www.sciencemag.org/cgi/content/full/328/5979/710/DC1

Supporting Online Material for

A Draft Sequence of the Neandertal Genome

Richard E. Green,* Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, Nancy F. Hansen, Eric Y. Durand, Anna-Sapfo Malaspinas, Jeffrey D. Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prüfer, Matthias Meyer, Hernán A. Burbano, Jeffrey M. Good, Rigo Schultz, Ayinuer Aximu-Petri, Anne Butthof, Barbara Höber, Barbara Höffner, Madlen Siegemund, Antje Weihmann, Chad Nusbaum, Eric S. Lander, Carsten Russ, Nathaniel Novod, Jason Affourtit, Michael Egholm, Christine Verna, Pavao Rudan, Dejana Brajkovic, Željko Kucan, Ivan Gušic, Vladimir B. Doronichev, Liubov V. Golovanova, Carles Lalueza-Fox, Marco de la Rasilla, Javier Fortea, Antonio Rosas, Ralf W. Schmitz, Philip L. F. Johnson, Evan E. Eichler, Daniel Falush, Ewan Birney, James C. Mullikin, Montgomery Slatkin, Rasmus Nielsen, Janet Kelso, Michael Lachmann, David Reich,* Svante Pääbo*

*To whom correspondence should be addressed. E-mail: green@eva.mpg.de (R.E.G.); reich@genetics.med.harvard.edu (D.R.); paabo@eva.mpg.de (S.P.)

Published 7 May 2010, *Science* **328**, 710 (2010)
DOI: 10.1126/science.1188021

This PDF file includes:

Supplemental Online Materials

Table of Contents

Supplemental Online Material 1

Restriction Enzyme Enrichment for Hominid DNA

Kay Prüfer* and Adrian Briggs*

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

To increase the fraction of endogenous Neandertal DNA in our sequencing libraries, we used restriction enzymes to deplete libraries of microbial DNA. A 454 adaptor-ligated library molecule cannot be amplified or sequenced if it is cut by a restriction enzyme because the two product molecules will carry only a single 454 adaptor each.

To identify appropriate restriction enzymes, we analyzed the relative abundance of different restriction sites in 454 data from Neandertal shotgun sequencing runs, using the database of restriction enzyme prototypes in REBase (S*1*). We first excluded all restriction enzymes in this database that would cut the 454 adaptor sequences. The remaining enzymes were matched against 423,076 reads from a 454 run (NT154) using PatMaN (S*2*). Reads were not trimmed and the 454 A-adaptor sequence was appended to each read to simulate the original library molecule. Based on the classification from our Neandertal pipeline (see SOM 3 for details), we divided reads further into two fractions: reads whose best alignment was to a primate sequence, and all other reads. For each restriction site we counted its abundance in each of these sequence fractions, allowing us to identify enzymes that would cut many non-primate fragments but few primate fragments. These enzymes all contained CpG dinucleotides in their recognition sequences, reflecting the particularly low abundance of this dinucleotide in mammalian DNA. Based on these results we designed two restriction enzyme mixes. Mix 1, containing three enzymes, was predicted to have moderate enrichment power, and was predicted to cut very few of the primate sequences. Mix 2, which contained 6 enzymes, was predicted to have a higher enrichment power than Mix 1, but was also predicted to cut more of the primate sequences than Mix 1. Table S1 shows the composition of these enzyme mixes and the predicted effect of the mixes on primate and non-primate reads in the sample Neandertal run NT154.

We experimentally tested how whole-library amplification and restriction enrichment affected endogenous DNA percentage, fragment length distribution and GC content in a Neandertal library. An aliquot of a Neandertal library, L72, was sequenced on the 454 platform (454 Life Sciences, Branford CT, USA) (NT376: 17,352 sequences). A second aliquot of the same library was whole-library-amplified for 7 cycles using the 454 adaptor priming sites as described (S*3*), treated with restriction enzyme Mix 1, and subsequently sequenced with the library name L127 (NT454: 619,088 sequences, see SOM 2 for further

details on experimental methods). We found that 3.1% of reads from the untreated library (L72) aligned best to primate, while 13.1% of reads from the amplified and enriched library (L127) aligned (Table S2, Fig. S1). This represents a greater than four-fold enrichment for endogenous DNA. The magnitude was close to the expected value given the sequence composition of the untreated L72 sequences and the enzymes used, indicating a high efficiency of restriction digestion.

Amplification and restriction enzyme treatment may affect library composition in ways other than percentage of endogenous DNA. Therefore we also measured the sequence lengths and GC content of the raw (L72) versus the amplified/restricted (L127) library. First, we analyzed the length distribution of primate and non-primate reads (Table S2, Figure S2). While the lengths of primate sequences were relatively consistent between the libraries, the non-primate reads were considerably shorter in the treated library L127 than the untreated library L72, as expected since longer fragments will have an increased likelihood of containing at least one restriction site and thus being cut. We next analyzed the GC contents of reads. Primate and non-primate reads showed a lower average GC content in the treated library L127 than in the untreated library L72, (see Table S2, Figure S2) which is expected as all three restriction enzymes in Mix 1 have GC-rich recognition sites. The effect was stronger for the non-primate reads, which is also expected since a higher proportion of non-primate sequences are cut by these restriction enzymes than primate sequences.

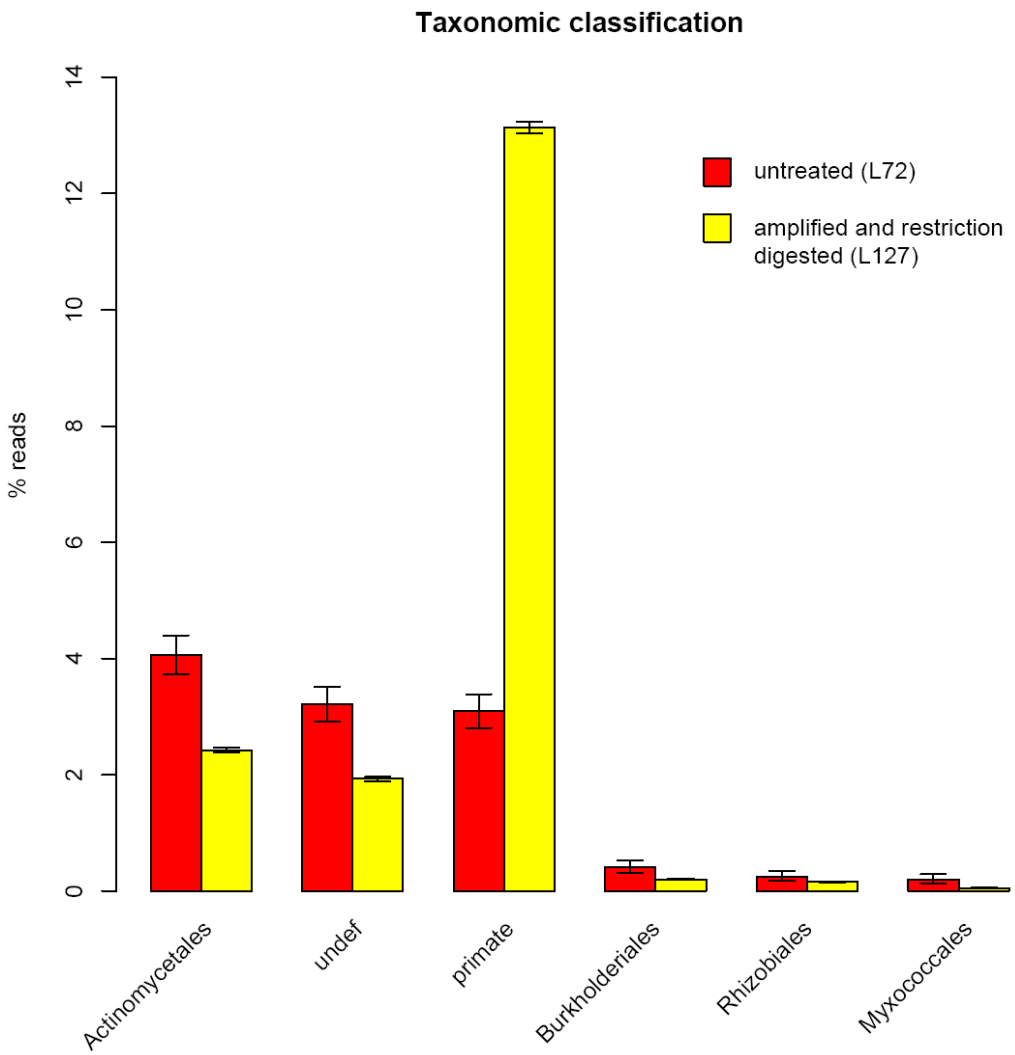


Figure S1. Taxonomic content of Neandertal library before and after restriction enzyme enrichment treatment. Read abundance in the six most commonly hit taxonomic orders in the untreated library L72 (Run: NT376) are shown, as well as in the treated library L127 (Run: NT454).

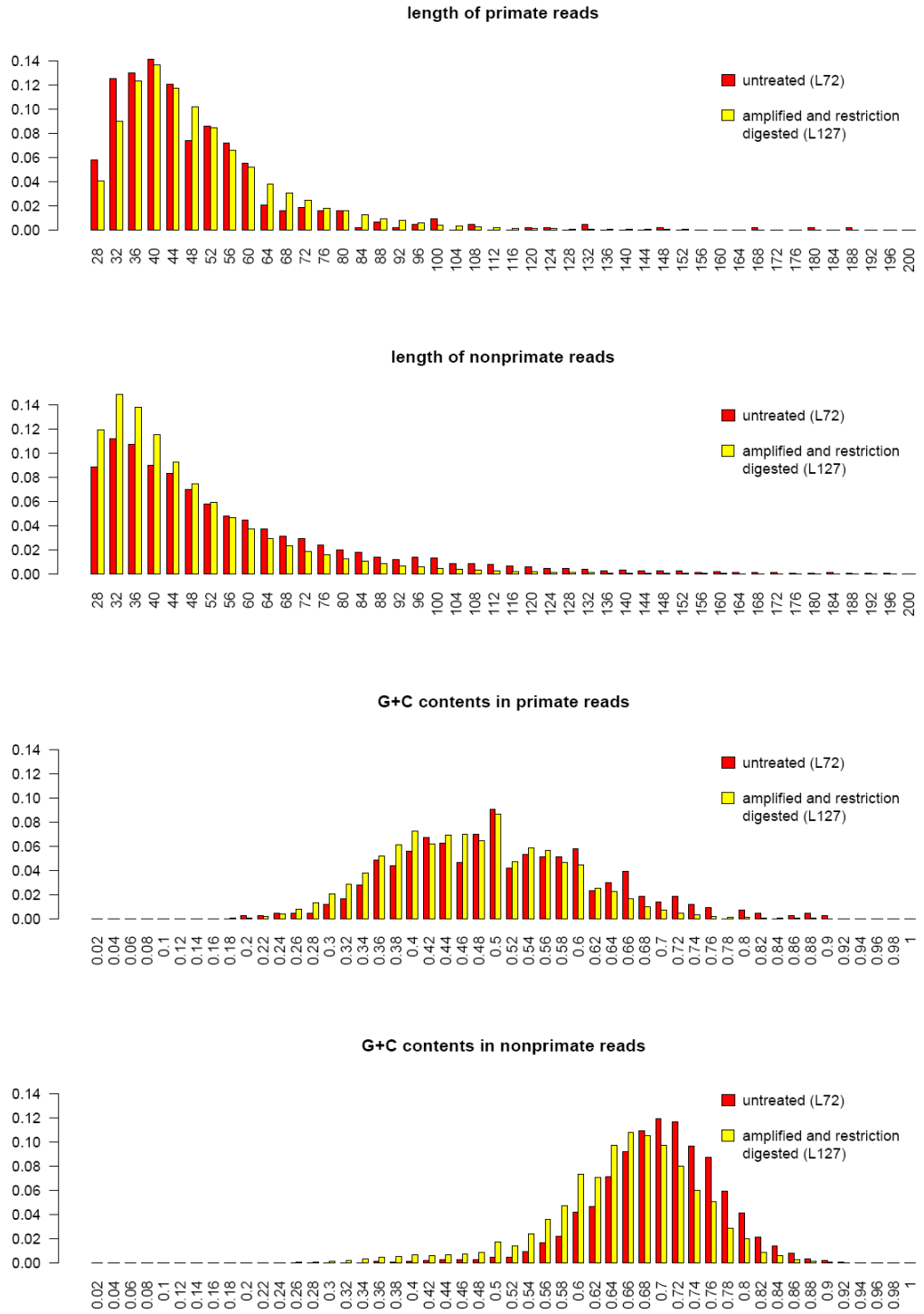


Figure S2. Effect of amplification and restriction enzyme enrichment treatment on fragment length distributions and GC contents of Neandertal (primate) vs. environmental (non-primate) DNA.

Enzyme mix	Enzymes	Restriction site	NT154 primate cut	NT154 non-primate cut	Predicted primate % after enrichment
Mix1	BstUI	CGCG	17.7%	73.0%	**11.0%**
	Hpy99I	CGWCG			
	BsiEI	CGRYCG			
Mix2	TaqI	TCGA	31.1%	91.1%	**23.1%**
	BstUI	CGCG			
	BsiEI	CGRYCG			
	MspI	CCGG			
	TauI	GCSGC			
	HinP1I	GCGC			

Table S1: Composition of the two restriction enzyme mixes used to enrich Neandertal DNA libraries for endogenous DNA. The predicted effect of these mixes on the endogenous DNA percentage was calculated as follows: the numbers of primate and non-primate sequences containing at least one relevant restriction site were counted from a sample of non-treated Neandertal sequencing run (NT154). A total of 3.74% of reads in run NT154 had a best hit to a primate DNA sequence, allowing a prediction of the primate percentage if the restriction mixes had been used on the library prior to sequencing.

Library (Run)	L72 (NT376)	L127 (NT454)
Amplified?	-	+
Restriction enriched?	-	+
Sequences (N)	17,352	619,088
primate sequences (%)	**3.09%**	**13.14%**
primate sequence length (nt)	47.2	48.7
primate GC (%)	49.8	46.5
non-primate sequences (%)	96.91%	86.86%
non-primate sequence length (nt)	56.6	46.2
non-primate GC (%)	68.8	64.4

Table S2: Taxonomic distribution and sequence lengths/GC contents in an untreated (L72) versus amplified restriction enzyme enriched (L127) Neandertal DNA library.

Supplemental Online Material 2

DNA extraction, initial contamination assays, library preparation and sequencing

Martin Kircher* and Johannes Krause*

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

Neandertal DNA extraction, library preparation and contamination assay

To find extracts suitable for sequencing larger parts of the Neandertal genome, 89 different Neandertal bones from 19 sites were analyzed. In total 201 DNA extracts were made from 5 – 560mg of bone (Table S3) as previously described (S*4*), with slight modifications. Specifically, the DNA elution from the silica particles was done using TE buffer with 0.05% Tween 20, and siliconized tubes were used for long time storage of the extracted DNA. In initial screening, the proportion of human contamination in each extract was tested using a PCR based approach as described in refs. (S*5, 6*) using up to three primer pairs in a multiplex 2-step-PCR (S*7*) that targeted conserved mtDNA regions containing informative positions where humans and Neandertals differ in sequence. The retrieved PCR products were tagged (S*8*), and sequenced on the Roche 454 platform.

We recently compared the approach for assessing contamination based on PCR to a targeted direct sequencing approach using Primer Extension Capture (PEC) (S*3*). Our analysis showed that PEC is more accurate for estimating the proportion of human contamination (S*9*). As the project evolved, we therefore changed the assessing human contamination to PEC. For the PEC analyses, 23 µl from each Neandertal DNA extract were made into 454 sequencing libraries as described in ref. (S*3*). Before PEC, each library was PCR amplified using 454-emulsion PCR primers (S*10*). The PEC protocol was performed as described in ref. (S*3*) using six biotinylated PEC primers that target informative differences between human and Neandertal mtDNA (S*6*). For Feldhofer 1 and 2, Mezmaiskaya 1 and Sidron 1253, contamination levels were estimated by capturing the entire mtDNA as described (S*6*). PEC products were quantified with qPCR and used directly as templates for emulsion PCR (S*11*). Results for PEC and PCR contamination assay can be found in Table S4.

As an additional estimate of the proportion of human contamination, the percentage of endogenous Neandertal DNA was estimated for each Neandertal extract by direct sequencing of amplified 454 libraries on a single 16th lane of the Roche 454 FLX platform. Amplification was carried out in a total volume of 100ul, containing 4 Units of *Taq* Gold polymerase, 2.5mM MgCl2, 1mg/ml BSA, 250µM of each dNTP, and 1uM of each 454-emulsion PCR primer (S*10*). The annealing temperature was 60°C and between 7-20 cycles were performed (Table S4).

Extracts that were estimated to contain more than 1.5% endogenous Neandertal DNA with less than 5% human contamination were considered for high throughput shotgun sequencing on the Roche 454 FLX/Titanium and Illumina GA_{II} platforms. The samples chosen for further shotgun sequencing are listed in Table S4.

To increase the proportion of endogenous Neandertal DNA vs. microbial DNA, amplified 454 libraries were treated with restriction enzymes as in SOM 1 (also Table S4).

Emulsion PCR and 454 sequencing

Emulsion PCR and 454 sequencing were carried out according to the manufacturer's instructions on the Roche 454 FLX and Titanium platforms. All Neandertal 454 libraries were PCR amplified for at least 7 cycles before sequencing, to compensate for losses in emulsion PCR. The total number of PCR cycles for each library is listed in Table S4.

454 to Illumina/Solexa library conversion

To sequence the prepared 454 Neandertal sequencing libraries on the Illumina GA_{II} platform, we converted them to Illumina/Solexa libraries. For this purpose, a PCR primer pair was constructed that is complementary to the 454 A and B primers on the 3'-end and has a tail carrying the Solexa p5 and p7 adapter sequences. The 454 library was than amplified in a 100µl reaction containing 50µl Phusion^{TM} High-Fidelity Master Mix, and 500nM of each 454-Solexa-conversion primer (S*9*). The annealing temperature was 60°C and a total of 6-14 cycles of PCR were performed. The amplified products were spin column purified and quantified on an Agilent 2100 Bioanalyzer DNA 1000 chip.

Illumina sequencing and primary data analysis for Neandertal libraries

In total, 33 Solexa/Illumina-converted 454-sequencing-libraries carrying the project specific adapters and a converted PhiX control library carrying the 454 standard adapters were sequenced according to the manufacturer's instructions for paired end sequencing on the Illumina GA_{II} platform. Instead of the standard Genomic R1 and Genomic R2 sequencing primers, project specific primers were used for the forward and the reverse sequencing read, which anneal to the 454 adaptor sequences and allow for sequencing a project specific key at the beginning of each read. For the PhiX control library, a sequencing primer covering the 454 standard key was used for the forward read. In total 33 flow cells have been sequenced (for details see Table S5), 24 with 2 x 51 cycles (FC-204-20xx sequencing chemistry) and nine with 2 x 76 cycles (FC-103-300x sequencing chemistry).

Sequencing runs were analyzed from raw images using the Illumina Genome Analyzer pipeline 1.0 and 1.3.2 (Table S5). To overcome analytical challenges introduced by identical key sequences at the beginning of the first read, we used the first five (instead of two) sequencing cycles for cluster identification (using the Genome Analyzer pipeline 1.3.2). For the earlier pipeline, the Firecrest algorithm (S*12*) was modified to perform cluster identification in cycle 4 and then extract intensities from the clusters identified starting with cycle 1. The source code modifications for the Firecrest 1.9.5 algorithm are available from the authors on request.

After standard base calling using Bustard (S*12*) with parameter estimation done on the PhiX control sequenced in lane 4 of each flow cell, the obtained PhiX 174 reads were aligned to the corresponding reference sequence for creating a training data set for the alternative base caller Ibis (S*13*). Raw sequences from Ibis for the two paired end reads of each sequencing cluster were filtered for the 3 bases of the project specific key ('GAC') at the beginning of both/the second read(s) (see Table S5). The two reads of each cluster were then merged (including adapter removal) after requiring at least an 11 nucleotide overlap. For bases in the overlapping sequence, the consensus sequence was called by considering the base with the higher quality score

or, in case of agreement, summing up the quality scores. For further analysis, only successfully merged sequences were considered.

Table S3: Information on sites, presence of Neandertal DNA, and number of Neandertal samples per site that we analyzed.

Site	Country	Samples	Neandertal DNA present in at least one sample per site
Type site, Feldhofer	Germany	2	yes
Mezmaiskaya	Russia	2	yes
Sima de los Palomas	Spain	7	no
Salzgitter Lebenstedt	Germany	1	no
Shanidar	Iraq	2	no
El Sidron	Spain	15	yes
Okladnikov	Russia	3	yes
Regourdou	France	2	no
Hohlenstein Stadel	Germany	1	yes
Mala Balanica	Serbia	1	no
Spy	Belgium	1	yes
Teshik Tash	Uzbekistan	1	yes
St Cesaire	France	1	yes
Vindija	Croatia	44	yes
Zaskalnaya	Ukraine	2	no
La Chapelle aux Saints	France	1	yes
Scladina	France	1	no
Zeeland ridges neandertal	Netherlands	1	no
Le Moustier	France	1	no

Table S4: Information on samples, extracts, contamination level, number of PCR amplification cycles, and enzyme enrichment treatment of all Neandertal sequencing libraries used in this study.

Phase of study	Sample	Extract	mg	PCR contamin.		PEC contamin.			Treatment	Lib
				Nea	Hsa	Nea	Hsa	Cyc.		
Phase 1	**Vi33.26**	E137	62	na	na	66	0	10	Mix2	L262
454 +								12	Mix2	L312
enzyme		E138	72	na	na	136	1	12	Mix2	L313
								10	Mix2	L263
		E139	69	na	na	66	0	12	Mix2	L314
		E141	78	na	na	59	1	12	Mix2	L315
	Vi33.16	E38	75	477	81	174	1	16	Mix2	L306
								7	Mix1	L127
								7	Mix1	L125
								16	Mix2	L305
								12	Mix2	L311
								16	Mix2	L310
								20	Mix2	L282
								16	Mix2	L308
								7	Mix1	L241
								7	Mix1	L241
								7	Mix1	L241
								20	Mix2	L281
								16	Mix2	L309
								7	Mix1	L240

								7	Mix1	L240
								7	Mix1	L240
								16	Mix2	L307
Phase 2	**Vi33.26**	E137	62	na	na	66	0	16	Mix2	SL9
Solexa +								18	Mix2	SL11
enzyme								18	Mix2	SL24
		E138	72	na	na	136	1	18	Mix2	SL12
								16	Mix2	SL10
		E139	69	na	na	66	0	18	Mix2	SL13
		E111	42	0	321	37	1	20	Mix2	SL31
		E141	78	na	na	59	1	18	Mix2	SL26
								18	Mix2	SL27
								18	Mix2	SL14
		E86	59	152	18	162	1	20	Mix2	SL30
	Vi33.25	E84	70			36	1	22	Mix2	SL6
								20	Mix2	SL7
								22	Mix2	SL8
		E118	50	251	94	94	1	20	Mix2	SL32
	Vi33.16	E38	75	477	81	174	1	23	Mix2	SL22
								33	Mix2	SL16
								27	Mix2	SL15
								18	Mix2	SL21
								27	Mix2	SL29
								22	Mix2	SL20
								23	Mix2	SL23
								22	Mix2	SL19
								22	Mix2	SL17
		E39	150	197	27	272	0	19	no	SL33
								19	no	SL42
								19	no	SL46
								19	no	SL47
	Mez1	E149	70	na	na	2959	10	20	no	SL39
								20	Mix2	SL61
	Feld1	E142, E143	190	na	na	1436	21	18	no	SL37
	Sid1253	120207 .1	223	107	1	2807	7	18	no	SL49

Table S5: Information on 33 flow cells with 34 different Solexa/Illumina-converted 454-Neandertal-sequencing-libraries libraries sequenced on the Illumina Genome Analyzer II platform. This table provides an assignment of lanes (L), library identifiers (Lib, cross reference to Table S4 for bone assignment), read length, sequencing chemistry (C, 1 FC-204-20xx, 2 FC-103-300x) and analysis pipeline version (Ver). For each lane the number of raw clusters, as well as merged and key passed sequences used as input for further analyses, is provided.

