic-human-specific catalog (p = 0.71).

We list the top 30 most disruptive SNCs and InDels in the modern-human-specific catalog (Table S18.8 for fixed changes, Table S18.9 for high-frequency changes). We also list the top 30 most disruptive SNCs and InDels in the archaic-human-specific catalog (Table S18.10 for fixed changes, Table S18.11 for high-frequency changes).

Ontology enrichment using disruption scores

To identify the phenotypes or functions that may be affected by the SNCs predicted using the C-score to be disruptive in the modern-human-specific or archaic-human-specific catalogs we used 2 different methods:

A) Average PHRED-scaled C-score method: For each gene in the catalog, we obtained the sum of the PHRED-scaled C-scores for all of the changes associated with that gene (intronic, exonic, 3' UTR, 5' UTR, upstream and downstream, as defined by the VEP), divided that number by the number of associated changes, and then used the resulting value to rank the genes.
B) Weighted maximum PHRED-scaled C-score method: For each gene in the catalog, we obtained the maximum PHRED-scaled C-score for all the changes associated with that gene, divided that number by the number of associated changes, and then used the resulting value to rank the genes.

In both cases, we performed a Wilcoxon rank sum test on the resulting ranks using FUNC[15]. We show the results ($p < 0.01$) for the modern-human-specific catalog (Table S18.12) and the archaic-human-specific catalog (Table S18.13) using the Gene Ontology (FWER < 0.01) and the Human Phenotype and Diseases ontologies (FWER < 0.5).

Modern-human-specific catalog

Among the fixed modern-human-specific changes (Table S18.8), the highest scoring variant is an insertion for which Denisova is heterozygous (ancestral/derived). The insertion is located in a canonical splice site in *TNFRSF9*, a gene that codes for a tumor necrosis factor receptor known to inhibit the proliferation of T lymphocytes and apoptosis[17]. The third most disruptive change is a frameshift deletion in CLTCL1, coding for clathrin protein involved in vesicle trafficking[18] that is selectively expressed in adult skeletal muscle[19] and that may be involved in meningiomas[18].

Thirteen of the top 30 disruptive fixed changes are missense SNCs. The two most disruptive missense SNCs appear to be in genes involved in neural biology. The most disruptive SNC is found in *ST6GAL2*: a sialyltransferase[20] that may be associated with differential drug responses in schizophrenia (RefSeq, 2012). The second most disruptive missense SNC is found in a transcript of

JAM3 (ENST00000531698), a gene that may be involved in maintaining the structure of myelinated peripheral nerves and Schwann cell junctional attachments[21]. The site is homozygous ancestral in the Altai Neandertal but not in the Denisovan. An SNC in *CNTNAP2*, one of the few known *FOXP2* interactors[22], is homozygous ancestral in the Denisovan[1], but homozygous derived in the Altai Neandertal. Another SNC is located in *ERCC5*, a gene associated with xeroderma pigmentosum, a recessive disorder that causes a deficiency in UV light damage repair[23,24], but is ancestral only in one of the Denisovan's chromosomes[1] and homozygous derived in Altai Neandertal. Other disruptive missense SNCs are located in *GTF3C5*, which contributes to the recruitment of RNA polymerase III to make certain types of small RNAs and adenovirus-associated RNAs[25], and in *HGS*, which codes for a growth-factor tyrosine kinase that has a role in tumor suppression[26].

Among the high-frequency (but not fixed) modern-human-specific changes, the seven most disruptive are all nonsense mutations (Table S18.9). One of these is the introduction of a STOP codon in *CASP12*, which had been previously identified as being ancestral in the Denisovan[1] and three Neandertal exomes, including the Altai Neandertal[27]. The change is also recorded as a ClinVar pathogenic variant, as the ancestral variant is associated with increased risk of sepsis[28]. The derived variant also has strong evidence for recent positive selection[29]. The six other high-frequency nonsense mutations are STOP losses. The most highly disruptive STOP loss is located in *RP11-625H11.1*, a gene coding for an uncharacterized protein that is highly expressed in the liver[30]. The second most disruptive STOP loss is found in BOLL, a highly conserved gene[31] which may code for an RNA-binding protein involved in spermatogenesis. It is expressed in the testes[32], is associated with azoospermia in men and may be involved in gamete formation[32,33]. The third most disruptive STOP loss is in *OPRM1*, which codes for an important opioid receptor[34,35]. This site had also been previously identified as being ancestral in the Denisovan[1], and is also homozygous ancestral in Altai Neandertal. We also find a STOP loss in *GTF3C2*, which has been shown to interact with *GTF3C5*[25], a gene that also contains a highly disruptive fixed SNC (see above). Like *GTF3C2*, *GTF3C5* is required for RNA polymerase III recruitment[25].

b) Archaic-human-specific catalog

The three most disruptive derived changes in archaic humans (Altai Neandertal+Denisova) that are fixed ancestral in modern humans are STOP gains (Table S18.10). The first of these is located in *LASP1*, which codes for a protein that may be involved in actin binding and cell adhesion[36,37]and that is particularly highly expressed in ovarian and breast cancer[36]. The second STOP gain is in the *RPL28* gene, which codes for a ribosomal protein[38], while the third STOP gain is in *OR5AC2*, an olfactory receptor.

We also find several highly disruptive archaic-specific fixed InDels in canonical splice sites. These include 1-bp deletions in *SLC35B4* and *CREB3L1*, as well as 1-bp insertions in *TTPA*, *TRPC4*, *NCKAP1*, *DNAH5* and *IFNG*. Of particular note are the deletions in *TTPA* and *TRPC4*. *TTPA* is involved in vitamin E homeostasis and associated with retinitis pigmentosa and ataxia[39,40], while *TRPC4* codes for a cation channel whose deletion is associated with decreased sociability[41] and strategic learning[42]. In addition, *CREB1L* may play a role in endoplasmic reticulum stress response of astrocytes in the central nervous system (UniProtKB by similarity), *IFNG* codes for an interferon that is essential for the immune response against pathogens and tumors[43] and *DNAH5* has been associated with primary ciliary dyskinesia: a respiratory tract disorder[44].

Discussion

The catalogs of changes unique to either the modern or archaic human lineages can serve as a first step in identifying those genes and biological systems that were most important in recent human evolutionary history. Among non-synonymous changes, we see enrichment for genes affecting melanocyte development in the modern-specific catalog (Table S18.5), and different musculoskeletal and hair morphologies in the archaic-specific catalog (Table S18.6). We explore some of these

enrichments in more detail in a companion study[27] with the addition of 2 Neandertal exomes, to obtain a higher resolution for the allelic state of these non-synonymous changes in archaic human groups. Among genes with changes in 3' UTR regions in the modern-specific catalog (Table S18.5), we also observe enrichments for particular skeletal morphologies, including limb length, as well as morphologies of the larynx and the epiglottis which are involved in speech production. We note, however, that Neandertals are known to have had hyoid bones that were virtually identical to those of modern humans, which makes it likely that the anatomical structures necessary for human speech had already evolved before the Neandertal and modern human lineages split[45].

We observe that certain types of genomic segments have significant over- and underrepresentation of changes in the modern-human and the archaic-human catalogs, relative to all changes that occurred in those segments in humans since the human-chimpanzee split (Figure S18.6). The most significant overrepresentation is in gene end regions in the modern human catalog. This excess could be due to differing selective pressures in a particular set of genic regions, but we also caution that the explanation for these patterns need not be adaptive, and could also be explained by neutral changes and re-arrangements in the human genomic landscape that occurred since the human-chimpanzee ancestor.

The newly developed C-score allows us to prioritize SNCs and InDels for future experimental studies (Table S18.8 and Table S18.9), by synthesizing disruptiveness properties of protein deleteriousness scores as well as conservation and experimental data genome-wide, and thereby rank obvious candidates – like a missense mutation classified by PolyPhen as "probably damaging" – along with changes one would not immediately consider disruptive – like a change in a highly conserved site in a regulatory region. Some of the most highly disruptive changes that are fixed or high-frequency derived in modern humans, like chr11:104763117 in CASP12, already have known pathogenic conditions associated with them, while others, like chr7:146825878 in CNTNAP2, are promising candidates for future experimental testing. We also observe interesting candidates in the archaic specific catalog (Table S18.10 and Table S18.11), like a deletion in an essential splice site of TRPC4, a gene known to affect sociability (Cooper et al. 2011). This scoring system also allows us to identify functions or phenotypes that have been affected by particularly disruptive changes. We observe that genes involved in DNA-binding, heart failure and muscle contraction have changes with significantly highly disruptive C-scores relative to all sites in the modern-human-specific catalog (Table S18.12), while genes associated with pregnancy complications (pre-eclampsia), DNA-binding, heart failure and carcinomas have significantly high C-scores in the archaic-human-specific catalog (Table S18.13).

The catalogs we provide here are the complete set of genetic changes that distinguish modern and archaic humans from their common ancestor, and will allow the genetic processes that led to the evolution of the modern human and archaic human forms to be explored. As a first step, we have identified changes on the modern human lineage that fall within regions that we identify as having been affected by selective sweeps on the modern human lineage (SI 19b). Future functional genomic analyses may serve to further expand the insights we can gain from archaic human sequence data, with respect to both our own recent evolution and the evolution of other extinct human groups.

References

1 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).
2 Castellano, S. *et al.* Patterns of coding variation in the complete exomes of three Neandertals. *in review*.
3 Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).
4 McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-2070, doi:10.1093/bioinformatics/btq330 (2010).
5 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7.20, doi:10.1002/0471142905.hg0720s76 (2013).
6 Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-864 (1974).
7 Ng, P. C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* **7**, 61-80, doi:10.1146/annurev.genom.7.080505.115630 (2006).
8 Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
9 Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-913, doi:10.1101/gr.3577405 (2005).
10 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).
11 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121, doi:10.1101/gr.097857.109 (2010).
12 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).
13 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *in review*.
14 Kircher, M. *et al.* ([in review]).
15 Prüfer, K. *et al.* FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* **8**, 41, doi:10.1186/1471-2105-8-41 (2007).
16 Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**, 473-476, doi:10.1038/nmeth.1937 (2012).
17 Schwarz, H., Tuckwell, J. & Lotz, M. A receptor induced by lymphocyte activation (ILA): a new member of the human nerve-growth-factor/tumor-necrosis-factor receptor family. *Gene* **134**, 295-298 (1993).
18 Kedra, D. *et al.* Characterization of a second human clathrin heavy chain polypeptide gene (CLH-22) from chromosome 22q11. *Hum Mol Genet* **5**, 625-631 (1996).
19 Long, K. R., Trofatter, J. A., Ramesh, V., McCormick, M. K. & Buckler, A. J. Cloning and characterization of a novel human clathrin heavy chain gene (CLTCL). *Genomics* **35**, 466-472, doi:10.1006/geno.1996.0386 (1996).
20 Takashima, S., Tsuji, S. & Tsujimoto, M. Characterization of the second type of human beta-galactoside alpha 2,6-sialyltransferase (ST6Gal II), which sialylates Galbeta 1,4GlcNAc structures on oligosaccharides preferentially. Genomic analysis of human sialyltransferase genes. *J Biol Chem* **277**, 45719-45728, doi:10.1074/jbc.M206808200 (2002).
21 Scheiermann, C. *et al.* Expression and function of junctional adhesion molecule-C in myelinated peripheral nerves. *Science* **318**, 1472-1475, doi:10.1126/science.1149276 (2007).
22 Vernes, S. C. *et al.* A functional genetic link between distinct developmental language disorders. *N Engl J Med* **359**, 2337-2345, doi:10.1056/NEJMoa0802828 (2008).

23 Volker, M. *et al.* Sequential assembly of the nucleotide excision repair factors in vivo. *Mol Cell* **8**, 213-224 (2001).

24 Nouspikel, T. & Clarkson, S. G. Mutations that disable the DNA repair gene XPG in a xeroderma pigmentosum group G patient. *Hum Mol Genet* **3**, 963-967 (1994).

25 Hsieh, Y. J., Wang, Z., Kovelman, R. & Roeder, R. G. Cloning and characterization of two evolutionarily conserved subunits (TFIIIC102 and TFIIIC63) of human TFIIIC and their involvement in functional interactions with TFIIIB and RNA polymerase III. *Mol Cell Biol* **19**, 4944-4952 (1999).

26 Sun, C. X. *et al.* Functional analysis of the relationship between the neurofibromatosis 2 tumor suppressor and its binding partner, hepatocyte growth factor-regulated tyrosine kinase substrate. *Hum Mol Genet* **11**, 3167-3178 (2002).

27 Castellano, S. *et al.* (in review).

28 Saleh, M. *et al.* Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature* **429**, 75-79, doi:10.1038/nature02451 (2004).

29 Xue, Y. *et al.* Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am J Hum Genet* **78**, 659-670, doi:10.1086/503116 (2006).

30 Hägg, S. *et al.* Multi-organ expression profiling uncovers a gene module in coronary artery disease involving transendothelial migration of leukocytes and LIM domain binding 2: the Stockholm Atherosclerosis Gene Expression (STAGE) study. *PLoS Genet* **5**, e1000754, doi:10.1371/journal.pgen.1000754 (2009).

31 Xu, E. Y. *et al.* Human BOULE gene rescues meiotic defects in infertile flies. *Hum Mol Genet* **12**, 169-175 (2003).

32 Xu, E. Y., Moore, F. L. & Pera, R. A. A gene family required for human germ cell development evolved from an ancient meiotic gene conserved in metazoans. *Proc Natl Acad Sci U S A* **98**, 7414-7419, doi:10.1073/pnas.131090498 (2001).

33 Luetjens, C. M. *et al.* Association of meiotic arrest with lack of BOULE protein expression in infertile men. *J Clin Endocrinol Metab* **89**, 1926-1933 (2004).

34 Wang, J. B. *et al.* Human mu opiate receptor. cDNA and genomic clones, pharmacologic characterization and chromosomal assignment. *FEBS Lett* **338**, 217-222 (1994).

35 Zubieta, J. K. *et al.* Regional mu opioid receptor regulation of sensory and affective dimensions of pain. *Science* **293**, 311-315, doi:10.1126/science.1060952 (2001).

36 Traenka, C. *et al.* Role of LIM and SH3 protein 1 (LASP1) in the metastatic dissemination of medulloblastoma. *Cancer Res* **70**, 8003-8014, doi:10.1158/0008-5472.CAN-10-0592 (2010).

37 Schreiber, V. *et al.* Lasp-1, a novel type of actin-binding protein accumulating in cell membrane extensions. *Mol Med* **4**, 675-687 (1998).

38 Frigerio, J. M., Dagorn, J. C. & Iovanna, J. L. Cloning, sequencing and expression of the L5, L21, L27a, L28, S5, S9, S10 and S29 human ribosomal protein mRNAs. *Biochim Biophys Acta* **1262**, 64-68 (1995).

39 Ouahchi, K. *et al.* Ataxia with isolated vitamin E deficiency is caused by mutations in the alpha-tocopherol transfer protein. *Nat Genet* **9**, 141-145, doi:10.1038/ng0295-141 (1995).

40 Gotoda, T. *et al.* Adult-onset spinocerebellar dysfunction caused by a mutation in the gene for the alpha-tocopherol-transfer protein. *N Engl J Med* **333**, 1313-1318, doi:10.1056/NEJM199511163332003 (1995).

41 Rasmus[1,] K. C., Wang, J.-G., Varnell, A., Ostertag, E. M. & Cooper, D. C. (Nature Precedings, 2011).

42 Klipec, W. D. *et al.* Deletion of the trpc4 gene and its role in simple and complex strategic learning. *Nature Precedings*, doi:10.1124/pr.57.4.6 (2012).

43 Schoenborn, J. R. & Wilson, C. B. Regulation of interferon-gamma during innate and adaptive immune responses. *Adv Immunol* **96**, 41-101, doi:10.1016/S0065-2776(07)96002-2 (2007).

44 Olbrich, H. *et al.* Mutations in DNAH5 cause primary ciliary dyskinesia and randomization of left-right asymmetry. *Nat Genet* **30**, 143-144, doi:10.1038/ng817 (2002).

45 Arensburg, B. *et al.* A Middle Palaeolithic human hyoid bone. *Nature* **338**, 758-760, doi:10.1038/338758a0 (1989).

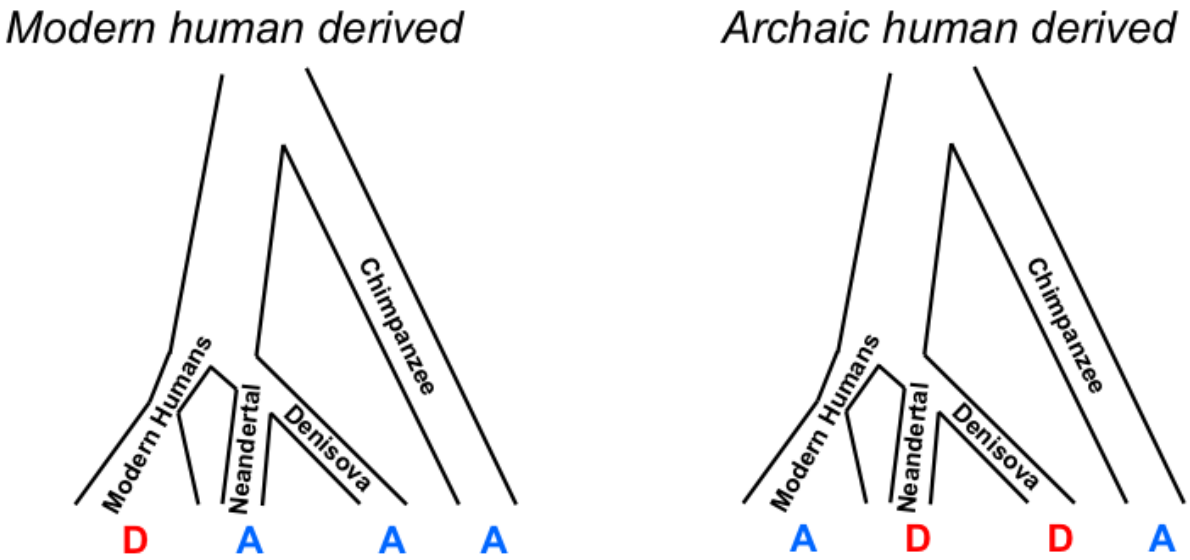


Figure S18.1. To look for modern human specific single nucleotide changes, we selected sites where either all or at least one of the archaic humans have an ancestral allele, and modern humans have a fixed or high-frequency derived allele. Conversely, to look for archaic human specific changes, we selected sites where both archaic humans are homozygous derived and modern humans have the ancestral allele fixed or at high frequency.

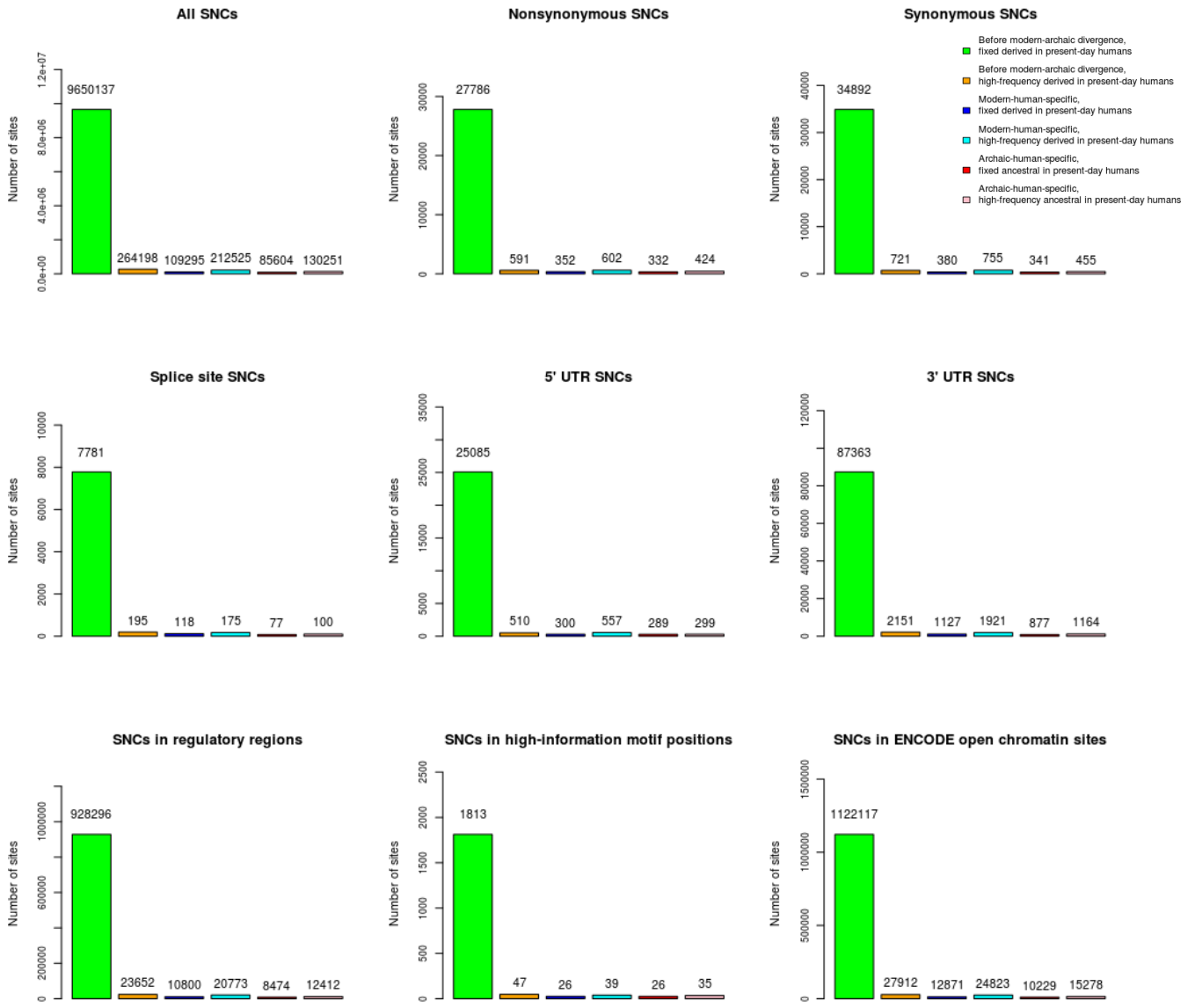


Figure S18.2. Number of single-nucleotide changes (SNCs) in each catalog (modern-human-specific, archaic-human-specific, before modern-archaic divergence). Modern-human-specific SNCs are SNCs where at least one archaic human (Denisova or Neandertal) has at least one ancestral allele, and the derived allele is fixed (blue) or at high-frequency (light blue) in present-day humans. Archaic-human-specific SNCs are SNCs where both archaic humans are homozygous for the same derived allele, and present-day humans have the ancestral allele globally fixed (red) or at high-frequency (pink). Before modern-archaic divergence SNCs are SNCs where both archaic humans are homozygous for the same derived allele and that allele is also fixed derived (green) or high-frequency derived (orange) in present-day humans. Regulatory regions and high-information motif positions are as defined by the Ensembl VEP, and ENCODE open chromatin sites were obtained from the ENCODE UCSC Open Chromatin track (Release 1) regions that were assigned some evidence for open chromatin (OC code = 1,2,3 or 4).