Run identifier	L	Lib	C	Cycle	Ver	Raw clusters	Key pass & Merged
080902_BIOLAB29_Run_PE51_1	1	SL7	1	2x51	1.0	7238465	4171912
080902_BIOLAB29_Run_PE51_1	2	SL7	1	2x51	1.0	12646584	6340911
080902_BIOLAB29_Run_PE51_1	3	SL6	1	2x51	1.0	8123010	4734598
080902_BIOLAB29_Run_PE51_1	5	SL6	1	2x51	1.0	11200209	6095960
080902_BIOLAB29_Run_PE51_1	6	SL6	1	2x51	1.0	14167561	7065201
080902_BIOLAB29_Run_PE51_1	7	SL8	1	2x51	1.0	6834065	4675106
080902_BIOLAB29_Run_PE51_1	8	SL8	1	2x51	1.0	11369083	6852186
081017_SOLEXA-GA02_JK_PE_SL10	1	SL10	1	2x51	1.0	8224599	6147835
081017_SOLEXA-GA02_JK_PE_SL10	2	SL10	1	2x51	1.0	8430708	6558573
081017_SOLEXA-GA02_JK_PE_SL10	3	SL10	1	2x51	1.0	8655595	6693209

Run identifier	L	Lib	C	Cycle	Ver	Raw clusters	Key pass & Merged
081017_SOLEXA-GA02_JK_PE_SL10	5	SL10	1	2x51	1.0	8373783	6464202
081017_SOLEXA-GA02_JK_PE_SL10	6	SL10	1	2x51	1.0	8393335	6489995
081017_SOLEXA-GA02_JK_PE_SL10	7	SL10	1	2x51	1.0	8431008	6502392
081017_SOLEXA-GA02_JK_PE_SL10	8	SL10	1	2x51	1.0	8184957	6154906
081021_SOLEXA-GA01_JK_PE_SL18	1	SL18	1	2x51	1.0	10738816	5908679
081021_SOLEXA-GA01_JK_PE_SL18	2	SL18	1	2x51	1.0	12835407	7376933
081021_SOLEXA-GA01_JK_PE_SL18	3	SL18	1	2x51	1.0	13997258	7870902
081021_SOLEXA-GA01_JK_PE_SL18	5	SL18	1	2x51	1.0	13759939	7321036
081021_SOLEXA-GA01_JK_PE_SL18	6	SL18	1	2x51	1.0	13797928	6750642
081021_SOLEXA-GA01_JK_PE_SL18	7	SL18	1	2x51	1.0	13841989	6213884
081021_SOLEXA-GA01_JK_PE_SL18	8	SL18	1	2x51	1.0	13444385	5161528
081023_SOLEXA-GA04_JK_PE_SL19_repeat	1	SL19	1	2x51	1.0	18260859	9407868
081023_SOLEXA-GA04_JK_PE_SL19_repeat	2	SL19	1	2x51	1.0	18885214	8025197
081023_SOLEXA-GA04_JK_PE_SL19_repeat	3	SL19	1	2x51	1.0	18653881	8342983
081023_SOLEXA-GA04_JK_PE_SL19_repeat	5	SL19	1	2x51	1.0	18245909	7248568
081023_SOLEXA-GA04_JK_PE_SL19_repeat	6	SL19	1	2x51	1.0	17897138	6284636
081023_SOLEXA-GA04_JK_PE_SL19_repeat	7	SL19	1	2x51	1.0	17876921	5499571
081023_SOLEXA-GA04_JK_PE_SL19_repeat	8	SL19	1	2x51	1.0	17012079	6776978
081028_SOLEXA-GA01_JK_PE_SL9repeat	1	SL9	1	2x51	1.0	11445239	6875455
081028_SOLEXA-GA01_JK_PE_SL9repeat	2	SL9	1	2x51	1.0	11828321	7084440
081028_SOLEXA-GA01_JK_PE_SL9repeat	3	SL9	1	2x51	1.0	11684064	7173997
081028_SOLEXA-GA01_JK_PE_SL9repeat	5	SL9	1	2x51	1.0	11828609	7170971
081028_SOLEXA-GA01_JK_PE_SL9repeat	6	SL9	1	2x51	1.0	11660081	6948609
081028_SOLEXA-GA01_JK_PE_SL9repeat	7	SL9	1	2x51	1.0	11669101	7051551
081028_SOLEXA-GA01_JK_PE_SL9repeat	8	SL9	1	2x51	1.0	11461566	6717434
081030_SOLEXA-GA02_JK_PE_SL14_SL18	1	SL14	1	2x51	1.0	9692908	7444547
081030_SOLEXA-GA02_JK_PE_SL14_SL18	2	SL14	1	2x51	1.0	11541159	9231751
081030_SOLEXA-GA02_JK_PE_SL14_SL18	3	SL14	1	2x51	1.0	11567491	9283862
081030_SOLEXA-GA02_JK_PE_SL14_SL18	5	SL18	1	2x51	1.0	1383433	129497
081030_SOLEXA-GA02_JK_PE_SL14_SL18	6	SL18	1	2x51	1.0	1187154	133582
081030_SOLEXA-GA02_JK_PE_SL14_SL18	7	SL18	1	2x51	1.0	1182456	133507
081030_SOLEXA-GA02_JK_PE_SL14_SL18	8	SL18	1	2x51	1.0	1210050	131471
081030_SOLEXA-GA04_JK_PE_SL24_SL26	1	SL24	1	2x51	1.0	12870400	8074548
081030_SOLEXA-GA04_JK_PE_SL24_SL26	2	SL24	1	2x51	1.0	10253521	7140097
081030_SOLEXA-GA04_JK_PE_SL24_SL26	3	SL24	1	2x51	1.0	13064897	8244647
081030_SOLEXA-GA04_JK_PE_SL24_SL26	5	SL26	1	2x51	1.0	11217773	8112897
081030_SOLEXA-GA04_JK_PE_SL24_SL26	6	SL26	1	2x51	1.0	8986478	6774573
081030_SOLEXA-GA04_JK_PE_SL24_SL26	7	SL26	1	2x51	1.0	12109517	8611132
081030_SOLEXA-GA04_JK_PE_SL24_SL26	8	SL26	1	2x51	1.0	13783056	9034466
081111_SOLEXA-GA02_JK_PE_SL11_SL12	1	SL11	1	2x51	1.0	10592656	8167195
081111_SOLEXA-GA02_JK_PE_SL11_SL12	2	SL11	1	2x51	1.0	10666540	8290272
081111_SOLEXA-GA02_JK_PE_SL11_SL12	3	SL12	1	2x51	1.0	10464053	8467538
081111_SOLEXA-GA02_JK_PE_SL11_SL12	5	SL12	1	2x51	1.0	10489021	8471243
081111_SOLEXA-GA02_JK_PE_SL11_SL12	6	SL12	1	2x51	1.0	10452535	8367764
081111_SOLEXA-GA02_JK_PE_SL11_SL12	7	SL12	1	2x51	1.0	10345766	8239727
081111_SOLEXA-GA02_JK_PE_SL11_SL12	8	SL12	1	2x51	1.0	9970310	7671772
081111_SOLEXA-GA04_JK_PE_SL11_SL13	1	SL13	1	2x51	1.0	11298316	8485008
081111_SOLEXA-GA04_JK_PE_SL11_SL13	2	SL13	1	2x51	1.0	11583259	8798865
081111_SOLEXA-GA04_JK_PE_SL11_SL13	3	SL13	1	2x51	1.0	12032150	8804600
081111_SOLEXA-GA04_JK_PE_SL11_SL13	5	SL13	1	2x51	1.0	11959088	9090601
081111_SOLEXA-GA04_JK_PE_SL11_SL13	6	SL13	1	2x51	1.0	11894634	9070207
081111_SOLEXA-GA04_JK_PE_SL11_SL13	7	SL13	1	2x51	1.0	11958527	9070050
081111_SOLEXA-GA04_JK_PE_SL11_SL13	8	SL11	1	2x51	1.0	9013643	6960717
081113_SOLEXA-GA01_JK_PE_SL27	3	SL27	1	2x51	1.0	8080585	5738813
081113_SOLEXA-GA01_JK_PE_SL27	5	SL27	1	2x51	1.0	8541184	5950190
081113_SOLEXA-GA01_JK_PE_SL27	6	SL27	1	2x51	1.0	8608986	6177051
081113_SOLEXA-GA01_JK_PE_SL27	7	SL27	1	2x51	1.0	8855976	6393413
081113_SOLEXA-GA01_JK_PE_SL27	8	SL27	1	2x51	1.0	8621448	5672384
081128_SOLEXA-GA01_JK_PE_SL27_22_15	1	SL22	1	2x51	1.0	7599144	3399476
081128_SOLEXA-GA01_JK_PE_SL27_22_15	2	SL27	1	2x51	1.0	10692712	6418888
081128_SOLEXA-GA01_JK_PE_SL27_22_15	3	SL27	1	2x51	1.0	11273563	6917575
081128_SOLEXA-GA01_JK_PE_SL27_22_15	5	SL15	1	2x51	1.0	9271297	5734136
081128_SOLEXA-GA01_JK_PE_SL27_22_15	6	SL15	1	2x51	1.0	8813734	5367642

Run identifier	L	Lib	C	Cycle	Ver	Raw clusters	Key pass & Merged
081128_SOLEXA-GA01_JK_PE_SL27_22_15	7	SL15	1	2x51	1.0	9111993	5503748
081128_SOLEXA-GA01_JK_PE_SL27_22_15	8	SL15	1	2x51	1.0	8691839	5019123
081128_SOLEXA-GA02_JK_PE_SL16_SL10	1	SL10	1	2x51	1.0	9835350	6528713
081128_SOLEXA-GA02_JK_PE_SL16_SL10	2	SL10	1	2x51	1.0	9957255	6962415
081128_SOLEXA-GA02_JK_PE_SL16_SL10	3	SL10	1	2x51	1.0	9981794	6213238
081128_SOLEXA-GA02_JK_PE_SL16_SL10	5	SL16	1	2x51	1.0	11280037	2891527
081128_SOLEXA-GA02_JK_PE_SL16_SL10	6	SL16	1	2x51	1.0	10468171	1874901
081128_SOLEXA-GA02_JK_PE_SL16_SL10	7	SL16	1	2x51	1.0	11378528	4881929
081128_SOLEXA-GA02_JK_PE_SL16_SL10	8	SL16	1	2x51	1.0	10930940	4051693
081128_SOLEXA-GA04_JK_PE_SL21	1	SL21	1	2x51	1.0	5188806	2136164
081128_SOLEXA-GA04_JK_PE_SL21	2	SL21	1	2x51	1.0	4800205	2224663
081128_SOLEXA-GA04_JK_PE_SL21	3	SL21	1	2x51	1.0	11772776	4171032
081128_SOLEXA-GA04_JK_PE_SL21	5	SL21	1	2x51	1.0	15633538	6005389
081128_SOLEXA-GA04_JK_PE_SL21	6	SL21	1	2x51	1.0	14953894	6008079
081128_SOLEXA-GA04_JK_PE_SL21	7	SL21	1	2x51	1.0	16062187	6181955
081128_SOLEXA-GA04_JK_PE_SL21	8	SL21	1	2x51	1.0	16604530	6061403
081209_SL-XAS_0004_30N17AAXX	1	SL6	2	2x76	1.0	7474775	5828456
081209_SL-XAS_0004_30N17AAXX	2	SL7	2	2x76	1.0	13214596	9738970
081209_SL-XAS_0004_30N17AAXX	3	SL8	2	2x76	1.0	7329405	6155613
081209_SL-XAS_0004_30N17AAXX	5	SL17	2	2x76	1.0	11436203	8542926
081209_SL-XAS_0004_30N17AAXX	6	SL20	2	2x76	1.0	13139768	9872399
081209_SL-XAS_0004_30N17AAXX	7	SL17	2	2x76	1.0	10842475	8390756
081209_SL-XAS_0004_30N17AAXX	8	SL20	2	2x76	1.0	14452715	8821377
081212_SOLEXA-GA04_JK_PE_SL32_2rep	1	SL32	1	2x51	1.0	20681265	3898532
081212_SOLEXA-GA04_JK_PE_SL32_2rep	2	SL32	1	2x51	1.0	21046307	1154954
081212_SOLEXA-GA04_JK_PE_SL32_2rep	3	SL32	1	2x51	1.0	20823021	1351138
081212_SOLEXA-GA04_JK_PE_SL32_2rep	5	SL32	1	2x51	1.0	21035450	1016032
081212_SOLEXA-GA04_JK_PE_SL32_2rep	6	SL32	1	2x51	1.0	20611253	1422546
081212_SOLEXA-GA04_JK_PE_SL32_2rep	7	SL32	1	2x51	1.0	20373022	1404815
081212_SOLEXA-GA04_JK_PE_SL32_2rep	8	SL32	1	2x51	1.0	20323894	768113
081217_SOLEXA-GA02_JK_PE_SL33_SL39	1	SL33	1	2x51	1.0	10277566	5361560
081217_SOLEXA-GA02_JK_PE_SL33_SL39	6	SL37	1	2x51	1.0	11806148	6528299
081217_SOLEXA-GA02_JK_PE_SL33_SL39	8	SL39	1	2x51	1.0	18173821	4641436
090102_SOLEXA-GA02_JK_PE51_SL29_SL30_SL31_rep	1	SL29	1	2x51	1.0	13063970	7065879
090102_SOLEXA-GA02_JK_PE51_SL29_SL30_SL31_rep	2	SL29	1	2x51	1.0	12527020	6612962
090102_SOLEXA-GA02_JK_PE51_SL29_SL30_SL31_rep	3	SL30	1	2x51	1.0	13175898	7472818
090102_SOLEXA-GA02_JK_PE51_SL29_SL30_SL31_rep	5	SL30	1	2x51	1.0	14379430	8480860
090102_SOLEXA-GA02_JK_PE51_SL29_SL30_SL31_rep	6	SL30	1	2x51	1.0	14030415	8128443
090102_SOLEXA-GA02_JK_PE51_SL29_SL30_SL31_rep	7	SL31	1	2x51	1.0	19223063	7627382
090102_SOLEXA-GA02_JK_PE51_SL29_SL30_SL31_rep	8	SL31	1	2x51	1.0	19276658	7214664
090107_SOLEXA-GA04_JK_PE_SL28titr	1	SL28	1	2x51	1.0	18037617	22516
090107_SOLEXA-GA04_JK_PE_SL28titr	2	SL28	1	2x51	1.0	17828775	10163
090107_SOLEXA-GA04_JK_PE_SL28titr	3	SL28	1	2x51	1.0	18580699	1187592
090107_SOLEXA-GA04_JK_PE_SL28titr	6	SL28	1	2x51	1.0	18608323	842982
090107_SOLEXA-GA04_JK_PE_SL28titr	7	SL28	1	2x51	1.0	18300059	2362698
090107_SOLEXA-GA04_JK_PE_SL28titr	8	SL28	1	2x51	1.0	16592082	4392867
090108_SOLEXA-GA03_JK_PE_SL32	1	SL32	1	2x51	1.0	18992400	6336508
090108_SOLEXA-GA03_JK_PE_SL32	2	SL32	1	2x51	1.0	18978525	5941173
090108_SOLEXA-GA03_JK_PE_SL32	3	SL32	1	2x51	1.0	19030120	5633488
090108_SOLEXA-GA03_JK_PE_SL32	5	SL32	1	2x51	1.0	19163886	3734368
090108_SOLEXA-GA03_JK_PE_SL32	6	SL32	1	2x51	1.0	19047216	4030399
090108_SOLEXA-GA03_JK_PE_SL32	7	SL32	1	2x51	1.0	19178346	3482734
090108_SOLEXA-GA03_JK_PE_SL32	8	SL32	1	2x51	1.0	19014971	4473651
090109_SL-XAQ_0002_FC30R0FAAXX	1	SL17	2	2x76	1.3.2	8756516	6831012
090109_SL-XAQ_0002_FC30R0FAAXX	2	SL17	2	2x76	1.3.2	9265834	7558798
090109_SL-XAQ_0002_FC30R0FAAXX	3	SL17	2	2x76	1.3.2	7748659	6321679
090109_SL-XAQ_0002_FC30R0FAAXX	5	SL17	2	2x76	1.3.2	7107537	5839285
090109_SL-XAQ_0002_FC30R0FAAXX	6	SL17	2	2x76	1.3.2	6968205	6018935
090109_SL-XAQ_0002_FC30R0FAAXX	7	SL17	2	2x76	1.3.2	6434045	5529228
090109_SL-XAQ_0002_FC30R0FAAXX	8	SL17	2	2x76	1.3.2	6364005	5180303
090109_SL-XAR_0002_FC30R2NAAXX	1	SL6	2	2x76	1.3.2	7001140	5203681
090109_SL-XAR_0002_FC30R2NAAXX	2	SL6	2	2x76	1.3.2	7036501	5675450
090109_SL-XAR_0002_FC30R2NAAXX	3	SL6	2	2x76	1.3.2	6490248	5118008

Run identifier	L	Lib	C	Cycle	Ver	Raw clusters	Key pass & Merged
090109_SL-XAR_0002_FC30R2NAAXX	5	SL6	2	2x76	1.3.2	6595036	5138386
090109_SL-XAR_0002_FC30R2NAAXX	6	SL6	2	2x76	1.3.2	6802654	4607428
090109_SL-XAR_0002_FC30R2NAAXX	7	SL6	2	2x76	1.3.2	6695451	4086408
090109_SL-XAR_0002_FC30R2NAAXX	8	SL6	2	2x76	1.3.2	6145008	3841395
090109_SL-XAU_0001_FC30R4RAAXX	1	SL6	2	2x76	1.3.2	6968812	3722393
090109_SL-XAU_0001_FC30R4RAAXX	2	SL6	2	2x76	1.3.2	7383657	3974975
090109_SL-XAU_0001_FC30R4RAAXX	3	SL6	2	2x76	1.3.2	7084598	3576474
090109_SL-XAU_0001_FC30R4RAAXX	5	SL6	2	2x76	1.3.2	9648357	1213971
090109_SL-XAU_0001_FC30R4RAAXX	6	SL6	2	2x76	1.3.2	10311430	764208
090109_SL-XAU_0001_FC30R4RAAXX	7	SL6	2	2x76	1.3.2	8633997	1559214
090109_SL-XAU_0001_FC30R4RAAXX	8	SL6	2	2x76	1.3.2	7607986	3652660
090111_SL-XAL_0003_30R1LAAXX	1	SL7	2	2x76	1.3.2	7776404	4968801
090111_SL-XAL_0003_30R1LAAXX	2	SL7	2	2x76	1.3.2	8100092	5860653
090111_SL-XAL_0003_30R1LAAXX	3	SL7	2	2x76	1.3.2	7940679	5488949
090111_SL-XAL_0003_30R1LAAXX	5	SL7	2	2x76	1.3.2	7649155	4953259
090111_SL-XAL_0003_30R1LAAXX	6	SL7	2	2x76	1.3.2	7913115	5073335
090111_SL-XAL_0003_30R1LAAXX	8	SL7	2	2x76	1.3.2	7841487	4695851
090111_SL-XBA_0002_30R4HAAXX	1	SL7	2	2x76	1.3.2	8183456	6025413
090111_SL-XBA_0002_30R4HAAXX	2	SL7	2	2x76	1.3.2	8026698	5969648
090111_SL-XBA_0002_30R4HAAXX	3	SL7	2	2x76	1.3.2	8018085	5750751
090111_SL-XBA_0002_30R4HAAXX	5	SL7	2	2x76	1.3.2	8292310	6143699
090111_SL-XBA_0002_30R4HAAXX	6	SL7	2	2x76	1.3.2	7863727	5939291
090111_SL-XBA_0002_30R4HAAXX	7	SL7	2	2x76	1.3.2	8114358	6148469
090111_SL-XBA_0002_30R4HAAXX	8	SL7	2	2x76	1.3.2	7931841	5815463
090112_SL-XAV_0004_FC30R45AAXX	1	SL7	2	2x76	1.3.2	6554477	4606341
090112_SL-XAV_0004_FC30R45AAXX	2	SL7	2	2x76	1.3.2	7245375	5022310
090112_SL-XAV_0004_FC30R45AAXX	3	SL7	2	2x76	1.3.2	7992019	5580123
090112_SL-XAV_0004_FC30R45AAXX	5	SL7	2	2x76	1.3.2	9036322	6389584
090112_SL-XAV_0004_FC30R45AAXX	6	SL7	2	2x76	1.3.2	9030499	6195778
090112_SL-XAV_0004_FC30R45AAXX	7	SL7	2	2x76	1.3.2	8870276	5048551
090112_SL-XAV_0004_FC30R45AAXX	8	SL7	2	2x76	1.3.2	8496174	4553319
090115_SL-XAK_0004_FC30R25AAXX	1	SL20	2	2x76	1.3.2	10902181	6963206
090115_SL-XAK_0004_FC30R25AAXX	2	SL20	2	2x76	1.3.2	10754137	8381707
090115_SL-XAK_0004_FC30R25AAXX	3	SL20	2	2x76	1.3.2	10412783	8015893
090115_SL-XAK_0004_FC30R25AAXX	5	SL20	2	2x76	1.3.2	10504427	7041408
090115_SL-XAK_0004_FC30R25AAXX	6	SL20	2	2x76	1.3.2	10453365	7432728
090115_SL-XAK_0004_FC30R25AAXX	7	SL20	2	2x76	1.3.2	10675743	8009394
090115_SL-XAK_0004_FC30R25AAXX	8	SL20	2	2x76	1.3.2	10765348	7456178
090115_SL-XAT_0004_FC30PMDAAXX	1	SL8	2	2x76	1.3.2	16421605	10517411
090115_SL-XAT_0004_FC30PMDAAXX	2	SL8	2	2x76	1.3.2	16324299	10454249
090115_SL-XAT_0004_FC30PMDAAXX	3	SL8	2	2x76	1.3.2	16114818	10366563
090115_SL-XAT_0004_FC30PMDAAXX	5	SL8	2	2x76	1.3.2	16299681	10433305
090115_SL-XAT_0004_FC30PMDAAXX	6	SL8	2	2x76	1.3.2	15836633	9825709
090115_SL-XAT_0004_FC30PMDAAXX	7	SL8	2	2x76	1.3.2	16895500	10744311
090115_SL-XAT_0004_FC30PMDAAXX	8	SL8	2	2x76	1.3.2	16553181	10480963
090115_SOLEXA-GA01_JK_PE_SL42_SL48	1	SL42	1	2x51	1.0	10233741	6485748
090115_SOLEXA-GA01_JK_PE_SL42_SL48	6	SL46	1	2x51	1.0	10919896	7287393
090115_SOLEXA-GA01_JK_PE_SL42_SL48	7	SL47	1	2x51	1.0	10993229	7304764
090116_SOLEXA-GA02_JK_PE_SL39	1	SL39	1	2x51	1.0	14270186	8865201
090116_SOLEXA-GA02_JK_PE_SL39	2	SL39	1	2x51	1.0	15117962	9998512
090116_SOLEXA-GA02_JK_PE_SL39	3	SL39	1	2x51	1.0	15090681	10089362
090116_SOLEXA-GA02_JK_PE_SL39	5	SL39	1	2x51	1.0	15067349	10058562
090116_SOLEXA-GA02_JK_PE_SL39	6	SL39	1	2x51	1.0	14820369	9761315
090116_SOLEXA-GA02_JK_PE_SL39	7	SL39	1	2x51	1.0	15106028	10051894
090116_SOLEXA-GA02_JK_PE_SL39	8	SL39	1	2x51	1.0	15024245	10010636
090116_SOLEXA-GA04_JK_PE_SL37_SL38	5	SL37	1	2x51	1.0	10084698	6742852
090116_SOLEXA-GA04_JK_PE_SL37_SL38	6	SL37	1	2x51	1.0	10033074	6803106
090116_SOLEXA-GA04_JK_PE_SL37_SL38	7	SL37	1	2x51	1.0	10271620	6899234
090116_SOLEXA-GA04_JK_PE_SL37_SL38	8	SL37	1	2x51	1.0	9034755	5975009
090126_SOLEXA-GA02_JK_PE_SL49	1	SL49	1	2x51	1.3.2	8510348	6033851
090126_SOLEXA-GA02_JK_PE_SL49	2	SL49	1	2x51	1.3.2	8812646	6445976
090126_SOLEXA-GA02_JK_PE_SL49	3	SL49	1	2x51	1.3.2	8574352	6289469
090126_SOLEXA-GA02_JK_PE_SL49	5	SL49	1	2x51	1.3.2	8788391	6484386

Run identifier	L	Lib	C	Cycle	Ver	Raw clusters	Key pass & Merged
090126_SOLEXA-GA02_JK_PE_SL49	6	SL49	1	2x51	1.3.2	8765173	6473555
090126_SOLEXA-GA02_JK_PE_SL49	7	SL49	1	2x51	1.3.2	8934107	6573421
090126_SOLEXA-GA02_JK_PE_SL49	8	SL49	1	2x51	1.3.2	8914034	6492071
090309_SOLEXA-GA03_JK_PE_SL61	1	SL61	1	2x51	1.3.2	10280804	7831441
090309_SOLEXA-GA03_JK_PE_SL61	2	SL61	1	2x51	1.3.2	10833871	8479809
090309_SOLEXA-GA03_JK_PE_SL61	3	SL61	1	2x51	1.3.2	10635445	8334614
090309_SOLEXA-GA03_JK_PE_SL61	5	SL61	1	2x51	1.3.2	10689037	8431149
090309_SOLEXA-GA03_JK_PE_SL61	6	SL61	1	2x51	1.3.2	10757033	8292442
090309_SOLEXA-GA03_JK_PE_SL61	7	SL61	1	2x51	1.3.2	10626179	8534606
090309_SOLEXA-GA03_JK_PE_SL61	8	SL61	1	2x51	1.3.2	10470192	8392562
090309_SOLEXA-GA02_JK_PE_SL56_SL60-57_SL61	3	SL61	1	2x51	1.3.2	11698496	9386929
090309_SOLEXA-GA02_JK_PE_SL56_SL60-57_SL61	5	SL61	1	2x51	1.3.2	12146794	9940651
090309_SOLEXA-GA02_JK_PE_SL56_SL60-57_SL61	6	SL61	1	2x51	1.3.2	12349468	10132157
090309_SOLEXA-GA02_JK_PE_SL56_SL60-57_SL61	7	SL61	1	2x51	1.3.2	12551581	10283728
090309_SOLEXA-GA02_JK_PE_SL56_SL60-57_SL61	8	SL61	1	2x51	1.3.2	11944314	9800758

Supplemental Online Material 3

Neandertal DNA sequence alignment and filtering

Richard E. Green*, Martin Kircher, and Udo Stenzel

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

Alignment of 454 Reads

Neandertal reads from the 454 instrument (see SOM 2) were processed as described previously (*S3*). All reads were aligned to the human and chimpanzee genomes (hg18 and panTro2) using *megablast (S14)***,** and any read that aligned with an E-value better than 0.001 to either genome was considered to be potentially of Neandertal origin.

It was necessary to remove from the analysis wells that appear to give a signal even though they contain no sequence template (*i.e.*: where the intensities measured are the result of bleed-over signal from the neighboring wells) and emulsion PCR duplicates (emulsion inclusions containing multiple beads). To this end, pairs of reads belonging to the same emulsion were aligned and joined into clusters if they were more than 90% similar. From each cluster, only the read aligning best to each genome was retained (number of reads and bases retained, see Table S8). For further analyses, the local alignments produced by *megablast* to the chimpanzee genome were extended into semi-global alignments using our own implementation of an adapted Smith-Waterman algorithm with affine gap scores. Here matches score 1, mismatches cost 3, and opening and extending gaps cost 5 and 2, respectively. We discarded alignments if the semi-global alignment score was negative. To remove the redundancy introduced by library amplification, reads were clustered by alignment coordinates. Two reads from the same original library were considered duplicates if they aligned to the same chromosome, strand, start and end coordinate, and from each set of such reads, the read with the best alignment was kept for further analyses (see Table S9).

Alignment of Illumina Reads

Neandertal Illumina reads (see SOM 2 for details) were mapped to the human genome (hg18), chimpanzee genome (pantro2) as well as to the single-copy aligned human and chimpanzee ancestor genome extracted from the 4-way Enredo-Pecan-Ortheus

(EPO) alignment (*S15-17*) of human, chimpanzee, orangutan, and macaque (see SOM section 10) using a custom mapper called *ANFO*. This custom alignment program was developed to take the characteristics of ancient DNA into account and is available from http://bioinf.eva.mpg.de/anfo. Briefly; *ANFO* builds an index of short words of the target genome, in a fashion similar to the method described by Morgulis et al. (*S18*). Query sequences are broken up into their constituent words to look up in the index. Adjacent or near adjacent words are combined into longer matches, and any match that is considered "long enough" serves as the seed for a semi-global alignment. We indexed every fourth word of length 12 nt and required seeds to be at least 16 bases nt long, which gives sensitivity better than *megablast* (*S14*) and a useful compromise between sensitivity and required computational effort.

When building alignments *ANFO* extends in both directions from all seeds using a best-first search. This takes advantage of the fact that we only need the two best alignments (not every alignment) for each query for calculating map quality scores (MAPQ). The search terminates when two alignments have been found, or when all remaining alignments are guaranteed to be of uninterestingly low quality. The score of an alignment is defined as its negative log-likelihood (i.e. better alignments have lower scores), so that actual differences and an understanding of damage patterns can be combined in a natural way. We define the mapping quality (MAPQ) to be the difference in score between the two best alignments. Following the observation and implementation by Briggs et al. (*S3*), *ANFO* uses different substitution matrices for DNA thought to be double stranded versus single stranded and changes between them if doing so affords a better score. The expected distribution of single stranded stretches is modeled as geometrically distributed. The full *ANFO* configuration file is given in Table S6 for the human and chimpanzee genome and in Table S7 for the single-copy common ancestor genome.

To separate hominid sequences from random similarities, we analyzed the distribution of alignment scores depending on read length. The scores are clearly a mixture of two distinct distributions, with the distinction becoming much less pronounced for shorter reads. We therefore required a minimum read length of 30 nt and a score no worse than 7.5 * (length - 20) to distinguish spurious alignments of bacterial, fungal and non-mammalian reads from actual hominid sequences.

For each library, consensus sequences were constructed from multiple reads of the same Neandertal molecule, defined as having the same orientation, read length, alignment length, and alignment start coordinates. All such clusters, regardless of their mapping quality, are replaced by their consensus sequence. For each observed base and each possible original base, we calculated the likelihood of the observation from its quality score. The base with the highest quality score (calculated by dividing each likelihood by the total likelihood) is used as the consensus. The resulting quality scores are limited to 40, the assumed error rate for polymerases. Table S11 summarizes the number of reads and bases obtained for each library when aligning to the human and chimpanzee genome.

For various analyses, different alignment sets were used. Alignments against the human genome (hg18) were used for determining the number of Neandertal individuals, estimating Y-chromosome and autosomal contamination, and the selective sweep screen.
Alignments to the chimpanzee (panTro2) were used for the pairwise population comparisons (as described in Supplemental methods section: Relationship to present-day humans). Finally, alignments to the inferred human-chimpanzee common ancestor sequence were used for estimating Neandertal-modern human genome divergence as described in Supplemental note: Neandertal-modern human genome divergence.

Downstream filtering of Illumina sequence alignments

Many of the downstream analyses may be sensitive to slight imbalances in base composition and our base caller, *Ibis* (*S13*) reports quality scores that reflect the underlying and variable accuracy for each of the four bases. Therefore, a single quality-score cutoff for all four bases would alter the background base composition and may therefore bias analyses. We implemented a simple scheme that allows both for quality-score filtering and preservation of base composition. We examined the distribution of quality scores for each of the four bases for all six bones and found a quality score cutoff that allows a given percentage of the base under consideration to be included. Provided that the overall base-composition in our reads, without regard to quality score, is an accurate depiction of the underlying base-composition in the

library, this scheme will retain that true base composition. To achieve higher resolution than the discrete Phred scores allow, we define both quality score cutoffs and probabilities of accepting a base with quality equal to the cutoff that will exactly maintain a given percentage of the bases. The cutoffs and probabilities for each base that retain 95% of the Neandertal data are given in Table S12. These cutoffs were used for filtering low quality bases in alignments against hg18, panTro2, and the inferred common ancestor sequence.

For some analyses we also filtered data based on coverage depth. Because unusually high depth can be indicative of mapping problems or copy differences relative to the reference sequence, we excluded sites whose coverage was above the 95th percentile. The genome-wide distribution of coverage from each Vindija bone, unfiltered for map-quality is shown in Figure S3. Peri-centromeric and peri-telomeric regions are noticeably higher covered, presumably due to the repeat content in these regions. For each of the Vindija bones, the coverage cutoff for further analysis used was 2 fold, i.e., no position with 3 or more observations was considered. We then randomly sampled at most a single base from each bone that passed the coverage and base-quality cutoffs, to represent that nucleotide.

Table S6: *ANFO* configuration file used for the alignment of Illumina Neandertal reads to the chimpanzee (pantro2) and human (hg18) genome.

```
policy {
    use_compact_index {
        name: "{hg18|pantro2}_12_4"
        cutoff: 500
    }

    repeat_threshold: 20000
    max_diag_skew: 2
    max_gap: 4
    min_seed_len: 16
    max_penalty_per_nuc: 15
}

aligner {
        rate_of_transversions: 0.001
        rate_of_transitions: 0.003
        gap_open_rate: 0.0001
        gap_extension_rate: 0.003
        rate_of_ds_deamination: 0.01
        rate_of_ss_deamination: 0.9
        mean_overhang_length: 0.5
}
```

Table S7: *ANFO* configuration file used for the alignment of Illumina Neandertal reads to the single-copy common ancestor genome of chimpanzee and human. Lines different to the configuration in Figure 3.1 are indicated by arrows.

```
policy {
    use_compact_index {
        name: "HCSCCA_12_4"
        cutoff: 125                     // <------
    }

    repeat_threshold: 5000              // <------
    max_diag_skew: 2
    max_gap: 4
    min_seed_len: 16
    max_penalty_per_nuc: 15
}

aligner {
        rate_of_transversions: 0.001
        rate_of_transitions: 0.003
        gap_open_rate: 0.0001
        gap_extension_rate: 0.003
        rate_of_ds_deamination: 0.01
        rate_of_ss_deamination: 0.9
        mean_overhang_length: 0.5
}
```

Table S8: Summary of 454 reads aligned to the human and chimpanzee genomes (hg18 and panTro2, resp.) using *megablast* after E-value filter and the removal of sequencing artifacts (ghost wells and emulsion PCR duplicates). Further pairs of reads were aligned and clustered together if they were over 90% identical (single linkage clustering) for all reads belonging to the same emulsion. The read aligning best to each genome was retained.

Library	Emul - sions	Pass filter wells	Aligned (hg18)	Bases (hg18)	Aligned (pantro2)	Bases (pantro2)
L125	14	7916385	989,700	46,080,849	964,170	44,804,908
L127	9	2237902	258,574	11,585,634	250,463	11,191,005
L240+ L307	34	21691346	3,341,180	151,718,742	3,249,588	147,059,911
L241+ L308	28	25437529	3,621,903	162,107,257	3,522,842	157,226,631
L281	13	14877272	2,542,581	116,223,630	2,469,052	112,459,792
L282	12	10944779	1,816,636	81,835,788	1,761,932	79,129,241
L305	8	4082906	705,554	32,251,523	684,194	31,179,223
L306	6	2879380	536,271	24,549,000	519,858	23,722,762
L309	14	11076061	1,610,850	69,869,328	1,561,828	67,533,697
L310	16	12392471	1,956,178	85,835,292	1,898,475	83,007,661
L311	16	6896388	1,375,906	63,704,384	1,332,857	61,517,137
L262	17	13936002	2,319,563	109,090,376	2,246,742	105,386,501
L263	13	11660364	1,947,181	94,737,946	1,887,846	91,522,132
L312	19	15072677	2,133,314	98,791,555	2,066,766	339,089,415
L313	18	14555842	2,418,789	111,119,308	2,340,748	107,184,523
L314	20	14550798	1,356,737	62,113,625	1,318,085	60,144,415
L315	20	15353546	1,330,477	57,284,202	1,288,032	55,329,910
Total	277	205,561,648	30,261,394	1,378,898,439	29,363,478	1,577,488,864

Table S9: Summary of 454 Neandertal data used for downstream analyses based on alignments to the chimpanzee genome (pantro2). *Megablast* alignments were extended into semi-global alignments and discarded if the semi-global alignment produced a negative score. Further, redundancy due to library amplification was removed by clustering reads based on alignment coordinates (same chromosome, strand, start and end coordinate) and propagating the read with the best alignment score.

Library	Bone	Reads aligned (pantro2)	Bases (pantro2)
L125	Vi33.16	342,898	18,004,896
L127	Vi33.16	87,803	4,516,506
L240.L307	Vi33.16	1,315,026	67,058,994
L241.L308	Vi33.16	1,340,553	67,614,304
L262	Vi33.16	1,059,689	56,380,897
L263	Vi33.16	978,450	53,677,554
L281	Vi33.16	1,007,678	51,214,626
L282	Vi33.16	773,569	38,764,436
L305	Vi33.16	334,600	17,119,200
L306	Vi33.16	260,021	13,294,743
L309	Vi33.16	713,525	34,465,580
L310	Vi33.26	816,065	39,961,222
L311	Vi33.26	647,297	33,437,711
L312	Vi33.26	799,171	41,572,033
L313	Vi33.26	1,091,140	55,791,699
L314	Vi33.26	562,407	29,917,692
L315	Vi33.26	414,880	20,770,545
Total	-	12,544,772	643,562,638

Table S10: Amount of DNA sequences obtained from six Neandertal bones.

	Fossil	Vi33.16	Vi33.25	Vi33.26	Sidron 1253	Mezmaiskaya 1	Feldhofer 1	Total
	Country	Croatia	Croatia	Croatia	Spain	Russia	Germany	-
	Age	38,310 [1]	-	-	38,790 [2]	65,000 [3]	39,900 [4]	-
454 data	Emulsions	107	-	170	-	-	-	277
	Key pass reads	85,129,229	-	120,432,419	-	-	-	205,561,648
	Aligned reads	11,506,061	-	18,755,333	-	-	-	30,261,394
	Aligned bases [nt]	**533,137,012**	**-**	**845,761,427**	**-**	**-**	**-**	***1,378,898,439***
	Genome coverage	**30.2 %**	**-**	**19.0 %**	**-**	**-**	**-**	***49.2 %***
Illumina data	Lanes	52	66	64	7	20	5	214
	Raw clusters	563,336,323	776,901,228	749,719,358	61,299,051	257,653,855	51,230,295	2,460,140,110
	Key pass & Merged	388,623,831	348,708,778	351,374,658	44,792,729	181,317,764	32,948,500	1,347,766,260
	Aligned reads	27,235,917	25,826,237	32,905,421	48,814	1,266,452	44,114	87,326,955
	Aligned bases [nt]	**1,265,369,366**	**1,306,019,506**	**1,515,063,376**	**2,234,572**	**56,405,304**	**2,228,645**	***4,147,320,769***
	Genome coverage	**54.1 %**	**46.6 %**	**45.2 %**	**0.1 %**	**2.0 %**	**0.1 %**	***1.5 x***

[1] Serre et al. 2004, [2] Lalueza-Fox et al. 2005, [3] Skinner et al. 2005, [4] Schmitz et al. 2002

Table S11: Summary of Illumina Neandertal data obtained for each library (see SOM section 2 for details) for *ANFO* mappings to the human (hg18) and chimp (pantro2) genome.

Bone	Library	Reads aligned (pantro2)	Reads aligned (hg18)	Bases (pantro2)	Bases (hg18)
SL6	Vi33.25	5,667,061	6,101,922	286,118,877	313,155,115
SL7	Vi33.25	9,485,209	10,159,845	507,690,512	550,909,507
SL8	Vi33.25	6,935,598	6,802,828	338,800,303	336,786,167
SL9	Vi33.26	3,972,346	3,848,222	174,712,624	170,093,379
SL10	Vi33.26	5,333,181	5,174,623	248,116,819	242,155,685
SL11	Vi33.26	1,772,112	1,717,062	81,189,777	79,145,279
SL12	Vi33.26	3,724,261	3,609,413	170,876,344	166,559,730
SL13	Vi33.26	2,426,391	2,356,063	112,220,644	109,682,763
SL14	Vi33.26	972,437	940,947	42,439,636	41,378,275
SL15	Vi33.16	2,279,429	2,213,391	103,579,624	101,106,139
SL16	Vi33.16	1,722,726	1,666,822	73,266,006	71,208,546
SL17	Vi33.16	5,884,950	5,780,413	281,736,583	278,685,434
SL18	Vi33.16	5,739,845	5,563,908	261,693,679	254,963,131
SL19	Vi33.16	4,720,118	4,602,472	210,731,145	206,296,549
SL20	Vi33.16	4,494,321	4,589,660	213,936,750	218,335,733
SL21	Vi33.16	4,506,450	4,391,179	207,554,885	203,170,779
SL22	Vi33.16	407,286	401,477	20,384,178	20,137,513
SL24	Vi33.26	1,889,692	1,842,389	90,313,875	88,479,463
SL26	Vi33.26	1,427,667	1,390,745	67,314,502	65,944,440
SL27	Vi33.26	3,154,498	3,074,570	150,812,120	147,807,763
SL28	Vi33.16	1,090,466	1,050,913	47,518,666	46,080,005
SL29	Vi33.16	1,747,057	1,703,324	79,650,411	78,029,652
SL30	Vi33.26	1,808,221	1,764,055	90,933,081	89,230,175
SL31	Vi33.26	755,111	744,064	36,439,944	36,076,219
SL32	Vi33.25	3,738,369	3,647,582	173,409,814	170,683,036
SL33	Vi33.16	89,329	88,914	4,265,107	4,292,210
SL37	Feld1	44,114	59,058	2,228,645	3,233,921
SL39	Mez1	625,768	611,431	27,484,145	27,164,933
SL42	Vi33.16	82,654	82,736	3,992,773	4,042,445
SL46	Vi33.16	77,773	77,869	3,787,041	3,832,306
SL47	Vi33.16	63,017	63,608	2,966,528	3,036,280
SL49	Sid1253	48,814	60,623	2,234,572	3,000,965
SL61	Mez1	640,684	628,243	28,921,159	28,607,274
Total	**-**	87,326,955	86,810,371	4,147,320,769	4,163,310,811

Table S12: Quality score filtering parameters for Neandertal sequence data.

Dataset	Acceptance rate	Base	Q-cutoff	P(accept|Q=cutoff)
Vi33.16	0.95	A	27	0.476
	0.95	C	27	0.662
	0.95	G	27	0.20
	0.95	T	27	0.43
Vi33.25	0.95	A	27	0.592
	0.95	C	26	0.449
	0.95	G	28	0.689
	0.95	T	26	0.190
Vi33.26	0.95	A	27	0.729
	0.95	C	26	0.601
	0.95	G	26	0.808
	0.95	T	27	0.563
Feldhofer1	0.95	A	22	0.595
	0.95	C	20	0.094
	0.95	G	21	0.543
	0.95	T	21	0.558
Mez.1	0.95	A	25	0.622
	0.95	C	24	0.201
	0.95	G	24	0.283
	0.95	T	25	0.999
Sid1253	0.95	A	27	0.629
	0.95	C	24	0.055
	0.95	G	25	0.912
	0.95	T	26	0.452

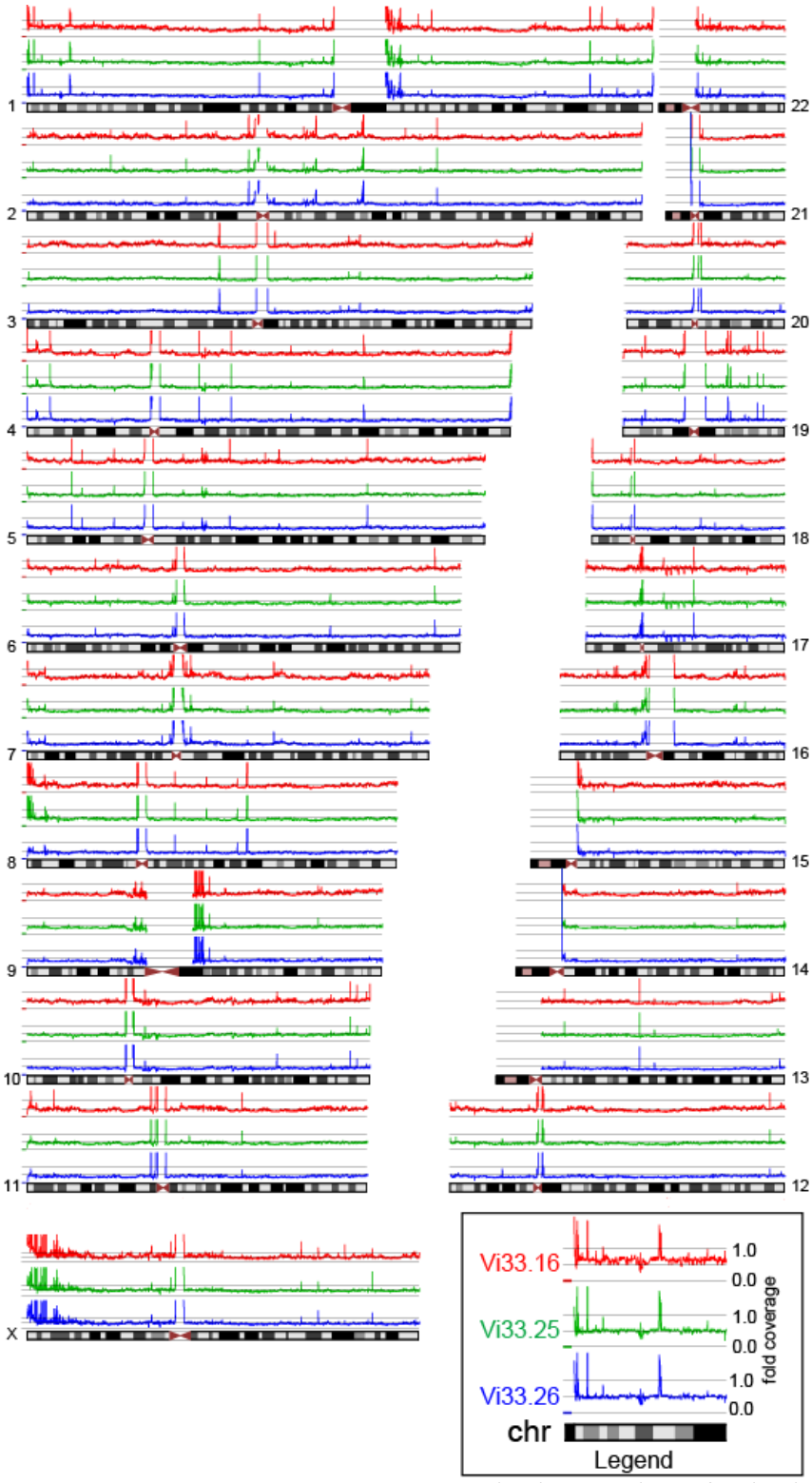


Figure S3: Aggregate sequence coverage depth across hg18 for the three Vindija Neandertal samples

Supplemental Online Material 4

Assembly of Vindija 33.26 mtDNA, sequencing error rate, and number of Neandertal individuals

Richard E. Green*, Adrian W. Briggs and Michael Lachmann

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

Vindija 33.26 mtDNA. In contrast to the two bones Vi33.16 and Vi33.25, for which complete mtDNA sequences have previously been determined (S*3*) and are known to differ from each other, the mtDNA sequence of Vi33.26 has not been determined. From all of the libraries of this bone used for shotgun sequencing, we identified 24,973 mitochondrial DNA fragments, after filtering for PCR duplicates, by mapping against the Neandertal reference mtDNA (AM948965) using a custom alignment program (S*3*). This mapping produced a near complete mtDNA sequence to an average of 68.4 fold coverage. At all 10 sites where Vi33.16 differs from Vi33.25, the Vi33.26 consensus matches Vi33.16. This suggests that Vi33.26 may be identical to Vi33.16. However, a complete comparison of Vi33.26 to Vi33.16 was not possible due to some gaps and regions of very low coverage in the Vi33.26 assembly. These regions of low coverage were situated close to recognition sites for the restriction enzymes we used to enrich libraries for Neandertal genomic DNA (see SOM1). In total the consensus sequence of Vi33.26 contained 51 N bases, most of which were situated within ten nucleotides of an enrichment restriction enzyme site. The same consensus sequence showed 35 substitution differences from the Vi33.16 sequence. Notably, however, all of these differences were in the form Vi33.16 = C/G, Vi33.26 = T/A, and are therefore likely to be due to deamination-derived errors in the Vi33.26 fragments. At all of these 35 sites of difference, at least one of the overlapping Vi33.26 fragments carried the C or G base, matching Vi33.16 (except for a single site of coverage = 1). In conclusion, since the Vi33.26 consensus sequence matches Vi33.16 at 99.58% of positions, with all non-matching sites most easily explained by gaps or cytosine deamination, we cannot distinguish the Vi33.26 from the Vi33.16 mtDNA sequences, and the two bones are likely, although not certain, to carry identical mtDNA.

Estimation of rate of sequencing errors in the Neandertal sequence

Because the mtDNA is a single, haploid locus and is recovered in high coverage from most samples, it is an ideal substrate with which to measure sequencing error. The

known mtDNA sequence can be used to compare all reads that map against it and measure the rate of differences from the true state. Reads from ancient DNA contain error from a combination of sources: deamination-derived misincorporations and the error produced by the sequencing instrument in use.

For each sequencing library of each bone, we compared all reads to the known (Vi33.16 and Vi33.25) or the newly assembled consensus (Vi33.26) sequence and measured the rate of each type of base mismatch. The results are shown in Figure S4. As expected for ancient DNA, the rate of C to T and G to A differences are much higher than any other category. For all other errors, the rate of errors is on the order of 0.001. This improvement over error rates typically seen for Illumina GA_{II} is attributable to the fact that many reads are, in fact, a consensus of several clonal copies of the same original starting molecule, merged to eliminate redundancy and reduce error (see SOM2).

Determination of numbers of individuals.

To determine whether the bones Vi33.26 and Vi33.16 derive from one or two individuals, we compared alleles at genomic positions covered by sequence from both bones. We then compared the rate in intra-bone allele matching to the rate of inter-bone allele matching. Since the rate of sequencing errors in the Neandertal sequence (~0.1% at transversions) is likely to exceed the rate of heterozygosity in Neandertals (0.03-0.04% for transversions in modern humans from the analysis presented here) we enriched for sites that carry genuine polymorphisms rather than sequencing errors by limiting the analysis to sites that are known to harbor single nucleotide polymorphisms (SNPs) in present-day humans (dbSNP, version 130), and to transversions as they are known to have a lower error rate than transitions. We further limited the analysis to sites where exactly three Neandertals fragments were sequenced and two came from the same bone. We used the snp130 annotation of dbSNP (*S19*) mapped to hg18 to identify the ~4.8 million bi-allelic SNP positions in which the two alleles are transversions. We then examined the Neandertal data mapped to hg18, filtered for map quality score >= 30 and base quality cutoffs as described in SOM3.