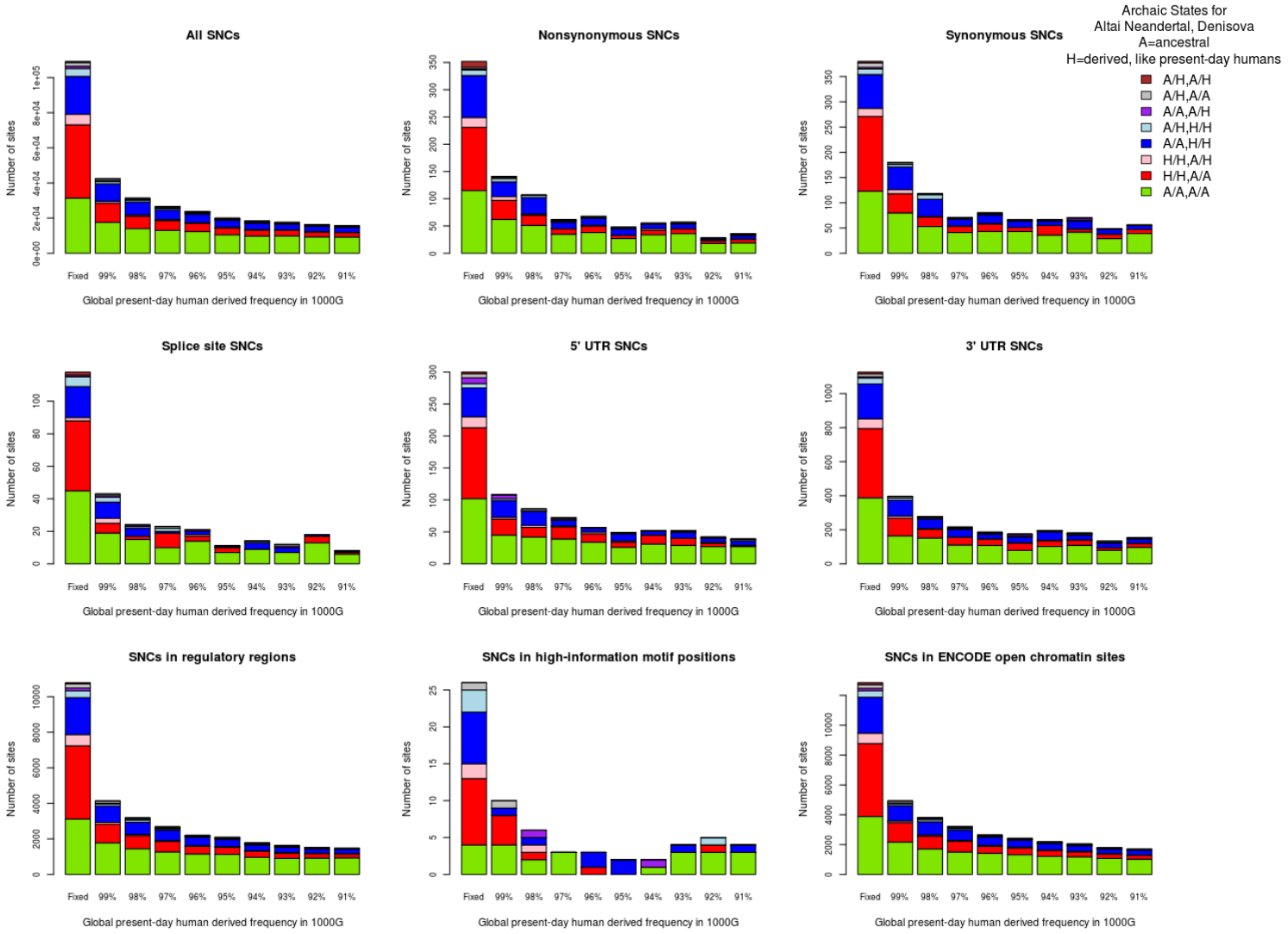


Figure S18.3. Present-day human derived single-nucleotide changes (SNCs) at different frequencies (91% to 100%) in sites that have at least one ancestral allele in either the Denisovan individual or the Altai Neandertal individual, partitioned by their genotype state in archaic humans, for different effect categories. Regulatory regions and high-information motif positions are as defined by the Ensembl VEP, and ENCODE open chromatin sites were obtained from the ENCODE UCSC Open Chromatin track (Release 1) regions that were assigned some evidence for open chromatin (OC code = 1,2,3 or 4). We excluded from these graphs all triallelic positions (where at least one of the archaic humans shows a second derived that is different from the present-day human derived site). Note the excess of Denisova

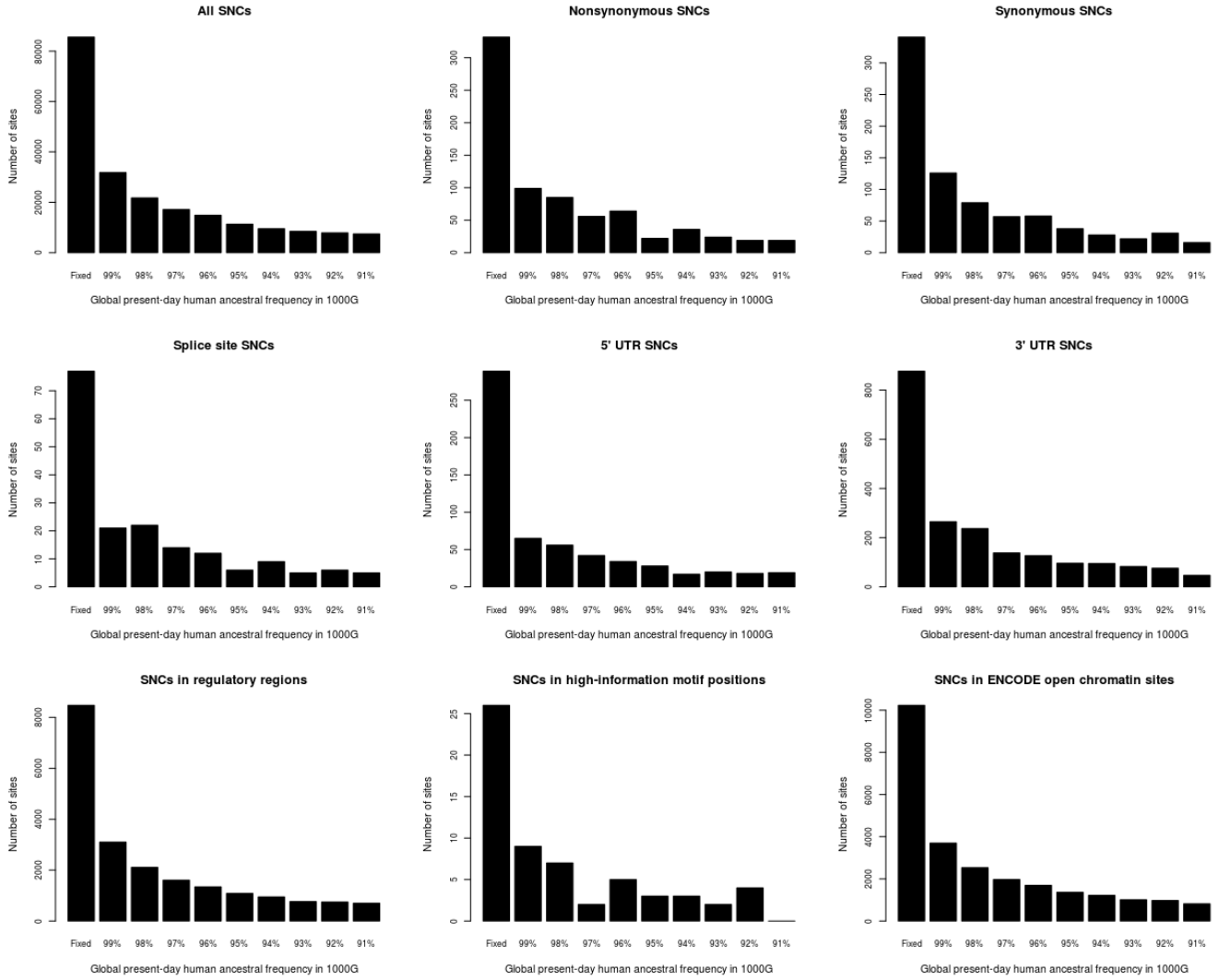
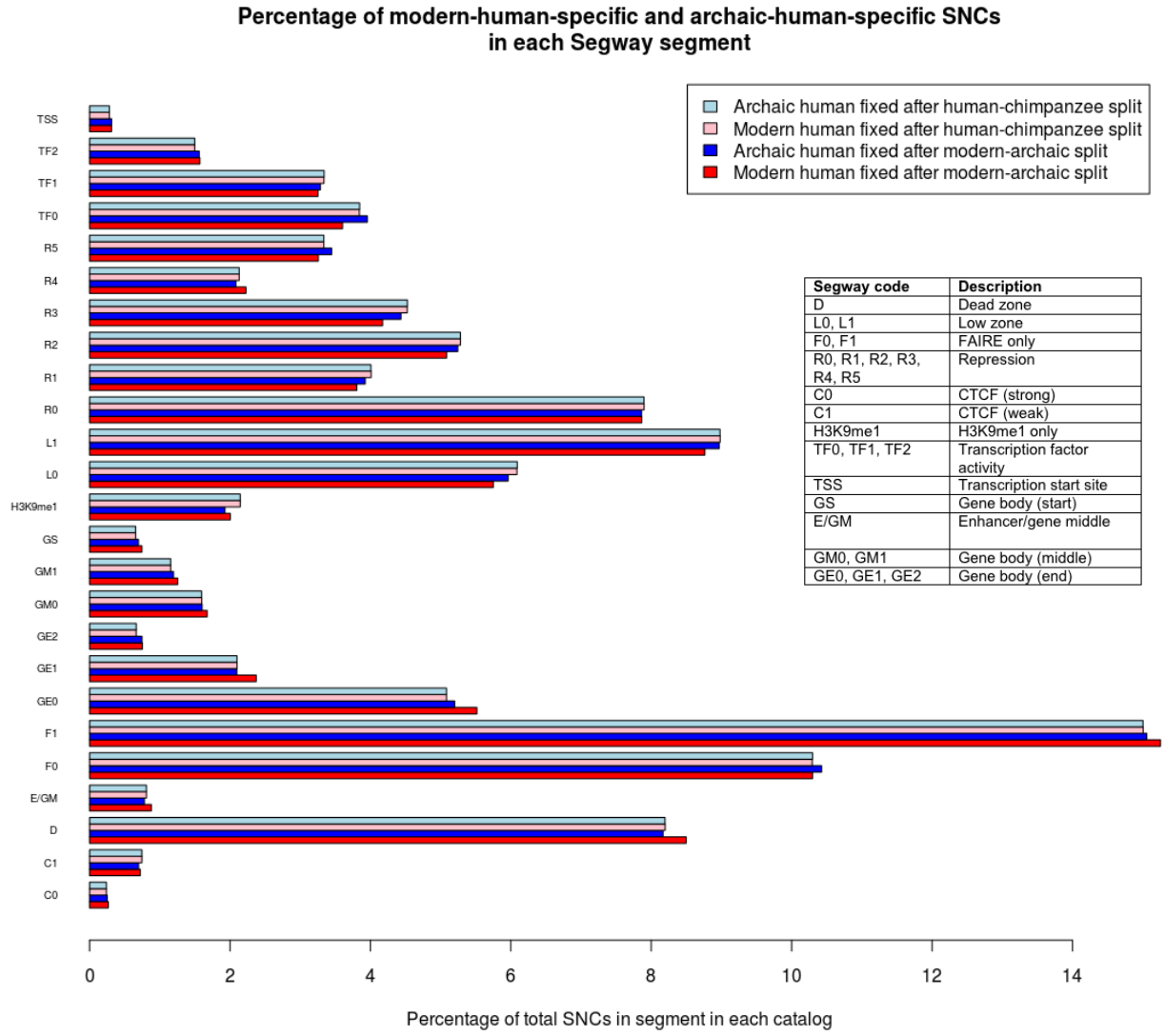


Figure S18.4. Archaic-human-specific (Altai Neandertal + Denisova homozygous derived) single-nucleotide changes (SNCs) in sites that are ancestral fixed or at high-frequency (91% to 100%) in present-day humans, for different effect categories. Regulatory regions and high-information motif positions are as defined by the Ensembl VEP, and ENCODE open chromatin sites were obtained from the ENCODE UCSC Open Chromatin track (Release 1) regions that were assigned some evidence for open chromatin (OC code = 1,2,3 or 4).



Segway code	Description
D	Dead zone
L0, L1	Low zone
F0, F1	FAIRE only
R0, R1, R2, R3, R4, R5	Repression
C0	CTCF (strong)
C1	CTCF (weak)
H3K9me1	H3K9me1 only
TF0, TF1, TF2	Transcription factor activity
TSS	Transcription start site
GS	Gene body (start)
E/GM	Enhancer/gene middle
GM0, GM1	Gene body (middle)
GE0, GE1, GE2	Gene body (end)

Figure S18.5. Percentage of SNCs in each Segway segment for each of four catalogs of changes having occurred in the human lineage. The upper table denotes the different types of catalogs. The lower tables shows the description corresponding to each Segway segment code.

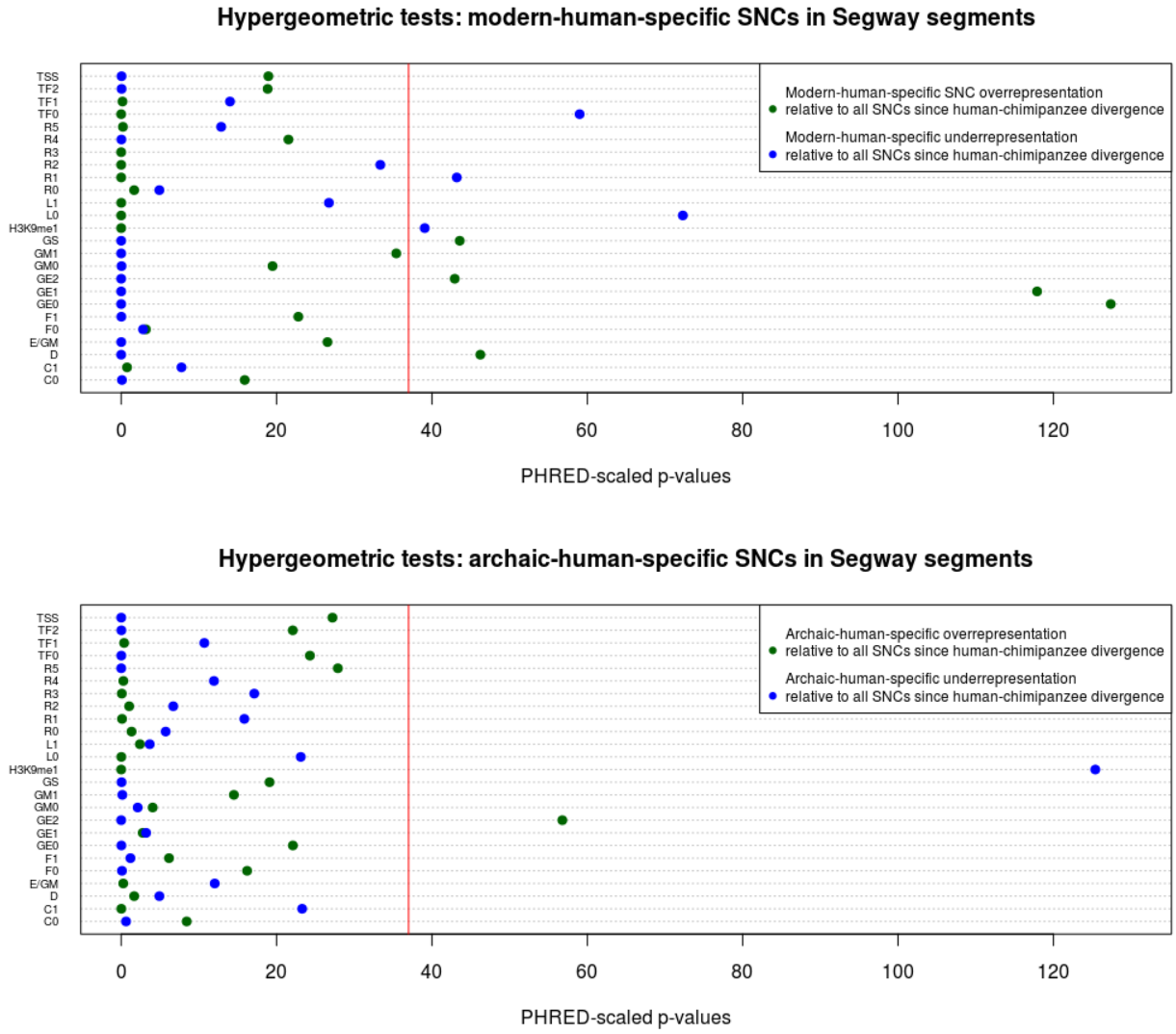


Figure S18.6. Hypergeometric tests to check for significant enrichments or depletions in particular Segway segments among A) modern-human-specific SNCs relative to all modern human SNCs that occurred since the human-chimpanzee split, and B) archaic-human-specific SNCs relative to all archaic human SNCs that occurred since the human-chimpanzee split. The red line denotes the Bonferroni-corrected 1% significance threshold.

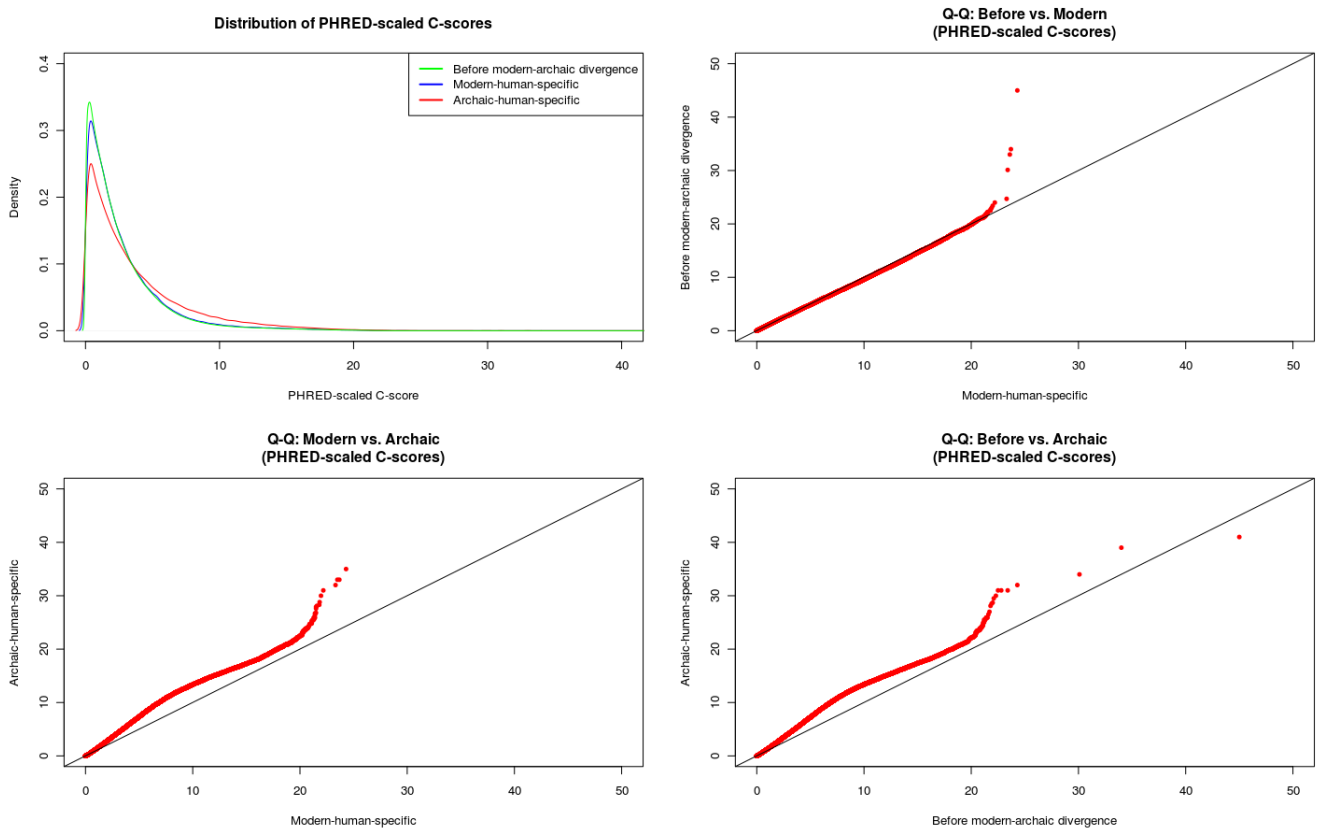


Figure S18.7. Distributions and pairwise Q-Q plots of PHRED-scaled C-scores for the modern-human-specific, archaic-human-specific and before modern-archaic divergence catalogs. We observe an archaic-specific excess of disruptive scores when compared to the modern-specific catalog (lower left panel) and the catalog of changes having occurred before the modern-archaic divergence (lower-right panel), which could be due to: a) an excess of fixed and high-frequency disruptive alleles in archaic humans due to their low effective population size, b) the fact that we only have two archaic humans to determine whether an allele is fixed or at high-frequency in all archaic humans or c) the fact that C-scores also aggregate annotations that include experimental data obtained from present-day humans.

Tables

Effect	Modern human state	Fixed derived				>90% derived, not fixed			
	Archaic human state	At least one archaic human has at least one ancestral allele	Both archaic humans are homozygous ancestral	Denisova is homozygous ancestral; Altai Neandertal is homozygous derived	Altai Neandertal is homozygous ancestral; Denisova is homozygous derived	At least one archaic human has at least one ancestral allele	Both archaic humans are homozygous ancestral	Denisova is homozygous ancestral; Altai Neandertal is homozygous derived	Altai Neandertal is homozygous ancestral; Denisova is homozygous derived
All SNCs		109,295	31,389	41,715	21,583	212,526	105,757	43,750	44,785
CCDS-verified genes	Missense	277	96	93	56	483	259	82	99
	Nonsense	0	0	0	0	3	2	1	0
	Synonymous	348	112	140	60	694	381	119	148
	Splice sites	87	32	32	15	141	83	22	20
	3' UTR	803	260	310	143	1,395	758	276	253
	5' UTR	130	47	46	20	267	131	56	60
All Ensembl genes	Missense	351	114	116	77	591	314	108	119
	Nonsense	1	1	0	0	11	6	3	1
	Synonymous	380	123	148	67	755	406	135	166
	Splice sites	118	45	43	19	175	100	28	28
	3' UTR	1,127	388	407	205	1,921	1,002	382	387
	5' UTR	300	102	111	45	557	300	112	102
Ensembl regulatory features	All	10,800	3,117	4,119	2,076	20,773	10,520	4,298	4,164
	Motif patterns	102	26	43	21	191	103	39	31
	High-information sites in motif patterns	26	4	9	7	39	19	7	8
Within mature miRNA		1	1	0	0	0	0	0	0
ENCODE open chromatin		12,871	3,891	4,882	2,415	24,823	12,653	5,086	5027

Table S18.1. Number of single-nucleotide changes in the modern-human-specific catalog, binned by their VEP-predicted effects. Genic changes with more than one effect were counted in the bin corresponding to the most severe effect as predicted by the Ensembl VEP. We also included an additional category for changes within ENCODE's open chromatin tracks.

Effect	Archaic human state	Both Denisova and Altai Neandertal are homozygous derived	
	Modern human state	Fixed ancestral	>90% ancestral, not fixed
All SNCs		85,604	130,252
CCDS-verified genes	Missense	274	332
	Nonsense	2	5
	Synonymous	318	422
	Splice sites	59	83
	3' UTR	606	872
	5' UTR	134	136
All Ensembl genes	Missense	327	417
	Nonsense	5	7
	Synonymous	341	455
	Splice sites	77	100
	3' UTR	877	1,164
	5' UTR	289	299
Ensembl regulatory features	All	8,474	12,412
	Motif patterns	115	129
	High-information sites in motif patterns	26	35
Within mature miRNA		1	0
ENCODE open chromatin		10,229	15,278

Table S18.2. Number of single-nucleotide changes in the archaic-human-specific catalog, binned by their VEP-predicted effects. Genic changes with more than one effect were counted in the bin corresponding to the most severe effect as predicted by the Ensembl VEP. We also included an additional category for changes within ENCODE's open chromatin tracks.

Effect	Modern human state / Archaic human state	Fixed derived				>90% derived, not fixed			
		At least one archaic human has at least one ancestral allele	Both archaic humans are homozygous ancestral	Denisova is homozygous ancestral; Altai Neandertal is homozygous derived	Altai Neandertal is homozygous ancestral; Denisova is homozygous derived	At least one archaic human has at least one ancestral allele	Both archaic humans are homozygous ancestral	Denisova is homozygous ancestral; Altai Neandertal is homozygous derived	Altai Neandertal is homozygous ancestral; Denisova is homozygous derived
All InDels		7,944	4,113	125	0	4,258	3,900	6	1
CCDS-verified genes	Frameshift	2	2	0	0	2	2	0	0
	Inframe	1	1	0	0	1	1	0	0
	Splice sites	9	2	0	0	2	2	0	0
	3' UTR	82	48	1	0	37	33	0	0
	5' UTR	11	9	0	0	5	5	0	0
All Ensembl genes	Frameshift	6	6	0	0	11	11	0	0
	Inframe	2	2	0	0	1	1	0	0
	Splice sites	12	3	0	0	3	3	0	0
	3' UTR	117	66	1	0	51	46	0	0
	5' UTR	20	16	0	0	9	9	0	0
Ensembl regulatory features	All	733	449	10	0	440	411	0	0
	Motif patterns	7	6	0	0	6	4	0	0
	High-information sites in motif patterns	2	2	0	0	2	0	0	0
Within mature miRNA		0	0	0	0	0	0	0	0

Table S18.3. Number of InDels in the modern-human-specific catalog, binned by their VEP-predicted effects. Genic changes with more than one effect were counted in the bin corresponding to the most severe effect as predicted by the Ensembl VEP.

Effect	Archaic human state / Modern human state	Both Denisova and Altai Neandertal are homozygous derived	
		Fixed ancestral	>90% ancestral, not fixed
All InDels		33,694	10,983
CCDS-verified genes	Frameshift	3	0
	Inframe	8	2
	Splice sites	37	18
	3' UTR	309	119
	5' UTR	20	9
All Ensembl genes	Frameshift	8	2
	Inframe	10	2
	Splice sites	44	20
	3' UTR	422	146
	5' UTR	51	23
Ensembl regulatory features	All	2,569	976
	Motif patterns	16	6
	High-information sites in motif patterns	2	0
Within mature miRNA		0	0

Table S18.4. Number of InDels in the archaic-human-specific catalog, binned by their VEP-predicted effects. Genic changes with more than one effect were counted in the bin corresponding to the most severe effect as predicted by the Ensembl VEP.