We then made each of the three pairwise bone comparisons by the following methodology. We first identified each of these sites that were exactly two-fold covered in one individual and one-fold covered in another, filtering for map quality of 30 and base qualities specific for each bone and base. We required that each of the Neandertal bases was one of the two alleles known in humans.

We then calculated the rate at which the two-fold covered bases matched one another within each bone. However, because these are sites already known to be heterozygous in humans, these numbers are not directly interpretable as the rate of heterozygosity within Neandertals.

We compared this to the rate at which two alleles matched when one was drawn from the 2-fold covered bone and the other from the 1-fold covered bone. Under the model that two bones come from the same individual, this rate of allele matching should be equivalent to the rate observed when both alleles come from the same bone. If we label the two-fold covered bone alleles A_1 and A_2 and the one-fold covered bone allele B_1, then we count the rate of allele matching within bone and between bones:

$$M_A = \mathrm{p}(A_1 = A_2) = h_A/2 + 2e_A$$
$$M_B = \mathrm{p}(B_1 = B_2) = h_B/2 + 2e_B$$

$$M_{AB} = \mathrm{p}(A_1 = B_1 \mid \text{bone A!=B}) = h_{AB} + e_A + e_B$$
$$M_{AB} = \mathrm{p}(A_1 = B_1 \mid \text{bone A=B}) = h_A/2 + e_A + e_B$$

where M_A and M_B are the observed rate of allele matching within bone A and B, respectively; M_{AB} is the rate of matching between bones A and B; h_A is the rate of heterozygosity (at these sites) within bone A; h_{AB} is the rate of heterozygosity between bone A and B (which may come from the same individual); e_A is the sequencing error rate in bone A and e_B is the sequencing error rate in bone B; A!=B is the model where bone A is not from the same individual as bone B and A=B is the model where they are from the same individual.

Under the model that bone A and bone B come from the same individual,

$$h_A = h_B = h_{AB}$$

and thus,

$$(M_A + M_B) / 2 = M_{AB}$$

Under the model that bone A and bone B come from different individuals A!=B, and thus:

$$M_{AB} - (M_A + M_B) / 2 = h_{AB} - (h_A/2 + h_B/2)/2 > 0$$

The key to the comparison is that half of all alleles sampled from within a bone will be from the same chromosome. Alleles sampled between bones will always be from different chromosomes if the bones come from different Neandertal individuals.

We measured M_A, M_B, and M_{AB} and compared the two models. The results are shown in the Table S13. For each pairwise comparison, we can reject that the two bones came from the same individual. Notice that bones Vi33.25 and Vi33.16 gave identical mitochondrial sequences, which was different from that of Vi33.25.

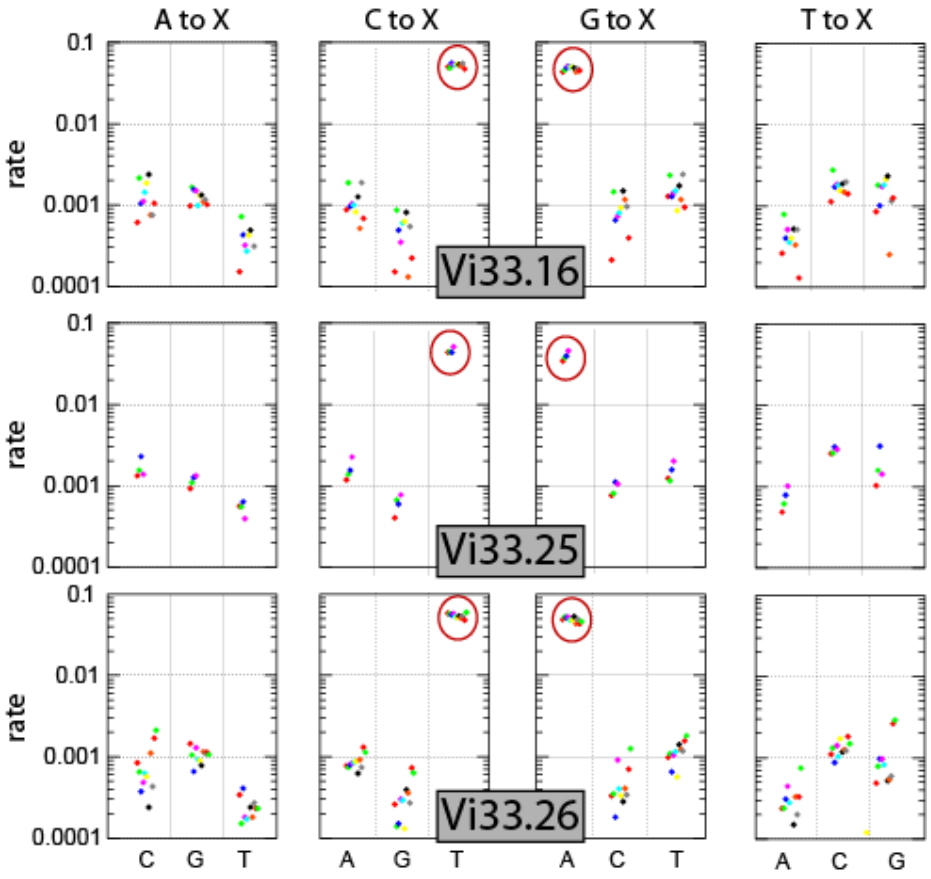


Figure S4: Error rates for the three Neandertal bones as determined from each mtDNA. Each dot represents the indicated rate of error in one sequencing library made from the bone indicated.

Table S13: Inter- and intra-bone allele matching for each pairwise bone comparison. The number of sites where two alleles match one another when drawn from within or between each bone are shown. If the two bones being compared come from the same individual, this rate should be equivalent. In each case, there is a significant increase in allelic differences between bones compared to within bones. Confidence intervals are one-sided, and given for 95% confidence.

A =Vi33.16; B=Vi33.26	**same**	**different**	**% different +(-) 95% conf**
M_A	29,515	406	1.36% + 0.12%
M_B	22,099	332	1.48%+ 0.14%
M_{AB}	**51,430**	**922**	**1.76%-0.09%**
M_{AB} **-(**M_A **+** M_B**) / 2**			**0.34%-0.16%**
P-value (M_A **+** M_B**) / 2 =** M_{AB}	$6x10^{-6}$		
A=Vi33.25; B=Vi33.26	**Same**	**different**	**% different**
M_A	25,421	399	1.55%+0.13%
M_B	24,420	393	1.58%+0.14%
M_{AB}	**49,707**	**926**	**1.83%-0.10%**
M_{AB} **-(**M_A **+** M_B**) / 2**			**0.26%-0.17%**
P-value (M_A **+** M_B**) / 2 =** M_{AB}	0.001		
A=Vi33.25; B=Vi33.16	**Same**	**different**	**% different**
M_A	24,799	395	1.57%+0.13%
M_B	32,364	442	1.35%+0.11%
M_{AB}	**56,947**	**1,053**	**1.82%-0.09%**
M_{AB} **-(**M_A **+** M_B**) / 2**			**0.36%-0.15%**
P-value (M_A **+** M_B**) / 2 =** M_{AB}	$6x10^{-7}$		

Supplemental Online Material 5

Mitochondrial DNA contamination estimates

Richard E. Green*

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

All sequence data from each library was assembled using *mia*, as described previously (S*3*). For the samples whose mtDNA sequences have been previously determined, we used the known sequences as the assembly reference (Table S14). For Vi33.26, we used the Neandertal reference mtDNA (AM948965), from which Vi33.26 was found to be indistinguishable (see SOM 4). The sequences built into each assembly were filtered to remove low-quality bases (Q < 20). The remove obvious nuclear-mtDNA (numt) sequences, we used *bwa* (*S20*) to map all putative mtDNA sequences against the human genome, hg18. Any read whose best match was against the mtDNA of hg18 or had no significant match was included in another *mia* assembly. In this next assembly, PCR duplicate sequences were collapsed. These duplicates are recognized on the basis of having identical start, end, and strand coordinates. This assembly was then used to assess the level of human mtDNA contamination. For each fragment that covered a diagnostic position in which the Neandertal sequence differ from at least 99% of a world-wide panel of 311 contemporary human mtDNAs, we examined the diagnostic position to infer if the sequence was Neandertal or modern human. The counts for each are shown in Table S15.

Table S14: Sequences used as reference for each mtDNA assembly. Each mtDNA, except Vi33.26, has been determined previously.

Bone	Sequence ID
Vi33.16	AM948965
Vi33.25	FM865410
Vi33.26	NA
Feld1	FM865407
Mez1	FM865411
Sidron1253	FM865409

Table S15: Human mtDNA contamination estimates for each library. The number of inferred human and Neandertal mtDNA fragments are shown for each library and summarized for each bone.

Bone	Library	# Neandertal	# Human	% human cont.
Vi33.16	SL15	1229	4	0.32
	SL16	866	1	0.12
	SL17	3868	11	0.28
	SL18	3410	5	0.15
	SL19	2649	14	0.53
	SL20	2668	9	0.34
	SL21	4036	7	0.17
	SL22	201	1	0.5
	SL28	635	2	0.31
	SL29	894	2	0.22
Total		**20456**	**56**	**0.27**
Vi33.25	SL6	326	2	0.61
	SL7	594	3	0.5
	SL8	460	2	0.43
	SL32	311	0	0
Total		**1691**	**7**	**0.41**
Vi33.26	SL9	490	0	0
	SL10	953	1	0.1
	SL11	343	0	0
	SL12	429	2	0.46
	SL13	509	2	0.39
	SL14	284	3	1.05
	SL24	224	0	0
	SL26	335	0	0
	SL27	680	1	0.15
	SL30	337	0	0
	SL31	186	1	0.53
Total		**4810**	**10**	**0.21**
Feld1	**SL37**	**43**	**0**	**0**
Mez1	**SL39**	**327**	**3**	**0.91**
Sidron1253	**SL49**	**41**	**0**	**0**

Supplemental Online Material 6

Sex determination and Y chromosome contamination estimates

Richard E. Green*

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

To determine the sex of the individual who carried each of the Neandertal bones, we looked for Y chromosome sequences. Identifying sequences that unambiguously derive from the Y chromosome is a challenge for several reasons. The main difficulty is that the Y chromosome is homologous to the X chromosome and therefore contains many regions of high similarity (*S21*). Further, the human Y chromosome is small and has a unique repeat structure of ampliconic regions within the Y chromosome (*S22*). With these difficulties in mind, we devised the following strategy for identifying regions of the Y chromosome where sequence hits can be confidently classified as Y chromosome.

First, for every 30mer sequence in the hg18 Y chromosome sequence, we identified all close matches within the rest of the human genome. We kept all Y chromosome 30mers that differ by at least 3 mismatches from any non-Y chromosome sequence. These Y chromosome unique 30mers were then further filtered to remove all that were within an annotated repetitive element using the rmsk327 table from the UCSC annotation of hg18. All remaining 30mers were scanned for contiguous stretches of at least 500 base pairs containing all unique Y-chromosome 30mers. The total length of these 157 segments is 111,132 base pairs.

We then counted all sequences that mapped within these regions. Under the assumption that sequence data derive from random sampling from a male individual, Y chromosome sequence should accumulate at this rate:

$$(111{,}132 / 2{,}800{,}000{,}000) \times 0.5 \times 0.4175 = 0.000008285$$

The factor of 0.5 reflects the fact that the Y chromosome is haploid and the nuclear genome is diploid. The factor of 0.4175 reflects the relative occurrence of the restriction enzymes sites (see SOM1) in these Y-unique regions (6.343 per kilobase) versus the rest of the nuclear genome (2.648 per kilobase). Note, however, that the

Sidron, Feldhofer and one of the Mezmaskaiya libraries were not treated with the restriction enzymes and thus this factor is not included. Thus, approximately 8 in 1,000,000 sequences should fall within these regions when sampling from a male individual. Sampling Neandertal sequences is heavily biased by local GC content (S*6*). These Y unique regions are overall 51.3% GC, higher than the human genome average of 40.9% or the Y chromosome average of 40.0%. This test should therefore over-estimate the presence of male-derived sequence and thus be conservative for determining if these are female bones.

For each Vindija bone, we calculated the total number of hits mapping to hg18 to arrive at an expected number, given that the individual is male, and test whether the observed number is different from the expectation. These data are shown in Table S16. In each case, we can reject that the Vindija individual is male. We note that this conclusion stands in opposition to that drawn previously for Vi33.16 (Vi80) from more limited sequence data (*S5, 23*). In these previous analyses, the presence of sequence that aligned anywhere within the Y-chromosome was taken as positive evidence of a male individual. The current analysis, restricted to regions that are unique within the Y-chromosome, illustrates the importance of requiring an unambiguous map position within the genome.

For the additional Neandertal individuals, there is less data and therefore this test is less sensitive. However, we see more than the expected number of Y unique region hits for the two bones previously known to be male, Sidron1253 (*S24*) and Feldhofer1 (*S25*).

Given that the Vindija individuals are female, we can assume that each of these hits comes from male contamination. If male modern human contamination, containing Y chromosome sequence, is equally likely to fall anywhere in the genome, the rate of accumulation of male contaminating sequences within these regions can be used to estimate contamination:

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

where y is the number of hits in the Y-unique regions, c is the percent of male contamination, Y is the fraction of the genome in the Y-unique regions, and n is the

number of reads. We note that this test provides a conservative estimate of male contamination for the same reason it provides a conservative estimate for the bone being a female. That is, GC bias in our sequence data will cause sequences in these regions of higher average GC content to be over-represented in our data. The male human contamination estimates are shown in Table 1.

To further test the maleness of each Neandertal, we also compared the rate of sequence mappings to the X chromosome versus a similarly sized autosome, chromosome 8. The rate of sequence recovery and identification is complicated by many factors, including GC content, sequence uniqueness of the regions to which it is mapped, and perhaps others factors. Nevertheless, the ratio of sequences mapping to these two chromosomes shows a clear bi-model distribution (Table S17) with the bones identified as female having a consistently lower ratio of chromosome X to chromosome 8 sequences.

Table S16: Y-unique region hit rate. The expected number of sequences mapping within the Y-chromosome unique regions is based on the fraction of the genome that these regions cover. For each of the three Vindija bones, the observed numbers are significantly lower that expected.

Bone	Total hg18 mapped sequences >= 30 nt	chrY hits expected from male	chrY hits observed	P-value male
Vi33.16	30,777,634	255.0	4	0
Vi33.25	24,275,777	201.1	0	0
Vi33.26	25,336,181	209.9	0	0
Mezmaskaiya1	1,188,309	23.6	0	1.0e-6
Sidron1253	44,936	0.9	5	1
Feldhofer1	43,508	0.9	1	0.5

Table S17: Ratio of autosomal to chrX sequences in Neandertal bones. The Vindija and Mexmaskaiya1 bones all have a similarly low ratio of autosomal to X chromosome sequence hits. These were previously identified as female by absence of Y unique region sequences. In contrast, Sidron1253 and Feldhofer1 both have a similarly high ratio of autosomal to chrX sequences, consistent with these bones deriving from male Neandertals carrying only 1 X chromosome.

Bone	Chr8 sequences	chrX sequences	8/X
Vi33.16	1,460,601	1,916,999	0.76
Vi33.25	1,228,228	1,528,493	0.80
Vi33.26	1,225,157	1,686,207	0.73
Mezmaskaiya1	53,370	73,953	0.72
Sidron1253	2,082	1,796	1.16
Feldhofer1	1,815	1,533	1.18

Supplemental Online Material 7

Estimation of nuclear contamination

Philip L. F. Johnson, Michael Lachmann, and Richard E. Green*

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

Introduction

Below we put a bound on the amount of contamination from present-day humans our sequences could contain. This is done using two methods. The first is a simple counting method that operates on sites assumed to be fixed-derived in contemporary humans and but does not make any further assumptions about the population genetics of humans or Neandertals. This method calculates an overestimate of contamination that also includes half the heterozygosity in Neandertals at the sites we test. The second method for estimating nuclear contamination builds a likelihood model that simultaneously estimates contamination as well as several nuisance parameters describing the relationship between the frequency of alleles in humans and in Neandertals. This method uses both segregating sites found in the HapMap Phase II as well the aforementioned list of fixed-derived sites.

We test our methods by artificially introducing contaminating sequence from a present-day human using the published genome of Craig Venter (*S26*). Both methods become biased downward under extremely high levels of contamination (>50%) because of difficulties with parameter identifiability. However, our extremely low estimates of contamination (method 1: $\hat{c} = 1.4\%$; method 2: $\hat{c} = 0.7\%$) in combination with the divergence estimates (SOM 10) and the low estimates from the mitochondrial DNA and Y chromosome contamination tests (SOM 5) allow us to eliminate this region of parameter space. In general, the simulation tests confirm the robustness of our methods to potential violations of assumptions.

``Simple'' estimate for contamination

To calculate an estimate of contamination, we looked at sites at which humans are fixed derived, and we see the ancestral allele in two sequences from two different Neandertals. We then look at a second sequence from one of these two Neandertals, and ask how often we see the derived allele. Since every time we hit a contaminating sequence, we will see the derived allele, this number is an estimate of contamination.

Therefore, we will also see the derived allele if the Neandertal is heterozygous derived and ancestral, and we sample the other copy of the site, we will have an overestimate of contamination. Because we base our characterization of a site as fixed in humans based on a limited sample size, our contamination estimate will be biased by sites that are fixed derived in our sample of humans but ancestral in the contaminating human.

Simple upper bound

In this method, we do not try to correct for the effect of heterozygosity. We therefore make all effort to reduce heterozygosity in the sites we examine.

We followed the following scheme:

1. All human sequences seen in our panel of 5 (see Supp Info 9) are derived, at least one from each individual, and hg18 is derived.
2. No sequence mapped to the site has an insertion or a deletion covering the site.
3. The change from human to chimpanzee is a transversion.
4. No bone has coverage higher than 2 sequences for the site.
5. One Neandertal has 2 sequences covering the site, and a second has at least 1.
6. None of the other sequences seen in any Neandertal is derived.
7. Quality cutoffs used are 32 for base quality in Neandertal, 30 for mapping quality in Neandertal, 25 for base quality in humans, 30 for mapping quality in humans.

Our test then asks:

Conditional on seeing an ancestral allele in one sequence each from the two Neandertals, we ask what the chance is to see a derived allele in the second sequence from the first Neandertal.

In this scheme, the chance to see the derived allele is $c+(1-c)h_{aa}/2$. Where h_{aa} stands for the heterozygosity conditional on seeing the ancestral allele in two sequences of Neandertal from different individuals, and on the allele being fixed derived in our human samples.

For each of the three individuals and each chromosome, we find the following total number of sites, n, and the subset of these in which we see the derived allele, $D < n$. The individual listed is the individual for which we looked at two sequences.

Table S18

	n 16	D 16	n 25	D 25	n 26	D 26
chr1	117	1	79	0	56	2
chr2	119	1	85	1	76	0
chr3	82	3	75	0	77	1
chr4	66	3	68	0	69	0
chr5	82	1	65	2	64	0
chr6	67	1	52	0	51	1
chr7	68	0	43	1	57	0
chr8	89	0	97	2	66	0
chr9	54	0	64	0	52	0
chr10	55	4	37	0	49	1
chr11	56	0	41	1	62	0
chr12	69	1	65	1	48	2
chr13	42	2	50	2	33	2
chr14	39	0	31	1	22	1
chr15	36	0	30	0	26	0
chr16	60	2	40	0	19	0
chr17	35	1	36	0	17	0
chr18	35	3	21	1	26	0
chr19	20	0	16	0	22	0
chr20	48	0	30	0	33	1
chr21	30	0	12	0	17	0
chr22	28	0	22	2	20	0

Table S19 Contamination estimates and 95% upper bounds

	n	D	contamination	upper	lower
Vi33.16	1297	23	1.8	2.6	1.1
Vi33.25	1059	14	1.3	2.2	0.7
Vi33.26	962	11	1.1	2.0	0.6
Total	3318	48	1.4	1.9	1.1

The contamination rates in the different bones are not significantly different. Thus we get an upper bound based on all three bones of 1.4%, with a 95% upper confidence interval of 1.9%.

Testing the method

In order to test our method, and our program, we used Craig Venter's published genome, and replaced a percentage p the Neandertal bases with bases taken from that genome at the homologous sites. Aside from potentially changing the base, we retained all the features for the original sequence such as quality, and

coverage. We then ran our procedure for p ranging from 2% to 40%, once for each point:

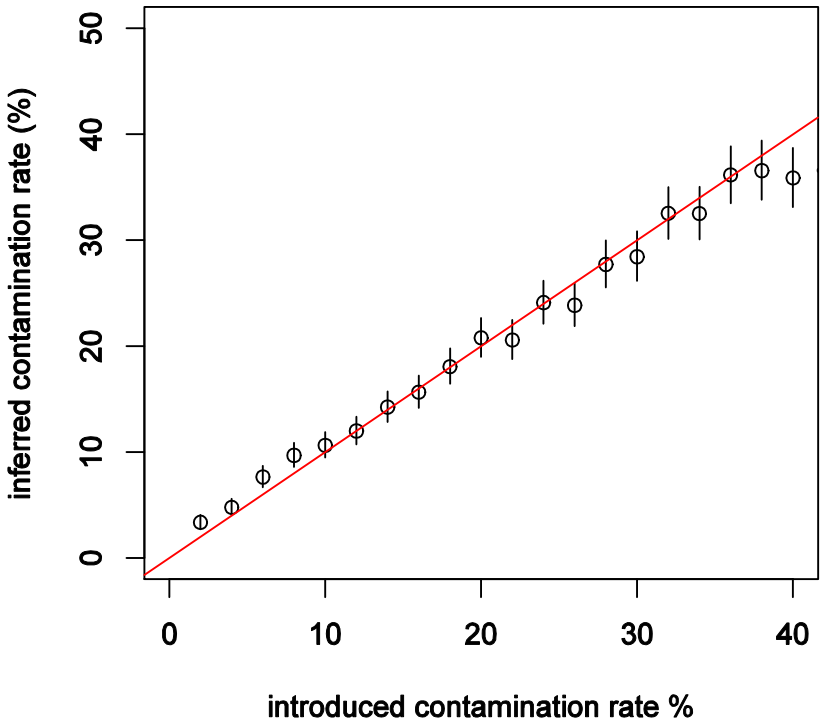


Figure S5: Introduced vs. inferred contamination. Notice that the estimate is conservative, so it starts above the $x = y$ line. The error increases with increasing contamination, because there are fewer and fewer sites at which we see two ancestral alleles. Introducing much higher levels of contamination will start to underestimate contamination, until at 100% introduced contamination we would estimate 0% contamination.

Likelihood method for contamination estimation

Looking only at sites known to be derived in at least some contemporary humans, we observe three types of sites in Neandertal: all ancestral, all derived, and mixed ancestral / derived. The data at each site $i \in \{1,\ldots,L\}$ from bone $j \in \{1,2,3\}$ can be summarized by three numbers: the total number of Neandertal reads ($n_{i,j}$), the number of these which are derived ($n_{i,j,d}$) and the frequency of the derived allele in humans (f_i).

Assumptions

1. All sites have the same rates of contamination and error

2. Sites are independent.

3. The probability of a particular derived allele frequency in Neandertals as a function of the allele frequency in contemporary humans can be approximated as linear for $f < 1$ and all segregating allele frequencies are equally probable. Simulations across a range of demographies including low levels of admixture support these approximations. A very high level of Neandertal to European admixture (higher than seen in our data) would break this approximation. Details in section 3.3 (`` $\Pr(N \mid H)$ '') below.

4. The probabilities in the previous assumption remain constant across all sites. Sex chromosomes would likely have different probabilities, so we focus exclusively on autosomes.

5. All Neandertal chromosomes sample from a single homogenous population (i.e. no inbreeding that might make the two chromosomes within a bone more similar, and no population subdivision separating the bones)

Parameters

Let $\Omega = \{c, p_d, p_s, a_d, a_s, \varepsilon\}$ denote the set of all parameters, where:

$c \rightarrow$ contamination rate. A given read will be from a contaminating human with probability c and from a Neandertal with probability $1-c$.

$p_d \rightarrow$ probability of Neandertal having all derived alleles, given derived allele is fixed in humans ($f_i = 1$).

$p_s \rightarrow$ probability of Neandertal being segregating, given derived allele is fixed in humans ($f_i = 1$).

$a_d \rightarrow$ probability of Neandertal having all derived alleles at position i is $a_d f_i$ for $f_i < 1$.

$a_s \rightarrow$ probability of Neandertal being segregating at position i is $a_s f_i$ for $f_i < 1$.

$\varepsilon \rightarrow$ probability of an error. We observe a derived allele when the truth is ancestral (or vice versa) with probability ε.

For ease of notation, we define π_i to be the derived allele frequency spectrum in Neandertal, conditional on the derived frequency in present-day humans, f_i:

$$\pi_{i,0} = \begin{cases} 1 - a_d f_i - a_s f_i & f_i < 1 \\ 1 - p_d - p_s & f_i = 1 \end{cases}$$

$$\pi_{i,2b} = \begin{cases} a_d f_i & f_i < 1 \\ p_d & f_i = 1 \end{cases}$$

$$\pi_{i,0<t<2b} = \begin{cases} a_s f_i/(n-1) & f_i < 1 \\ p_s/(n-1) & f_i = 1 \end{cases}$$

where b is the number of bones (each of which is diploid, so $2b$ chromosomes).

Likelihood

We write the probability of the observed numbers of derived alleles as the product of the probabilities of the individual sites, conditional on the number of reads sampled from each bone at each site, the allele frequencies in present-day humans and the parameters:

$$\text{lik}(\Omega) = \Pr(n_{1,1,d},\ldots,n_{L,3,d} \mid n_{1,1},\ldots,n_{L,3},f_1,\ldots,f_L,\Omega) = \prod_i \Pr(n_{i,\{1,2,3\},d} \mid n_{i,\{1,2,3\}},f_i,\Omega) \quad (1)$$

Dropping the subscript i for ease of notation, we condition on the true derived allele frequency in our Neandertal sample, t:

$$\Pr(n_{\{1,2,3\},d} \mid n_{\{1,2,3\}},f,\Omega) = \sum_{t=0}^{n} \pi_t \Pr(n_{\{1,2,3\},d} \mid t,n_{\{1,2,3\}},f,\Omega) \quad (2)$$

Now we further condition on the number of derived alleles in each of $b = 3$ bones (0 if homozygous ancestral, 1 if heterozygous, 2 if homozygous derived) denoted with subscripts as t_1, t_2 and $t_3 = t - t_1 - t_2$. The probability of the two free parameters follows a hypergeometric distribution (binomial sampling without replacement):

$$\Pr(n_{\{1,2,3\},d} \mid t,n_{\{1,2,3\}},f,\Omega) = \quad (3)$$
$$\sum_{t_1=0}^{\min[2,t]} \frac{\binom{t}{t_1}\binom{2b-t}{2-t_1}}{\binom{2b}{2}} \Pr(n_{1,d} \mid t_1,n_1,f,\Omega) \sum_{t_2=0}^{\min[2,t-t_1]} \frac{\binom{t-t_1}{t_2}\binom{2b-t+t_1-2}{2-t_2}}{\binom{2b-2}{2}}$$
$$\cdot \Pr(n_{2,d} \mid t_2,n_2,f,\Omega) \Pr(n_{3,d} \mid t_3,n_3,f,\Omega)$$

Note that extending to more than three bones would simply require adding more summations with the appropriate bounds to ensure that the total number of derived alleles sums to t.

After the above conditioning, the probability of the observed number of derived alleles in each bone (e.g., $n_{1,d}$ for bone 1) follows a binomial distribution where the probability of drawing a derived allele depends on the true state of that site (e.g., t_1 for bone 1) and the parameters c, f and ε:

$$\Pr(n_{1,d} \mid t_1,n_1,f,\Omega) = \binom{n_1}{n_{1,d}} q_{t_1}^{n_{1,d}} (1-q_{t_1})^{n_1-n_{1,d}} \quad (4)$$

$$q_2 = cf(1-\varepsilon) + c(1-f)\varepsilon + (1-c)(1-\varepsilon) \quad (5)$$
$$q_1 = cf(1-\varepsilon) + c(1-f)\varepsilon + (1-c)(1-\varepsilon)/2 + (1-c)\varepsilon/2$$
$$q_0 = cf(1-\varepsilon) + c(1-f)\varepsilon + (1-c)\varepsilon$$

The binomial probabilities for the other bones follow identically to equation 4.

We arrive at the final likelihood by assemblying all equations together -- starting from 1 and substituting in 2, 3, 4 and 5.

Finally we estimate our parameter of interest (c) by maximizing the likelihood of the data over all parameters $\Omega = \{c, p_d, p_s, a_d, a_s, \varepsilon\}$. However, we can fix one parameter, ε, by separately estimating it from sites in which three different bases are observed. Confidence intervals for c can be generated using a likelihood ratio test of the global maximum likelihood to the profile likelihood ($\ell(c) = \max_{p_d, p_s, a_d, a_s}[\text{lik}(\Omega)]$) and comparing to a χ^2 distribution with 1 degree of freedom.

Data and results

We begin by identifying a list of sites in present-day humans that are fixed-derived (using the above notation, $f = 1$) by comparing the sequences of five humans from the HGDP project sequenced (SOM 9) and selecting those sites where all five humans were identical to each other but different from the homologous chimpanzee base. Then we supplement this list of $f = 1$ sites with the list of SNPs from Phase II of the HapMap project (*S27*) using frequencies found in the CEU population ($0 < f < 1$). In all cases, we only retain sites in which the two alleles form a transversion rather than a transition, since the Neandertal data will frequently contain false transitions due to ancient DNA damage (*S28*).

Given this list of human-derived sites, we extract the homologous Neandertal bases from the Illumina data produced from the Vindja 33.16, 33.25, 33.26 bones (SOM 3). Then we apply the following filters:

Any site is discarded that either
- has a single bone with more than 3 reads
- has fewer than 2 or greater than 4 reads in total among all the bones (lower coverage yields little information, greater coverage selects duplicated regions)

We only examine reads placed with map-quality ≥ 30

Bases with quality < 32 are discarded

From these filtered data, we first estimate the error rate from triallelic sites to be $\varepsilon = 0.0005$ and then calculate the profile likelihood of the contamination rate (c) by maximizing over all other parameters (p_d, p_s, a_d, a_s). In Figure S6, we plot the resulting curves from using fixed sites only ($f = 1$; panel A) and from using all sites ($f \leq 1$; panel B).

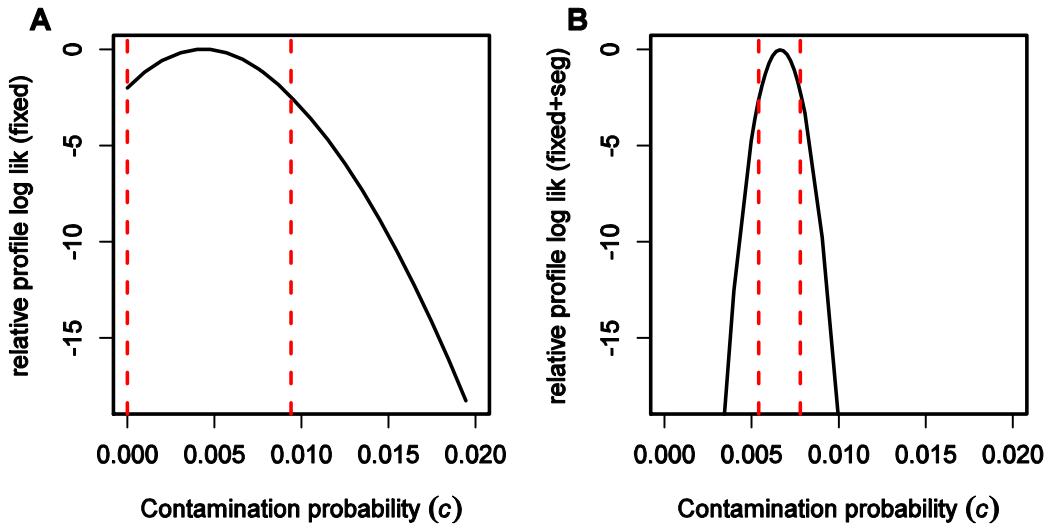


Figure S6: Profile likelihood curves for c using fixed sites only (A) or all sites (B). Y-axis shows log likelihood relative to maximum. Dashed vertical lines indicate approximate 95% confidence interval.

Our final contamination estimate for all data (S6B) is $\hat{c} = 0.66\%$, with an upper bound of 0.76% and a lower bound of 0.56%. The global maximum likelihood also produced the following estimates for the nuisance parameters:

$\hat{p}_d = 0.99$, $\hat{p}_s = 0.0017$, $\hat{a}_d = 0.47$, $\hat{a}_s = 0.55$

Testing the method

We tested our method by adding known amounts of "contamination" from Craig Venter's genome to the Neandertal data and examining the effect on $\hat{c}$. For $c < 0.5$, our estimate increased monotonically with the amount of the simulated contamination and linearly for $c < 0.3$) (Figure S7), suggesting that we can accurately estimate contamination in this parameter range from an individual with background similar to Venter. However, as the simulated contamination exceeds 50%, our maximum likelihood estimate actually declines. Note that, even ignoring the mitochondrial and Y chromosome estimates of contamination, our estimate of 0.66%

contamination is only consistent with either 0.66% or essentially 100% contamination -- and the latter can be eliminated by many other aspects of the data such as the divergence estimates (SOM 10).

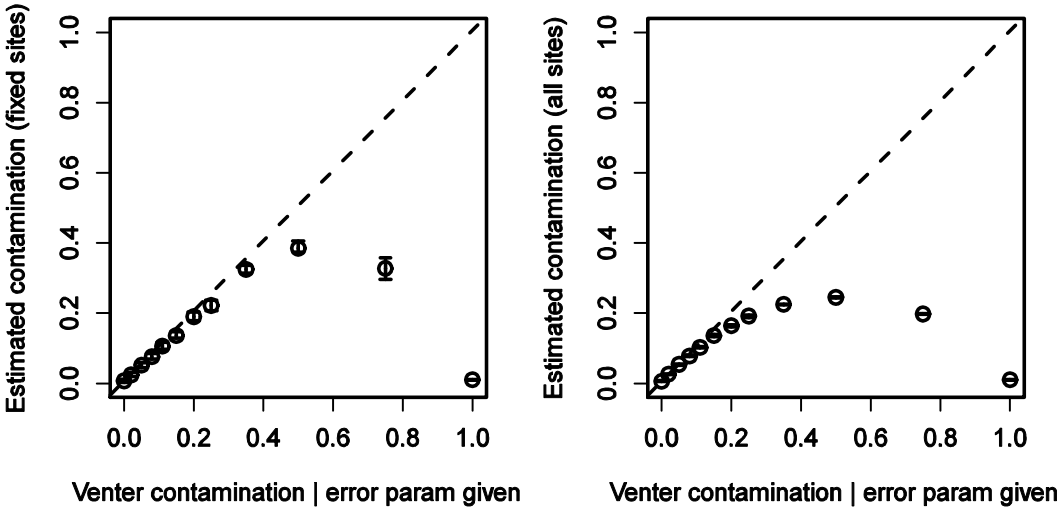


Figure S7: Performance of estimator on simulationed contamination using varying amounts of Craig Venter's genome. Confidence intervals denoted by error bars; dashed line has slope=1 and intercept equal to estimate from data without simulated contamination.

Futher investigation reveals that this underestimation arises from two sources. First, the likelihood must estimate both the evolutionary relationship between Neandertals and modern humans as well as the contamination rate. At sufficiently high doses of contamination, Venter is indistinguishable from a Neandertal -- both will mostly match the derived allele but occasionally each will have the ancestral allele instead (even for $f = 1$ sites; recall that we identify these solely on the basis of being fixed in our sample of 5 HGDP humans). Thus instead of estimating the evolutionary parameters (p_d, p_s, a_d, a_s) relating Neanderthal to modern humans, we are estimating these relating Venter to modern humans. This can be seen from the improved performance of the estimators in Figure S8 below, in which the evolutionary parameters are fixed at their no-added-contamination levels.

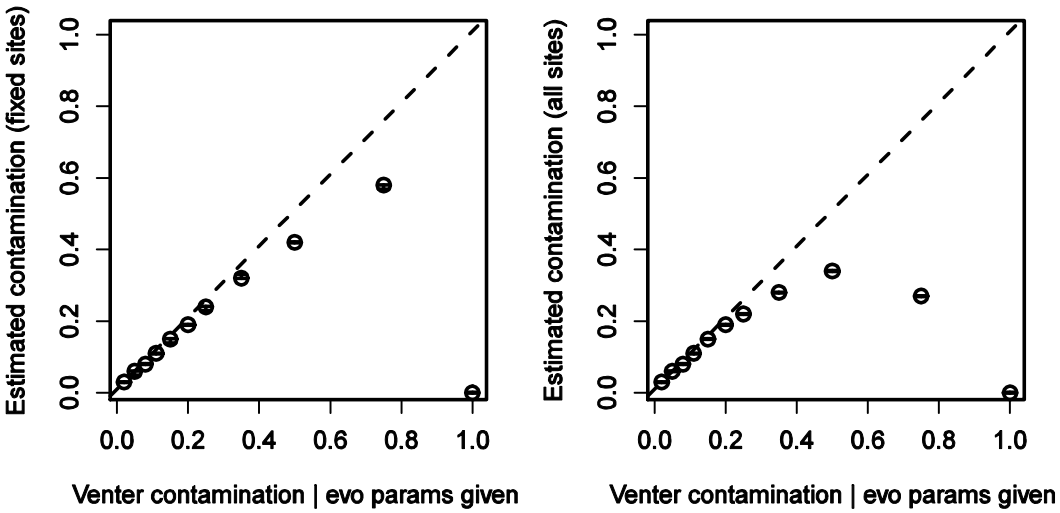


Figure S8: Performance of estimator with evolutionary parameters (p_d , p_s , a_d , a_s) fixed at the values estimated for no added contamination.

The second source of underestimation arises from our assumption that human contaminating will exactly match the frequencies specified by f . Of course, no single human genome will ever exactly match these frequencies at segregating sites, and anytime the contaminator has an ancestral allele at a $f = 1$ site, the likelihood will interpret this as non-contamination. If, instead of pulling bases from Venter's genome, we simulate ``ideal'' contamination by inserting the derived allele with probability cf , then we find the perfect estimator seen Figure S9.

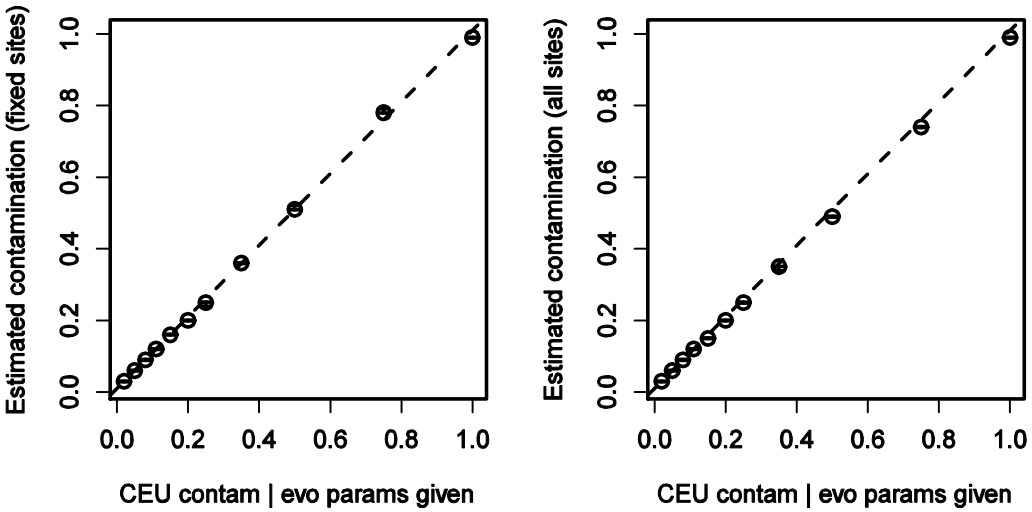


Figure S9: Performance of estimator with simulated contamination from the CEU population and evolutionary parameters (p_d , p_s , a_d , a_s) fixed at the values estimated for no added contamination.

$\Pr(N \mid H)$

The likelihood method described above requires the calculation of the evolutionary probability of the frequency of a derived allele in Neandertal given the frequency of that allele in present-day humans. Exact analytic computation of these probabilities has so far proven intractable, so we instead turn to coalescent simulations using the program *ms* (*S29*). The following plots give these probabilities for a human sample size of 100 chromosomes (50 West Africans, 50 Europeans), and a Neandertal sample size of 6 chromosomes (i.e. a three non-inbred diploid individuals from a single population) in the following four demographic scenarios:

1. simple (all populations the same size and constant across time)
2. model used in Wall et al. (*S30*), without admixture
3. model used in Wall et al. + Neandertal to European admixture from 40kya to 45kya with 4Nm = 1000
4. model used in Wall et al. + European to Neandertal admixture from 40kya to 45kya with 4Nm = 1000

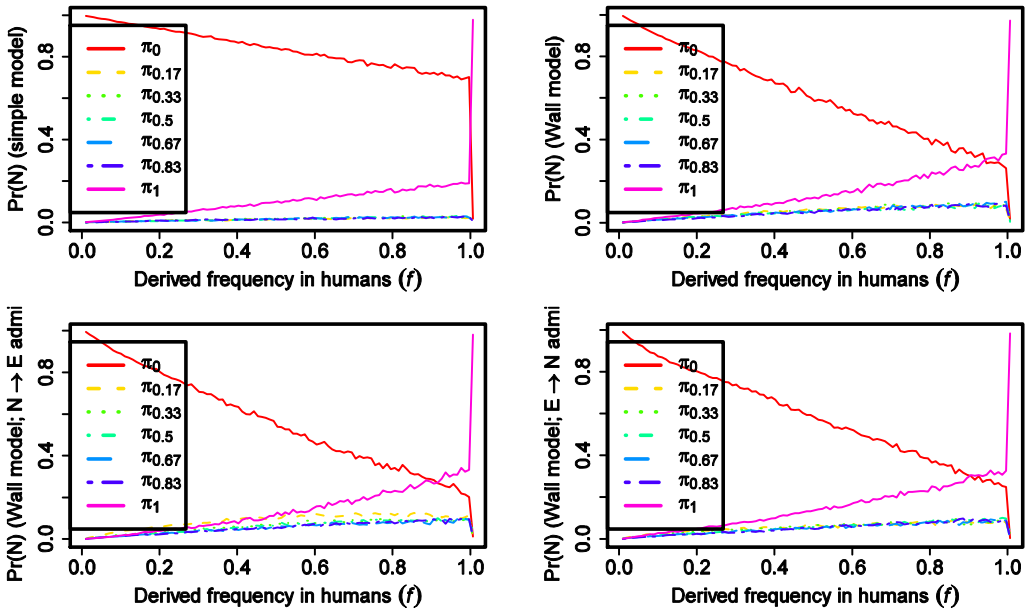


Figure S10

While these plots are all different, the functional form of each curve is close to linear with a y-intercept of 0. Further, the slope of the lines for different segregating frequencies in the Neandertal ($\pi_{0<x<1}$) are almost identical. Thus we need only to estimate the slopes of these lines to arrive at a good approximation. The $\pi_{1/6}$ line in the Neandertal to European admixture simulations (lower left panel, yellow dashed line) illustrates where how this approximation might break down; however, in our case, the approximation works well, as demonstrated in the simulated contamination results above.

Supplemental Online Material 8

Additional Neandertal Individuals

Martin Kircher, Johannes Krause, and Richard E. Green*

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

To further explore the genetic relationship between Neandertals and modern humans, three additional Neandertal bones were identified that are largely devoid of mtDNA contamination, but unfortunately show low amounts of Neandertal DNA relative to microbial sequences. DNA libraries from these bones are therefore uneconomical for genome-scale sequencing. However, as these bones cover much of the geographic range of Neandertals, they are useful for generating more modest amounts of data for comparison to the Vindija samples that all come from a single site.

We generated sequencing libraries from DNA extracts from Sidron1253 (*S31*), Feldhofer1 (*S25*), and Mezmaiskaya1 (*S32*) (Supplementary Table S20). The five sequencing libraries described in this table have been sequenced on 36 Illumina lanes (Supplementary Table S21). Library preparation and sequencing was performed as described in SOM Section 2 *DNA extraction, contamination assays, library preparation and sequencing of Neandertal libraries* and the alignment to the human and chimpanzee genome as described in SOM 3 *Neandertal DNA sequence alignment and filtering*. The total amount of raw sequence data generated and the fraction of Neandertal data in each is shown in Supplementary Table S22.

Using these data, we calculated average human-Neandertal genome divergence as discussed in the SOM Section: 10. Neandertal-modern human genome divergence.

Table S20: Information on samples, extracts, contamination level as well as number of PCR amplification cycles for the four Neandertal sequencing libraries

			PCR contamination		PEC contamination		Amplification	
Sample	**Extract**	**mg**	**Nea**	**Human**	**Nea**	**Human**	**Cycles (PCR)**	**Lib**
Mezmaiskaya 1	E149	70	na	na	2959	10	20	SL39
							20	SL61
Feldhofer 1	E142, E143	190	na	na	1436	21	18	SL37
Sidron 1253	120207.1	223	107	1	2807	7	18	SL49

Table S21: Sequencing runs for the four Neandertal sequencing libraries. This table provides an assignment of lanes, library identifiers (cross reference to Table S20 for bone assignment), read length, chemistry and analysis pipeline used. For each lane the number of raw clusters obtained, as well as the number of merged and key passed sequences used as input for further analyses is provided.

Run identifier	**Lane**	**Library identifier**	**Chemistry** 1 FC-204-20xx, 2 FC-103-300x	**Cycles**	**GA Pipeline version**	**Raw clusters**	**Key pass & Merged**
090116_SOLEXA-GA02_JK_PE_SL39	1	SL39	1	2x51	1.0	14270186	8865201
090116_SOLEXA-GA02_JK_PE_SL39	2	SL39	1	2x51	1.0	15117962	9998512
090116_SOLEXA-GA02_JK_PE_SL39	3	SL39	1	2x51	1.0	15090681	10089362
090116_SOLEXA-GA02_JK_PE_SL39	5	SL39	1	2x51	1.0	15067349	10058562
090116_SOLEXA-GA02_JK_PE_SL39	6	SL39	1	2x51	1.0	14820369	9761315
090116_SOLEXA-GA02_JK_PE_SL39	7	SL39	1	2x51	1.0	15106028	10051894
090116_SOLEXA-GA02_JK_PE_SL39	8	SL39	1	2x51	1.0	15024245	10010636
090116_SOLEXA-GA04_JK_PE_SL37_SL38	5	SL37	1	2x51	1.0	10084698	6742852
090116_SOLEXA-GA04_JK_PE_SL37_SL38	6	SL37	1	2x51	1.0	10033074	6803106
090116_SOLEXA-GA04_JK_PE_SL37_SL38	7	SL37	1	2x51	1.0	10271620	6899234
090116_SOLEXA-GA04_JK_PE_SL37_SL38	8	SL37	1	2x51	1.0	9034755	5975009
090126_SOLEXA-GA02_JK_PE_SL49	1	SL49	1	2x51	1.3.2	8510348	6033851
090126_SOLEXA-GA02_JK_PE_SL49	2	SL49	1	2x51	1.3.2	8812646	6445976
090126_SOLEXA-GA02_JK_PE_SL49	3	SL49	1	2x51	1.3.2	8574352	6289469
090126_SOLEXA-GA02_JK_PE_SL49	5	SL49	1	2x51	1.3.2	8788391	6484386
090126_SOLEXA-GA02_JK_PE_SL49	6	SL49	1	2x51	1.3.2	8765173	6473555
090126_SOLEXA-GA02_JK_PE_SL49	7	SL49	1	2x51	1.3.2	8934107	6573421
090126_SOLEXA-GA02_JK_PE_SL49	8	SL49	1	2x51	1.3.2	8914034	6492071
090309_SOLEXA-GA03_JK_PE_SL61	1	SL61	1	2x51	1.3.2	10280804	7831441
090309_SOLEXA-GA03_JK_PE_SL61	2	SL61	1	2x51	1.3.2	10833871	8479809
090309_SOLEXA-GA03_JK_PE_SL61	3	SL61	1	2x51	1.3.2	10635445	8334614
090309_SOLEXA-GA03_JK_PE_SL61	5	SL61	1	2x51	1.3.2	10689037	8431149
090309_SOLEXA-GA03_JK_PE_SL61	6	SL61	1	2x51	1.3.2	10757033	8292442
090309_SOLEXA-GA03_JK_PE_SL61	7	SL61	1	2x51	1.3.2	10626179	8534606
090309_SOLEXA-GA03_JK_PE_SL61	8	SL61	1	2x51	1.3.2	10470192	8392562
090309_SOLEXA-GA02_JK_PE_SL56_SL60-57_SL61	3	SL61	1	2x51	1.3.2	11698496	9386929
090309_SOLEXA-GA02_JK_PE_SL56_SL60-57_SL61	5	SL61	1	2x51	1.3.2	12146794	9940651

Run identifier	Lane	Library identifier	Chemistry 1 FC-204-20xx, 2 FC-103-300x	Cycles	GA Pipeline version	Raw clusters	Key pass & Merged
090309_SOLEXA-GA02_JK_PE_SL56_SL60-57_SL61	6	SL61	1	2x51	1.3.2	12349468	10132157
090309_SOLEXA-GA02_JK_PE_SL56_SL60-57_SL61	7	SL61	1	2x51	1.3.2	12551581	10283728
090309_SOLEXA-GA02_JK_PE_SL56_SL60-57_SL61	8	SL61	1	2x51	1.3.2	11944314	9800758

Table S22: Amount of sequencing data obtained for the three additional Neandertal samples

Neandertal bone	Lanes	Raw reads	Merged reads	Total unique Neandertal reads (mapped to hg18)	Neandertal sequence (bp, mapped to hg18)
Mezmaiskaya1	20	257,653,855	181,317,764	1,266,452	56,405,304
Sidron 1253	7	61,299,051	44,792,729	48,814	2,234,572
Feldhofer 1	5	51,230,295	32,948,500	44,114	2,228,645

Supplemental Online Material 9

Sequencing of five present-day humans

Matthias Meyer, Nancy Hansen, and Martin Kircher

Library Preparation for HGDP individuals

One microgram of DNA was obtained for each of the five individuals HGDP00778 (Han Chinese), HGDP00542 (Papuan), HGDP00927 (Yoruba), HDGP01029 (San) and HGDP00521 (French) from the HGDP-CEPH panel. DNA was sheared into small fragments (200 – 400 bp) using the Bioruptor UCD-200 (Diagenode). Shearing was performed four times for seven minutes at "HIGH" setting with an ON/OFF interval of 30 seconds. Illumina sequencing libraries were prepared from the sheared samples according to the 454 protocol described in Supplementary Material 2 with the following modifications for creating Illumina sequencing libraries without conversion: (i) Blunt end ligation of short adapters was followed by amplification with 5'-tailed primers, but different adapter and PCR primer sequences were used (see Table S23); (ii) for each sample a narrow band around 300 bp was excised from a 2 % agarose gel after adapter ligation to obtain inserts of optimal size for sequencing. DNA was extracted from the gel slices using the QIAquick Gel Extraction kit (Qiagen).