Modern-human-specific catalog		Both archaic humans are homozygous ancestral	
Ontology	**Effect**	**Modern human fixed derived**	**Modern human fixed and high-frequency derived**
Human Phenotype Ontology	**Nonsynonymous**	None	- Giant melanosomes in melanocytes (p-6.77e-6; FWER=0.091; FDR=0.091)
	Splice sites	None	None
	3' UTR	None	- 1-3 toe syndactyly (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - 1-5 toe syndactyly (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Aplasia/Hypoplasia of the distal phalanx of the thumb (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Bifid or hypoplastic epiglottis (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Central polydactyly (feet) (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Distal shortening of limbs (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Distal urethral duplication (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Dysplastic distal thumb phalanges with a central hole (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Hypothalamic hamartoma (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Laryngeal cleft (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Midline facial capillary hemangioma (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Preductal coarctation of the aorta (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Radial head subluxation (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Short distal phalanx of the thumb (p=1.34288e-05; FWER=0.538; FDR=0.0887928)
	5' UTR	None	None
Gene Ontology	**Nonsynonymous**	- Spindle microtubule (p=1.39e-5; FWER=0.061; FDR=0.061) - Attachment of spindle microtubules to kinetochore (p=2.29e-6; FWER=0.075; FDR=0.075)	None
	Splice sites	None	- Carbon-nitrogen ligase activity, with glutamine as amido-N-donor (p=2.44e-5; FWER=0.21; FDR=0.069)
	3' UTR	- L-cystine transmembrane transporter activity (p=12.88e-8; FWER=0.3; FDR=0.089)	None
	5' UTR	None	None
Disease Ontology	**Nonsynonymous**	None	None
	Splice sites	None	None
	3' UTR	None	None
	5' UTR	None	None

Table S18.5. Overrepresented terms ($p < 0.01$ and $FDR < 0.1$) in each effect category in the modern human lineage for 3 different ontologies using a binomial test in FUNC. FWER = family-wise error rate. FDR = false discovery rate.

Archaic-human-specific catalog		Both archaic humans are homozygous derived	
Ontology	**Effect**	**Modern human fixed ancestral**	**Modern human fixed and high-frequency ancestral**
Human Phenotype Ontology	**Nonsynonymous**	None	- Abnormality of the thumb (p=3.01e-5; FWER=0.025; FDR=0.02) - Aplasia/Hypoplasia of the thumb (p=6.31-5; FWER=0.054; FDR=0.024) - Facial cleft (p=0.0004; FWER=0.36; FDR=0.098) - Wide pubic symphysis (p=0.0004; FWER=0.36; FDR=0.098) - Abnormality of the frontal hairline (p=0.00042; FWER=0.39; FDR=0.096) - Abnormality of the tongue (p=0.00045; FWER=0.4; FDR=0.09) - Abnormality of the scalp hair (p=0.00045; FWER=0.42; FDR=0.084) - Abnormality of the scalp (p=0.00049; FWER=0.42; FDR=0.084) - Abnormality of the finger (p=0.0005; FWER=0.44; FDR=0.08) - Brachydactyly syndrome (p=0.00062; FWER=0.48; FDR=0.088)
	Splice sites	None	None
	3' UTR	None	None
	5' UTR	None	None
Gene Ontology	**Non-synonymous**	None	None
	Splice sites	None	None
	3' UTR	None	- Mitochondrial intermembrane space protein transporter complex (p=2.34e-9; FWER=0.083; FDR=0.083)
	5' UTR	None	None
Disease Ontology	**Nonsynonymous**	None	None
	Splice sites	None	None
	3' UTR	None	None
	5' UTR	None	None

Table S18.6. Overrepresented terms ($p < 0.01$ and FDR < 0.1) in each effect category in the archaic human lineage for 3 different ontologies using a binomial test in FUNC. FWER = family-wise error rate. FDR = false discovery rate.

	Position	dbSNP	Ancestral / derived alleles	Archaic state: Altai, Denisova (A=ancestral, D=derived)	1000G derived frequency	Risk allele in positive strand	Gene	Effect	SNP disease association (OMIM)
Modern-human-specific	chr8:18080001	rs4987076	A/G	A/A,G/G	98%	A (ancestral)	*NAT1*	missense	Affects acetylation activity of a gene involved in drug susceptibility (Doll et al. 1997). Gene is known to have undergone positive selection in recent human history (Patin et al. 2006).
	chr8:21976710	rs7014851	C/T	C/C,C/C	92%	C (ancestral)	*HR*	missense	May contribute to alopecia universalis congenita (Ahmad et al. 1998).
	chr9:34649442	rs2070074	G/A	G/G,A/A	95%	G (ancestral)	*GALT*	missense	Ancestral variant causes reduced activity of a gene associated with galactosemia (Elsas et al. 1994).
	chr17:42338945	rs5036	C/T	T/T,C/C	93%	C (ancestral)	*SLC4A1*	missense, splice site	Abnormalities of erythrocyte shape (Ranney et al. 1990, Wilder et al. 2009).
	chr14:61924239	rs2230500	A/G	A/A,A/A	94%	A (ancestral)	*PRKCH*	missense	Susceptibility to cerebral infarction (Kubo et al. 2007).
	chr11:104763117	rs497116	G/A	G/G,G/G	96%	G (ancestral)	*CASP12*	STOP gained	Response to bacterial infection (Saleh et al. 2004, Shimoke et al. 2011). Known to have undergone positive selection before the Neolithic (Kachapati et al. 2006, Hervella et al. 2012).
	chr11:59612859	rs35211634	C/T	C/C,C/C	92%	C (ancestral)	*GIF*	missense	Serves as marker for inheritance of susceptibility of congenital intrinsic factor deficiency (Gordon et al. 2004).
Archaic-human-specific	chr19:15291576	rs35769976	C/G	G/G,G/G	8%	G (archaic derived)	*NOTCH3*	missense, regulatory feature	May be associated with cerebral arteriopathy (Scheid et al. 2008) but see Quattrone and Mazzei (2009)

Table S18.7. Clinically pathogenic variants retrieved from ClinVar for the modern human and archaic human catalogs. We note that all variants are polymorphic in modern humans, as expected.

Position (hg19)	Ancestral / derived alleles	Present-day human derived frequency	dbSNP ID	Altai Nea / Denisova state (A=ancestral, H=like modern human major derived allele)	PHRED-scaled C-score	GERP rejected substitution score	Consequence	Gene
chr1:7980985	T-/TA	fixed	-	H/H,A/H	32	6.17	CANONICAL SPLICE SITE	*TNFRSF9*
chr5:135513083	*A-/AT*	*fixed*	-	*A,A/A/A*	*31*	*5.83*	*FRAMESHIFT*	*SMAD5*
chr22:19189003	AC/A-	fixed	-	A,A/A/A	31	2.29	FRAMESHIFT	*CLTCL1*
chr2:107429680	T/C	fixed*	rs140361370	H/H,A/A	24.3	3.77	NONSYNONYMOUS	*ST6GAL2*
chr2:24402127	CAA/C---	fixed	-	A,A/A/A	24.1	4.75	FRAMESHIFT	*C2orf84*
chr11:134019534	G/A	fixed	-	A/A,H/H	23.7	-1.08	NONSYNONYMOUS	*JAM3*
chr15:75117900	C/T	fixed	-	H/H,A/A	23.6	2.68	NONSYNONYMOUS	*LMAN1L*
chr13:103527849	G/C	fixed*	rs9514066	H/H,A/H	23.4	4.28	NONSYNONYMOUS	*ERCC5*
chr1:228555619	T/A	fixed	-	H/H,A/A	23.3	3.86	NONSYNONYMOUS	*OBSCN*
chr1:228699943	G/A	fixed*	rs185286334	A/A,H/H	22.2	2.64	NONSYNONYMOUS	*BTNL10*
chr7:146825878	A/G	fixed	-	H/H,A/A	22	2.18	NONSYNONYMOUS	*CNTNAP2*
chr21:34169317	CT/C-	fixed	-	A,A/A/A	22	1.85	FRAMESHIFT	*C21orf49*
chr1:245582905	G/A	fixed	-	A/A,A/A	21.9	3.71	NONSYNONYMOUS	*KIF26B*
chr10:105437845	C/G	fixed	-	A/A,H/H	21.8	-0.08	NONSYNONYMOUS	*SH3PXD2A*
chr9:35706519	T/G	fixed*	-	A/A,A/A	21.8	-1.78	SPLICE SITE	*TLN1*
chr9:135930371	C/G	fixed*	rs191292694	A/A,A/A	21.6	-1.58	NONSYNONYMOUS	*GTF3C5*
chr8:8869129	G/T	fixed*	rs112570397	A/A,A/A	21.5	1.69	NONSYNONYMOUS	*ERI1*
chr17:79668341	A/G	fixed	-	A/A,H/H	21.5	1.65	NONSYNONYMOUS	*HGS*
chr6:149918766	C/T	fixed*	rs73781249	A/A,A/A	21.5	5.48	UPSTREAM	*RP1-12G14.7*
chr10:104474107	G/A	fixed*	rs2248679	A/A,A/A	21.5	5.47	UPSTREAM	*SFXN2*
chr5:60866462	G/A	fixed	-	A/A,H/H	21.4	5.32	INTERGENIC	N/A
chr16:79245573	T/C	fixed	-	A/A,H/H	21.4	-8.54	NONSYNONYMOUS	*WWOX*
chr8:92916612	C/T	fixed	-	A/A,H/H	21.4	6.06	INTRONIC	*RP11-122C21.1*
chr10:112660260	G/A	fixed	-	A/A,A/A	21.4	5.3	DOWNSTREAM	*PDCD4*
chr5:102512570	CA/C-	fixed	-	A/A,A/A	21.4	2.23	FRAMESHIFT	*PPIP5K2*
chr2:55795378	G/A	fixed	-	H/H,A/A	21.3	4.6	DOWNSTREAM	*SNORA12*
chr5:75591644	A/C	fixed	-	A/A,A/A	21.3	5.87	INTRONIC	*RP11-466P24.6*
chr17:55553444	C/T	fixed	-	H/H,A/H	21.3	5.88	INTRONIC	*MSI2*
chr16:24199400	G/A	fixed*	rs145337227	H/H,A/A	21.3	5.15	INTRONIC	*PRKCB*
chr2:60534334	G/T	fixed	-	H/H,A/A	21.3	5.29	INTERGENIC	N/A

Table S18.8. Top 30 fixed modern-human-specific SNCs and InDels, ranked by their corresponding PHRED-scaled C-scores. InDels nearby other InDels (+/- 5bp) are shown in cursive, as they could be the result of mismapping in the 1000G data, and so should be treated with caution.

Position (hg19)	Ancestral / derived alleles	Present-day human derived frequency	dbSNP ID	Altai Nea / Denisova state (A=ancestral, H=like modern human major derived allele)	PHRED-scaled C-score	GERP rejected substitution score	Consequence	Gene
chr15:62932556	C/G	96%	rs35757182	A/H,A/A	42	1.78	STOP LOST	*RP11-625H11.1*
chr2:198593260	C/A	99%	rs74375706	H/H,A/A	39	4.76	STOP LOST	*BOLL*
chr6:154360569	T/C	97%	rs17174638	A/A,A/A	39	-3.43	STOP LOST	*OPRM1*
chr11:104763117	G/A	96%	rs497116	A/A,A/A	36	-5.99	STOP GAINED	*CASP12*
chr12:57003964	A/T	96%	rs2230580	A/A,A/A	34	-2.82	STOP LOST	*BAZ2A*
chr2:27551325	A/G	93%	rs6721927	A/A,A/A	33	4.07	STOP LOST	*GTF3C2*
chr14:74763086	C/G	94%	rs36031534	H/H,A/A	33	0.67	STOP LOST (NMD TRANSCRIPT)	*ABCD4*
chr13:97639827	G/T	95%	rs9300380	A/A,A/A	33	3.34	UNKNOWN	*OXGR1*
chr12:40815007	G/C	96%	rs80212515	H/H,A/A	32	5.11	CANNONICAL SPLICE SITE	*MUC19*
chr11:111853105	*A-/AG*	*95%*	*rs200882091*	*A/A,A/A*	*31*	*6.17*	*UNKNOWN*	*DIXDC1*
chr9:34648088	A/G	96%	rs41274867	A/A,H/H	31	4.47	CANNONICAL SPLICE SITE	*GALT*
chr8:48805814	*G-/GA*	*95%*	-	*A/A,A/A*	*31*	*2.78*	*FRAMESHIFT*	*PRKDC*
chr19:50879835	C/T	98%	rs73932483	A/A,A/A	31	3.7	CANNONICAL SPLICE SITE	*NR1H2*
chr3:183353583	C/T	92%	rs112664695	A/A,A/A	28.8	3.47	CANNONICAL SPLICE SITE	*KLHL24*
chr6:20212410	C/T	94%	rs61737148	A/A,H/H	28.1	3.29	CANNONICAL SPLICE SITE	*RP11-239H6*
chr10:11789382	G/A	93%	rs4750090	H/H,A/A	27.2	5.53	NONSYNONYMOUS	*ECHDC3*
chr22:45810275	*G-/GA*	*98%*	*rs202120654*	*A/A,A/A*	*26.8*	*4.99*	*FRAMESHIFT*	*RIBC2*
chr14:50298962	T/A	93%	rs3100906	A/A,A/A	26.6	5.33	NONSYNONYMOUS	*NEMF*
chr9:6328947	T/C	94%	rs3847262	A/A,A/A	26.2	4.41	NONSYNONYMOUS	*TPD52L3*
chr19:2340153	*T-/TG*	*97%*	-	*A/A,A/A*	*25.7*	*4.32*	*FRAMESHIFT*	*SPPL2B*
chr20:590541	*A-/AG*	*91%*	*rs71212728*	*A/A,A/A*	*25.2*	*4.35*	*FRAMESHIFT*	*TCF15*
chr11:104761921	C/T	99%	rs115100183	H/H,A/A	24.7	0.22	NONSYNONYMOUS	*CASP12*
chr17:43318777	*G-/GC*	*96%*	-	*A/A,A/A*	*24.2*	*3.09*	*FRAMESHIFT*	*FMNL1*
chr12:85277615	G/A	96%	rs79063785	A/A,A/A	24.2	-2.67	NONSYNONYMOUS	*SLC6A15*
chr5:65118738	T/C	93%	rs6860508	A/A,A/A	24	-0.75	NONSYNONYMOUS	*NLN*
chr2:128381861	A/G	99%	rs61743523	A/A,H/H	23.9	-0.06	NONSYNONYMOUS	*MYO7B*
chr16:48149467	G/A	94%	rs9302750	A/A,A/A	23.7	-6.61	NONSYNONYMOUS	*ABCC12*
chr6:108076801	C/T	98%	rs117914882	A/A,H/H	23.7	-1.02	NONSYNONYMOUS	*SCML4*
chr12:72050332	A/G	98%	rs75740654	A/A,A/A	23.3	-2.59	NONSYNONYMOUS	*ZFC3H1*
chr1:171178090	C/T	96%	rs6661174	H/H,A/A	23.2	5.99	STOP GAINED	*FMO2*

Table S18.9. Top 30 high-frequency modern-human-specific SNCs and InDels, ranked by their corresponding PHRED-scaled C-scores. InDels nearby other InDels are shown in cursive, as they could be the result of mismapping in the 1000G data, and so should be treated with caution.

Position (hg19)	Ancestral / derived alleles	Present-day human ancestral frequency	dbSNP ID	Altai Nea / Denisova state (N=derived, different from present-day human major allele)	PHRED-scaled C-score	GERP rejected substitution score	Consequence	Gene
chr17:37034365	C/T	fixed*	rs141320621	N/N,N/N	35	-0.224	STOP GAINED	*LASP1*
chr19:55898080	G/A	fixed	-	N/N,N/N	33	1.2	STOP GAINED	*RPL28*
chr3:97806515	C/T	fixed	-	N/N,N/N	33	1.27	STOP GAINED	*OR5AC2*
chr12:6422334	TG/T-	fixed	-	N/N,N/N	33	-4.28	UNKNOWN	*PLEKHG6*
chr7:122338819	CAT/C--	fixed	-	N/N,N/N	33	4.25	UNKNOWN	*RNF133*
chr7:133985015	T-/TA	fixed	-	N/N,N/N	33	5.78	CANONICAL SPLICE SITE	*SLC35B4*
chr8:63978658	TA/T-	fixed	-	N/N,N/N	32	5.54	CANONICAL SPLICE SITE	*TTPA*
chr13:38320594	TA/T-	fixed	-	N/N,N/N	32	5.7	CANONICAL SPLICE SITE	*TRPC4*
chr18:74208485	CG/C-	fixed	-	N/N,N/N	32	0.666	UNKNOWN	*RP11-17M16.1*
chr2:183806893	TA/T-	fixed	-	N/N,N/N	32	4.82	CANONICAL SPLICE SITE	*NCKAP1*
chr6:49494627	GC/G-	fixed	-	N/N,N/N	32	2.25	UNKNOWN	*GLYATL3*
chr12:16377347	C/T	fixed*	rs117974895	N/N,N/N	32	4.7	CANONICAL SPLICE SITE	*SLC15A5*
chr5:13868103	TA/T-	fixed	-	N/N,N/N	32	5.12	CANONICAL SPLICE SITE	*DNAH5*
chr11:46342259	A-/AG	fixed	-	N/N,N/N	32	4.08	CANONICAL SPLICE SITE	*CREB3L1*
chr1:156354347	TC/T-	fixed	-	N/N,N/N	32	4.03	UNKNOWN	*RHBG*
chr12:68552041	TA/T-	fixed	-	N/N,N/N	31	5.2	CANONICAL SPLICE SITE	*IFNG*
chr7:104946931	CA/C-	fixed	-	N/N,N/N	31	2.66	UNKNOWN	*SRPK2*
chr11:111853106	G-/GC	fixed	-	N/N,N/N	31	6.17	UNKNOWN	*DIXDC1*
chr10:126673560	G-/GA	fixed	-	N/N,N/N	31	4.22	FRAMESHIFT	*ZRANB1*
chr13:100517195	CTG/C--	fixed	-	N/N,N/N	31	4.36	UNKNOWN	*CLYBL*
chr7:11676519	G/C	fixed	-	N/N,N/N	31	6.02	NONSYNONYMOUS	*THSD7A*
chr17:56692633	G/A	fixed	-	N/N,N/N	30	0.971	NONSYNONYMOUS	*TEX14*
chr11:126432775	C/T	fixed*	rs182260035	N/N,N/N	28.8	5.62	NONSYNONYMOUS	*KIRREL3*
chr8:88365925	T/A	fixed*	rs186699266	N/N,N/N	28.3	4.98	NONSYNONYMOUS	*CNBD1*
chr11:123994464	C/T	fixed	-	N/N,N/N	28.1	4.72	NONSYNONYMOUS	*VWA5A*
chr12:16347327	G/A	fixed*	rs117728539	N/N,N/N	27.9	4.46	NONSYNONYMOUS	*SLC15A5*
chr4:186815528	C/T	fixed	-	N/N,N/N	27.6	3.54	CANONICAL SPLICE SITE	*SORBS2*
chr22:20779973	C-/CG	fixed	-	N/N,N/N	27.5	3.01	FRAMESHIFT	*SCARF2*
chr2:26705331	C/T	fixed	-	N/N,N/N	26.8	3.36	NONSYNONYMOUS	*OTOF*

Table S18.10. Top 30 fixed archaic-human-specific SNCs and InDels, ranked by their corresponding PHRED-scaled C-scores.

Position (hg19)	Ancestral / derived alleles	Present-day human ancestral frequency	dbSNP ID	Altai Nea / Denisova state (N=derived, different from present-day human major allele)	PHRED-scaled C-score	GERP rejected substitution score	Consequence	Gene
chr9:21481483	G/A	94%	rs2039381	N/N,N/N	41	3.84	STOP GAINED	*IFNE*
chr6:25969631	C/T	97%	rs76757832	N/N,N/N	39	3.95	STOP GAINED	*TRIM38*
chr4:77296802	C/T	96%	rs7686674	N/N,N/N	36	4.07	STOP GAINED	*CCDC158*
chr9:4626449	C/T	99%	rs28548276	N/N,N/N	34	1.88	STOP GAINED	*SPATA6L*
chr20:44511257	G/A	99%	rs35972756	N/N,N/N	33	0.395	STOP GAINED	*ZSWIM1*
chr11:56310356	A/T	96%	rs17547284	N/N,N/N	33	-0.628	STOP GAINED	*OR5M11*
chr22:32643460	C/A	99%	rs62239058	N/N,N/N	33	0.571	STOP GAINED	*SLC5A4*
chr11:62569105	T--/TAC	91%	-	N/N,N/N	32	5.46	CANONICAL SPLICE SITE	*NXF1*
chr7:126890902	C/T	96%	rs17866749	N/N,N/N	32	5.68	CANONICAL SPLICE SITE	*GRM8*
chr5:6377364	T-/TG	94%	rs60432194	N/N,N/N	32	5.25	CANONICAL SPLICE SITE	*MED10*
chr2:218577813	AC/A-	91%	-	N/N,N/N	32	1.76	UNKNOWN	*DIRC3*
chr1:248129415	TTG/T--	99%	-	N/N,N/N	31	2.08	UNKNOWN	*OR2AK2*
chr16:89829196	G/C	92%	rs12598276	N/N,N/N	31	0.122	STOP GAINED	*FANCA*
chr17:45447802	G/A	94%	rs76299620	N/N,N/N	31	3.86	CANONICAL SPLICE SITE	*C17orf57*
chr15:49926993	TA/T-	92%	rs111446752	N/N,N/N	31	2.89	CANONICAL SPLICE SITE	*DTWD1*
chr13:39422624	C/T	91%	rs9548505	N/N,N/N	31	5.66	NONSYNONYMOUS	*FREM2*
chr1:169799880	G/T	95%	rs2272920	N/N,N/N	31	6.05	NONSYNONYMOUS	*C1orf112*
chr1:186134246	A/T	92%	rs41317507	N/N,N/N	30	5.27	NONSYNONYMOUS	*HMCN1*
chr18:67721492	G/C	99%	rs34717557	N/N,N/N	30	4.44	NONSYNONYMOUS	*RTTN*
chr11:6503335	G/T	99%	rs60702727	N/N,N/N	30	5.83	NONSYNONYMOUS	*FXC1*
chr20:19956311	G/A	99%	rs181298473	N/N,N/N	30	4.88	NONSYNONYMOUS	*RIN2*
chr7:42004062	G/A	98%	rs35364414	N/N,N/N	29.9	6.03	NONSYNONYMOUS	*GLI3*
chr4:155241735	C/T	97%	rs79535970	N/N,N/N	29.6	5.59	NONSYNONYMOUS	*DCHS2*
chr15:101551007	C/A	99%	rs55739947	N/N,N/N	29.5	2.05	NONSYNONYMOUS	*LRRK1*
chr1:46874246	C/T	99%	rs77101686	N/N,N/N	28.8	5.4	NONSYNONYMOUS	*FAAH*
chrX:117732044	A/T	95%	rs16995229	N/N,N/N	28.7	4.36	NONSYNONYMOUS	*DOCK11*
chr19:7437765	A/G	98%	rs78885217	N/N,N/N	28.5	2.05	NONSYNONYMOUS	*CTB-133G6.1*
chr5:1232408	G/A	96%	rs113861454	N/N,N/N	28.5	5.45	NONSYNONYMOUS	*SLC6A18*
chr22:32628900	C/T	99%	rs62239049	N/N,N/N	28.4	4.74	NONSYNONYMOUS	*SLC5A4*
chr7:102448819	T/G	99%	rs79250295	N/N,N/N	28.1	4.84	NONSYNONYMOUS	*FAM185A*

Table S18.11. Top 30 high-frequency archaic-human-specific SNCs and InDels, ranked by their corresponding PHRED-scaled C-scores.

C-score Wilcoxon rank-sum test: modern-human-specific catalog (p < 0.01)	**Both archaic humans are homozygous ancestral**			
Ontology	**Modern human fixed derived**		**Modern human fixed and high-frequency derived**	
Ranking method	**METHOD A**	**METHOD B**	**METHOD A**	**METHOD B**
Gene Ontology (FWER <= 0.01)	- Positive regulation of biological process (p=1.2e-5; FWER=0.01; FDR=0.0062)	None	- Sequence-specific DNA binding (p=5.58e-6; FWER=0.02; FDR=0.00067) - Protein binding (p=7.98e-6; FWER=0.003; FDR=0.00075) - Contractile fiber part (p=9.85e-6; FWER=0.003; FDR=0.0017) - Cellular developmental process (p=1.93e-6; FWER=0.004; FDR=0.00083) - Myofibril (p=4.33e-5; FWER=0.005; FDR=0.0018) - Sarcomere (p=3.86e-5; FWER=0.005; FDR=0.0022) - Signaling (p=3.51e-6; FWER=0.007; FDR=0.001) - Regulation of multicellular organismal process (p=3.1e-6; FWER=0.007; FDR=0.0011) - Contractile fiber (p=7.02e-5; FWER=0.01; FDR=0.0027)	None
Human Phenotype Ontology (FWER <= 0.5)	None	None	None	None
Disease Ontology (FWER <= 0.5)	None	None	- Abdominal aortic aneurysm (p=0.00087; FWER=0.21; FDR=0.21) - Congestive heart failure (p=0.0009; FWER=0.22; FDR=0.22)	- Abdominal aortic aneurysm (p=0.0004; FWER=0.103; FDR=0.103)

Table S18.12. Terms with significantly high-ranking C-scores in the modern-human-specific catalog (Wilcoxon rank-sum test) for 3 different ontologies (see main text for description of methods). FWER = family-wise error rate. FDR = false discovery rate.