Table S23: Oligonucleotide sequences used in library preparation (PTO bonds indicated by *)

Name	Sequence
Adapter P5 oligo 1	5'-A*C*A*C*TCTTTCCCTACACGACGCTCTTCCG*A*T*C*T-3'
Adapter P5 oligo 2	5'-A*G*A*T*CGGAA*G*A*G*C-3'
Adapter P7 oligo 1	5'-Biotin-G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCG*A*T*C*T-3'
Adapter P7 oligo 2	identical to Adapter P5 oligo 2
PCR primer P5	5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT-3'
PCR primer P7	5'-CAAGCAGAAGACGGCATACGAGATgatgctGTGACTGGAGTTCAGACGTGT-3'

Illumina sequencing of HGDP individuals

Each of the five human Illumina libraries has been sequenced with 2 x 76 cycles on one flow cell according to the manufacturer's instructions for Multiplex sequencing on the Genome Analyzer II platform (FC-103-300x sequencing chemistry). The protocol was followed with one exception being that no index read was performed. Due to low cluster densities for the run of the French individual (HGDP00521),

another 4 lanes have been sequenced from this library on an additional sequencing run (same chemistry and protocols) for which the index read was performed, but not evaluated for these four lanes.

The runs were analyzed starting from intensity files (CIF format) using the Illumina Genome Analyzer Pipeline 1.4.0. Base calling parameter estimation for the Illumina base caller Bustard (S*12*) was done on a PhiX control lane sequenced in lane 4 of each run. The Bustard-called PhiX 174 reads were aligned to the corresponding reference sequence to obtain a training data set for the alternative base caller Ibis (S*13*). Raw sequences called from Ibis for the two paired end reads of each sequencing cluster were aligned separately with BWA (*S33*) to the human (hg18) and chimpanzee (pantro2) genome with default parameters. Using BWA's sampe command the alignments for two reads were combined and converted to SAM/BAM format (*S20*). In this step, missing paired alignments were searched within a window of 800nt around one aligned read (BWA sampe parameter -a 800). Subsequently, the BAM output files of all lanes from the same library were merged and the resulting files filtered by removing read pairs for which either the forward or reverse read failed one of the following criteria: (a) Missing the "properly paired" bit in the BAM file. (b) Mapping quality < 30. (c) "Duplicated", i.e. read pairs for which another, higher or equal quality, read pair had boundaries that map to the same outer coordinates (samtools rmdup command). (d) Reads with sequence entropy < 1.0, where entropy is calculated by summing $-p*\log_2(p)$ for each of the four nucleotides. Table S24 summarizes the number of raw reads obtained for each library, the fraction aligned as well as the fraction passing the described filter.

Table S24: Sequence data obtained for the five HGDP individuals. The table summarizes the number of raw reads for each library, the fraction aligned (*S33*) to the human (hg18) and chimp reference (pantro2) genome as well as the fraction of paired end reads obtained after removing pairs for which at least one of the reads failed one of the following filter criteria: (1) Missing the "properly paired" bit in the BAM file. (2) Mapping quality < 30. (3) "Duplicated", i.e. read pairs for which another, higher or equal quality, read pair had boundaries that map to the same outer coordinates. (4) Reads with sequence entropy < 1.0, where entropy is calculated by summing $-p*\log_2(p)$ for each of the four nucleotides.

Sample	HGDP00778	HGDP00542	HGDP00927	HDGP01029	HGDP00521
Name	Han Chinese	Papuan	Yoruba	San	French
Raw reads	203,092,890	201,695,232	234,562,514	322,490,746	245,570,296
Chimpanzee (pantro2)	174,666,290 (86.0%)	164,630,156 (81.6%)	198,199,303 (84.5%)	266,555,406 (82.7%)	205,523,864 (83.7%)
filtered	138,448,786 (68.2%)	131,635,040 (65.3%)	159,019,706 (67.8%)	215,603,094 (66.9%)	162,032,716 (66.0%)
Human (hg18)	184,070,514 (90.6%)	177,444,185 (88.0%)	210,052,921 (89.6%)	282,364,120 (87.6%)	217,075,840 (88.4%)
filtered	154,826,008 (76.2%)	150,620,614 (74.7%)	177,997,474 (75.9%)	240,522,810 (74.6%)	181,488,714 (73.9%)

Supplemental Online Material 10

Neandertal – modern human genome divergence

Richard E. Green*

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

To estimate the average genome divergence between Neandertals and modern humans, we generated 3-way alignment between these two and the chimpanzee (panTro2). We then counted the number of differences that are unique to each of the three lineages: modern human (n_H), chimpanzee (n_C), and Neandertal (n_N). Assuming constant evolutionary rates along lineages, modern human-Neandertal divergence can be reported as a ratio of the branch length leading to modern humans since Neandertal divergence and half the total distance between chimpanzees and modern humans:

$$n_H / ((n_C + n_H)/2)$$

The accuracy of such an estimate depends critically on correctly identifying orthologous segments and accurately aligning them (*S34*). Because the amount of sequence data available is not limiting for such an analysis, we stringently filtered the data to ensure 3-way orthology using the following methodology.

We began by extracting all regions of single-copy aligned human and chimpanzee sequence from the 4-way alignment of human, chimpanzee, orangutan, and macaque as provided by the Enredo-Pecan-Ortheus (EPO) whole-genome multiple sequence alignment of these genomes (*S15-17*). We took only aligned blocks that were represented by a single human and chimpanzee segment to avoid human-lineage or chimpanzee-lineage duplicated segments. We also required at least one segment from orangutan or macaque to be present to ensure reasonable inference of the human-chimpanzee ancestral state. We used the inferred human-chimpanzee single copy ancestral sequence (HCSCCA) for all such regions to align the Neandertal reads using *ANFO*, (see SOM 3). We note that *ANFO* makes use of DNA ambiguity codes during alignment, minimizing the loss in sensitivity that occurs when mapping against a more distant reference, such as the common ancestor sequence. For each read mapping to an unambiguous position within the HCSCCA (map-quality >= 30), we

compared the separate *ANFO* mapping coordinates to the human and chimpanzee genomes. We required unambiguous best alignment positions within each of these genomes (map-quality >=30) that are compatible with one another via the hg18topanTro2 whole genome alignment and the panTro2tohg18 whole genome alignment from UCSC. Because each aligned segment in the HCSCCA comes from a discrete segment of the human and chimpanzee genome, we also required that these alignment positions were equivalent to the position in the HCSCCA. In summary, there were independent, direct mappings of Neandertal sequence to three genome sequences: the human genome, N(hg18), the chimpanzee genome, N(pt2), and the inferred single-copy common ancestor sequence, N(HCSCCA). There were three further, indirect mappings inferred via whole genome alignments, N(hg18-via-pt2), N(pt2-via-hg18), and N(hg18-via-HCSCCA). Our unambiguous orthology filter required equivalent coordinates between:
N(hg18) and N(hg18-via-pts);
N(hg18) and N(hg18-via-HCSCCA);
N(pt2) and N(pt2-via-hg18)

This strategy is shown schematically in Figure S14.

We used a similar strategy for the five HGDP current human samples. However, in this case we used the *bwa(S20)* mappings instead of *ANFO* mappings. Furthermore, we filtered these reads for consistent paired-end coordinates on the hg18 mappings.

For all reads passing this filter, we then generated a human/Neandertal/chimpanzee 3-way alignment. At each HCSCCA genomic position, we randomly sampled a single allele, if any, passing the base- and library-specific quality cutoffs. We ignored sites where the data coverage was higher than the 95th percentile for that sample, as coverage outlier sites could be symptomatic of regions of mapping problems. These coverage cutoffs are shown in Table S25. We then used the HCSCCA-chimpanzee-human (EPO) alignment to extract the corresponding human and chimpanzee base at each position. This strategy circumvents the bias inherent in progressive multiple sequence alignment (*S16*).

From these alignments, we counted the number of lineage-specific differences between chimpanzee, the reference human genome, and the Neandertal. We summed over all autosomes and considered the X chromosome separately. Because of the high rate of misincorporation-induced transitions in the Neandertal sequence, we restricted the divergence estimation to only transversion sites. This has the additional benefit of avoiding parallel substitutions in any of the three lineages as transversions accumulate more slowly than transitions.

The reference human genome is a mosaic of BAC sequences whose ancestry can be inferred (S*35*), which presents an opportunity to explore the genetic relationship between Neandertals and current human populations. We focused on regions of the reference human genome contributed by the BAC library RPCI-11 as it has been shown to be contributed by an individual with both African and European ancestry (S*35*). Using only regions of the reference human genome derived from confidently assigned ancestry, we partitioned the human reference into regions of inferred African and European ancestry. We then calculated the Neandertal/reference human divergence in each partition separately. The divergence data for each individual at all sites, and by these partitions is shown in Figure S11 and Table S26. The substitution spectra for the three, lower-coverage Neandertals is shown in Figure S12. Note that the extreme excess of C to T and G to A substitutions, typical for ancient DNA, is seen in each of these Neandertal datasets. In contrast, Figure S13 shows the divergence estimate and substitution spectra for the five HGDP individuals.

This partitioning of the reference human genome into regions of European and African ancestry provides a convenient way to cross-check our human divergence estimates. Because the filtering and analysis methodology presented here was developed specifically to deal with the peculiarities of ancient DNA sequence, its applicability to modern DNA sequence is not obvious.

Prior estimates of Yoruban divergence have been presented (S*36*). An estimate from the Illumina sequencing of about 40 fold coverage of a Yoruban individual (S*37*) and further estimates generate Yoruban/Yoruba divergence estimates that range between 8.00% and 8.30%. The divergence estimate presented here between the Yoruban sample and the inferred African portions of the reference genome, 9.17%, we note is

10 to 15% higher than these previous estimates. What follows is an exploration of the effects of differences in the methodology applied to arrive at these estimates.

The discrepancy between the estimate here and these external estimates are partially explicable by methodological differences used to arrive at these numbers. The divergence estimate methodology designed for ancient Neandertal DNA is necessarily restricted to transversion differences because of the high rate of erroneous transitions in ancient DNA (S*38*). We apply this restriction to the HGDP individual divergences as well, in the interests of generating comparable numbers. However, previous estimates of Yoruba divergence do not include this restriction. Therefore, we removed this filter and examined both transitions and transversions, but excluding CpG sites and estimate a Yoruba divergence of 9.09%, lower than the 9.17% when examining only transversions.

Another difference between then input sequence data analyzed here and that used for previous estimates concerns sequencing error rates. Sampling a single base at each position renders the data effectively 1x coverage. We estimate the error rate of the human sequencing data in this analysis by considering the excess of substitutions localized to the HGDP individual lineage as shown in Figure S13.

Sequencing error in the HGDP individuals will bias the divergence. The basis for this effect can be demonstrated by considering the effect of sequencing error in the HGDP sequence when aligned to the reference human and reference chimpanzee, which are considered here to be error-free. Each position in the reference chimpanzee (*c)*, reference human (*h*), and HGDP human (*s*) alignment can be classified into one of five possible categories:

a_1 = sites where *c, h,* and *s* are identical, i.e., no substitutions have occurred

a_2 = sites where *h* and *s* are identical but different from *c*, (n_C sites)

a_3 = sites where *c* and *h* are identical, *s* is different (combination of sequencing error and sites of divergence in the HGDP individual)

a_4 = sites where *c* and *s* are identical, *h* is different (n_H sites)

a_5 = sites where *c, h,* and *s* are all different

Assuming equal branch lengths of the reference human and the HGDP individual, the relative number of each category of sites is expected to be:

$$a_1 >>> a_2 >> a_3 > a_4 > a_5$$

Note, however, that only a_2 and a_4 sites are used to calculate divergence. Therefore, we consider the effect of sequencing error in falsely inflating or reducing these numbers. If we assume that sequencing error will cause sites to be misclassified randomly the overall effect is a function only of the overall sequencing error rate, e, and the number of actual counts in each category. The largest overall effect of sequencing error will be that a_1 sites are falsely classified as a_3 sites. That is, sites where no actual divergence has occurred will appear as diverged in the HGDP sample. The largest effect of consequence for divergence estimation occurs when sites that are actually a_2 have a sequencing error in the HGDP base. In this case, two-thirds of the errors will cause this position to be falsely classified as a_5. One-third, however, will cause these sites to be falsely classified as a_4. In other words, sites where an actual substitution separating chimpanzees from the humans occurred will appear as sites on the reference human lineage. This misclassification increases the numerator and decreases the denominator of the divergence equation:

$$n_H / ((n_C + n_H)/2)$$

To formalize the effect of sequence error in our counts, we set up this matrix transformation to learn the actual numbers in each category (a_1, a_2, a_3, a_4, a_5) given a rate of sequencing error, e, and the observed numbers in each category (b_1, b_2, b_3, b_4, b_5):

$$\begin{bmatrix} 1-e & 0 & e/3 & 0 & 0 \\ 0 & 1-e & 0 & e/3 & e/3 \\ e & 0 & 1-e/3 & 0 & 0 \\ 0 & e/3 & 0 & 1-e & e/3 \\ 0 & 2e/3 & 0 & 2e/3 & 1-2e/3 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix}$$

We can then solve this matrix for (a_1, a_2, a_3, a_4, a_5) to neutralize the effect of misclassification due to sequencing error. Using the rate of sequencing error estimated from the increased number of Yoruban lineage differences relative to that of the reference human, 0.005059, we arrive at a Yoruban divergence estimate of 8.78%.

We note that this estimate, 8.78% still represents a 7%-10% increase relative to the mean of previous Yoruba divergence estimates. The source of this residual increase may lie in differences in filtering for unambiguous orthology that are necessary for accurate Neandertal divergence estimation that less tractable to explore using comparison, high coverage datasets. In any case, we note that the HGDP divergence estimates presented here may be systematically biased to be deeper than previous estimates. They are presented for comparison to the Neandertal divergence estimates, using the Neandertal divergence estimate methodology. Table S27 shows the divergence estimates for each of the five HGDP individuals relative to the reference human genome and to the inferred European and African portions under these two modifications.

To investigate the regional variation in divergence, we calculated divergence, $n_H / ((n_C + n_H)/2)$, between Neandertals and each HGDP individual and the reference human in non-overlapping 100kb segments across the genome. The distribution of this regional divergence for each 100kb segment for which there were at least 50 counts of $n_H + n_C$ is shown in Figure 3.

Table S25: Sequence coverage cutoffs at the 95th percentile of the distribution of coverage depth. These cutoffs were used to exclude sites of high-coverage.

Individual	Coverage cutoff (95th percentile)
Vi33.16	2
Vi33.25	2
Vi33.26	2
Feldhofer1	1
Mezmaiskaya1	1
Sidron1253	1
French	10
San	12
Han	8
Papuan	8
Yoruban	9

Table S26: Divergence to the reference human genome. 95% confidence intervals are calculated using a block-jackknife over 10Mb segments of the indicated segments of the genome.

Individual	category	n_C	n_N	n_H	$n_H / ((n_C + n_H)/2)$ [95% CI] %
Vi33.16	autosomes	449,619	129,103	30,413	12.67 [12.45-12.90]
	X chromosome	20,106	7,158	1,297	12.12 [11.49-12.75]
	Eur. autosomes	121,614	35,567	7,863	12.15 [11.78-12.51]
	Afr. autosomes	66,603	19,062	4,688	13.15 [12.68-13.62]
Vi33.25	autosomes	478,270	204,845	32,347	12.67 [12.45-12.89]
	X chromosome	22,823	12,392	1,507	12.39 [11.53-13.25]
	Eur. autosomes	129,882	56,995	8,510	12.30 [11.93-12.66]
	Afr. autosomes	72,346	31,015	5,091	13.15 [12.67-13.62]
Vi33.26	autosomes	451,459	111,215	30,548	12.68 [12.45-12.90]
	X chromosome	21,661	6,343	1,367	11.87 [11.12-12.62]
	Eur. autosomes	122,611	30,561	8,097	12.39 [12.02-12.76]
	Afr. autosomes	68,540	16,841	4,767	13.01 [12.45-13.56]
Feldhofer1	autosomes	1,022	693	66	12.13 [5.94-18.33]
	X chromosome	15	24	5	
	Eur. autosomes				
	Afr. autosomes				
Mez1	autosomes	22,497	9,499	1,527	12.71 [12.08-13.35]
	X chromosome	937	541	71	
	Eur. autosomes	6,058	2,589	416	
	Afr. autosomes	3,172	1,386	217	
Sidron1253	autosomes	950	700	81	15.71 [9.05-22.38]
	X chromosome	34	121	3	
	Eur. autosomes				
	Afr. autosomes				
French	autosomes	1,600,826	872,504	66,709	8.00 [7.78-8.22]
	X chromosome	20,106	7,158	1,297	5.28 [4.63-5.94]
	Eur. autosomes	459,724	252,265	16,741	7.03 [6.68-7.37]
	Afr. autosomes	232,630	126,336	11,325	9.28 [8.93-9.64]
Han	autosomes	449,619	129,103	30,413	8.45 [8.23-8.67]
	X chromosome	74,286	49,247	2,015	5.84 [5.03-6.65]
	Eur. autosomes	121,614	35,567	7,863	7.69 [7.38-8.01]
	Afr. autosomes	66,603	19,062	4,688	9.59 [9.21-10.00]
Papuan	autosomes	1,620,290	3,163,531	79,201	9.32 [9.09-9.55]
	X chromosome	72,844	173,944	2,255	6.01 [5.53-6.48]
	Eur. autosomes	464,691	922,939	20,913	8.61 [8.26-8.97]
	Afr. autosomes	239,988	458,472	12,820	10.14 [9.77-10.52]
Yoruban	autosomes	1,660,528	1,570,486	82,123	9.43 [9.22-9.63]
	X chromosome	77,606	88,963	3,097	7.68 [7.04-8.31]
	Eur. autosomes	474,635	544,854	23,075	9.27 [8.94-9.59]
	Afr. autosomes	245,851	225,165	11,817	9.17 [8.85-9.50]
San	autosomes	1,729,241	1,573,773	94,126	10.32 [10.12-10.53]
	X chromosome	85,438	95,905	4,055	9.06 [8.48-9.64]
	Eur. autosomes	494,310	457,706	26,225	10.08 [9.73-10.42]
	Afr. autosomes	256,784	227,963	13,787	10.19 [9.86-10.53]

Table S27: Divergence estimates of HGDP samples using all sites except CpG sites and error neutralization. Confidence intervals are calculated here using the binomial variance.

Individual	category	error rate for error neutralization	n_C	n_H	$n_H / ((n_C + n_H)/2)$ [95% CI] %
French	autosomes	none	3,588,873	147,691	7.91 [7.87-7.94]
	autosomes	0.003007	3,599,550	144,515	7.72 [7.68-7.76]
	European autosomes	none	1,039,766	36,921	6.86 [6.79-6.93]
	European Autosomes	0.003007	1,042,870	35,983	6.67 [6.60-6.74]
	African Autosomes	none	527,189	25,258	9.14 [9.03-9.25]
	African autosomes	0.003007	528,754	24,802	8.96 [8.85-9.07]
Han	autosomes	none	3,693,544	160,908	8.35 [8.31-8.39]
	autosomes	0.004931	3,711,590	155,571	8.05 [8.01-8.09]
	European autosomes	none	1,068,160	42,150	7.59 [7.52-7.66]
	European Autosomes	0.004931	1,073,390	40,585	7.29 [7.22-7.36]
	African Autosomes	none	548,542	27,313	9.49 [9.38-9.60]
	African autosomes	0.004931	551216	26,537	9.19 [9.08-9.29]
Papuan	autosomes	none	3,668,521	176,932	9.20 [9.16-9.24]
	autosomes	0.010793	3,707,940	165,366	8.54 [8.50-8.58]
	European autosomes	none	1,061,023	46,598	8.41 [8.34-8.49]
	European Autosomes	0.010793	1,072,440	43,204	7.75 [7.67-7.82]
	African Autosomes	none	548,732	28,914	10.01 [9.90-10.12]
	African autosomes	0.010793	554,619	27,211	9.35 [9.24-9.46]
Yoruban	autosomes	No	3,742,642	183,373	9.34 [9.30-9.38]
	autosomes	0.005205	3,761,910	177,767	9.02 [8.98-9.07]
	European autosomes	none	1,078,460	51,874	9.18 [9.10-9.26]
	European Autosomes	0.005205	1,084,010	50,254	8.86 [8.79-8.94]
	African Autosomes	none	559,961	26,657	9.09 [8.98-9.20]
	African autosomes	0.005205	562,845	25,814	8.77 [8.67-8.88]
San	autosomes	none	3,903,203	209,704	10.20 [10.15-10.24]
	autosomes	0.004960	3,922,310	204,266	9.90 [9.86-9.94]
	European autosomes	none	1,125,410	58,709	9.92 [9.83-9.99]
	European Autosomes	0.004960	1,130,920	57,121	9.62 [9.54-9.69]
	African Autosomes	none	585,168	31,114	10.10 [9.99-10.21]
	African autosomes	0.004960	588,034	30,291	9.80 [9.69-9.91]

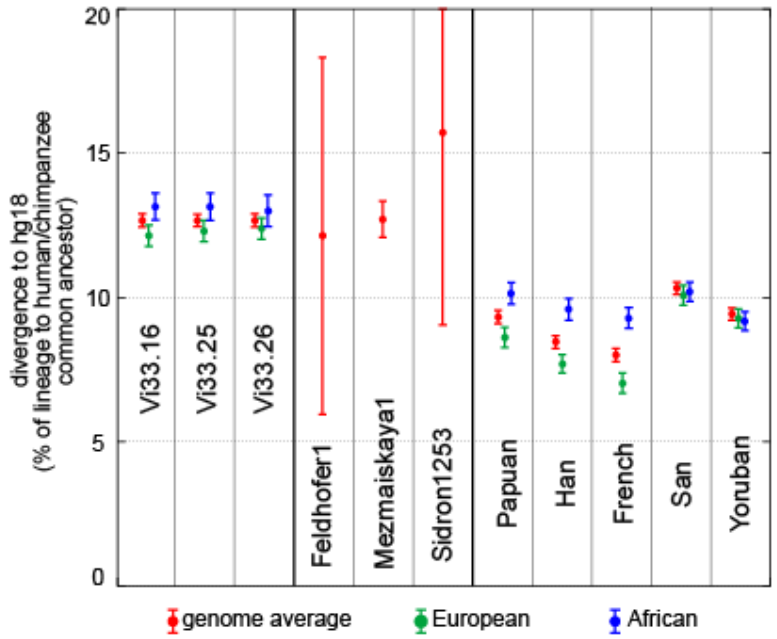


Figure S11: Divergence of Neandertal and human genomes. The average genome divergence is shown between the reference human genome (hg18) and the genomes of Neandertals from Croatia (Vindija), Germany (Feldhofer1), Russia (Mezmaiskaya1), and Spain (Sidron1253), as well as the five present-day humans sequenced in this study. For each individual, divergence is expressed as a percentage of the hominin lineage, i.e., the time since human-chimpanzee divergence. The genome-wide average is shown in red. The average for segments of the genome with inferred ancestry of European (green) and African (blue) ancestry are shown separately. Error bars are 95% confidence intervals calculated using a block-jackknife over 10 megabase segments of the genome and are especially large for the Feldhofer1 and Sidron1253 samples for which little data was collected

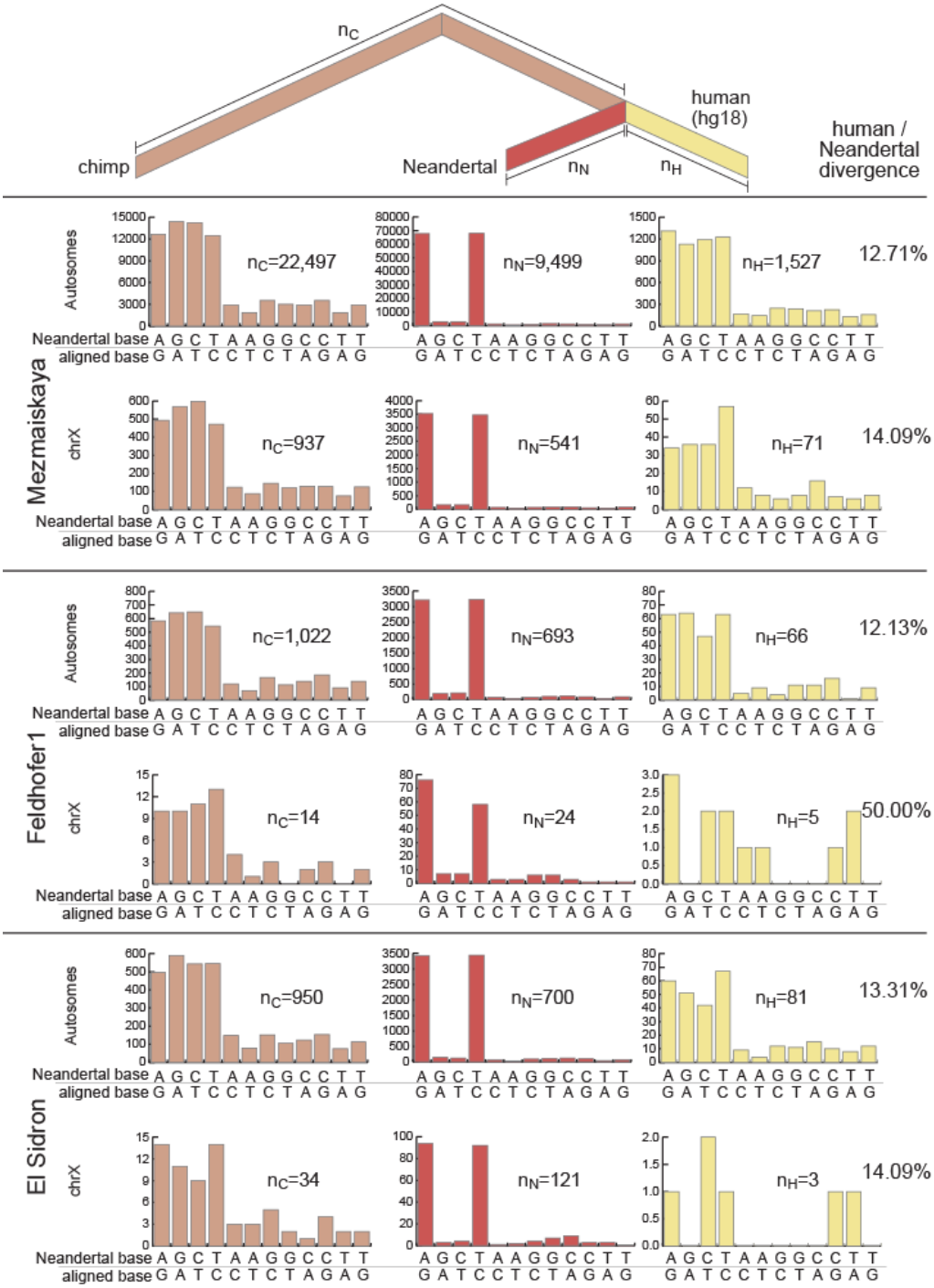


Figure S12: Divergence estimate and lineage substitution spectra of three Neandertals from Russia, Germany and Spain. n_C, n_N, and n_H values are for transversion sites only, i.e., those used to calculate divergence.

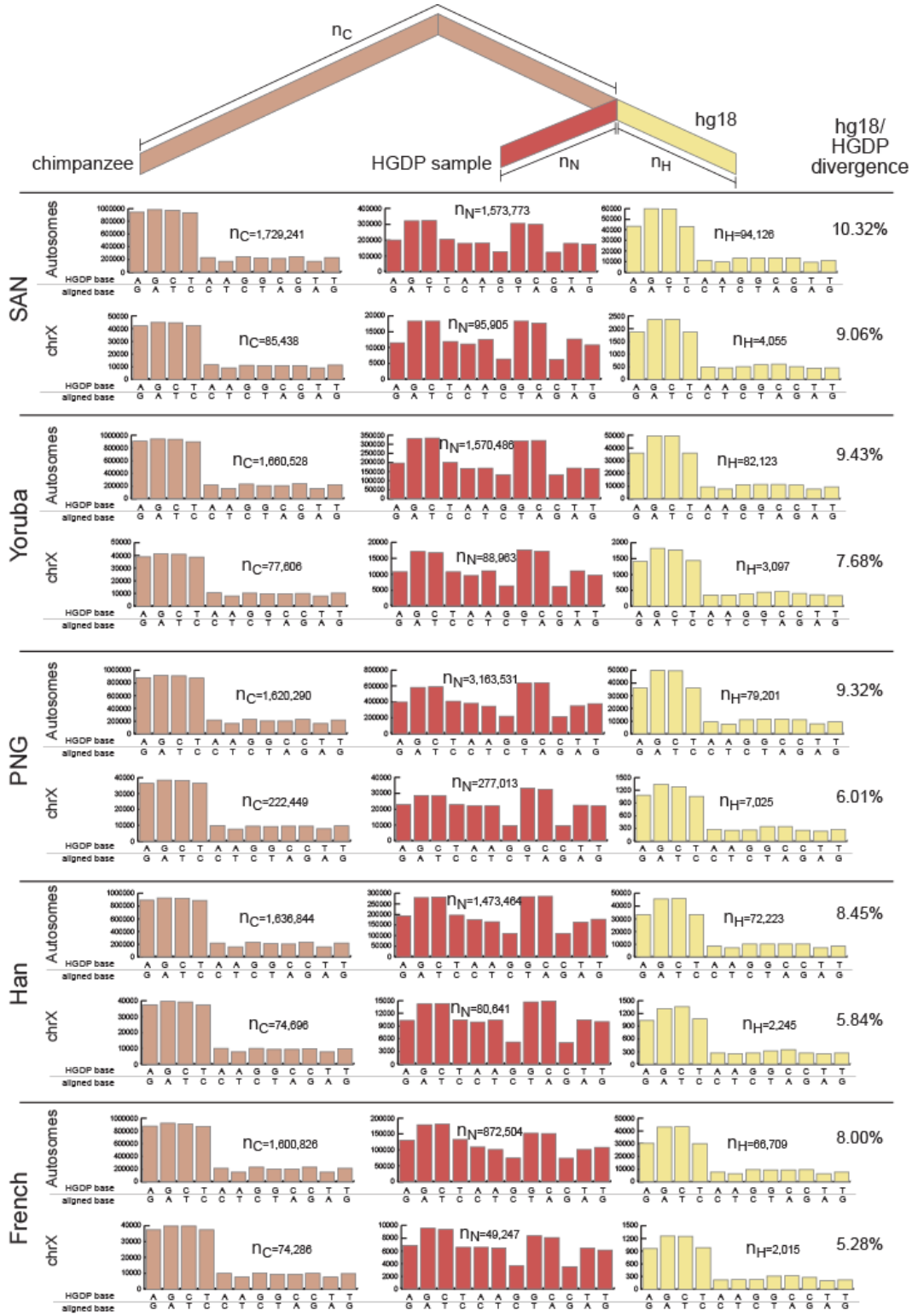


Figure S13: Divergence estimate and lineage substitution spectra of five HGDP individuals relative to the reference human sequence.

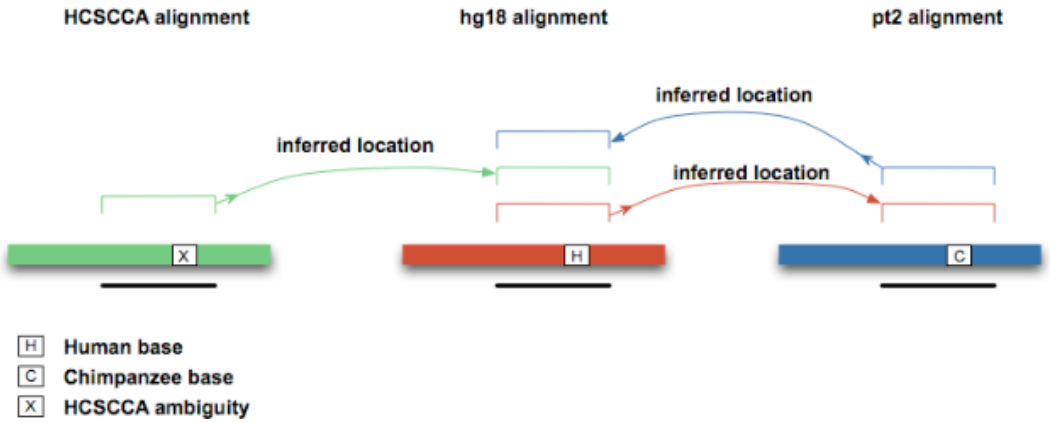


Figure S14: Filter for unambiguous orthology mapping of reads. Each Neandertal read was independently mapped to the human genome (hg18), the chimpanzee genome (pt2), and the inferred common ancestor sequence (HCSCCA). The concordance of all alignment positions was checked via the whole-genome alignments of genomes. Any discordantly mapped read was not considered in divergence calculations.

Supplemental Online Material 11

A catalogue of features unique to the human genome

Martin Kircher, Udo Stenzel, Hernán A. Burbano, and Janet Kelso*

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

Identification of changes on the whole human lineage

We identified positions that have changed on the human lineage using whole genome alignments of human (hg18), chimpanzee (pantro2), orangutan (ponabe2) and rhesus macaque (rhemac2). Multi-species whole genome alignments were created from pair-wise alignments available from the UCSC Genome browser (http://hgdownload.cse.ucsc.edu/downloads.html) using autoMultiZ from the UCSC Kent-tools (http://hgdownload.cse.ucsc.edu/admin/exe/). Whole genome alignments differentiate between a reference and a target genome. While each position of the reference genome appears at most once in the alignment, sequences from the target genome can be used multiple times. This causes whole genome alignments to be non-symmetrical. We therefore used two multi-species whole genome alignments, one based on hg18 as reference and the other based on pantro2 as reference. These alignments were screened for differences between the human and chimpanzee sequences, and the lineage on which the change occurred was assigned based on two out-group sequences (orangutan and rhesus macaque). We identified 15,216,383 single nucleotide changes (SNCs) and 1,364,433 insertion or deletion changes (InDels) from the human-based alignment and 15,523,445 SNCs and 1,507,910 InDels from the chimpanzee-based alignment. We filtered these datasets, requiring that positions be identified in both human-based and chimpanzee-based alignments. We also required (1) that no gaps be present within a 5nt-window of the event; (2) that there is sequence available for both out-groups and that they are identical; (3) that InDel length does not vary between species; and (4) that an InDel sequence is not marked as a repeat. This generates a set of 10,535,445 SNCs and 479,863 InDels inferred to have occurred on the human lineage.

Identification of positions with Neandertal sequence coverage

We aligned all Neandertal sequence reads from Vi33.16, Vi33.25 and Vi33.26 to the human (hg18) and chimpanzee genomes (pantro2) using ANFO (SOM 3). To reduce

the effects of sequencing error, we used the alignments of Neandertal reads to the human and chimpanzee reference genomes to construct human-based and chimpanzee-based consensus "minicontigs" (SOM File 11.1). To generate the consensus, we selected uniquely placed, overlapping alignments (ANFO MAPQ ≥ 90) and merged these into a single multi-sequence alignment using the common reference genome sequence. At each position in the resulting alignment, for each observed base, and for each possible original base, we calculate the likelihood of the observation, estimate the likely length of single stranded overhangs, and consider the potential for ancient DNA damage using the Briggs-Johnson model (*S28*). If most observations in a given position show a gap, the consensus becomes a gap, otherwise the base with the highest quality score (calculated by dividing each likelihood by the total likelihood) is used as the consensus. At the current coverage, heterozygous sites will appear as low quality bases with the second base having a similar likelihood to the consensus base. Likewise, heterozygous InDels are included only by chance or may show up as stretches of low quality bases.

We extracted the Neandertal sequence for the identified human-lineage-specific changes from minicontig alignments to both the human and the chimpanzee genomes. To ensure consistency in the alignment process, we further filtered the data:

(1) The Neandertal sequence at the same position in both human and chimpanzee-based alignments was required to be identical and to have a PHRED quality score > 30.

(2) All positions that fall within 5 nucleotides of the ends of minicontigs were excluded to minimize alignment errors and substitutions due to the nucleotide misincorporations, which are frequent close to the ends of ancient DNA molecules.

(3) Positions that fall within 5 nucleotides of insertions or deletions (i.e. gaps) in the minicontig alignments were excluded.

Using this filtered dataset, we have Neandertal sequence coverage for 3,202,190 of the 10,535,445 substitutions and 69,029 of 479,863 InDels inferred to have occurred on the human lineage, respectively (SOM File 11.2).

Annotation

We annotated all SNC and InDel events using the Ensembl v54 annotation for hg18 and Ensembl v55 for pantro2 (in cases where no human annotation was available). A

set of 16,762 human CCDS genes, each representing the longest annotated coding sequence for the respective gene, was used for downstream analyses.

Using the latter annotation, we identified SNC and InDel events in CCDS for which we have reliable Neandertal sequence coverage, and assigned them to the lineage on which they occurred (Figure S15). We discuss below each class of substitutions and highlight features of interest.

Changes in protein coding sequence

Amino acid substitutions

We identified 19,780 SNCs in the coding regions of the human CCDS set. Six of these occur in two overlapping transcripts, while one occurs in three overlapping transcripts. These positions result in 11,337 synonymous substitutions and 8,451 non-synonymous substitutions. Non-synonymous amino acid substitutions that have become fixed in modern humans since the separation from Neandertals might be of special interest. We therefore excluded all non-synonymous substitutions where current humans are known to vary (dbSNP v130), and identified 78 fixed, non-synonymous amino substitutions from a total of 2,910 positions where the Neandertal carries the ancestral (chimpanzee) allele (Table S28 and SOM File 11.2).

We identify five genes affected by two substitutions that either change amino acids or introduce a stop codon, and that have become fixed among humans since the divergence from Neandertals:

(1) *DCHS1* (CCDS7771), which encodes fibroblast cadherin-1, a calcium-dependent cell-cell adhesion molecule that may be involved in wound healing;

(2) *RPTN* (CCDS41397), which encodes repetin, an epidermal matrix protein that is expressed in the epidermis and particularly strongly in eccrine sweat glands, the inner sheaths of hair roots and the filiform papilli of the tongue.

(3) *SPAG17* (CCDS899) sperm-associated antigen-17 that is thought to be important for the structural integrity of the central apparatus of the sperm axoneme, which is important for flagellar movement.

(4) *TTF1* (CCDS6948), a terminator of ribosomal gene transcription and regulator of RNA polymerase I transcription, and

(5) *SOLH* (CCDS10410), which encodes a protein of unknown function.

It is striking that two of these genes are expressed primarily in the skin. This may suggest that modern humans and Neandertals differed with respect skin morphology and physiology.

Besides the number of changes in each gene, the potential physicochemical impact of exchanging an amino acid in a protein is relevant for prioritizing these 78 positions. We have categorized the amino acid replacements into classes of chemical similarity (Table S28 and Table 2 of the main text) using Grantham scores (S*39*). Based on the classification proposed by Li (S*40*) scores from 0-50 are considered conservative, 51-100 are moderately conservative, 101-150 moderately radical and >151 are considered radical.

On this basis, only one of the substitutions in the five genes with multiple SNCs is considered radical, resulting in the change of codon 431 in *sperm associated antigen 17* from the ancestral aspartic acid to the derived tyrosine.

A further four of the complete list of 78 nucleotide substitutions result in radical amino acid changes, 7 in moderately radical changes, 33 in moderately conservative, 32 in conservative changes and a single one affects a stop-codon (Table S28). The genes showing radical amino acid substitutions are involved in reproduction, hormone response, olfaction, and immunity - groups which have been shown in human-chimpanzee genome comparisons to have undergone positive selection (S*41*).

(1) *GREB1* (CCDS42655, *Gene regulated in breast cancer 1*), an estrogen-responsive gene which is an early response gene in the estrogen receptor-regulated pathway. The amino acid substitution in *GREB1* occurs in a serine-rich region of the protein.

(2) *OR1K1* (CCDS35132, *olfactory receptor, family 1, subfamily K, member 1*), an olfactory receptor, has an exchange of arginine to cysteine in one of the extracellular domains of the protein.

(3) *NLRX1* (CCDS8416, *nucleotide-binding oligomerization domain, leucine rich repeat containing X1*) acts as a modulator of the innate immune response elicited from the mitochondria in response to viral challenge. Expression of *NLRX1* results in inhibition of the *RLH* and *MAVS*-mediated interferon-beta promoter activity and in the disruption of virus-induced *RLH-MAVS* interactions. The amino acid substitution is in the *NACHT* domain of the protein which is thought to interact with *MAV* in order to bring about the innate viral response.

(4) *NSUN3* (CCDS2927, *NOL1/NOP2/Sun domain family 3*) is a protein of unknown function which seems to have S-adenosyl-L-methionine-dependent methyl-transferase activity.

Each of the rather small number of amino acid substitutions that have become fixed in humans since the divergence from Neandertals, particularly those with potentially radical effects on the protein structure, may be of sufficient interest to investigate functionally. For example, in the 19,780 SNCs falling in coding sequences in the present genome analysis, we find 175 changes in signal peptides, 106 are non-synonymous. For 91 of these positions Neandertal shows the derived state, and for 15 sites Neandertal shows the ancestral state. All the latter changes are known to be polymorphic in modern humans (dbSNP 130) and are therefore possibly functionally equivalent and may not have been relevant in modern human evolution.

Stop/Start codon substitutions

We identified only one gene (*RPTN*, CCDS41397) in which a fixed, non-synonymous substitution introduces a stop codon in the human protein which is not seen in Neandertal. We examined the Neandertal minicontigs in the region surrounding the position of the human stop codon and identified only one stop codon - 108 amino acids downstream of the position at which the human stop codon is observed. The earlier mentioned human *RPTN* protein is thus shortened by 108 amino acids - from 892 amino acids in Neandertal to 784 amino acid residues in humans. A second gene (*KIAA1751*, CCDS3097) carries a stop codon showing a non-synonymous change which is known to be polymorphic in modern humans.

We identified no fixed, non-synonymous changes in start codons where Neandertal shows the ancestral allele. Just one non-synonymous change which is not fixed in modern humans was identified in the melastatin gene (*TRPM1*). Functional variants of the *TRPM1* (CCDS10024) that use alternative start positions have been described in human tissues and may be able to compensate for the loss of the specific transcript variant (S*42*). *TRPM1* encodes an ion channel with the function of maintaining normal melanocyte pigmentation.

Indels in coding sequence

We identified 36 insertion/deletion events within coding sequences. In four cases the Neandertal is ancestral, and in all of these cases are modern humans known to be polymorphic for the position (SOM File 11.2).

Gene ontology analysis

We tested the set of the 78 genes with fixed, human-specific amino acid changes for enrichment in specific categories of the Biological Process division of the Gene Ontology (S*43*). We used the GOstats package from the Bioconductor project (S*44*) together with the statistical software package R (S*45*). The Bioconductor project includes GO annotation data for Entrez Gene identifiers as well as mappings to Ensembl Gene identifiers, which have been used as input. As the background set, we used all 16,762 CCDS annotated Ensembl Genes. Significant associations ($p < 0.01$) were found to genes associated with mesoderm development, transcriptional pre-initiation, sterol and lipoprotein metabolism (Table S29).

Changes in non protein-coding sequences

5'-UTR substitutions and insertion/deletions

We have reliable Neandertal sequence data for 2,616 of the 12,045 substitutions in 5'-UTRs occurring on the human lineage. Of these, 42 affect positions where the ancestral allele is observed in Neandertals, and humans are fixed derived. Two genes show multiple changes. *TMEM105* (CCDS11781), a transmembrane protein with no known function, has three changes in the 5'UTR, and *SLC25A2* (CCDS4258), which is thought to have a role in metabolism as a mitochondrial transport protein, has two such changes. Neandertal state information was also obtained for 71 of 810 InDels in 5'UTRs; three of them show the ancestral state retained in Neandertals (*ARHGEF11*/CCDS1163, *ZNF564*/CCDS42505, *RIBC2*/CCDS14066) while present-day humans are fixed derived.

When testing all 42 genes with substitution or InDel changes for enrichment in GO terms of the Biological Process category (as described before), a diverse set of terms associated with development, gene expression regulation, cell division and DNA replication/repair, lipid metabolism and other metabolic processes was obtained (Table S30).

3'-UTR substitutions and insertion/deletions

We have reliable Neandertal sequence data for 18,909 of 55,883 substitutions in 3'-UTRs. Among these, there are 190 positions where Neandertal shows the ancestral state and modern humans are fixed derived. Twelve genes show multiple substitutions, with one gene having four substitutions (*CCDC117*/CCDS13846) and three genes having three substitutions each (*ATP9A*/CCDS33489, *LMNB2*/CCDS12090, *RCOR1*/CCDS9974). We also identify 784 of 5,972 InDels in 3'-UTRs, 33 of which show the ancestral state in Neandertals while modern humans are fixed derived. Each of the 33 InDels is found in a different gene.

When testing all 173 genes with substitution or InDel changes in their 3'-UTR sequence for enrichment in GO terms of the Biological Process category (as described above), dendrite morphogenesis, mitochondrial membrane organization, adult locomotory behavior, insulin receptor signaling pathway and camera-type eye development are over-represented (Table S31).

miRNAs

MiRNAs are small non-coding RNAs that regulate gene expression by mRNA cleavage or repression of mRNA translation. MiRNAs have been shown to have important role in mammalian brain and embryonic development. We have Neandertal sequence data for 103 of the 357 single nucleotide changes in 1,685 miRNAs annotated in Ensembl 54 (including 670 miRBase-derived microRNAs). In 88 cases the Neandertal carries the derived allele. The remaining 15 alleles are ancestral. In only one case, ENSG00000221170 (*hsa-mir-1304*), is the Neandertal state ancestral and human state fixed derived (Figure S16A). For *hsa-mir-1304*, the substitution occurs in the seed region of the mature miRNA, suggesting that it is likely to have altered target specificity in present-day humans relative to Neandertals and the outgroups. However, folding is unlikely to be changed since base pairing is unaffected by the substitution. Reliable Neandertal sequence data is also available for two of the seventeen insertion/deletion events in miRNAs that occurred on the human lineage. In ENSG00000211530 (*AL354933.8*) the Neandertal has the derived allele, while in ENSG00000212045 (*AC109351.3*) the allele is ancestral while modern humans are fixed derived. ENSG00000211530 has a large loop that is one base shorter in Neandertal than in human (Figure S16B). This change is not expected to alter folding of the miRNA or to change target specificity of the putative miRNA.

Human Accelerated Regions

Human Accelerated Regions (HARs) are defined as regions of the genome conserved throughout vertebrate evolution, but which have changed radically since humans and chimpanzees split from their common ancestor. Whether the acceleration is functionally relevant and driven by positive selection or a byproduct of biased gene conversion is a matter of intense debate (S*46-50*). In order to determine whether the acceleration took place before or after the human-Neandertal split, we examined a total set of 2,613 Human Accelerated Regions (HARs) identified in five different studies (S*51-55*).

We are restricted in our coverage of the HAR regions due to the filtering of the whole-genome alignments (see above) and by the Neandertal coverage within each HAR region. We identified a total of 8,949 single nucleotide changes and 213 InDels on the human lineage in these HARs. Reliable Neandertal sequence was available for 3,259 (3,226 substitutions + 33 InDels). If we calculate the percentage of positions showing the derived state (2,977; 91.35% [90.32%, 92.28%] - 95% Wilson two-sided confidence interval for a proportion including continuity correction), we observe that this is considerably higher than for the complete set (87.86% [87.82%, 87.90%]) of all derived substitutions (2,813,802) and all derived deletions (60,248). When we count the percentage of positions in which Neandertal shows the derived state only at the positions which may be the sites of biased gene conversion (A/T in chimp to G/C in human, 62% of the positions under consideration), we find that this effect is even more extreme (derived 96.84% [97.49%, 96.03%]) whereas the incidence of changes from G/C in human to A/T in Neandertal is 77.6% derived [80.54%, 74.50%]. The fact that the vast majority of A/T to G/C changes are shared between Neandertal and human suggests that positions affected by biased gene conversion probably predate the human/Neandertal split considerably.

However, there are 51 positions (in 45 HARs) where Neandertal is ancestral and humans are derived and not known to vary (SOM File 11.3). These are likely to represent very recent changes that have occurred since the human-Neandertal split.

Highly-cited genes

A number of genes and genomic features are proposed to have been important in the evolution of human-specific traits. These have been identified through various methods including functional analyses (*S24, 31, 56, 57*), genome-wide comparisons

with other primates (S*51-55*) and variation in present-day human populations (S*58*). Genes frequently identified include those for brain size, speech and language, olfaction, pigmentation, skeletal development and metabolism.

We have previously shown that it is possible to target specific loci in order to determine their state in Neandertals (*S24, 57, 59*). However, probing for individual variants is sample and time consuming. Whole genome sequencing data is an efficient means to address this challenge genome-wide.