C-score Wilcoxon rank-sum test: archaic-human-specific catalog (p < 0.01)	**Both archaic humans are homozygous derived**			
Ontology	**Modern human fixed ancestral**		**Modern human fixed and high-frequency ancestral**	
Ranking method	**METHOD A**	**METHOD B**	**METHOD A**	**METHOD B**
Gene Ontology (FWER <= 0.01)	- Sequence-specific DNA binding transcription factor activity (p=2.7e-5; FWER=0.007; FDR=0.0029) - Nucleic acid binding transcription factor activity (p=2.62e-5; FWER=0.007; FDR=0.004)	None	None	- Immunoglobulin-mediated immune response (p=7.43e-3; FWER=0.009; FDR=0.0053)
Human Phenotype Ontology (FWER <= 0.5)	- Mode of inheritance (p=0.00026; FWER=0.1; FDR=0.1) - Psychosis (p=0.00044; FWER=0.2; FDR=0.2)	None	- Mode of inheritance (p=0.00026; FWER=0.15; FDR=0.15)	None
Disease Ontology (FWER <= 0.5)	None	- Pre-eclampsia (p=0.00014; FWER=0.027; FDR=0.016) - Heart valve disease (p=0.0017; FWER=0.34; FDR=0.34)	- Lung small cell carcinoma (p=0.002; FWER=0.46; FDR=0.46)	- Pre-eclampsia (p=0.00013; FWER=0.028; FDR=0.017) - Nervous system cancer (p=0.0014; FWER=0.33; FDR=0.33) - Uveal disease (p=0.0016; FWER=0.37; FDR=0.37) - Central nervous system cancer (p=0.0018; FWER=0.41; FDR=0.41)

Table S18.13. Terms with significantly high-ranking C-scores in the archaic-human-specific catalog (Wilcoxon rank-sum test) for 3 different ontologies (see main text for description of methods). FWER = family-wise error rate. FDR = false discovery rate.

Supplementary Information 19a

Selective Sweeps on the Human Lineage

Kay Prüfer*, Christoph Theunert, Matthias Ongyerth, Gabriel Renaud, Michael Dannemann and Michael Lachmann*

* To whom correspondence should be addressed (pruefer@eva.mpg.de, lachmann@eva.mpg.de)

We used the Neandertal and Denisova genome sequence to identify long regions where Yoruban and Luhya genomes from the 1000 genomes project show a surprisingly recent coalescence by testing whether Neandertal or Denisova fall outside of the human variation. The identified regions were scored by their genetic length and the informative sites that led to the classification as "external region". We observed that regions that score high in Neandertal and Denisova overlapped more often than expected at random. Based on the excess of overlap, we define a cutoff on the score to limit further analyses to regions that likely underwent positive selection on the human lineage.

Method

We use a hidden markov model (HMM) to identify regions where an outgroup (Neandertal or Denisova) falls outside of the variation of modern humans (Fig. S19a.1). These are genomic regions where the time to most recent common ancestor (tMRCA) of all humans is more recent than the tMRCA of all humans plus the outgroup. The HMM has two hidden states, *internal* and *external*, corresponding to positions where the outgroup falls within and outside of the human variation. The *internal* and *external* states are inferred from observing whether the outgroup carryies the *derived* or the *ancestral* allele at a polymorphic position in modern humans, respectively.

Given a tree where the outgroup falls external to all humans, we expect that the outgroup will not carry the derived variant. (We give a small probability of 0.01 to accommodate a small fraction of sequencing error or misassignment of the ancestral state). However, if the outgroup is *internal*, the probabilities of observing *ancestral* or *derived* depend on the age of the derived variant: even when falling internal, a very young variant is not expected to be shared by the outgroup, while older variants are more likely to be shared. The frequency of the derived variant within humans can serve as a proxy for the age of a SNP, since a longer time is needed for a neutral variant to rise to higher frequency. Due to this, our emission probabilities for *ancestral* and *derived* are dependent on the allele frequency when the hidden state is *internal*.

The transitions between *internal* and *external* regions depend on the genetic distance between neighboring SNPs, since the local phylogeny is broken by recombination. In our model, the genetic lengths of internal and external regions follow an exponential distribution and the transition probabilities correspond to this distribution.

A very similar HMM was used for detecting sweeps in chimpanzees using bonobo as an outgroup[1]. A

more detailed description of the model can be found in the supplementaries accompanying that publication.

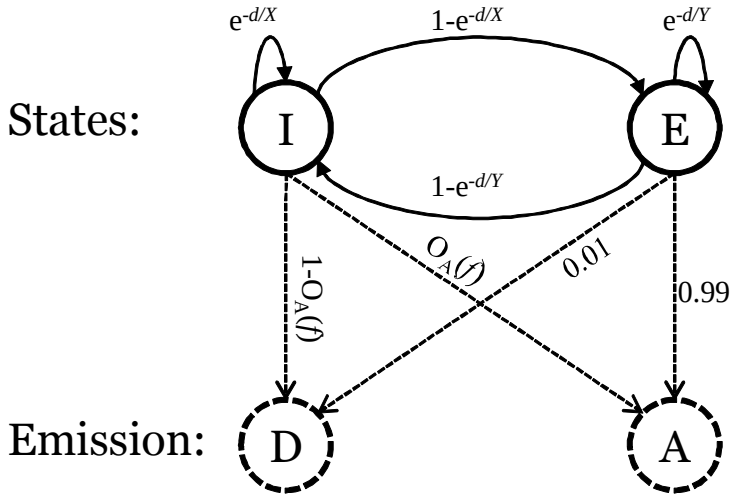


Figure S19a.1: Schematic description of the hidden Markov model used to detect regions where an outgroup (Altai Neandertal or Denisova) falls outside the variation of modern humans. Hidden states are shown as solid circles (I for *internal*; E for *external*), dashed circles give the emissions (D for outgroup *derived*; A for outgroup *ancestral*). Transition probabilities have parameters *X* and *Y* for average length of *internal* and *external* regions, respectively, and *d* for the distance to the previous SNP position. $O_A(f)$ gives the probability of emitting *ancestral* given the derived allele frequency *f* in modern humans.

Datasets and Input Preparation

Denisova and Neandertal VCFs were prepared as described earlier (SI 3). Positions in these VCFs were filtered according to our set of minimal filters (map35_100%, mapping quality, coverage, tandem repeats; see also SI 5b) and by removing LOW_QUAL genotype calls. Only sites that passed filters in both Denisova and Altai Neandertal were considered further.

Additionally, phased and imputed low-coverage genome data of 185 Luhya and Yoruba individuals from the 1000 genomes phase I[2] were used. Both datasets were combined by tabulating the frequency of the derived variant in the 1000 genomes individuals for each SNP and extracting the corresponding Neandertal and Denisova allele. If Neandertal or Denisova were polymorphic, one allele was chosen at random. The chimpanzee-human common ancestor inferred from EPO alignment (prepared for the 1000 Genomes project (phase I)) was used to assign ancestral and derived states for SNPs.

Genetic distances between the extracted SNPs were added using an African American recombination map[3]. Over all autosomes, 11.9 million SNPs pass the filters.

Estimating the Expected Size of External and Internal Regions and Emission Probabilities

We run coalescent simulations using *ms*[4] for one outgroup and 370 chromosomes, corresponding to the 185 Luhya and Yoruba individuals. The simulation assumed a generation time of 28 years, a per generation mutation rate of 1e-8, a uniform recombination rate of 1Mb/cM, an effective population size of 10,000 for the outgroup, an expansion from 10,000 to 17,000 individuals in Luhya and Yoruba over the last 200,000 years, and a time to the most recent common ancestor with Neandertals of 375,000 years. Simulations with

these parameters give a good match to our observed data when plotting derived allele frequency in modern humans against frequency of derived observed in the outgroup (see Fig. S19a.2).

Based on the simulations, we then estimate several parameters for the HMM. First, we classify simulated regions in external and internal to estimate the expected length of regions. We find that the average external region is expected to stretch 3.5 kilobases and the average internal region 26kb under a genome average recombination rate. These values give the lengths under neutral expectation and are used as parameters *X* and *Y* in Figure S19a.1. Second, we tabulate how often our simulated outgroup falls internal or external (according to the simulated tree) at SNP positions with a certain derived allele frequency in our simulation (using a best fit line to the simulations to reduce noise). These values are used to subtract the fraction of ancestral calls from our ancestral/derived counts per allele frequency observed in real data. The values calculated this way give an estimate of the probability with which ancestral or derived allele is observed at a given derived allele frequency when being in an internal state (Parameter $O_A(f)$ in Figure S19a.1).

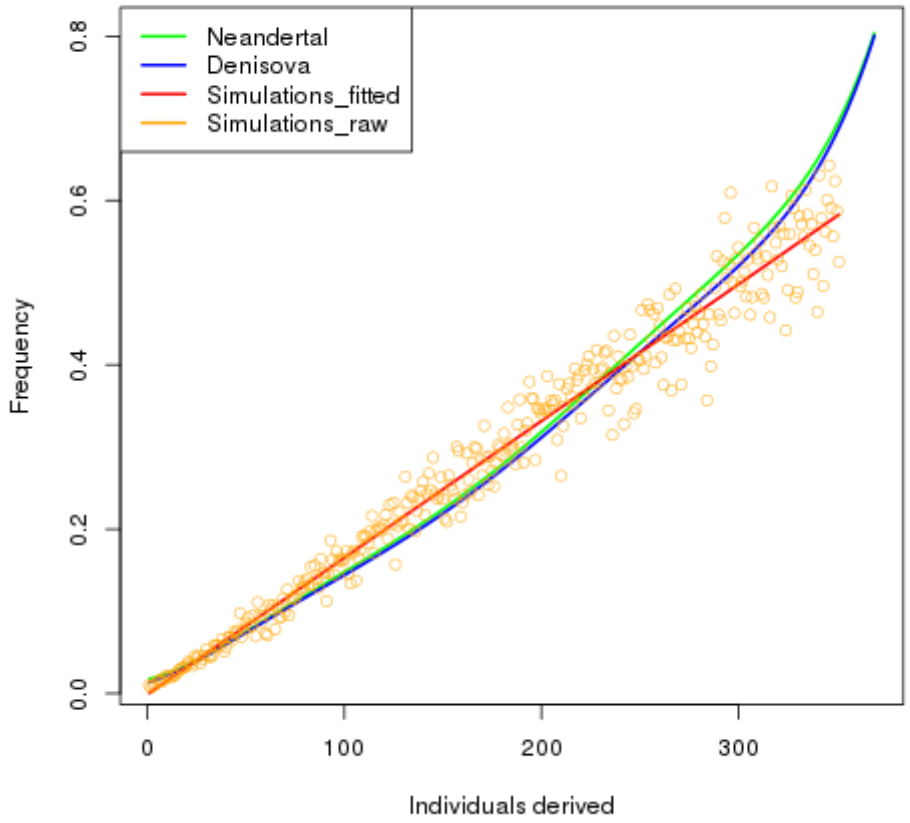


Figure S19a.2: Frequency of Denisova derived (blue) and Neandertal derived (green) given the derived allele frequency in Luhya and Yoruba. Simulated data is shown as yellow points with a best fit line shown in red.

Scoring of regions

The posterior decoding algorithm to determine the hidden states (*internal* or *external*) was executed independently for Denisova and Neandertal. In each run, we first identified regions that gave consistently high posterior probabilities for being external ($p \geq 0.8$). Some of these regions appear long because they contain large stretches of missing data, while others are long and well-supported by many SNPs. If we were to score solely based on genetic length, the regions with missing data would rank equally high. In order to avoid this issue, we took the spacing and number of SNPs in the region into account when calculating the score. For this, a small window of 3,500 SRR (standard recombination rate; i.e. the genetic size that would correspond to 3,500 basepairs at genome wide average recombination rate) was put around each SNP position in an external region. The window-size of 3,500 SRR corresponds to the average size of external regions we expect under neutrality. Overlapping neighboring windows of SNPs were merged. The total genetic length of the resulting windows was assigned as score for each region.

Comparison of Denisova and Neandertal Regions

Denisova and Neandertal were analysed separately, providing an opportunity to compare the external regions identified with each archaic genome. Table S19a.1 gives a comparison of summary statistics that show that more external regions were identified using Denisova as compared to Neandertal. Denisova external regions were also found to be longer on average. However, this difference is not significant (Wilcoxon rank test, one sided for shift towards higher values in Denisova: p=0.057 for physical length; p=0.126 for genetic length).

Measure	Denisova	Neandertal
Number regions	5713	5305
Total span of external regions	144.4 Mb	127.6 Mb
Average physical length	25284 bp	24056 bp
Average genetic length	2809 SRR	2762 SRR

Table S19a.1: Characteristics of Neandertal and Denisova external regions

We separated regions into three categories: external regions that are found only in Denisova, regions that are found only in Neandertal, and regions that are overlapping between Denisova and Neandertal (at least 1 basepair overlap). Table S19a.2 shows summary statistics for the three types of regions.

Measure	Denisova only	Neandertal only	Shared (Denisova / Neandertal)
Number of external regions	2584	2169	3124 / 3136
Total span of external regions	45.8 Mb	34.2 Mb	98.4 / 93.3
Average physical length	17710 bp	15799 bp	31498 / 29768 bp
Average genetic length	2384 SRR	2333 SRR	3163 / 3058 SRR

Table S19a.2: Characteristics of external regions that are found exclusively in Neandertal or in Denisova respectively, and external regions detected in both (shared). Values are given independently for the Denisova and Neandertal runs for regions that are shared.

We further examined the distributions of physical and genetic length by testing for a significant shift between the distributions using the Wilcoxon rank test. We found that overlapping regions are significantly longer and have a significantly higher score than Neandertal specific or only Denisova specific regions ($p<1.1e-12$ for all comparisons, test: Wilcoxon rank, one-sided for higher values for shared).

Interestingly, Denisova specific regions are significantly longer (physical length) compared to Neandertal specific regions ($p=0.034$, Wilcoxon rank test, one-sided for Denisova longer). However, this signal is not observed when comparing genetic length ($p=0.240$). Similarly, no signal is observed when comparing the length for overlapping regions between Denisova and Neandertal (physical length: $p=0.126$; genetic length: $p=0.060$).

Divergence of External Regions

We calculated divergence for the three classes of external regions using divergence triangulation with a Yoruban individual (see SI7 for details on the method). Divergence was calculated on the Yoruban lineage, eliminating the potential influence from different error rates or branch-shortening between Neandertal and Denisova. Overlapping external regions gave the lowest divergence (Denisova: 16.8%, 95% resampling confidence interval 15.8-17.8%; Neandertal: 17.1%, CI: 16.0-18.1%) as compared to Neandertal-specific (19.9%, CI: 18.8-21.1%) and Denisova-specific regions (19.6%, CI: 18.6-20.7%). When we divide the data in bins of physical sizes, we observe that overlapping and specific regions give a lower divergence with increasing physical size of the region (Figure S19a.3). The same effect is observed for increasing genetic sizes of overlapping regions and to a smaller degree for lineage-specific regions (Figure S19a.4).

More Overlap than Expected at Random

In order to assess how much overlap between Neandertal external regions and Denisova external regions would be expected, we carry out simulations using two simplifying assumptions. First, we treat the part of the genome sequence that passes filtering in Denisova and Neandertal as one contiguous sequence (~1.6Gb). Second, we assume that external regions are independently and randomly placed on this contiguous sequence. Regions have identical sizes to those found in Neandertal and Denisova.

In total, we carry out 200 simulations under these assumptions. We estimate that on average 932 regions are expected to overlap at random (95% confidence interval over 200 simulations: 884-984), much less than the more than 3,000 regions overlapping between Neandertal and Densiova (see Table S19a.2). The simulations also yields an expected length of overlapping regions if all overlap were random (71kb; CI: 63-80kb). The overlapping regions are thus much larger than observed in real data (~30kb; see Table S19a.2). However, when we restrict the sizes of regions to only the overlapping part, the average size is 12.3kb (CI: 10.9-13.6kb) in simulations, while we observe a significantly larger overlap of 26.2kb in our real data.

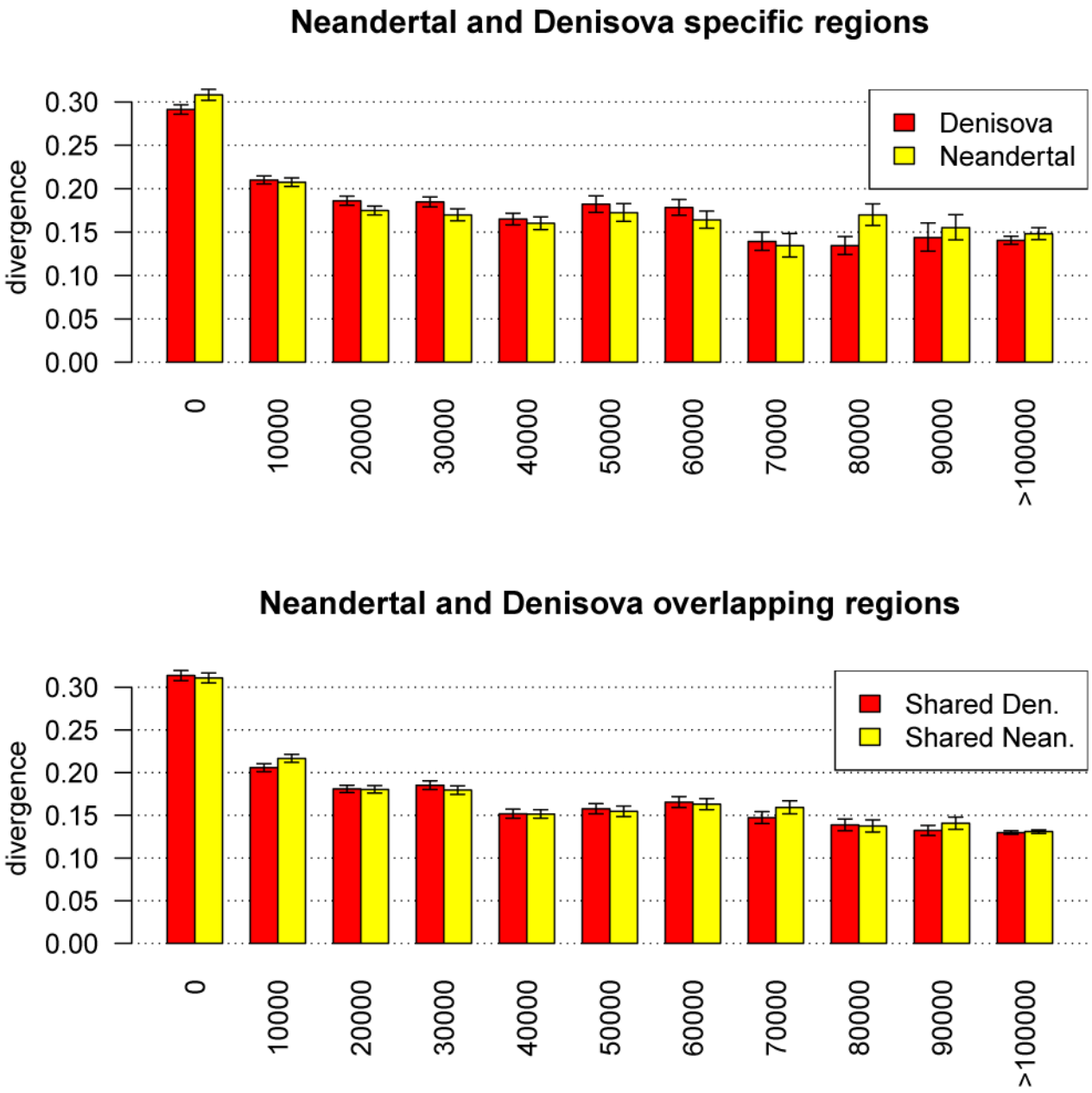


Figure S19a.3: Divergence of external regions in 11 bins of 10000 basepairs physical length. Error bars give the 95% binomial confidence interval.

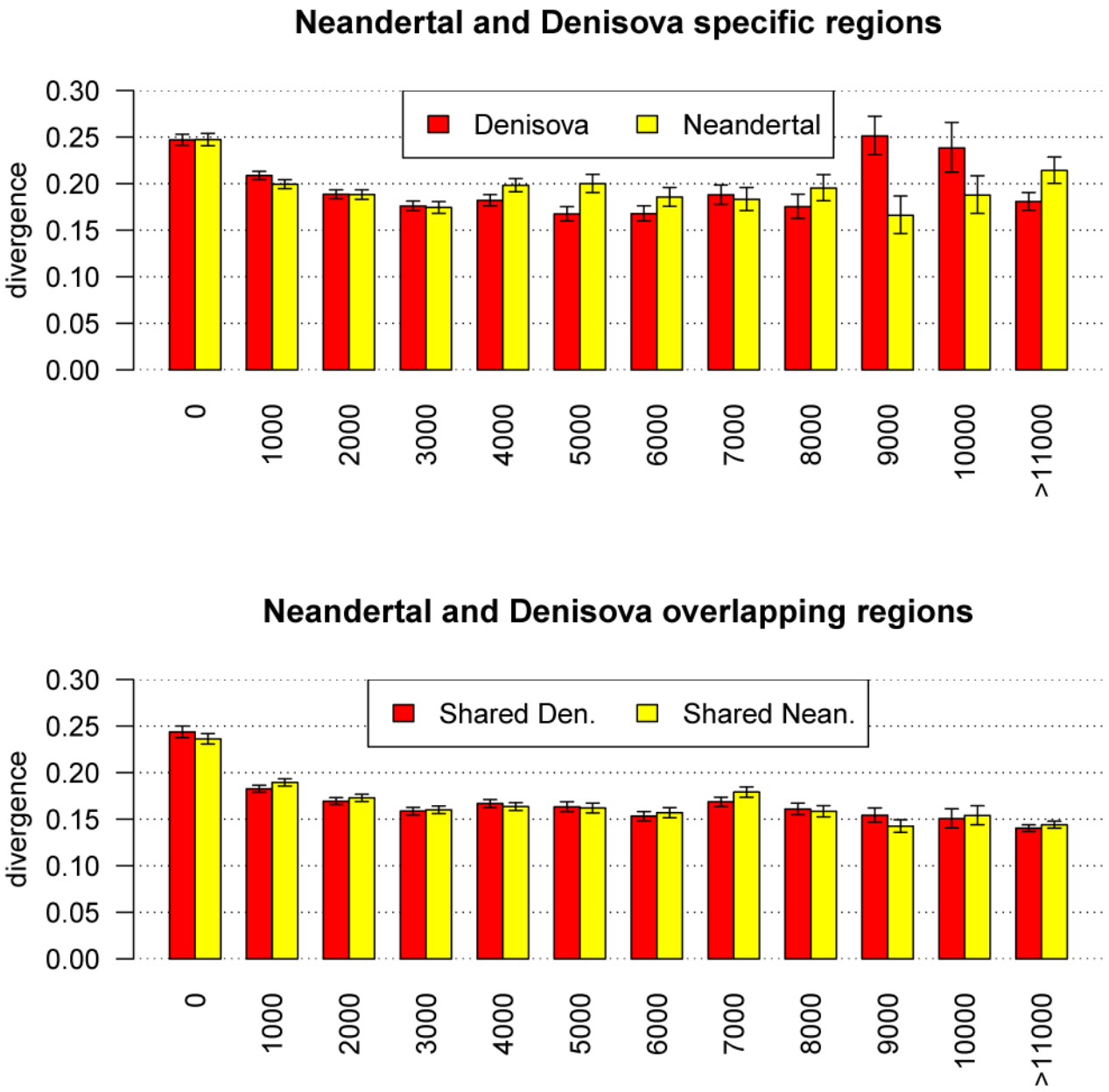


Figure S19a.4: Divergence of external regions in 12 bins of 1000 SRR genetic length. Error bars give the 95% binomial confidence interval.

Excess of Matching for Long Regions and Selection Candidates

Overlapping regions between Denisova and Neandertal rank significantly higher (according to the previously described score per external region) than regions that are found with only one of the two archaic genomes. It stands to reason that this excess is driven by true cases of positive selection that will generate long regions that are shared between Neandertal and Denisova. In line with this hypothesis, shared external regions show an excess of long regions (Figure S19a.4).

We use this excess of overlap in the top regions to define a cutoff. For this, Denisova and Neandertal regions are each sorted according to score and assigned a rank. The cumulative number of overlapping regions up to a certain rank in both Neandertal and Denisova are then tabulated (e.g.: there are 5 regions that overlap with rank<=10 in both archaics, as shown in Table S19a.3). Regions in one archaic that overlap more

than one region in the other archaic are counted only once with the lowest ranks in both lists. We then randomize the rank assignments 1000 times and tabulate the cumulative number of overlaps in each iteration. The randomized ranks give us an expected number of overlaps up to a certain rank if the score would be uninformative (Figure S19a.5). Based on these, we calculate the fold excess of true overlap for each rank cutoff by dividing the number of overlapping regions in out real data by the average number of overlaps in the 1000 random rank assignments. We find that up to rank 120 (corresponds to the top 63 regions overlapping regions), a 20-fold excess over random assignment is observed (Table S19a.3).

Since high-scoring regions tend to have a larger physical size and since longer regions have a higher chance to overlap with each other, an excess of overlaps in high-scoring regions may be expected by chance. We test whether this is the case by simulation (100 simulations placing Neandertal- and Denisova-sized regions on a contiguous sequence of 1.6Gb size). We find that 5 out of 100 simulations yielded exactly 1 region with a 20-fold excess over randomized ranks. This corresponds to a point estimate of 0.05 regions expected by chance, far less than the 63 regions detected in our real dataset (see Figure S19a.6). We conclude that random overlap cannot explain the observed excess and regard the 63 regions as candidates for positive selection (Table S19a.3).

Chr	Neandertal start	Neandertal end	Denisova start	Denisova end	Neandertal rank	Denisova rank	fold-increase over random
7	48716745	48844151	48716745	48844151	1	3	99.95
6	140479537	140818535	140479537	140814218	2	4	99.95
9	115216115	115313373	115209171	115313373	10	6	99.55
8	38022368	38201651	38022368	38210591	8	9	99.55
8	47888545	48984630	48086484	48988810	3	10	99.55
4	61374585	61396609	61374585	61396609	9	11	99.55
16	46923655	47779223	46766792	47777599	13	2	99.42
2	235781118	235789355	235780311	235789355	18	16	99.14
10	85515457	85564927	85515457	85564927	14	17	99.14
10	121525578	121712069	121525578	121712069	16	18	99.14
7	106803800	107247365	106803800	107247365	17	19	99.12
3	36420559	36481209	36420559	36481209	21	20	98.95
2	198233377	198480855	198233087	198512849	22	15	98.9
2	27106625	27125297	27106625	27125297	23	21	98.87
10	105001058	105131571	105001058	105134080	28	26	98.63
11	102219237	102265467	102218518	102265467	27	28	98.63
11	66382800	66647455	66382800	66636034	5	29	98.63
8	99564059	99874891	99757016	99878713	6	30	98.63
18	75970721	75981573	75970721	75979606	19	35	98.46
12	59503064	59550821	59465583	59602786	39	1	98.31
2	104354111	104396471	104354111	104396471	31	39	98.31
3	164871338	164915989	164842189	164916417	40	7	98.3
4	145738779	146120625	145738779	146153702	42	22	98.27
13	60969094	61131503	60969094	61131503	37	43	98.27
6	50416490	50492334	50416490	50492334	38	44	98.24

9	87106535	87121705	87106535	87121705	41	47	98.1
18	19092914	19462669	19092914	19470682	52	40	98
3	33707430	33825811	33707430	33825811	43	50	98
1	46664390	46782192	46664390	46782192	44	51	98
4	119721635	119786141	119721440	119786141	51	52	98
6	127498282	127528962	127498282	127529130	48	53	98
11	32902932	33016145	32902932	33016145	46	54	98
10	84328222	84379232	84328222	84380431	59	57	97.71
15	70210192	70232099	70210192	70232099	55	58	97.71
4	93760250	93794317	93760250	93794317	56	59	97.71
8	79636331	79747674	79636331	79747674	60	66	97.23
17	26348559	26498612	26295127	26498612	69	32	97.09
11	119046887	119180415	119046887	119180415	72	71	96.91
3	13925816	14091908	13925816	14108543	75	70	96.71
1	13997067	14111906	13997067	14111979	66	76	96.7
7	131010522	131187677	131008394	131187677	24	77	96.68
13	58032390	58227576	58057785	58228837	67	79	96.58
8	49736528	49866059	49736528	49866059	70	80	96.55
6	143214782	143246590	143214782	143246590	71	81	96.49
7	113400982	113496358	113400982	113495857	74	82	96.45
11	46348663	46721745	46348663	46721745	81	86	96.29
2	193850849	193958114	193850849	193958114	82	89	96.18
16	66523879	66606016	66523879	66606016	84	91	96.13
10	106238504	106244934	106238504	106244934	86	95	96.11
2	143435835	143448494	143435756	143448494	90	96	96.11
1	41464219	41576611	41464219	41574852	87	97	96.09
5	126794107	126807416	126794107	126807416	92	98	96.09
20	44668704	44724837	44668704	44724837	93	99	96.09
13	45496492	45616911	45496492	45616911	100	102	95.97
14	83546414	83603591	83546414	83603591	103	106	95.69
8	126898700	126939906	126898700	126939906	105	107	95.64
2	63138963	63302464	63138963	63302464	110	108	95.5
6	143426452	143496530	143426452	143496530	108	113	95.3
6	136762335	136802362	136762081	136802362	112	114	95.29
8	53473872	53665865	53458478	53666849	115	78	95.25
1	186184398	186214962	186190163	186214962	65	117	95.18
2	71479539	71676366	71479539	71676366	119	119	95.03
3	44478332	44570479	44477861	44570700	120	118	95

Table S19a.3: Candidate regions for positive selection with their coordinates according to the Denisova and Neandertal scans.

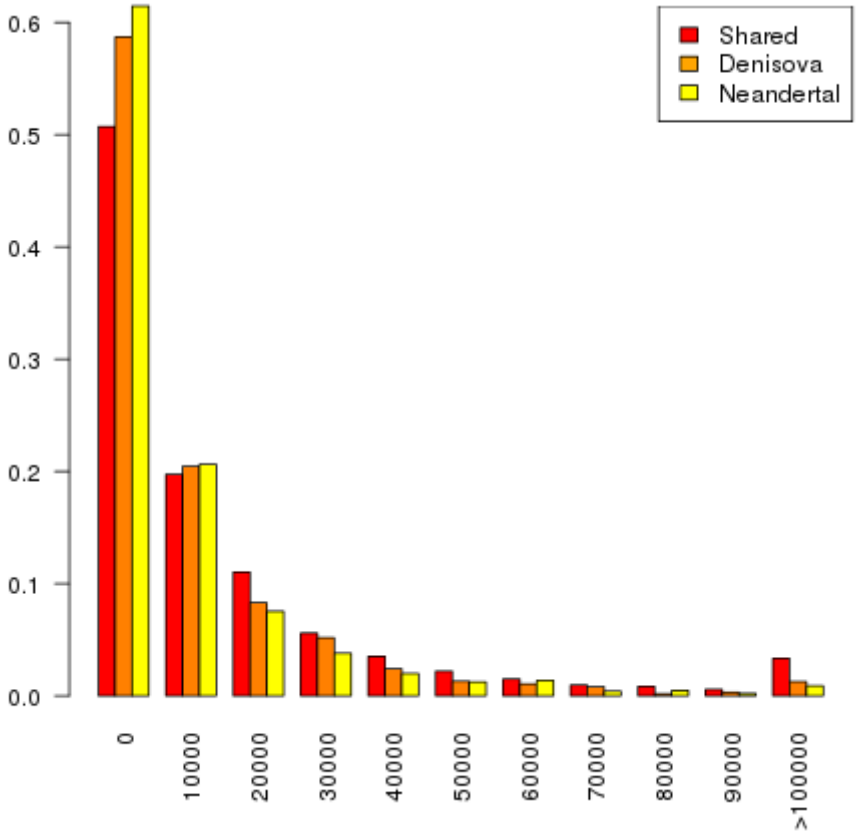


Figure S19a.4: Fraction of external regions in 11 bins of 10000 basepairs physical length.

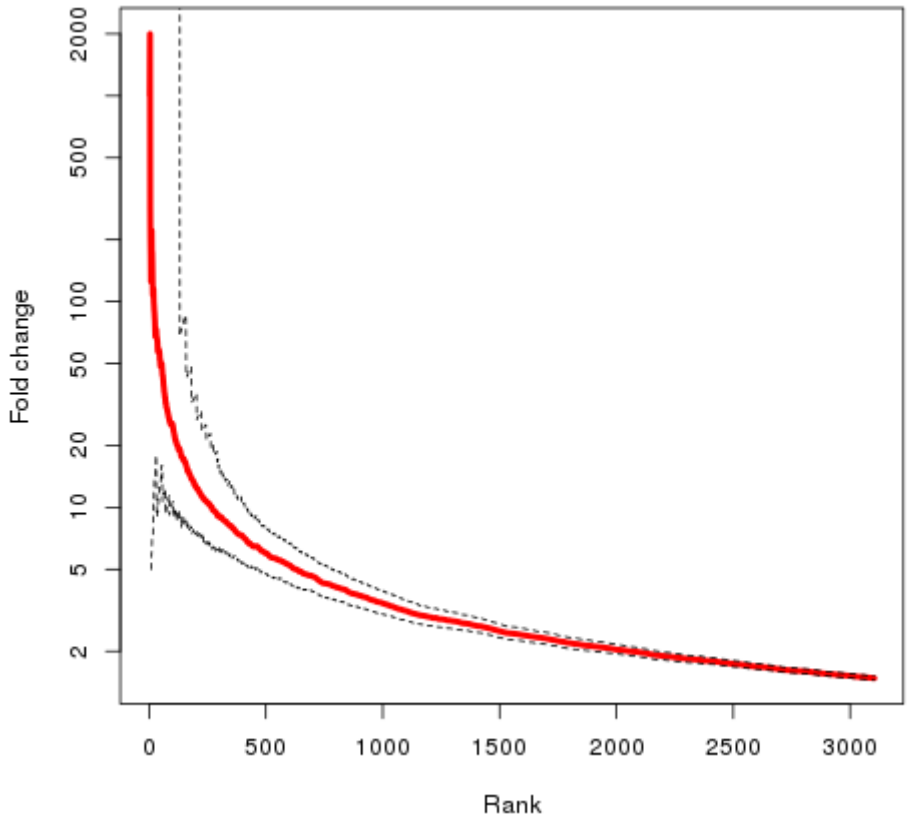


Figure S19a.5: For each rank cutoff, the number of matching regions in real data is divided by the number of matching regions when ranks are assigned randomly. Dashed lines give the lowest and highest estimate for the fold-change when comparing over 1000 random assignments; red line gives the mean.

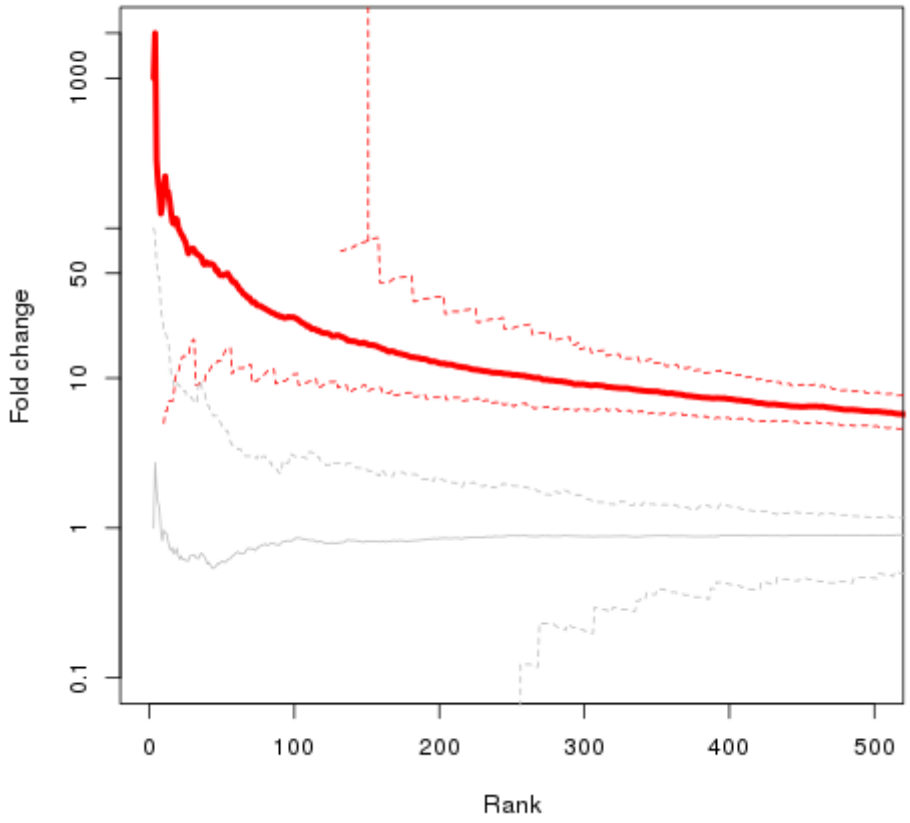


Figure S19a.6: For each rank cutoff, the fold-change of matching regions in real data (red) and simulated data (gray) is shown. Dashed lines give the lowest and highest estimate for the fold-change when comparing over 1000 random assignments; solid lines show the mean.

Choice of Recombination Map

We used the African American genetic map[3] to score our results. We have chosen this recombination map for two reasons: first, the African American map was generated from recombination events between regions of European and African ancestry and thus constructed from recombination events that happened up to a few generations ago[3]. This procedure is different from a LD-based recombination map, such as the HapMap recombination map[5], which averages over recombination events in the coalescent history of the population and may be influenced by selective sweeps in the past; sweeps can partly erase the history of recombination preceding the sweep and may lead to an underestimate of the true recombination rate for LD-based maps and may thus limit our power to detect selected regions. Second, the African American map has a higher resolution compared to previous maps that are based on recent recombination events[6].

In order to test how much our results are influenced by the choice of recombination map, we repeated our analysis using the HapMap II map[5] and DeCode recombination map[6]. Table S19a.4 gives the number of regions that overlap in various comparisons between recombination maps. While a similar order of regions are found with each map (54 – 84 regions with a 20-fold enrichment over random), the intersection of these top-scoring regions between the African American, the HapMap and the DeCode maps

yields only 15 regions (Table S19a.5). Since these regions are found as outliers independent of the tested recombination maps, they may represent a subset of regions with higher confidence.

Recombination Map / Intersection	Number of Regions
African American	63
DeCode	84
HapMap	54
African American & Decode	23
African American & HapMap	23
Decode & HapMap	24
African American & Decode & HapMap	15

Table S19a.4: Number and overlap of top scoring regions for three recombination maps: HapMap[2], Decode[5] and African American map[3].

Chromosome	Start	End
7	48716745	48804068
6	140480464	140814218
8	48086484	48960824
4	61374585	61378164
10	121600100	121711232
7	106803800	107246893
12	59503064	59550821
18	19092914	19409406
4	119724632	119785381
15	70217239	70231010
11	119046887	119176499
1	14000581	14111906
8	49736528	49866059
11	46348663	46717673
8	53473872	53662944

Table S19a.5: Top scoring regions found independently with the HapMap, Decode and African American maps. The coordinates give the intersecting part of the regions found in all three maps.

References

1 Prufer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**, 527-531, doi:10.1038/nature11128 (2012).

2 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

3 Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170-175, doi:10.1038/nature10336 (2011).

4 Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-338 (2002).

5 Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861, doi:10.1038/nature06258 (2007).

6 Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-1103, doi:10.1038/nature09525 (2010).

Supplementary Information 19b

Characterization of Selective Sweep Screen

Fernando Racimo*, Martin Kuhlwilm, Martin Kircher, Kay Prüfer and Janet Kelso

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

Table of contents

1. Pathway and disease annotation of genes in screen

We focused on the regions of the selective sweep screen that have have a 20-fold excess over randomized rank assignments (SI 19a). The chromosomal distribution of these 63 regions is plotted in Figure 1. We annotated the 112 genes that lie within these regions and that have Entrez gene and Ensembl gene IDs, using KOBAS 2.0[1], which combines annotations from a variety of sources, including GAD[2], KEGG[3], FunDO[4,5], PID[6], PANTHER[7] and the NHGRI GWAS catalog[8]. The disease and pathway annotations for each of the genes in the screen are listed in Table S19b.1.

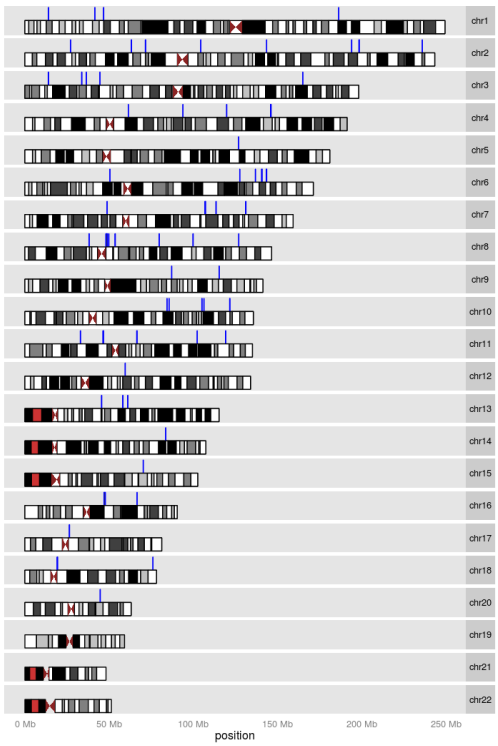


Figure S19b.1. Top 63 regions (blue markers) identified by the selective sweep screen (SI 19a).

Region rank	Entrez Gene ID	Ensembl ID	HGNC	Pathway annotation	Disease annotation
3	84263	ENSG00000119471	*HSDL2*	N/A	N/A
3	158405	ENSG00000165185	*FLJ39294*	N/A	N/A
4	23259	ENSG00000085788	*DDHD2*	N/A	N/A
4	84513	ENSG00000147535	*PPAPDC1B*	N/A	N/A
4	54904	ENSG00000147548	*WHSC1L1*	Lysine degradation (KEGG PATHWAY)	Leukemia, acute myeloid (OMIM)
4	9530	ENSG00000156735	*BAG4*	Apoptosis signaling pathway (PANTHER); tnfr1 signaling pathway (PID BioCarta); ceramide signaling pathway (PID BioCarta); sodd/tnfr1 signaling pathway (PID BioCarta); tnf/stress related signaling (PID BioCarta); hiv-1 nef: negative effector of fas and tnf (PID BioCarta); TNF receptor signaling pathway (PID Curated); Ceramide signaling pathway (PID Curated); HIV-1 Nef: Negative effector of Fas and TNF-alpha (PID Curated)	Depression (FunDO); Hypertension (FunDO)
4	27257	ENSG00000175324	*LSM1*	RNA degradation (KEGG PATHWAY); Metabolism of RNA (Reactome); Gene Expression (Reactome)	Prostate cancer (FunDO); Breast cancer (FunDO)
5	4173	ENSG00000104738	*MCM4*	DNA replication (KEGG PATHWAY); Cell cycle (KEGG PATHWAY); cdk regulation of dna replication (PID BioCarta); C-MYB transcription factor network (PID Curated); Switching of origins to a post-replicative state (PID Reactome); Removal of licensing factors from origins (PID Reactome); DNA Replication (Reactome); Cell Cycle (Reactome)	Hemolytic-Uremic syndrome (FunDO)
5	23514	ENSG00000164808	*FLJ35017*	N/A	N/A
5	7336	ENSG00000169139	*UBE2V2*	Ubiquitin proteasome pathway (PANTHER); Immune System (Reactome)	N/A
5	1052	ENSG00000221869	*CEBPD*	Validated transcriptional targets of deltaNp63 isoforms (PID Curated); IL6-mediated signaling events (PID Curated); Regulation of retinoblastoma protein (PID Curated); Validated targets of C-MYC transcriptional repression (PID Curated); FOXA2 and FOXA3 transcription factor networks (PID Curated); C-MYB transcription factor network (PID Curated); Developmental Biology (Reactome)	Prostate cancer (FunDO); Breast cancer (FunDO); Leukemia (FunDO)
5	5591	ENSG00000253729	*PRKDC*	Non-homologous end-joining (KEGG PATHWAY); Cell cycle (KEGG PATHWAY); cell cycle: g2/m checkpoint (PID BioCarta); hiv-1 nef: negative effector of fas and tnf (PID BioCarta); fas signaling pathway (cd95) (PID BioCarta); BARD1 signaling events (PID Curated); Coregulation of Androgen receptor activity (PID Curated); Class I PI3K signaling events mediated by Akt (PID Curated); DNA-PK pathway in nonhomologous end joining (PID Curated); Nonhomologous End-joining (NHEJ) (PID Reactome); Processing of DNA ends prior to end rejoining (PID Reactome); DNA Repair (Reactome)	N/A
7	10294	ENSG00000069345	*DNAJA2*	Protein processing in endoplasmic reticulum (KEGG PATHWAY)	N/A
7	5257	ENSG00000102893	*PHKB*	Insulin signaling pathway (KEGG PATHWAY); Calcium signaling pathway (KEGG PATHWAY); Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway (PANTHER); Glycogen breakdown (glycogenolysis) (PID Reactome); Metabolism (Reactome)	Congenital disorders of carbohydrate metabolism (KEGG DISEASE); Congenital disorders of metabolism (KEGG DISEASE); Glycogen storage disease (GSD) (KEGG DISEASE); Phosphorylase kinase deficiency of liver and muscle, autosomal recessive (OMIM); liver glycogenosis caused by Phk deficiency (GAD)
7	81533	ENSG00000129636	*ITFG1*	N/A	N/A
7	84706	ENSG00000166123	*GPT2*	Alanine, aspartate and glutamate metabolism (KEGG PATHWAY); Metabolic pathways (KEGG PATHWAY); alanine biosynthesis II (BioCyc); alanine degradation III (BioCyc); Amino acid synthesis and interconversion (transamination) (PID Reactome); Metabolism (Reactome)	N/A
7	81831	ENSG00000171208	*NETO2*	N/A	N/A
10	11196	ENSG00000107651	*SEC23IP*	N/A	N/A
10	79892	ENSG00000197771	*MCMBP*	N/A	N/A
10	22876	ENSG00000198825	*INPP5F*	superpathway of D-myo-inositol (1,4,5)-trisphosphate metabolism (BioCyc); D-myo-inositol (1,4,5)-trisphosphate degradation (BioCyc); 3-phosphoinositide degradation (BioCyc)	N/A
11	55973	ENSG00000075790	*BCAP29*	N/A	N/A
11	26959	ENSG00000105856	*HBP1*	Validated transcriptional targets of deltaNp63 isoforms (PID Curated); E2F transcription factor network (PID Curated); Signaling mediated by p38-alpha and p38-beta (PID Curated); Validated transcriptional targets of TAp63 isoforms (PID Curated); Regulation of nuclear beta catenin signaling and target gene transcription (PID Curated); C-MYC pathway (PID Curated)	Cancer (FunDO)
11	11062	ENSG00000105865	*DUS4L*	N/A	N/A
11	10466	ENSG00000164597	*COG5*	N/A	Congenital disorder of glycosylation, type IIi (OMIM)
11	2845	ENSG00000172209	*GPR22*	N/A	N/A
12	6769	ENSG00000144681	*STAC*	N/A	N/A

13	80219	ENSG00000115520	*COQ10B*	N/A	N/A
13	23451	ENSG00000115524	*SF3B1*	Spliceosome (KEGG PATHWAY); mRNA Processing (Reactome); Gene Expression (Reactome)	N/A
13	25843	ENSG00000115540	*MOB4*	N/A	N/A
13	100529241	ENSG00000115540	*HSPE1-MOB4*	N/A	N/A
13	3336	ENSG00000115541	*HSPE1*	N/A	Metastasis to lymph nodes (FunDO); Embryoma (FunDO); Ischemia (FunDO); Spastic paraplegia, Hereditary (FunDO); Endometrial cancer (FunDO); Colon cancer (FunDO); ovarian cancer (GAD); CANCER (GAD)
13	3329	ENSG00000144381	*HSPD1*	Tuberculosis (KEGG PATHWAY); Legionellosis (KEGG PATHWAY); RNA degradation (KEGG PATHWAY); Type I diabetes mellitus (KEGG PATHWAY); Validated targets of C-MYC transcriptional activation (PID Curated); Endogenous TLR signaling (PID Curated)	Congenital disorders of metabolism (KEGG DISEASE); Musculoskeletal and skin diseases (KEGG DISEASE); Pelizaeus-Merzbacher disease (KEGG DISEASE); Skin and soft tissue diseases (KEGG DISEASE); Congenital disorders of lipid/glycolipid metabolism (KEGG DISEASE); Hereditary spastic paraplegia (SPG) (KEGG DISEASE); Leukodystrophy, hypomyelinating, 4 (OMIM); Spastic paraplegia-13 (OMIM); Yersinia infection (FunDO); Cervical cancer (FunDO); Rheumatoid arthritis (FunDO); Lyme disease (FunDO); Prostate cancer (FunDO); Metastasis to lymph nodes (FunDO); Spastic paraplegia, Hereditary (FunDO); Atherosclerosis (FunDO); Down syndrome (FunDO); Dental plaque (FunDO); Liver tumor (FunDO); Tic disorder (FunDO); Liver cancer (FunDO); Colon cancer (FunDO); Heart failure (FunDO); Leukodystrophy NOS (FunDO); Systemic infection (FunDO); Dermatitis (FunDO); Prion disease (FunDO)
13	130132	ENSG00000162944	*RFTN2*	N/A	N/A
14	56896	ENSG00000157851	*DPYSL5*	Axon guidance (KEGG PATHWAY); Axon guidance mediated by semaphorins (PANTHER); Pyrimidine Metabolism (PANTHER); Developmental Biology (Reactome)	Neuropathy (FunDO); Eye disease (FunDO); Lung cancer (FunDO); Bone marrow disease (FunDO)
15	9118	ENSG00000148798	*INA*	N/A	Neuroblastoma (FunDO); Cytomegalovirus infection (FunDO)
15	6877	ENSG00000148835	*TAF5*	Basal transcription factors (KEGG PATHWAY); Herpes simplex infection (KEGG PATHWAY); Transcription (Reactome); Disease (Reactome); Gene Expression (Reactome)	N/A
15	84108	ENSG00000156374	*PCGF6*	N/A	N/A
16	329	ENSG00000110330	*BIRC2*	Focal adhesion (KEGG PATHWAY); Small cell lung cancer (KEGG PATHWAY); Pathways in cancer (KEGG PATHWAY); Toxoplasmosis (KEGG PATHWAY); Apoptosis (KEGG PATHWAY); Ubiquitin mediated proteolysis (KEGG PATHWAY); HTLV-I infection (KEGG PATHWAY); NOD-like receptor signaling pathway (KEGG PATHWAY); Apoptosis signaling pathway (PANTHER); role of mitochondria in apoptotic signaling (PID BioCarta); keratinocyte differentiation (PID BioCarta); hiv-1 nef: negative effector of fas and tnf (PID BioCarta); TNF receptor signaling pathway (PID Curated); Canonical NF-kappaB pathway (PID Curated); CD40/CD40L signaling (PID Curated); FAS (CD95) signaling pathway (PID Curated); Caspase cascade in apoptosis (PID Curated); Apoptotic cleavage of cellular proteins (PID Reactome); Immune System (Reactome); Apoptosis (Reactome)	Multiple sclerosis (FunDO); Cancer (FunDO); Schizophrenia (FunDO)
17	5091	ENSG00000173599	*PC*	Citrate cycle (TCA cycle) (KEGG PATHWAY); Pyruvate metabolism (KEGG PATHWAY); Metabolic pathways (KEGG PATHWAY); Pyruvate metabolism (PANTHER); Metabolism (Reactome)	Congenital disorders of carbohydrate metabolism (KEGG DISEASE); Congenital disorders of metabolism (KEGG DISEASE); Pyruvate carboxylase deficiency (KEGG DISEASE); Pyruvate carboxylase deficiency (OMIM)
17	78999	ENSG00000173621	*LRFN4*	N/A	N/A
17	9986	ENSG00000173653	*RCE1*	N/A	N/A
17	79703	ENSG00000173715	*C11orf80*	N/A	N/A
17	6712	ENSG00000173898	*SPTBN2*	Developmental Biology (Reactome)	Nervous system diseases (KEGG DISEASE); Neurodegenerative diseases (KEGG DISEASE); Spinocerebellar ataxia (KEGG DISEASE); Spinocerebellar ataxia-5 (OMIM)
17	83759	ENSG00000173914	*RBM4B*	N/A	N/A
17	5936	ENSG00000173933	*RBM4*	N/A	Nephroblastoma (FunDO); Alzheimer's disease (FunDO); Down syndrome (FunDO)
17	10432	ENSG00000239306	*RBM14*	N/A	N/A
17	100526737	ENSG00000248643	*RBM14-RBM4*	N/A	N/A
18	6788	ENSG00000104375	*STK3*	MAPK signaling pathway (KEGG PATHWAY)	N/A
22	22865	ENSG00000121871	*SLITRK3*	N/A	N/A
23	10393	ENSG00000164162	*ANAPC10*	Progesterone-mediated oocyte maturation (KEGG PATHWAY); Ubiquitin mediated proteolysis (KEGG PATHWAY); HTLV-I infection (KEGG PATHWAY); Cell cycle (KEGG PATHWAY); Oocyte meiosis (KEGG PATHWAY); Cell cycle (PANTHER); APC/C-mediated degradation of cell cycle proteins (PID Reactome); Regulation of APC/C activators between G1/S and early anaphase (PID Reactome); Autodegradation of Cdh1 by Cdh1:APC/C (PID Reactome); Phosphorylation of the APC/C (PID Reactome); APC-Cdc20 mediated degradation of Nek2A (Reactome); Immune System (Reactome); Cell Cycle (Reactome); Cdc20:Phospho-APC/C mediated degradation of Cyclin A (Reactome)	N/A
23	6059	ENSG00000164163	*ABCE1*	N/A	Prostate cancer (FunDO); Colon cancer (FunDO)
23	54726	ENSG00000164164	*OTUD4*	N/A	N/A
24	81550	ENSG00000083544	*TDRD3*	N/A	N/A
27	57534	ENSG00000101752	*MIB1*	Notch signaling pathway (PID Curated)	Brain tumor (FunDO); Brain disease (FunDO); Central nervous system disease (FunDO); Nervous system tumor (FunDO)

27	114799	ENSG00000141446	*ESCO1*	N/A	N/A
27	80000	ENSG00000141449	*GREB1L*	N/A	N/A
27	171586	ENSG00000158201	*ABHD3*	N/A	N/A
27	6632	ENSG00000167088	*SNRPD1*	Systemic lupus erythematosus (KEGG PATHWAY); Spliceosome (KEGG PATHWAY); snRNP Assembly (PID Reactome); Metabolism of RNA (Reactome); mRNA Processing (Reactome); Gene Expression (Reactome)	N/A
28	23122	ENSG00000163539	*CLASP2*	Developmental Biology (Reactome); DNA Replication (Reactome); Cell Cycle (Reactome)	N/A
29	55624	ENSG00000085998	*POMGNT1*	Other types of O-glycan biosynthesis (KEGG PATHWAY)	Muscular diseases (KEGG DISEASE); Congenital disorders of metabolism (KEGG DISEASE); Musculoskeletal and skin diseases (KEGG DISEASE); Dystroglycanopathy (KEGG DISEASE); Congenital disorders of glycan/glycoprotein metabolism (KEGG DISEASE); Congenital muscular dystrophies (CMD/MDC) (KEGG DISEASE); Limb-girdle muscular dystrophy (LGMD) (KEGG DISEASE); Muscular dystrophy-dystroglycanopathy (OMIM)
29	8438	ENSG00000085999	*RAD54L*	Homologous recombination (KEGG PATHWAY)	Adenocarcinoma, colonic, somatic (OMIM); Breast cancer, invasive ductal (OMIM); Lymphoma, non-Hodgkin, somatic (OMIM)
29	10489	ENSG00000132128	*LRRC41*	Immune System (Reactome)	N/A
29	541468	ENSG00000171357	*C1orf190*	N/A	N/A
29	7388	ENSG00000173660	*UQCRH*	Parkinson's disease (KEGG PATHWAY); Oxidative phosphorylation (KEGG PATHWAY); Cardiac muscle contraction (KEGG PATHWAY); Alzheimer's disease (KEGG PATHWAY); Metabolic pathways (KEGG PATHWAY); Huntington's disease (KEGG PATHWAY); Metabolism (Reactome)	N/A
29	440567	ENSG00000173660	*UQCRHL*	Parkinson's disease (KEGG PATHWAY); Oxidative phosphorylation (KEGG PATHWAY); Cardiac muscle contraction (KEGG PATHWAY); Alzheimer's disease (KEGG PATHWAY); Metabolic pathways (KEGG PATHWAY); Huntington's disease (KEGG PATHWAY)	N/A
30	9871	ENSG00000150961	*SEC24D*	Protein processing in endoplasmic reticulum (KEGG PATHWAY); Membrane Trafficking (Reactome); Immune System (Reactome); Metabolism of proteins (Reactome)	N/A
31	84870	ENSG00000146374	*RSPO3*	N/A	HEMATOLOGICAL (GAD); serum markers of iron status (GAD)
32	79832	ENSG00000060749	*QSER1*	N/A	Parkinson's disease (age of onset) (GAD); NEUROLOGICAL (GAD)
33	10718	ENSG00000185737	*NRG3*	ErbB signaling pathway (KEGG PATHWAY); g-secretase mediated erbb4 signaling pathway (PID BioCarta); ErbB receptor signaling network (PID Curated); ErbB4 signaling events (PID Curated); Signal Transduction (Reactome)	Breast cancer (FunDO); response to iloperidone treatment (QT prolongation) (GAD); ADHD (GAD); parental expressed emotion (GAD); PSYCH (GAD); PHARMACOGENOMIC (GAD); schizophrenia (GAD)
35	2895	ENSG00000152208	*GRID2*	Long-term depression (KEGG PATHWAY); Neuroactive ligand-receptor interaction (KEGG PATHWAY)	N/A
36	3574	ENSG00000104432	*IL7*	Jak-STAT signaling pathway (KEGG PATHWAY); Hematopoietic cell lineage (KEGG PATHWAY); Cytokine-cytokine receptor interaction (KEGG PATHWAY); Interleukin signaling pathway (PANTHER); Immune System (Reactome)	Common variable immunodeficiency (FunDO); Immunologic deficiency syndrome (FunDO); Lymphopenia (FunDO); Rheumatoid arthritis (FunDO); Ovarian cancer (FunDO); Atherosclerosis (FunDO); Polyarthritis (FunDO); Leukemia (FunDO); Pulmonary fibrosis (FunDO); Hemolytic-Uremic syndrome (FunDO)
37	51701	ENSG00000087095	*NLK*	MAPK signaling pathway (KEGG PATHWAY); Adherens junction (KEGG PATHWAY); Wnt signaling pathway (KEGG PATHWAY); Wnt signaling pathway (PANTHER); wnt signaling pathway (PID BioCarta); Presenilin action in Notch and Wnt signaling (PID Curated); Noncanonical Wnt signaling pathway (PID Curated); C-MYB transcription factor network (PID Curated)	N/A
38	4162	ENSG00000076706	*MCAM*	N/A	Melanoma (FunDO); Prostate cancer (FunDO); Embryoma (FunDO); Ovarian cancer (FunDO); Kidney failure (FunDO); Diabetes mellitus (FunDO); Atherosclerosis (FunDO)
38	867	ENSG00000110395	*CBL*	Jak-STAT signaling pathway (KEGG PATHWAY); Insulin signaling pathway (KEGG PATHWAY); ErbB signaling pathway (KEGG PATHWAY); Pathways in cancer (KEGG PATHWAY); Bacterial invasion of epithelial cells (KEGG PATHWAY); Chronic myeloid leukemia (KEGG PATHWAY); T cell receptor signaling pathway (KEGG PATHWAY); Endocytosis (KEGG PATHWAY); Ubiquitin mediated proteolysis (KEGG PATHWAY); EGF receptor signaling pathway (PANTHER); il-2 receptor beta chain in t cell activation (PID BioCarta); sprouty regulation of tyrosine kinase signals (PID BioCarta); cbl mediated ligand-induced downregulation of egf receptors pathway (PID BioCarta); EPHA forward signaling (PID Curated); TCR signaling in naive CD4+ T cells (PID Curated); Signaling events mediated by Stem cell factor receptor (c-Kit) (PID Curated); EPO signaling pathway (PID Curated); IL8- and CXCR1-mediated signaling events (PID Curated); CDC42 signaling events (PID Curated); Internalization of ErbB1 (PID Curated); FGF signaling pathway (PID Curated); VEGFR1 specific signals (PID Curated); Signaling events mediated by VEGFR1 and VEGFR2 (PID Curated); Reelin signaling pathway (PID Curated); TCR signaling in naive CD8+ T cells (PID Curated); Integrins in angiogenesis (PID Curated); EPHA2 forward signaling (PID Curated); Fc-epsilon receptor I signaling in mast cells (PID Curated); IL4-mediated signaling events (PID Curated); Notch signaling pathway (PID Curated); IFN-gamma pathway (PID Curated); PDGFR-beta signaling pathway (PID Curated); Insulin Pathway (PID Curated); IL8- and CXCR2-mediated signaling events (PID Curated); Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met) (PID Curated); EGFR downregulation (PID Reactome); Immune System (Reactome); Signal Transduction (Reactome); Disease (Reactome)	Noonan syndrome (KEGG DISEASE); Congenital disorders of development (KEGG DISEASE); Other congenital disorders (KEGG DISEASE); Noonan syndrome-like disorder (OMIM); Esotropia (FunDO); Stroke (FunDO); Leukemia (FunDO)
38	79671	ENSG00000160703	*NLRX1*	Influenza A (KEGG PATHWAY); RIG-I-like receptor signaling pathway (KEGG PATHWAY); Immune System (Reactome)	N/A
38	79849	ENSG00000172367	*PDZD3*	N/A	N/A

38	283152	ENSG00000248712	*CCDC153*	N/A	N/A
40	7799	ENSG00000116731	*PRDM2*	N/A	CANCER (GAD); overall effect (GAD)
41	5420	ENSG00000128567	*PODXL*	N/A	Cancer (FunDO)
41	4289	ENSG00000128585	*MKLN1*	N/A	N/A
42	27253	ENSG00000118946	*PCDH17*	N/A	N/A
43	6591	ENSG00000019549	*SNAI2*	Adherens junction (KEGG PATHWAY); Signaling events mediated by Stem cell factor receptor (c-Kit) (PID Curated); Direct p53 effectors (PID Curated); Regulation of nuclear beta catenin signaling and target gene transcription (PID Curated)	Congenital disorders of metabolism (KEGG DISEASE); Piebaldism (KEGG DISEASE); Other congenital disorders of metabolism (KEGG DISEASE); Waardenburg syndrome (WS) (KEGG DISEASE); Congenital disorders of amino acid metabolism (KEGG DISEASE); Waardenburg syndrome, type 2D (OMIM); Piebaldism (OMIM); Cancer (FunDO)
44	3097	ENSG00000010818	*HIVEP2*	N/A	N/A
46	4192	ENSG00000110492	*MDK*	Beta1 integrin cell surface interactions (PID Curated); Syndecan-4-mediated signaling events (PID Curated); Alpha4 beta1 integrin signaling events (PID Curated); Glypican 2 network (PID Curated)	Brain ischemia (FunDO); Rheumatoid arthritis (FunDO); Endometriosis (FunDO); Diabetes mellitus (FunDO); Cancer (FunDO); Stroke (FunDO); Neurofibromatosis (FunDO); colorectal cancer (GAD); CANCER (GAD)
46	55626	ENSG00000110497	*AMBRA1*	N/A	N/A
46	8525	ENSG00000149091	*DGKZ*	Phosphatidylinositol signaling system (KEGG PATHWAY); Glycerolipid metabolism (KEGG PATHWAY); Glycerophospholipid metabolism (KEGG PATHWAY); Metabolic pathways (KEGG PATHWAY); Signal Transduction (Reactome); Hemostasis (Reactome)	N/A
46	392	ENSG00000175220	*ARHGAP1*	VEGF signaling pathway (PANTHER); Angiogenesis (PANTHER); PDGF signaling pathway (PANTHER); Cytoskeletal regulation by Rho GTPase (PANTHER); rac1 cell motility signaling pathway (PID BioCarta); adp-ribosylation factor (PID BioCarta); rho cell motility signaling pathway (PID BioCarta); t cell receptor signaling pathway (PID BioCarta); Regulation of CDC42 activity (PID Curated); Regulation of RAC1 activity (PID Curated); Signal Transduction (Reactome)	Bone mineral density (hip) (GAD); METABOLIC (GAD)
46	9776	ENSG00000175224	*ATG13*	mTOR signaling pathway (PID Curated)	N/A
46	283254	ENSG00000180423	*HARBI1*	N/A	N/A
46	1132	ENSG00000180720	*CHRM4*	Cholinergic synapse (KEGG PATHWAY); Regulation of actin cytoskeleton (KEGG PATHWAY); Neuroactive ligand-receptor interaction (KEGG PATHWAY); Alzheimer disease-amyloid secretase pathway (PANTHER); Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway (PANTHER); Muscarinic acetylcholine receptor 2 and 4 signaling pathway (PANTHER); Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway (PANTHER); Signal Transduction (Reactome)	Supranuclear palsy, progressive (FunDO); Schizophrenia (FunDO)
48	113540	ENSG00000089505	*CMTM1*	N/A	N/A
48	146227	ENSG00000166546	*BEAN1*	N/A	Nervous system diseases (KEGG DISEASE); Neurodegenerative diseases (KEGG DISEASE); Spinocerebellar ataxia (KEGG DISEASE); Spinocerebellar ataxia 31 (OMIM)
48	7084	ENSG00000166548	*TK2*	Drug metabolism - other enzymes (KEGG PATHWAY); Metabolic pathways (KEGG PATHWAY); Pyrimidine metabolism (KEGG PATHWAY); salvage pathways of pyrimidine deoxyribonucleotides (BioCyc); Pyrimidine salvage reactions (PID Reactome); Metabolism (Reactome); Cell Cycle (Reactome)	Congenital disorders of metabolism (KEGG DISEASE); Mitochondrial DNA depletion syndrome (MDS) (KEGG DISEASE); Other congenital disorders of metabolism (KEGG DISEASE); Mitochondrial DNA depletion syndrome 2 (OMIM); Squamous cell cancer (FunDO); Myopathy (FunDO); Lung cancer (FunDO); mitochondrial myopathy (GAD); MITOCHONDRIAL (GAD)
48	51192	ENSG00000217555	*CKLF*	N/A	N/A
48	100529251	ENSG00000254788	*CKLF-CMTM1*	N/A	N/A
51	22955	ENSG00000010803	*SCMH1*	N/A	height (GAD); DEVELOPMENTAL (GAD)
51	200172	ENSG00000171790	*SLFNL1*	N/A	N/A
51	1503	ENSG00000171793	*CTPS*	Metabolic pathways (KEGG PATHWAY); Pyrimidine metabolism (KEGG PATHWAY); pyrimidine ribonucleotides interconversion (BioCyc); pyrimidine ribonucleotides de novo biosynthesis (BioCyc); De novo pyrimidine ribonucleotides biosythesis (PANTHER); Synthesis and interconversion of nucleotide di- and triphosphates (PID Reactome); Metabolism (Reactome)	N/A
52	84466	ENSG00000145794	*MEGF10*	N/A	N/A
53	57468	ENSG00000124140	*SLC12A5*	GABAergic synapse (KEGG PATHWAY); Transmembrane transport of small molecules (Reactome)	N/A
53	57727	ENSG00000124160	*NCOA5*	N/A	N/A
54	26747	ENSG00000083635	*NUFIP1*	N/A	N/A
54	55425	ENSG00000133114	*LSR7*	N/A	N/A
57	23301	ENSG00000115504	*EHBP1*	N/A	Prostate cancer, hereditary, 12 (OMIM); prostate cancer (GAD); CANCER (GAD)
57	5013	ENSG00000115507	*OTX1*	N/A	N/A
58	51390	ENSG00000146416	*AIG1*	N/A	N/A
59	9053	ENSG00000135525	*MAP7*	N/A	N/A
60	9821	ENSG00000023287	*RB1CC1*	mTOR signaling pathway (PID Curated)	Breast cancer, somatic (OMIM); Alzheimer's disease (FunDO); Embryoma (FunDO); Breast cancer (FunDO)
60	389658	ENSG00000196711	*FAM150A*	N/A	N/A
62	27332	ENSG00000075292	*ZNF638*	N/A	N/A
63	285346	ENSG00000178917	*ZNF852*	N/A	N/A

63	353274	ENSG00000185219	*ZNF445*	Gene Expression (Reactome)	N/A

Table S19b.1. Pathway and disease annotation of genes in regions of the selective sweep screen which have a 20-fold enrichment over random (Table S19a.3 in SI 19a). Annotations were obtained using KOBAS 2.0. The screen was performed using an African-American genetic map[9]. Highlighted in red are genes in the 15 regions that were also found to have a 20-fold enrichment over random when using two other recombination maps: the HapMap[10] and the deCode[11] maps, and may thereby have higher probability to have been under selection (Table S19a.5 in SI 19a).

2. Ontology enrichment

For the analysis of functional or phenotypic enrichment within the putatively selected regions we first created 100 sets of 63 random regions of the genome; each region having the same size as each the 63 putatively selected regions. To control for differences in purifying selection and gene content, we required each random region in the simulated sets to have a B score[12] within 100 points of the B score of the selected region to which it corresponded in size, as well as the same number of genes with Entrez Gene IDs. The R package *biomaRt*[13] was used to retrieve information from the Ensembl database.

To test for enrichment of Gene Ontology (GO) categories, we used the software FUNC[14]. Genes within a sweep region (or within 5,000 bp of its borders) were scored according to their rank, and genes within a simulated region (or within 5,000 bp of its borders) were scored with a rank of "0". No GO category with fewer than 2 genes was considered for the enrichment analysis. Wilcoxon rank sum tests were performed using as background each of the 100 sets of simulated regions. Categories with p-values < 0.1 and false discovery rate (FDR) < 0.1 were merged for each of these tests, and an adjustment for the p-value over 100 tests was performed using the *stats* package in R and the Benjamini-Hochberg correction[15]. Only one category with adjusted p-value of < 0.1, "single-stranded DNA binding" (adjusted p-value = 0.06), was identified. The region contains the genes HSPD1 and MCM4.

3. Interaction partners

We sought to find whether the putatively selected regions were enriched for genes that interact with a large number of other genes. Interaction partners for the proteins encoded by genes in the putatively selected regions were retrieved from the NCBI PubMed gene articles. We retrieved physical interaction information from three different sources (HPRD, BioGRID, BIND) and removed duplicate interaction partners. For each gene, we counted the number of interaction partners. We compared genes in the selective sweep regions to genes in the 100 simulated sets of similar regions (Figure S19b.2A), and found that 93 of the simulations show a lower average number of interaction partners than the average number found for the genes in the putatively selected regions (= 17.5). This is a slight but not significant enrichment for genes with many interactions.

For each interaction partner, we also counted the number of genes within the putatively selected regions that it interacts with, which we defined as "shared interactions." We compared the average number of shared interactions within the selective sweep regions (1.49) and within the 100 simulated sets of regions. We observe only one case with a higher number of shared interactions than the putatively selected regions (Figure S19b.2B), and the average number of shared interactions for the putatively selected regions is 2.9 standard deviations above the median for the 100 simulated sets. This suggests that proteins encoded by genes in the putatively selected regions may be significantly connected through their interaction partners, and may thus influence the same pathways.

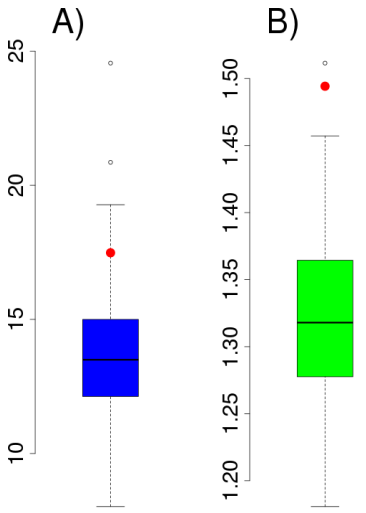


Figure S19b.2. A) Average number of interaction partners for 100 simulations (boxplot) and the selective sweep regions (red dot). B) Average number of shared interactions within 100 simulations (boxplot) and the selective sweep regions (red dot).

4. Overlap with expression changes on the human lineage

Some of the changes in gene expression on the human lineage may have been driven by selection after the split between Neandertals and modern humans. Brawand et al.[16] performed high throughput RNA sequencing for six tissues in a set of mammals. Among the genes that were found to be significantly differentially expressed in humans compared to other great apes and macaques, eight overlap with genes in the putatively selected regions (Table S19b.2). This is not a significant enrichment when compared to the 100 sets of simulated regions ($p > 0.05$). Among these genes, *MCAM* and *ARHGAP1* may have contributed to morphological differences in the modern human lineage, since they are involved in anatomical structure morphogenesis[17] and associated with bone mineral density[18], respectively. *PC* codes for pyruvate carboxylase, which plays a role in the synthesis of the neurotransmitter glutamate[19] and in glucose metabolism[20].

Gene ID	Gene name	Tissue	Direction
ENSG00000076706	*MCAM*	Testis	up
ENSG00000083544	*TDRD3*	Testis	down
ENSG00000110497	*AMBRA1*	Cerebellum	up
ENSG00000164162	*ANAPC10*	Testis	up
ENSG00000173599	*PC*	Cerebellum	up
ENSG00000175220	*ARHGAP1*	Testis	up
ENSG00000180423	*HARBI1*	Testis	down
ENSG00000197771	*MCMBP*	Testis	up

Table S19b.2. Genes in the top 63 regions of the selective sweep screen that differ significantly in expression on the human lineage. Direction indicates whether the expression is up-regulated or down-regulated in humans compared to other great apes.

5. Overlap with catalog of modern human changes

We examined the overlap between the top 63 putatively selected regions of the genome, and the catalog of fixed and high-frequency derived changes on the modern human lineage (SI 18). We identified 2,123 fixed and high-frequency (>90%) derived single-nucleotide changes (SNCs) and 61 derived fixed and high-

frequency InDels within the sweep regions. Of these, four are non-synonymous SNCs, three are splice site SNCs, twenty-two are 3' UTR SNCs, nine are 5' UTR SNCs and one is an SNC located in a high-information position within a transcription factor binding motif in a regulatory region. In addition, we find a modern-human-specific InDel causing a frame-shift. However, this InDel is 2bp away from a second 1000G InDel, which suggests it may be a mapping artifact in the 1000G data. We list the missense, splice site, frame-shift and high-information changes in Table S19b.3.

Position (hg19)	Ancestral / derived alleles	Modern human derived frequency	Altai Neandertal / Denisova state (A=ancestral, H=like modern human major derived allele)	Consequence	Gene
chr11:66568548	G/T	99%	A/A,A/A	missense	*C11orf80*
chr11:119059199	C/G	99%	A/A,A/A	missense	*PDZD3*
chr1:14105639	A/G	99%	A/A,A/A	missense	*PRDM2*
chr8:53568742	T/C	fixed	A/A,A/A	missense	*RB1CC1*
chr4:119736176	C/T	97%	A/A,A/A	splice site	*SEC24D*
chr20:44676727	A/T	96%	A/A,A/A	splice site	*SLC12A5*
chr2:71607348	A/G	fixed*	A/A,A/A	splice site	*ZNF638*
chr8:48805814	*G-/GA*	*95%*	*A/A,A/A*	*splice site, frameshift*	*PRKDC*
chr8:53608138	C/G	fixed	A/A,A/A	high-information position in regulatory motif feature	In an intron of *RB1CC1*

Table S19b.3. Non-synonymous, splice site, frame-shift and regulatory high-information derived changes that are fixed or at high-frequency (>90%) in modern humans (where Denisova or Altai Neandertal have the ancestral state) and that lie within the top 63 regions of the selective sweep screen. Changes labeled "fixed*" are fixed in 1000G but have a dbSNP ID. Highlighted in red are changes in the 15 top-scoring regions that were also found independently using two other recombination maps (Table S19a.5 in SI 19a). The PRKDC InDel is shown in cursive because it lies nearby another 1000G InDel, and may be the result of mapping artifacts.

All four of the missense mutations are predicted by PolyPhen[21] and SIFT[22] to be benign / tolerated. One of them (chr1:14105639) is located in PRDM2, a tumor suppressor gene that codes for a zinc finger protein and may play a role in retinoblastoma and neuron differentiation[23]. Another mutation is in a splice site of SLC12A5, a gene coding for a neuron-specific K/Cl co-transporter[24] that is found in the human cortex[25] and has a critical role in the modulation of neuronal plasticity[26]. We also find a missense mutation in PDZD3: a gene expressed in the kidney and intestinal tract whose protein product regulates the activity of the enterotoxin receptor GUCY2C[27] and the pH regulator SLC9A3[28].

The only regulatory high-information SNC found in the screen (chr8:53608138) is in a motif feature that overlaps an ENCODE DNaseI hypersensitivity cluster and an H3K27Ac mark within an intron of RB1CC1 (Figure S19b.3), which also contains a modern-human-specific missense mutation. This gene codes for a tumor suppressor that may be involved in breast cancer[29] and muscoskeletal differentiation[30], and its insufficiency may induce neuronal atrophy in Alzheimer's brain tissues[31].

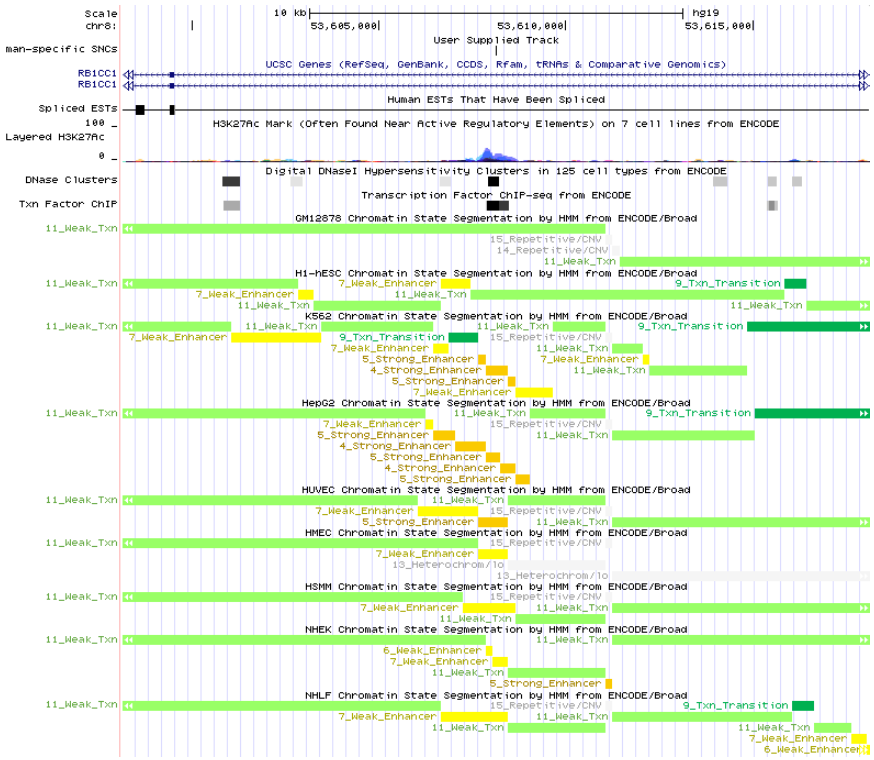


Figure S19b.3. UCSC genome browser views of the only modern-human-specific change lying within the top 63 selective sweep regions that is predicted by the VEP to be in a high-information position of a motif in an Ensembl regulatory region. The site lies within an H3K27Ac mark and a DNaseI hypersensitivity clusters, and is predicted by ChromHMM to be in a region of enhancer activity. Color codes for ChromHMM are: light green = weak transcription, dark green = transcriptional transition/elongation, yellow = weak enhancer, orange = strong enhancer. USCS Genome browser[32] screenshots were obtained from http://genome.ucsc.edu.

6. Disruptive changes in screen

We used the combined-annotation "C-scores" described in Supplementary Information 18 to rank the predicted disruptive effect of all SNCs and small InDels that lie within the selective sweep screen and that are specific to the modern human lineage, being globally fixed or at a high derived frequency (> 90%). In Figure S19b.4, we show the distribution of these scores for SNCs in the top 63 putatively selected regions compared to the 100 sets of regions with similar properties but that are not in the screen, sampled as described above. We do not observe a particularly strong proportion of disruptive or benign changes in the top 63 regions of the screen, which suggests that, as a whole, the changes in these regions are not particularly more or less disruptive than other regions of similar genomic characteristics. In Table S19b.4, we list the top 20 most disruptive fixed single-nucleotide changes (SNCs) and short InDels in the screen, and we list the top 20 most disruptive high-frequency SNCs and short InDels in the screen in Table S19b.5. For each of these changes, we show both the PHRED-scaled C-score and the GERP rejected substitution score. We exclude any InDels that are nearby other 1000G InDels, as they could be the result of mapping artifacts in the 1000G data.

The large majority of the most disruptive changes are in non-coding sites, and several of them are found in regulatory regions near genes. In Figure S19b.5, we present UCSC genome browser views of the top three most disruptive fixed modern-human-specific changes in the top 63 regions, which are found overlapping particularly strong signals for regulatory activity – including strong ENCODE H3K27Ac marks, UCSC genome browser ChromHMM track segments[33,34] corresponding to strong enhancer activity and clusters of DNase I hypersensitivity – as well as regions of high mammalian conservation scores[35].

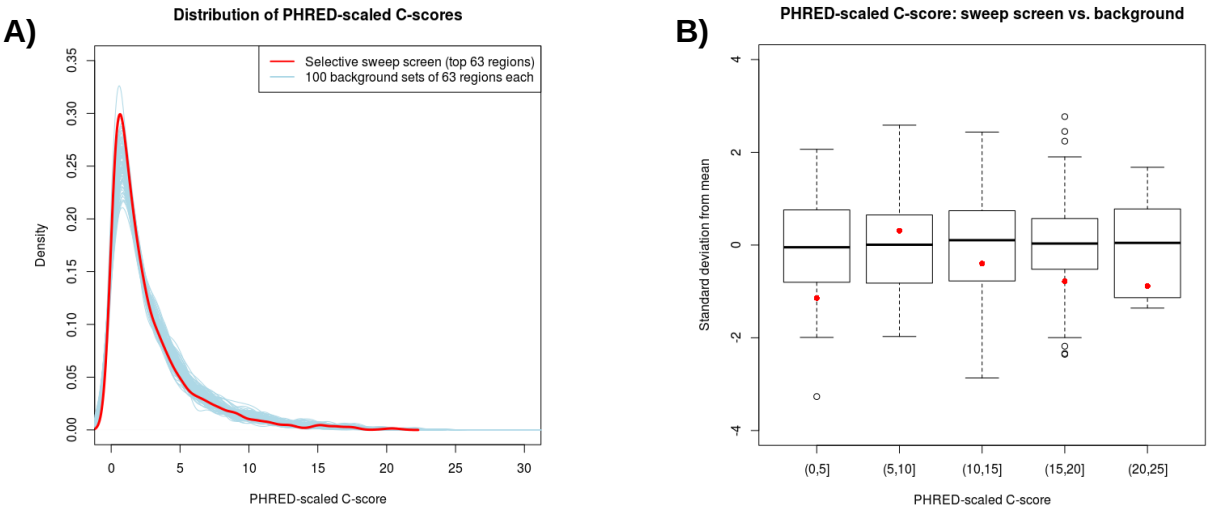


Figure S19b.4. A) Distribution of PHRED-scaled C-scores for SNCs in the putatively selected top 63 regions (red) and in 100 sets of 63 regions that are not ranked high in the screen but have similar genomic characteristics to the top 63 regions in the screen (blue). B) Boxplots showing deviations of the mean of each background set from the mean PHRED-scaled C-score across background sets for different bins of scores. The red dots in each bin correspond to the mean C-score of the top 63 regions of the selective sweep screen.

Many of the highly disruptive changes are found near genes affecting the cell cycle. Among the top 20 fixed changes (Table S19b.4), the most disruptive change is located in a regulatory region within an intron of *STK3*, which codes for a serine/threonine kinase involved in apoptotic activity[36]. A Drosophila homolog is known to regulate heart development[37]. The third most disruptive change also affects a regulatory region located in an intron of a serine/threonine kinase: *NLK*, a highly conserved gene whose expression leads to induced apoptosis in colon cancer cells[38]. Two highly disruptive fixed changes are located in *ANAPC10* (chr4:145981908 and chr4:145905234). This gene codes for an important subunit of the cyclosome, which plays an essential role in cell cycle regulation, by triggering the separation of sister chromatids and the termination of mitosis[39,40]. It is also one of the genes that are differentially expressed in humans (Table S19b.2).

A particularly interesting candidate is a fixed derived regulatory change (chr6:140783784) upstream of *CITED2*, in a strong enhancer region (as predicted by ChromHMM) that is ancestral in Denisova but derived in the Altai Neandertal. The gene CITED2 is involved in chromatin remodeling[41] and in the development of the heart and neural tube[42], and it appears to be a regulatory target of *FOXP2*[43,44], a gene involved in speech and language development. *CITED2* also contains a nearby intergenic change (chr6:140736356) that is highly disruptive, 99% derived in modern humans and ancestral in both Altai Neandertal and Denisovan (Table S19b.5).

Some of the changes are found near or inside genes related to sugar metabolism. We find a highly disruptive SNC (chr16:46963043) in the 3' UTR region of *GPT2* that is homozygous ancestral in Denisova but homozygous derived in Altai Neandertal. *GPT2* codes for an enzyme that catalyzes the transamination between 2-oxoglutarate and alanine, a key step in gluconeogenesis[45]. We also find a disruptive SNC in an intronic region of *PHKB*. Various coding, splice site and InDel mutations in this gene have been associated with glycogen storage diseases[46], resulting in hepatomegaly and reduced height and weight.

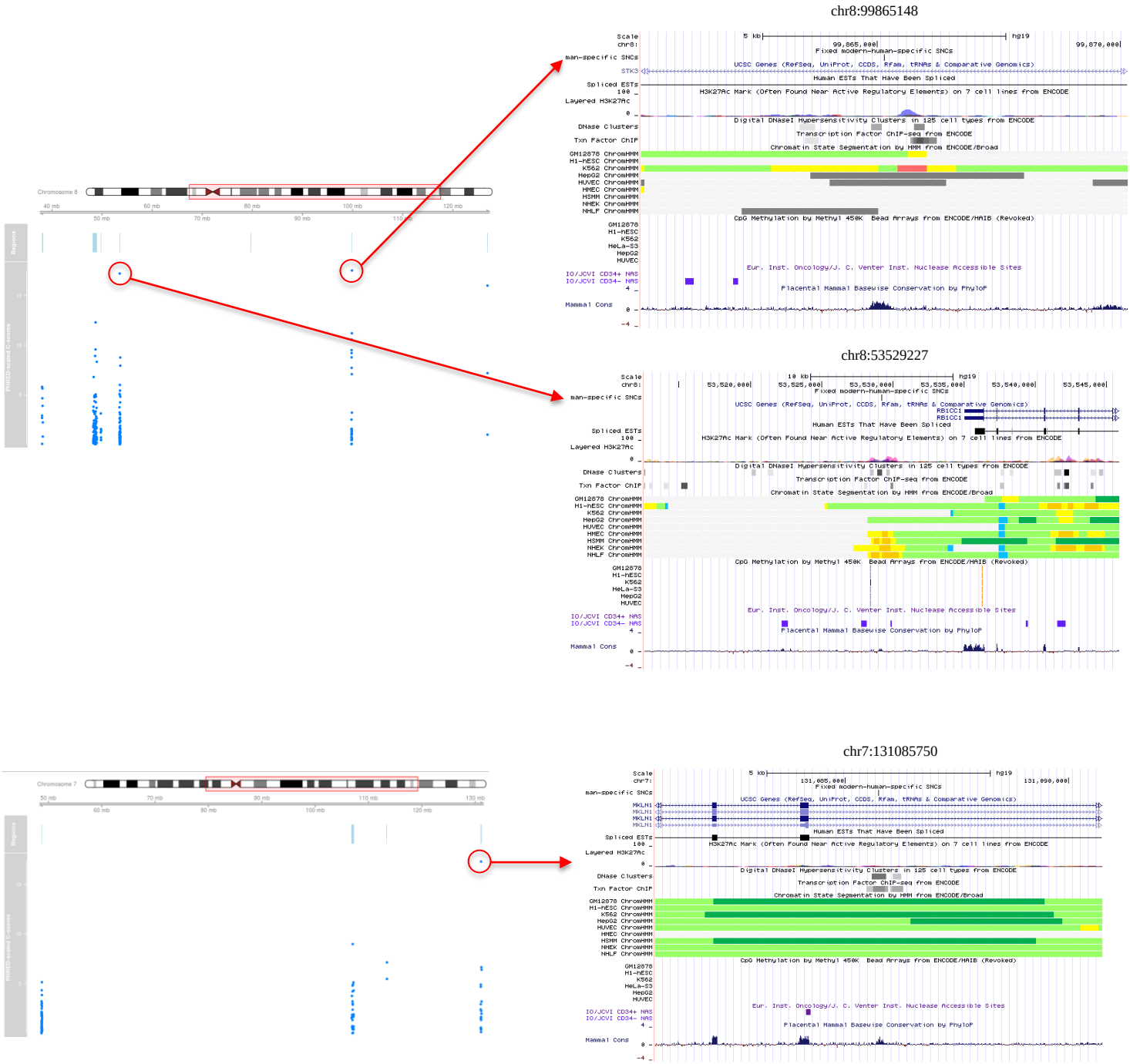


Figure S19b.5. UCSC genome browser views of the top 3 fixed modern-human-specific changes with highest C-scores that lie within the selective sweep screen: chr8:99865148 (both archaic humans homozygous ancestral), chr8: 53529227 (both archaic humans homozygous ancestral), chr7:131085750 (Altai Neandertal homozygous ancestral, Denisova homozygous derived). Each of these changes lie in genomic segments showing strong evidence for regulatory activity, based on various ENCODE regulatory marks (H3K27Ac, DNase clusters, transcription factor Chip-Seq) and genomic segmentation algorithms (ChromHMM[33]), and are located in clusters of high PhyloP[35] mammalian conservation scores. Color codes for ENCODE's ChromHMM are: light green = weak transcription, dark green = transcriptional transition/elongation, blue = insulator, yellow = weak enhancer, orange = strong enhancer, light red = weak promoter. Left images were created using the Gviz package in R Bioconductor. USCS Genome browser[32] screenshots were obtained from http://genome.ucsc.edu.

Also of note are two fixed changes in or nearby genes involved in nervous system activity or development. *GRID2*, which codes for a glutamate neurotransmitter receptor, contains a change that is fixed in 1000G but has a dbSNP ID (rs17019944), and ranks seventeenth among all the fixed changes in the screen. *GRID2* is specifically expressed in Purkinje cell synapses[47]. It has been associated with motor control and long-term depression[48], and its protein product may be a component of a synaptic organizing complex[49]. Additionally, the ninth most disruptive fixed change (chr13:58091727) is in an intergenic region whose closest gene is *PDH17*. This gene has a role in synaptic activity[50] and is specifically expressed in the cerebral cortex and superior temporal gyrus during fetal development[51]. We also find a high-frequency (not fixed) change near this gene that is highly disruptive (Table S19b.5).

Position (hg19)	Ancestral / derived alleles	Modern human derived frequency	dbSNP ID	Altai Neandertal / Denisova state (A=ancestral, H=like modern human major derived allele)	PHRED-scaled C-score	GERP rejected substitution score	Consequence	Gene [nearest]	Gene description / gene product function (Entrez Gene, UniProt, OMIM)
chr8:99865148	T/C	fixed*	rs184755172	A/A,A/A	17.53	5.5	INTRONIC / REGULATORY	STK3	Serine/threonine kinase involved in apoptosis
chr7:131085750	C/T	fixed	-	A/A,H/H	17.29	4.14	INTRONIC / REGULATORY	MKLN1	Mediator of cell spreading
chr8:53529227	T/C	fixed	-	A/A,A/A	17.22	0.825	INTERGENIC / REGULATORY	[RB1CC1]	Tumor suppressor, involved in muscoskeletal differentiation
chr8:126899188	C/T	fixed*	rs190223007	A/A,A/A	16	-2.6	INTERGENIC	[TRIB1]	May be involved in myelodysplastic syndromes and acute myeloid leukemia
chr17:26439781	T/C	fixed*	rs113795836	A/A,A/A	15.5	2.96	INTRONIC / REGULATORY / DOWNSTREAM	NLK	Serine/threonine kinase involved in apoptosis
chr16:46963043	C/T	fixed	-	H/H,A/A	15.46	1.63	3' UTR	GPT2	Glutamic pyruvate transaminase
chr16:47655884	C/T	fixed	-	A/A,H/H	15.38	-0.52	INTRONIC	PHKB	Subunit of phosphorylase kinase, involved in glycogen storage
chr16:47223245	AC/A-	fixed	-	A/A,A/A	15.36	1.36	INTRONIC	ITFG1	T-cell modulator
chr13:58091727	A/G	fixed	-	A/A,A/A	15.06	0.51	INTERGENIC	[PCDH17]	Protocadherin, differentially expressed in cortex of fetal brains
chr6:140783784	A/G	fixed	-	H/H,A/A	14.92	4.56	INTERGENIC / REGULATORY	[CITED2]	Transactivator associated with heart defects
chr16:47762604	T/C	fixed	-	A/A,H/H	14.58	0.87	INTRONIC	RP11-523L20.2	lincRNA
chr11:66505194	G/C	fixed	-	A/A,A/A	13.95	3.13	INTERGENIC	[C11orf80]	N/A
chr2:63206488	G/C	fixed*	rs146025524	A/A,A/A	13.1	4.06	INTRONIC	EHBP1	Endocyte trafficking, associated with prostate cancer
chr10:106243544	T/C	fixed*	rs143641704	A/A,A/A	12.91	0.945	UPSTREAM	RP11-127O4.3	lincRNA
chr8:48718321	C/A	fixed*	rs147652268	A/A,A/A	12.3	-0.42	INTRONIC / UPSTREAM	PRKDC	Catalytic subunit of DNA-PK, involved in transcriptional modulation and V(D)J recombination
chr4:145981908	T/A	fixed	-	A/A,A/A	12.12	0.051	INTRONIC	ANAPC10	Core subunit

									of cyclosome
chr4:93781683	G/A	fixed*	rs17019944	A/A,A/A	11.85	-1.72	INTRONIC	GRID2	Glutamate neurotransmitter receptor
chr4:145905234	A/T	fixed	-	H/H,A/A	11.76	2.06	INTRONIC	ANAPC10	Core subunit of cyclosome
chr18:19102852	T/C	fixed	-	A/A,A/A	11.6	5.76	3' UTR	GREB1L	N/A
chr11:66573255	A/G	fixed	-	A/A,A/A	11.38	0.94	INTRONIC	C11orf80	N/A

Table S19b.4. Top 20 most disruptive single-nucleotide changes (SNCs) and InDels that are fixed derived in modern humans, that are ancestral in Denisova or Altai Neandertal and that lie in the top 75 regions of the selective sweep screen. Changes labeled "fixed*" are fixed in 1000G but have a dbSNP ID. InDels nearby other InDels are shown in cursive, as they could be the result of mismapping or miscalling either in an archaic genome or in the 1000G Project data, and so should be treated with caution. A gene name in brackets refers to the gene nearest to a change in an intergenic region. Highlighted in red are changes in the 15 top-scoring regions that were also found independently using two other recombination maps (Table S19a.5 in SI 19a).

Three of the top 20 most disruptive high-frequency SNCs (Table S19b.5) in the screen are in regulatory regions near *SEC24D* (rs114019902, rs116514715, rs114678295), which codes for a transport protein involved in vesicle trafficking[52]. Two of these SNCs (rs116514715 and rs114678295) are adjacent to each other and lie within the peak of an ENCODE H3K27Ac mark and a DNAseI hypersensitivity cluster, which is suggestive of regulatory activity in the region, though their close proximity may also suggest mapping errors in the 1000G data.

The second and eighteenth most disruptive high-frequency SNCs are located in an intergenic region nearby ABCA13 (chr7:48720361 and chr7:48720277), an ATP-binding cassette transporter[53]. Variants in this gene have been associated with schizophrenia, depression and bipolar disorder[54], suggesting it may play a role in brain activity. The fourth most disruptive high-frequency change is located in an intronic region of AIG1 (chr6:143448406), a gene whose expression is sensitive to androgen and that may be involved in the regulation of hair growth[55]. There is also a disruptive high-frequency SNC in the CMTM1 gene, which is highly expressed in the testis[56]. Read-through transcripts exist containing both this gene and CKLF, which regulates the proliferation of skeletal muscle cells[57].

Position (hg19)	**Ancestral / derived alleles**	**Human derived frequency**	**dbSNP ID**	**Altai Neandertal / Denisova state (A=ancestral, H=like modern human major derived allele)**	**PHRED-scaled C-score**	**GERP rejected substitution score**	**Consequence**	**Gene [nearest]**	**Gene description / gene product activity (Entrez Gene, UniProt, OMIM)**
chr11:66508128	G/T	99%	rs115469135	A/A,A/A	21	5.71	UPSTREAM	C11orf80	N/A
chr7:48720361	A/G	97%	rs78071946	A/A,A/A	20.6	4.51	INTERGENIC	[ABCA13]	Associated with schizophrenia and bipolar disorder
chr18:75979445	T/C	98%	rs73495877	A/A,A/A	20.3	3.79	INTERGENIC	[SALL3]	Zinc finger protein
chr6:143448406	C/T	96%	rs9403446	A/A,A/A	20.2	4.79	INTRONIC	AIG1	May be involved in hair follicle growth
chr17:26457614	A/T	97%	rs75032756	A/A,A/A	19.57	5.25	INTRONIC	NLK	Serine/threonine kinase involved in apoptosis
chr4:119755564	A/G	97%	rs114019902	A/A,A/A	17.92	4.05	INTRONIC / REGULATORY	SEC24D	Involved in vesicle trafficking
chr14:83547201	A/G	95%	rs77074144	A/A,A/A	17.86	2.47	INTERGENIC	[SEL1L]	May be involved in ER-associated degradation
chr6:127498350	A/G	99%	rs73771610	A/A,A/A	17.34	0.842	INTRONIC / REGULATORY	RSPO3	Involved in proliferation of epithelium in gastrointestinal tract, associated with colon cancer

chr14:83547431	C/T	95%	rs77892409	A/A,A/A	17.18	-0.249	INTERGENIC	[SEL1L]	Involved in dislocation of misfolded proteins in the ER
chr1:14020411	CAG/C--	99%	-	A/A,A/A	16.82	2.89	DOWNSTREAM	SCARNA11	Small nucleolar RNA (snoRNA) specific to Cajal bodies
chr4:119756280	T/G	97%	rs116514715	A/A,A/A	16.67	4.52	INTRONIC / REGULATORY	SEC24D	Involved in vesicle trafficking
chr16:66605008	TG/T-	92%	-	A/A,A/A	16.6	1.93	INTRONIC / DOWNSTREAM	CMTM1	Chemokine-like factor
chr4:119756281	T/C	97%	rs114678295	A/A,A/A	16.59	4.52	INTRONIC / REGULATORY	SEC24D	Involved in vesicle trafficking
chr6:140736356	T/C	99%	rs73777324	A/A,A/A	16.42	3.39	INTERGENIC	[CITED2]	Transactivator associated with heart defects
chr11:66478115	C/A	99%	rs34275473	A/A,A/A	16.34	4.06	DOWNSTREAM / SYNONYMOUS	Metazoa SRP	Signal recognition particle RNA
chr13:58183927	G/C	96%	rs112860980	A/A,A/A	16.28	3.27	INTERGENIC	[PCDH17]	Protocadherin, differentially expressed in cortex of fetal brains
chr2:143441362	A/G	92%	rs11893266	A/A,A/A	16.22	-0.028	INTERGENIC	[KYNU]	Kynureninase, involved in tryptophan metabolism
chr7:48720277	C/T	97%	rs77014876	A/A,A/A	15.95	1.76	INTERGENIC	[ABCA13]	Associated with schizophrenia and bipolar disorder
chr9:115251691	C/A	93%	rs4434682	A/A,A/A	15.5	0.721	INTRONIC / REGULATORY / UPSTREAM	KIAA1958	N/A
chr7:131113586	G/A	99%	rs114153487	A/A,A/A	15.24	-1.2	INTRONIC	MKLN1	Mediator of cell spreading

Table S19b.5. Top 20 most disruptive single-nucleotide changes (SNCs) and small InDels that are high-frequency derived (>90%) but not fixed in modern humans, that are ancestral in Denisova or Altai Neandertal and that lie in the top 75 regions of the selective sweep screen. InDels nearby other InDels are shown in cursive, as they could be the result of mismapping or miscalling either in an archaic genome or in the 1000G Project data, and so should be treated with caution. A gene name in brackets refers to the gene nearest to a change in an intergenic region. Highlighted in red are changes in the 15 top-scoring regions that were also found independently using two other recombination maps (Table S19a.5 in SI 19a).

7. Conclusion

Among the genes in the screen, we observe an overrepresentation of genes related to DNA binding. We also find a significant number of shared interactions among the genes in the screen. We believe this may suggest that regulatory changes affecting similar processes could have swept to fixation or near fixation in recent human evolution since the split between modern and archaic humans. The catalog of modern human-specific differences allows us to prioritize those genes for experimental testing based on the phenotypes or biological functions that may have been affected by these changes. Among the changes ranked highest in their probability for leading to some form of disruption, we observe a number of fixed and high-frequency derived SNCs in regions with strong evidence for regulatory activity in or nearby genes involved in biological activities, including cell cycle regulation, sugar metabolism, neurotransmitter receptor activity and hair growth, among others. Experimental testing will be required to assess whether any of these changes are actually causative of particular changes in human phenotypes, and whether or not they played a role during our evolutionary history.

8. References

1 Xie, C. *et al.* KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res* **39**, W316-322, doi:10.1093/nar/gkr483 (2011).

2 Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The genetic association database. *Nat Genet* **36**, 431-432, doi:10.1038/ng0504-431 (2004).

3 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30 (2000).

4 Osborne, J. D. *et al.* Annotating the human genome with Disease Ontology. *BMC Genomics* **10 Suppl 1**, S6, doi:10.1186/1471-2164-10-S1-S6 (2009).

5 Du, P. *et al.* From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics* **25**, i63-68, doi:10.1093/bioinformatics/btp193 (2009).

6 Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res* **37**, D674-679, doi:10.1093/nar/gkn653 (2009).

7 Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**, 2129-2141, doi:10.1101/gr.772403 (2003).

8 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-9367, doi:10.1073/pnas.0903103106 (2009).

9 Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170-175, doi:10.1038/nature10336 (2011).

10 Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861, doi:10.1038/nature06258 (2007).

11 Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-1103, doi:10.1038/nature09525 (2010).

12 McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics* **5**, e1000471, doi:10.1371/journal.pgen.1000471 (2009).

13 Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439-3440, doi:10.1093/bioinformatics/bti525 (2005).

14 Prufer, K. *et al.* FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC bioinformatics* **8**, 41, doi:10.1186/1471-2105-8-41 (2007).

15 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289-300 (1995).

16 Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343-348, doi:10.1038/nature10532 (2011).

17 Sers, C., Kirsch, K., Rothbächer, U., Riethmüller, G. & Johnson, J. P. Genomic organization of the melanoma-associated glycoprotein MUC18: implications for the evolution of the immunoglobulin domains. *Proc Natl Acad Sci U S A* **90**, 8514-8518 (1993).

18 Rivadeneira, F. *et al.* Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat Genet* **41**, 1199-1206, doi:10.1038/ng.446 (2009).

19 Gamberino, W. C., Berkich, D. A., Lynch, C. J., Xu, B. & LaNoue, K. F. Role of pyruvate carboxylase in facilitation of synthesis of glutamate and glutamine in cultured astrocytes. *J Neurochem* **69**, 2312-2325 (1997).

20 Liu, Y. Q., Han, J., Epstein, P. N. & Long, Y. S. Enhanced rat beta-cell proliferation in 60% pancreatectomized islets by increased glucose metabolic flux through pyruvate carboxylase

pathway. *Am J Physiol Endocrinol Metab* **288**, E471-478, doi:10.1152/ajpendo.00427.2004 (2005).
21 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7.20, doi:10.1002/0471142905.hg0720s76 (2013).
22 Ng, P. C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* **7**, 61-80, doi:10.1146/annurev.genom.7.080505.115630 (2006).
23 Buyse, I. M., Shao, G. & Huang, S. The retinoblastoma protein binds to RIZ, a zinc-finger protein that shares an epitope with the adenovirus E1A protein. *Proc Natl Acad Sci U S A* **92**, 4467-4471 (1995).
24 Rivera, C. *et al.* The K+/Cl- co-transporter KCC2 renders GABA hyperpolarizing during neuronal maturation. *Nature* **397**, 251-255, doi:10.1038/16697 (1999).
25 Szabadics, J. *et al.* Excitatory effect of GABAergic axo-axonic cells in cortical microcircuits. *Science* **311**, 233-235, doi:10.1126/science.1121325 (2006).
26 Blaesse, P., Airaksinen, M. S., Rivera, C. & Kaila, K. Cation-chloride cotransporters and neuronal function. *Neuron* **61**, 820-838, doi:10.1016/j.neuron.2009.03.003 (2009).
27 Scott, R. O., Thelin, W. R. & Milgram, S. L. A novel PDZ protein regulates the activity of guanylyl cyclase C, the heat-stable enterotoxin receptor. *J Biol Chem* **277**, 22934-22941, doi:10.1074/jbc.M202434200 (2002).
28 Zachos, N. C. *et al.* Elevated intracellular calcium stimulates NHE3 activity by an IKEPP (NHERF4) dependent mechanism. *Cell Physiol Biochem* **22**, 693-704, doi:10.1159/000185553 (2008).
29 Chano, T., Kontani, K., Teramoto, K., Okabe, H. & Ikegawa, S. Truncating mutations of RB1CC1 in human breast cancer. *Nat Genet* **31**, 285-288, doi:10.1038/ng911 (2002).
30 Chano, T., Saeki, Y., Serra, M., Matsumoto, K. & Okabe, H. Preferential expression of RB1-inducible coiled-coil 1 in terminal differentiated musculoskeletal cells. *The American journal of pathology* **161**, 359-364, doi:10.1016/S0002-9440(10)64190-9 (2002).
31 Chano, T., Okabe, H. & Hulette, C. M. RB1CC1 insufficiency causes neuronal atrophy through mTOR signaling alteration and involved in the pathology of Alzheimer's diseases. *Brain Res* **1168**, 97-105, doi:10.1016/j.brainres.2007.06.075 (2007).
32 Meyer, L. R. *et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* **41**, D64-69, doi:10.1093/nar/gks1048 (2013).
33 Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-216, doi:10.1038/nmeth.1906 (2012).
34 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906 (2011).
35 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121, doi:10.1101/gr.097857.109 (2010).
36 O'Neill, E., Rushworth, L., Baccarini, M. & Kolch, W. Role of the kinase MST2 in suppression of apoptosis by the proto-oncogene product Raf-1. *Science* **306**, 2267-2270, doi:10.1126/science.1103233 (2004).
37 Heallen, T. *et al.* Hippo pathway inhibits Wnt signaling to restrain cardiomyocyte proliferation and heart size. *Science* **332**, 458-461, doi:10.1126/science.1199010 (2011).
38 Yasuda, J. *et al.* Nemo-like kinase induces apoptosis in DLD-1 human colon cancer cells. *Biochem Biophys Res Commun* **308**, 227-233 (2003).
39 Grossberger, R. *et al.* Characterization of the DOC1/APC10 subunit of the yeast and the human anaphase-promoting complex. *J Biol Chem* **274**, 14500-14507 (1999).

40 Wendt, K. S. *et al.* Crystal structure of the APC10/DOC1 subunit of the human anaphase-promoting complex. *Nat Struct Biol* **8**, 784-788, doi:10.1038/nsb0901-784 (2001).

41 Bhattacharya, S. *et al.* Functional role of p35srj, a novel p300/CBP binding protein, during transactivation by HIF-1. *Genes Dev* **13**, 64-75 (1999).

42 Yin, Z. *et al.* The essential role of Cited2, a negative regulator for HIF-1alpha, in heart development and neurulation. *Proc Natl Acad Sci U S A* **99**, 10488-10493, doi:10.1073/pnas.162371799 (2002).

43 Vernes, S. C. *et al.* A functional genetic link between distinct developmental language disorders. *N Engl J Med* **359**, 2337-2345, doi:10.1056/NEJMoa0802828 (2008).

44 Nelson, C. S. *et al.* Microfluidic affinity and ChIP-seq analyses converge on a conserved FOXP2-binding motif in chimp and human, which enables the detection of evolutionarily novel targets. *Nucleic Acids Res* **41**, 5991-6004, doi:10.1093/nar/gkt259 (2013).

45 Yang, R. Z., Blaileanu, G., Hansen, B. C., Shuldiner, A. R. & Gong, D. W. cDNA cloning, genomic structure, chromosomal mapping, and functional expression of a novel human alanine aminotransferase. *Genomics* **79**, 445-450, doi:10.1006/geno.2002.6722 (2002).

46 Burwinkel, B. *et al.* Autosomal glycogenosis of liver and muscle due to phosphorylase kinase deficiency is caused by mutations in the phosphorylase kinase beta subunit (PHKB). *Hum Mol Genet* **6**, 1109-1115 (1997).

47 Landsend, A. S. *et al.* Differential localization of delta glutamate receptors in the rat cerebellum: coexpression with AMPA receptors in parallel fiber-spine synapses and absence from climbing fiber-spine synapses. *J Neurosci* **17**, 834-842 (1997).

48 Kashiwabuchi, N. *et al.* Impairment of motor coordination, Purkinje cell synapse formation, and cerebellar long-term depression in GluR delta 2 mutant mice. *Cell* **81**, 245-252 (1995).

49 Matsuda, K. *et al.* Cbln1 is a ligand for an orphan glutamate receptor delta2, a bidirectional synapse organizer. *Science* **328**, 363-368, doi:10.1126/science.1185152 (2010).

50 Suzuki, S. T. Recent progress in protocadherin research. *Exp Cell Res* **261**, 13-18, doi:10.1006/excr.2000.5039 (2000).

51 Abrahams, B. S. *et al.* Genome-wide analyses of human perisylvian cerebral cortical patterning. *Proc Natl Acad Sci U S A* **104**, 17849-17854, doi:10.1073/pnas.0706128104 (2007).

52 Pagano, A. *et al.* Sec24 proteins and sorting at the endoplasmic reticulum. *J Biol Chem* **274**, 7833-7840 (1999).

53 Prades, C. *et al.* The human ATP binding cassette gene ABCA13, located on chromosome 7p12.3, encodes a 5058 amino acid protein with an extracellular domain encoded in part by a 4.8-kb conserved exon. *Cytogenet Genome Res* **98**, 160-168, doi:69852 (2002).

54 Knight, H. M. *et al.* A cytogenetic abnormality and rare coding variants identify ABCA13 as a candidate gene in schizophrenia, bipolar disorder, and depression. *Am J Hum Genet* **85**, 833-846, doi:10.1016/j.ajhg.2009.11.003 (2009).

55 Seo, J., Kim, J. & Kim, M. Cloning of androgen-inducible gene 1 (AIG1) from human dermal papilla cells. *Mol Cells* **11**, 35-40 (2001).

56 Han, W. *et al.* Identification of eight genes encoding chemokine-like factor superfamily members 1-8 (CKLFSF1-8) by in silico cloning and experimental validation. *Genomics* **81**, 609-617 (2003).

57 Han, W. *et al.* Molecular cloning and characterization of chemokine-like factor 1 (CKLF1), a novel human cytokine with unique structure and potential chemotactic activity. *Biochem J* **357**, 127-135 (2001).

Supplementary Information 20

Expression pattern enrichment among genes with protein coding changes in humans

Trygve Bakken and Ed Lein

Correspondence should be addressed (trygveb@alleninstitute.org or EdL@alleninstitute.org)

The expression patterns of the 87 genes in which protein coding changes are seen in present-day humans, but which differ from Neandertal, Denisovan and the great apes, are of particular interest to understanding the potential functional impact that recent human-specific changes might have on the human phenotype. To determine whether there is evidence for enrichment of the 87 genes with fixed derived changes in humans in specific regions or layers of the brain, or with a particular developmental expression patterns we compared these genes to a background set of 108 genes containing fixed derived synonymous variants (11 of these genes also contain non-synonymous derived variants).

We looked for spatiotemporal patterns of gene expression in developing and adult human and macaque monkey brains using data from four brain atlases. These data sets included the Allen Human Brain Atlas (adult) (Hawrylycz et al. 2012), BrainSpan Atlas of the Developing Human Brain (Kang et al. 2011), NIH Blueprint Non-Human Primate Atlas (http://www.blueprintnhpatlas.org/macrodissection/index), and adult macaque monkey microarray atlas (http://www.blueprintnhpatlas.org/nhp/download.html). In particular, we focused on the portion of the BrainSpan atlas that includes expression data from proliferative and post-mitotic cell layers of mid-fetal human brain because genes expressed in these regions are critical in determining the size and distribution of cortical areas in the adult human brain. These data included ~20 laser microdissected (LMD) regions of neocortex from two pairs of mid-fetal human brains of similar ages (15-16 and 21 post-conceptional weeks). LMD was used to capture 9 cytoarchitectonically distinct layers from the apical to basal side of developing neocortex, from the ventricular zone to the subgranular zone. Gene expression was assayed in each layer by genome-wide microarray.

For each data set, we first looked for enrichment in genes with non-synonymous and synonymous changes for a general spatiotemporal feature. Then, for each general feature, we tested for enrichment for the most common specific feature seen in non-synonymous and synonymous gene sets. We tested for the significance of both general and specific spatiotemporal patterns for 7 categories of patterns. Several of these tests looked for enrichment for Weighted Gene Co-expression Network Analysis (WGCNA) modules since WGCNA is an effective statistical method for capturing the major correlated trends in gene expression across brain regions and over time (Langfelder et al. 2008). In all, we performed 28 hypergeometric enrichment tests (14 non-synonymous and 14 synonymous) and report corresponding Bonferroni corrected p-values (Table S20.1).

We found that 81/87 genes with non-synonymous changes and 108/108 genes with synonymous changes are expressed sometime during cortical development in mid-fetal human brain. Greater than 90% of all genes are expressed somewhere in the developing cortex, and approximately 40% of genes have a spatiotemporal pattern of some sort. We did not find significant enrichment for general patterning (i.e. all types of patterns, including laminar, temporal and areal gradient expression) relative to background for non-synonymous (37/81; 45%) or synonymous (44/108; 41%) genes.

On the other hand, we did find a significant difference between the non-synonymous and synonymous mutations when considering specific laminar patterns, and particularly enrichment in the ventricular zone. 14/81 (17%) genes with non-synonymous changes showed laminar expression

in the ventricular zone as compared to 1497/20268 genes genome-wide. This gave an enrichment (hypergeometric) nominal p-value = 0.0023 (Bonferroni corrected p = 0.06). Only 10/108 (9%) genes with synonymous changes showed laminar expression in the ventricular zone (nominal p = 0.28; Bonferroni corrected p = 1).

This enriched expression in the ventricular zone led us to test for a significant difference in expression between genes with non-synonymous and synonymous changes in all cortical layers of mid-fetal human brain. For each set of genes, we calculated the average proportion of gene expression in each cortical layer across all genes in that set (Figure S20.1). We calculated average proportional expression by layer in each of four mid-fetal brains and used ANOVA to show that, among genes with non-synonymous changes, there was significantly higher expression in proliferating cell layers and significantly lower expression in the subplate (Bonferroni corrected $p < 0.05$). These differences were consistent across the four brains (15-21 post-conceptional weeks). In order to convince ourselves that these differences did not occur by chance, we selected 1000 random sets of 100 genes that are expressed in developing neocortex and calculated the average proportion of expression by layer for each random set of genes. Genes with non-synonymous changes have, on average, a greater proportion of their expression in both ventricular and subventricular zones than 99% of these random sets of genes.

In addition to enriched expression in proliferative cell layers, we also found nominally significant enrichment for areal patterning across the developing cortex in genes with non-synonymous but not synonymous changes. 6/81 non-synonymous genes expressed in developing cortex of mid-fetal human brain showed frontotemporal gradient expression (rostral-caudal: *SLITRK1*, *TKTL1*; caudal-rostral: *C21orf62*, *GLDC*, *IFI44L* and *ZNF185*) as compared to 468/20268 genes genome-wide (nominal p = 0.011; Bonferroni corrected p = 0.31). In contrast, only 1/108 synonymous genes (*CTNNBL1*) showed a rostral-caudal gradient (nominal p = 0.92).

The premise of this analysis was that many of the genes with non-synonymous changes would target the same biological process or cell types. Single gene changes could also have profound effects, and some genes with human-specific amino acid substitutions have known relationships to cortical development and neuropsychiatric disease. For example, *CASC5* is expressed in early fetal human brain in proliferating cortical layers, is involved in cell cycle control (Kiyomitsu et al. 2011), and gene mutations can cause primary microcephaly (Genin et al. 2012). *SLITRK1* is enriched in the cortical subplate of mid-fetal human, shows frontotemporal gradient expression, is thought to be involved in neurite outgrowth (Marteyn et al. 2011), and gene mutations are associated with Tourette's syndrome (Abelson et al. 2005).

References

Abelson, J. F., Kwan, K. Y., O'Roak, B. J., Baek, D. Y., Stillman, A. A., Morgan, T. M., … State, M. W. (2005). Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science*, *310*, 317–20. doi:10.1126/science.1116502

Genin, A., Desir, J., Lambert, N., Biervliet, M., Van Der Aa, N., Pierquin, G., … Abramowicz, M. (2012). Kinetochore KMN network gene CASC5 mutated in primary microcephaly. *Human molecular genetics*, *21*(24), 5306–17. doi:10.1093/hmg/dds386

Hawrylycz, M. J., Lein, E. S., Guillozet-bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., … Riley, Z. L. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, *489*, 391–399. doi:10.1038/nature11405

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., … Sestan, N. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, *478*(7370), 483–9. doi:10.1038/nature10523

Kiyomitsu, T., Murakami, H., & Yanagida, M. (2011). Protein interaction domain mapping of human kinetochore protein Blinkin reveals a consensus motif for binding of spindle assembly checkpoint proteins Bub1 and BubR1. *Molecular and cellular biology*, *31*(5), 998–1011. doi:10.1128/MCB.00815-10

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559. doi:10.1186/1471-2105-9-559

Marteyn, A., Maury, Y., Gauthier, M. M., Lecuyer, C., Vernet, R., Denis, J. A., … Martinat, C. (2011). Mutant human embryonic stem cells reveal neurite and synapse formation defects in type 1 myotonic dystrophy. *Cell stem cell*, *8*, 434–444.

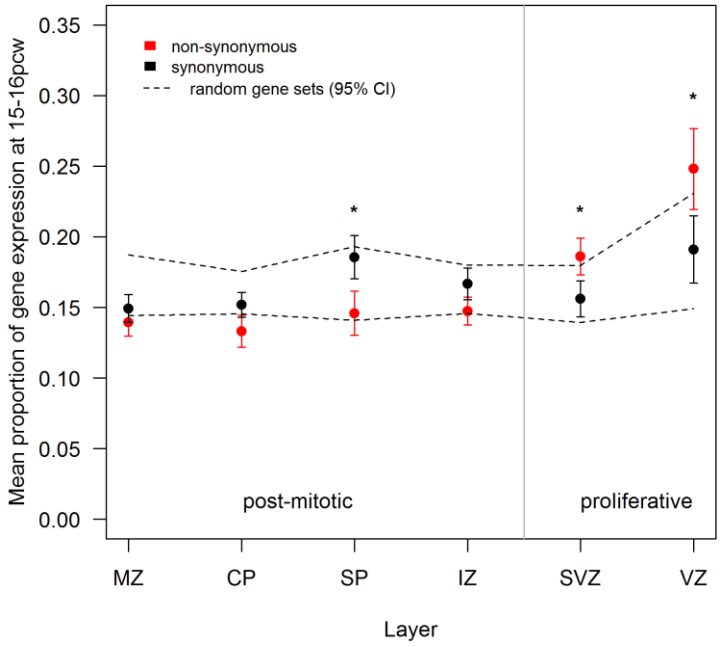


Figure S20.1. Proportion of gene expression (mean ± SEM) in each layer in mid-fetal (15-16 post-conceptional week) human brain. For each group of genes, the proportional expression in each layer was averaged across all genes that showed a significant difference in expression across layers (as assessed by a Bonferroni corrected ANOVA p-value < 0.0005) and that showed a high degree of expression correlation ($r > 0.8$) between 15 and 16 pcw brains. Dotted lines indicate the 95% confidence interval for the mean proportional expression in each layer for 1000 sets of 100 genes selected at random from all genes expressed during human brain development. * Bonferroni corrected $p < 0.05$. MZ – marginal zone; CP – cortical plate; SP – subplate zone; IZ – intermediate zone; SVZ – subventricular zone; VZ – ventricular zone.

doi:10.1038/nature12886

Gene list	Allen Atlas data set	Test Category	# Annotated Genes in data set	# Genes in test category	# Genes in list that are annotated in data set	# Genes in list that are in category	Nominal p-value	Bonf corr p	Most common category match	# Genes in data set in most common category	# Genes in list in most common category	Nominal p-value	Bonf corr p
Non-syn	Dev human	Spatiotemporal pattern (WGCNA module)	8485	8237	37	37	0.3329	1	M1	4655	21	0.4756	1
Non-syn	Dev human LMD	Laminar distribution (15/16pcw)	20268	2304	81	20	0.0006	0.016	Ventricular zone	1497	14	0.0023	0.064
Non-syn	Dev human LMD	Laminar distribution (21/21pcw)	20268	2316	81	9	0.5877	1	Cortical plate	776	3	0.6042	1
Non-syn	Dev human LMD	Areal gradient (15-21pcw)	20268	468	81	6	0.0111	0.311	Caudal	180	4	0.0059	0.165
Non-syn	Adult human	Spatial pattern (WGCNA module)	21337	15347	76	52	0.7926	1	M2 (turquoise)	5110	16	0.7630	1
Non-syn	Dev macaque	Spatiotemporal pattern (WGCNA module)	11102	10089	44	41	0.4207	1	turquoise	2062	8	0.5878	1
Non-syn	Adult macaque	Spatial pattern (WGCNA module)	10665	8991	43	34	0.8740	1	turquoise	3154	10	0.8603	1
Syn	Dev human	Spatiotemporal pattern (WGCNA module)	8485	8237	42	39	0.9665	1	M1	4655	20	0.8644	1
Syn	Dev human LMD	Laminar distribution (15/16pcw)	20268	2304	108	21	0.0095	0.267	Ventricular zone	1497	10	0.2751	1
Syn	Dev human LMD	Laminar distribution (21/21pcw)	20268	2316	108	14	0.3501	1	Cortical plate	776	6	0.2331	1
Syn	Dev human LMD	Areal gradient (15-21pcw)	20268	468	108	1	0.9203	1	none	0	0	1.0000	1
Syn	Adult human	Spatial pattern (WGCNA module)	21337	15347	103	71	0.7864	1	M2 (turquoise)	5110	22	0.7653	1
Syn	Dev macaque	Spatiotemporal pattern (WGCNA module)	11102	10089	69	58	0.9788	1	blue	1609	16	0.0354	0.990
Syn	Adult macaque	Spatial pattern (WGCNA module)	10665	8991	64	48	0.9825	1	turquoise	3154	19	0.5393	1

Table S20.1. Summary of tests for enrichment of spatiotemporal expression patterns among genes with non-synonymous (non-syn) and synonymous (syn) substitutions. Data sets: Dev human = BrainSpan developmental atlas (early fetal through adulthood) microarray (http://www.brainspan.org/rnaseq/search/index.html); Dev human LMD = BrainSpan atlas mid-fetal human laser microdissected microarray (http://www.brainspan.org/lcm/search/index.html); Adult human = Allen Human Brain Atlas microarray (http://human.brain-map.org/). Dev macaque = NIH Blueprint NHP Atlas – postnatal (0-48 months) macaque monkey microarray (http://www.blueprintnhpatlas.org/macrodissection/index). Adult macaque = Adult macaque monkey laser microdissected microarray (http://www.blueprintnhpatlas.org/nhp/download.html). Test categories: WGCNA module = sets of genes with correlated expression over brain structures and/or time discovered using a data-driven method called Weighted Gene Co-expression Network Analysis; Laminar distribution = laminar expression in mid-fetal human cortex (i.e. expression correlation r > 0.5 with a binary layer template); Areal gradient = frontotemporal gradient expression in mid-fetal human cortex (i.e. expression correlation r > 0.5 with an angular vector indicating rostral to caudal position on the cortical surface).