The literature contains many such lists of genes and regions. We compiled a set of 253 frequently cited CCDS genes and examined their state in the Neandertal sequence data. None of these genes is included in the 78 fixed amino acid substitutions described above. We note that there may be some bias towards these genes having more SNPs in dbSNP than randomly selected genes, due to the intense interest in these as important and interesting genes for modern human evolution. We may therefore be less likely to identify those genes with fixed, derived changes on the human lineage.

When screening the total list of 1,254 segregating non-synonymous changes (which have been excluded for the set of 78 fixed non-synonymous changes) for which Neandertal shows the ancestral state, we identify 69 genes which have a radical non-synonymous change based on their Grantham score (Table S32). Eleven of the 253 "interesting genes" are found in this list. Among these are positions in two genes, ASPM and MCPH1, involved in microcephaly. In ASPM, we observe that the Neandertal carries the ancestral alleles at both rs61249253 and rs62624968 SNP positions. This is also the case for the position identified in MCPH1. The human-derived D-haplogroup in MCPH1 has been hypothesized (S*56*) to have entered modern humans by gene flow from Neandertals based on it's present-day pattern of variability. Our whole genome data suggests that introgression from Neandertals seems unlikely as we do not see the derived allele in the Vindija Neandertals at the diagnostic SNP position (rs930557). However, only a high coverage Neandertal genome or multiple individual data can resolve whether these sites might also have been polymorphic in Neandertals.

Conclusion

Further work is needed to elucidate the physiological consequences of each of these changes. We note that once the Neandertal genome is sequenced to higher coverage,

we expect to approximately triple the number of amino acid sequence changes that rose to fixation in the modern human lineage after divergence from Neandertals. All such changes may be of sufficient interest to investigate functionally in the future.

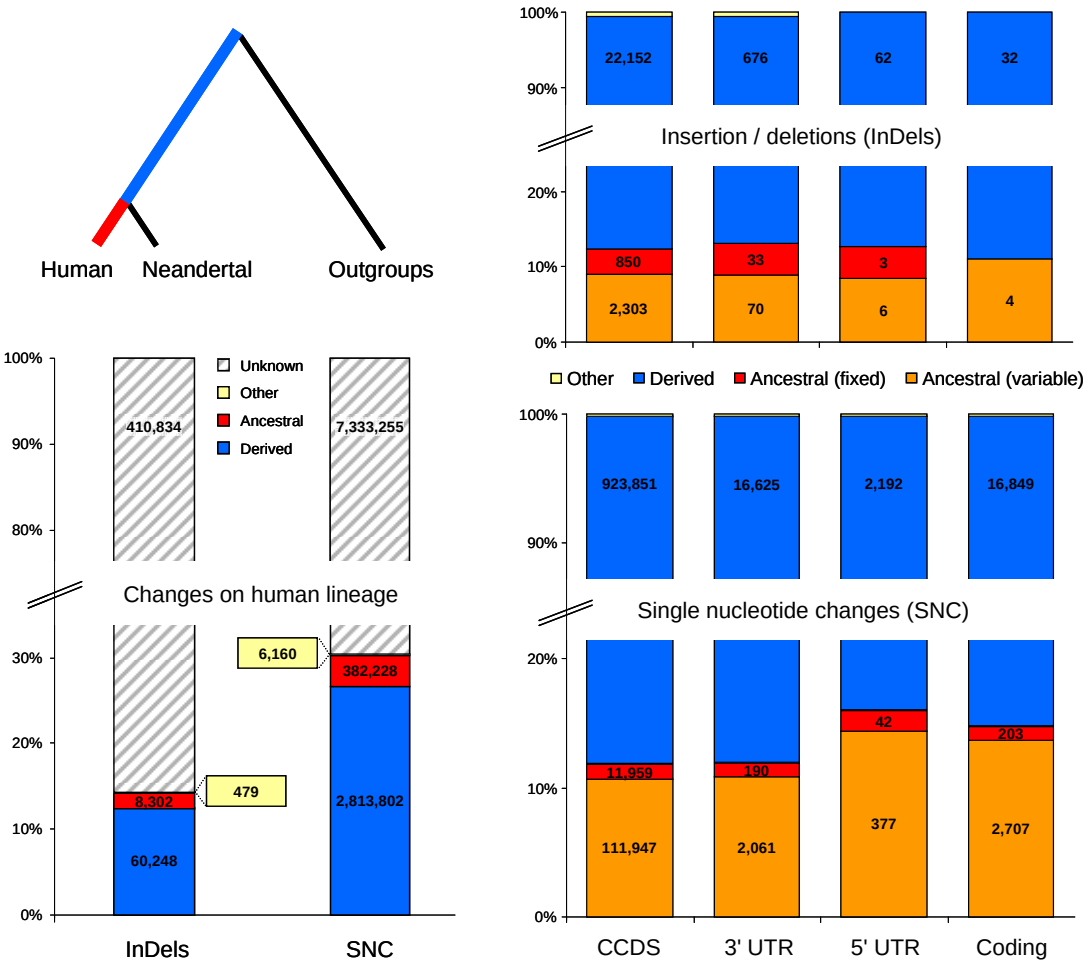


Figure S15: Single nucleotide changes (SNC) and insertion/deletion changes (InDels) on the human lineage inferred from whole genome alignments and their state obtained from the minicontigs created from all Vi33.16, Vi33.25 and Vi33.26 reads.

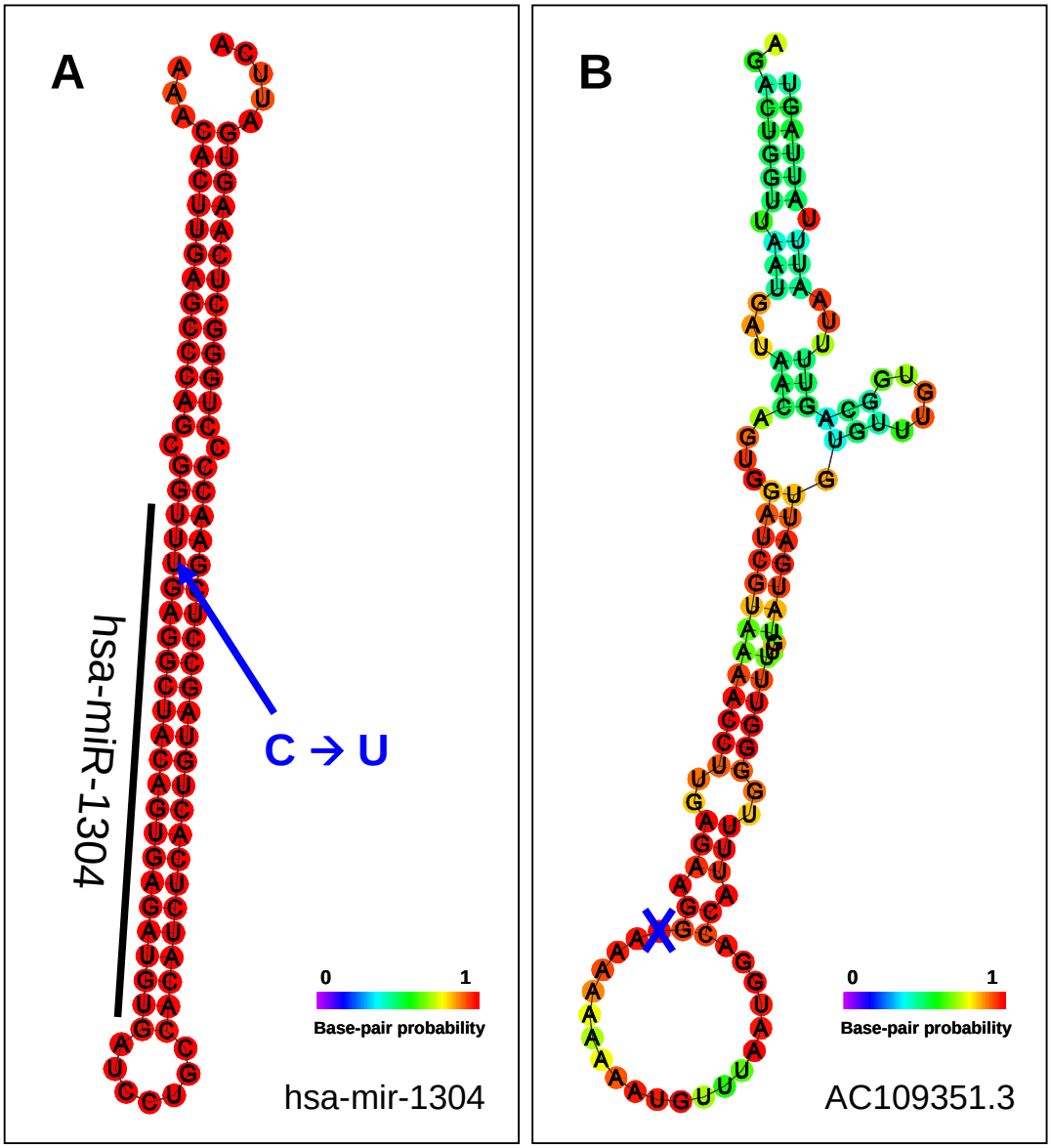


Figure S16: RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) output for the two microRNAs showing the ancestral state in Neandertal while being fixed derived in modern humans with respect to dbSNP. For hsa-mir-1304 (A) the ancestral cytosine is observed in Neandertal while modern humans are fixed derived for thymine (uracil in the microRNA transcript). The change is located in the seed of the mature microRNA, thus it likely to alter target specificity of derived version present in modern humans. In AC109351.3 (B), Neandertal shows an ancestral adenosine as an additional base in the big bulge. This change is not likely to effect folding and function of this putative microRNA.

Table S28: Table of 78 single nucleotide differences in coding sequences of CCDS genes for which Neandertal shows the ancestral state while modern humans are fixed for the derived state. The table is sorted by Grantham scores (GS). Based on the classification proposed by Li (*S40*), 5 amino acid substitutions are considered radical (> 150), 7 moderately radical (101-150), 33 moderately conservative (51-100), 32 conservative (1-50) and one change in stop-codon falls out of this scoring scheme. Genes showing multiple substitution changes are marked with colored database identifiers.

Human (derived)				Chimpanzee (ancestral)				Nea.	Database identifier			Amino acid information			
Base	Chr	+/-	Pos	Base	Chr	+/-	Pos	Base	EnsemblTranscript	CCDS ID	SwissProt	Pos	Base	AA	GS
A	chr1	+	150393846	G	chr1	+	131234541	G	ENST00000316073	CCDS41397	RPTN	785	1	*/R	-
C	chr2	+	11675942	T	chr2a	+	11846259	T	ENST00000381486	CCDS42655	GREB1	1164	1	R/C	180
C	chr9	+	124603021	T	chr9	+	122451445	T	ENST00000277309	CCDS35132	OR1K1	267	1	R/C	180
A	chr1	+	118435820	C	chr1	-	119478811	C	ENST00000336338	CCDS899	SPG17	431	1	Y/D	160
T	chr11	+	118550510	G	chr11	+	118054004	G	ENST00000292199	CCDS8416	NLRX1	330	1	Y/D	160
C	chr3	+	95285751	T	chr3	+	97827124	T	ENST00000314622	CCDS2927	NSUN3	78	2	S/F	155
T	chr1	+	180836069	G	chr1	+	162274446	G	ENST00000367558	CCDS1348	RGS16	197	2	D/A	126
C	chr4	+	13206636	G	chr4	+	13452091	G	ENST00000040738	CCDS3411	BOD1L	2684	1	G/R	125
T	chr6	+	121644484	A	chr6	+	123350109	A	ENST00000368464	CCDS43501	CF170	505	1	S/C	112
G	chr7	+	89631971	C	chr7	+	89789078	C	ENST00000297205	CCDS5614	STEA1	336	2	C/S	112
C	chr11	+	6195544	A	chr11	+	6063773	A	ENST00000265978	CCDS7760	F16A2	630	3	R/S	110
G	chr15	+	39585013	T	chr15	+	38515020	T	ENST00000263800	CCDS10077	LTK	569	1	R/S	110
A	chrX	+	18131667	C	chrX	+	18244953	C	ENST00000380033	CCDS14184	BEND2	261	2	V/G	109
C	chr11	+	6177181	T	chr11	+	6045584	T	ENST00000311352	CCDS31407	O52W1	51	2	P/L	98
T	chr16	+	538119	C	chr16_r.	+	5709532	C	ENST00000219611	CCDS10410	CAN15	427	2	L/P	98
A	chr3	+	47444153	G	chr3	+	48489458	G	ENST00000265565	CCDS2755	SCAP	140	2	I/T	89
A	chr9	+	134265413	G	chr9	+	132451336	G	ENST00000334270	CCDS6948	TTF1	474	2	I/T	89
C	chr3	+	99555910	G	chr3	+	102255730	G	ENST00000354924	CCDS33802	OR5K4	175	1	H/D	81
C	chrX	+	17678236	T	chrX	+	17782496	T	ENST00000380041	CCDS35210	SCML1	202	2	T/M	81
A	chr1	+	1110351	C	chr1	+	1106337	C	ENST00000400931	CCDS8	TTL10	394	2	K/T	78
A	chr2	+	99577084	G	chr2a	+	100577578	G	ENST00000356421	CCDS33258	AFF3	516	1	S/P	74
T	chr20	+	45078304	C	chr20	+	44410094	C	ENST00000360649	CCDS42887	EYA2	131	1	S/P	74
G	chr4	+	2919072	C	chr4	+	3060926	C	ENST00000314262	CCDS33945	NOP14	493	2	T/R	71
T	chr11	+	129277503	G	chr11	+	128968019	G	ENST00000360871	CCDS8484	PRDM10	1129	2	N/T	65
T	chr3	+	113671299	G	chr3	+	116666154	G	ENST00000334529	CCDS33819	BTLA	197	2	N/T	65
A	chr11	+	74477736	G	chr11	+	73459069	G	ENST00000305159	CCDS31639	O2AT4	224	2	V/A	64
T	chr16	+	537906	C	chr16_r.	+	5709319	C	ENST00000219611	CCDS10410	CAN15	356	2	V/A	64
T	chr2	+	220087788	C	chr2b	+	225458007	C	ENST00000358078	CCDS33384	ACCN4	160	2	V/A	64
T	chr22	+	39090924	C	chr22	+	39366824	C	ENST00000216194	CCDS14001	PUR8	429	2	V/A	64
G	chr6	+	100475589	A	chr6	+	101531965	A	ENST00000281806	CCDS5044	MCHR2	324	2	A/V	64

T	chr7	+	17341917	C	chr7	+	17496236	C	ENST00000242057	CCDS5366	AHR	381	2	V/A	64
C	chr1	+	46650472	G	chr1	+	47202157	G	ENST00000243167	CCDS535	FAAH1	476	2	A/G	60
T	chr1	+	118360155	C	chr1	-	119555018	C	ENST00000336338	CCDS899	SPG17	1415	1	T/A	58
C	chr15	+	40529604	T	chr15	+	39548722	T	ENST00000263805	CCDS32208	ZF106	697	1	A/T	58
T	chr16	+	65504565	C	chr16	+	66617267	C	ENST00000299752	CCDS10823	CAD16	342	1	T/A	58
T	chr17	+	37020864	C	chr17	-	15934889	C	ENST00000301653	CCDS11401	K1C16	306	1	T/A	58
T	chr2	+	128113346	C	chr2b	+	128489293	C	ENST00000324938	CCDS2147	LIMS2	360	1	T/A	58
A	chr3	+	44737863	G	chr3	+	45716980	G	ENST00000296091	CCDS2719	ZN502	184	1	T/A	58
G	chr4	+	88986215	A	chr4	+	90763412	A	ENST00000361056	CCDS3625	MEPE	391	1	A/T	58
T	chr5	+	132562844	C	chr5	+	134834351	C	ENST00000265342	CCDS34238	FSTL4	791	1	T/A	58
C	chr8	+	51628204	G	chr8	+	48568603	G	ENST00000338349	CCDS6147	SNTG1	241	2	T/S	58
T	chr1	+	150393996	C	chr1	+	131234775	C	ENST00000316073	CCDS41397	RPTN	735	1	K/E	56
T	chr11	+	118278035	C	chr11	+	117781441	C	ENST00000334801	CCDS8403	BCL9L	543	1	S/G	56
T	chr17	+	24983160	C	chr17	-	27682886	C	ENST00000269033	CCDS11253	SSH2	1033	1	S/G	56
T	chr19	+	62017061	C	chr19	+	62668946	C	ENST00000326441	CCDS12948	PEG3	1521	1	S/G	56
T	chr21	+	33782703	G	chr21	+	33236795	G	ENST00000381947	CCDS13626	DJC28	290	1	K/Q	53
T	chr13	+	48179102	G	chr13	+	48595018	G	ENST00000282018	CCDS9412	CLTR2	50	1	F/V	50
A	chr3	+	44831503	G	chr3	+	45821006	G	ENST00000326047	CCDS33744	KIF15	827	2	N/S	46
T	chr1	+	32052458	C	chr1	+	32221622	C	ENST00000360482	CCDS347	SPOC1	355	2	Q/R	43
C	chr9	+	134267344	T	chr9	+	132453251	T	ENST00000334270	CCDS6948	TTF1	229	2	R/Q	43
T	chr9	+	139259702	G	chr9	+	137500097	G	ENST00000344774	CCDS35186	F166A	134	1	T/P	38
C	chr12	+	62873951	G	chr12	-	25278113	G	ENST00000398055	CCDS41803	CL066	426	1	V/L	32
C	chr11	+	6611387	G	chr11	+	6490781	G	ENST00000299441	CCDS7771	PCD16	763	1	E/Q	29
T	chr11	+	2383887	C	chr11	+	2449712	C	ENST00000155858	CCDS31340	TRPM5	1088	1	I/V	29
T	chr11	+	92535568	C	chr11	+	91690970	C	ENST00000326402	CCDS8291	S36A4	330	2	H/R	29
C	chr14	+	104588537	G	chr14	+	105590623	G	ENST00000392585	CCDS9997	GP132	328	1	E/Q	29
T	chr14	+	67344044	C	chr14	+	67361749	C	ENST00000347230	CCDS9788	ZFY26	237	2	H/R	29
A	chr7	+	134293531	G	chr7	+	135457097	G	ENST00000361675	CCDS5835	CALD1	671	1	I/V	29
A	chr8	+	25416950	G	chr8	+	21955896	G	ENST00000330560	CCDS6049	CDCA2	606	1	I/V	29
G	chr8	+	145211313	C	chr8	+	144064958	C	ENST00000355091	CCDS43776	GPAA1	275	1	E/Q	29
A	chrX	+	3012475	G	chrX_r.	+	1786421	G	ENST00000381127	CCDS14123	ARSF	200	1	I/V	29
G	chr11	+	59039869	A	chr11	+	57731876	A	ENST00000329328	CCDS31564	OR4D9	303	2	R/K	26
G	chr18	+	2880589	A	chr18	-	13843773	A	ENST00000254528	CCDS11828	EMIL2	155	2	R/K	26
T	chr9	+	124622444	C	chr9	+	122471604	C	ENST00000259467	CCDS6845	PHLP	216	2	K/R	26
G	chrX	+	153196802	A	chrX	+	153627492	A	ENST00000369915	CCDS35448	TKTL1	317	2	R/K	26
C	chr1	+	12012533	G	chr1	+	12228263	G	ENST00000235332	CCDS143	MIIP	280	3	H/Q	24
T	chr1	+	156914834	C	chr1	+	137934885	C	ENST00000368148	CCDS41423	SPTA1	265	1	N/D	23
C	chr11	+	6611345	T	chr11	+	6490739	T	ENST00000299441	CCDS7771	PCD16	777	1	D/N	23
C	chr19	+	3498315	G	chr19	+	3591211	G	ENST00000398558	CCDS42464	CS028	326	3	L/F	22
G	chr3	+	198158892	A	chr3	+	202594008	A	ENST00000238138	CCDS3324	PIGZ	425	1	L/F	22

G	chr1	+	221244597	A	chr1	+	203711670	A	ENST00000284476	CCDS1536	DISP1	1079	1	V/M	21
A	chr14	+	20581121	G	chr14	+	19941316	G	ENST00000298690	CCDS41914	RNAS7	44	1	M/V	21
C	chr21	+	30576509	T	chr21	+	30082864	T	ENST00000340345	CCDS42915	KR241	205	1	V/M	21
A	chr20	+	31275867	G	chr20	+	30217610	G	ENST00000375454	CCDS13216	SPLC3	108	3	I/M	10
A	chr20	+	32801190	C	chr20	+	31822769	C	ENST00000374796	CCDS13241	NCOA6	823	3	I/M	10
G	chr4	+	184423847	T	chr4	+	187919923	T	ENST00000281445	CCDS34109	WWC2	479	3	M/I	10
T	chr10	+	73557900	A	chr10	+	71258470	A	ENST00000394919	CCDS31219	ASCC1	301	3	E/D	0
C	chr2	+	95309419	G	chr2a	+	96196093	G	ENST00000317668	CCDS2012	PROM2	458	3	D/E	0

Table S29: GO analysis (biological process) of genes with non-synonymous coding sequence changes on the human lineage for which Neandertal shows positions with the ancestral state, while modern humans are fixed derived.

	GOBPID	P-value	Odds Ratio	Exp Count	Count	Size	Term
1	GO:0006994	0.004518938	Inf	0.004518938	1	1	positive regulation of sterol regulatory element binding protein target gene transcription involved in sterol depletion response
2	GO:0007501	0.004518938	Inf	0.004518938	1	1	mesodermal cell fate specification
3	GO:0032933	0.004518938	Inf	0.004518938	1	1	SREBP-mediated signaling pathway
4	GO:0045716	0.004518938	Inf	0.004518938	1	1	positive regulation of low-density lipoprotein receptor biosynthetic process
5	GO:0006506	0.007029479	17.54717	0.126530277	2	28	GPI anchor biosynthetic process
6	GO:0006991	0.009017826	224.35185	0.009037877	1	2	response to sterol depletion
7	GO:0045541	0.009017826	224.35185	0.009037877	1	2	negative regulation of cholesterol biosynthetic process
8	GO:0045899	0.009017826	224.35185	0.009037877	1	2	positive regulation of transcriptional preinitiation complex assembly
9	GO:0045939	0.009017826	224.35185	0.009037877	1	2	negative regulation of steroid metabolic process

Table S30: GO analysis (biological process) of genes with 5' UTR changes on the human lineage for which Neandertal shows positions with the ancestral state, while modern humans are fixed derived.

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	0001701	0.001418183	15.195378	0.224139348	3	88	in utero embryonic development
2	0006683	0.002547038	Inf	0.002547038	1	1	galactosylceramide catabolic process
3	0015919	0.002547038	Inf	0.002547038	1	1	peroxisomal membrane transport
4	0033595	0.002547038	Inf	0.002547038	1	1	response to genistein
5	0033600	0.002547038	Inf	0.002547038	1	1	negative regulation of mammary gland epithelial cell proliferation
6	0000910	0.003156008	26.938821	0.084052255	2	33	cytokinesis
7	0051345	0.004029231	10.382488	0.323473831	3	127	positive regulation of hydrolase activity
8	0007200	0.005556965	19.865353	0.112069674	2	44	activation of phospholipase C activity by G-protein coupled receptor protein signaling pathway coupled to IP3 second messenger
9	0006350	0.005961565	2.842535	5.651877414	12	2219	transcription
10	0006936	0.006288117	8.801859	0.379508668	3	149	muscle contraction

11	0010468	0.007375581	2.753573	5.79960562	12	2277	regulation of gene expression
12	0006978	0.007622293	202.3	0.007641114	1	3	DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator
13	0019374	0.007622293	202.3	0.007641114	1	3	galactolipid metabolic process
14	0048478	0.007622293	202.3	0.007641114	1	3	replication fork protection
15	0035264	0.007691322	16.675862	0.132445978	2	52	multicellular organism growth
16	0010863	0.008274197	16.03183	0.137540054	2	54	positive regulation of phospholipase C activity
17	0010517	0.008876384	15.435504	0.14263413	2	56	regulation of phospholipase activity

Table S31: GO analysis (biological process) of genes with 3' UTR changes on the human lineage for which Neandertal shows positions with the ancestral state, while modern humans are fixed derived.

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	0048813	0.003042538	32.475676	0.08627064	2	7	dendrite morphogenesis
2	0007006	0.003071339	11.661808	0.29578506	3	24	mitochondrial membrane organization
3	0008344	0.005837668	9.06576	0.36973133	3	30	adult locomotory behavior
4	0008286	0.007636895	8.157143	0.40670446	3	33	insulin receptor signaling pathway
5	0043010	0.007944817	5.461644	0.78876017	4	64	camera-type eye development

Table S32: Table of 11 non-synonymous and non-fixed substitution changes in coding sequences of genes hypothesized to be important in human evolution for which Neandertal (Nt) shows the ancestral state, i.e., matches the chimpanzee, while the human reference genome shows the derived state. The table is sorted by Grantham scores (GS).

Human (derived, but not fixed)				Nt	Database identifier			Amino acid information			
	Chr	+/-	Pos		Ensembl	CCDS ID	Swiss Prot	Codon	Pos	AA	G Score
A	chr1	+	195339182	G	00000367409	1389	ASPM	1940	1	C/R	180
T	chr21	+	46660549	C	00000359568	33592	PCNT	2096	2	L/P	98
A	chr8	+	6289825	G	00000344683	43689	MCPH1	391	2	D/G	94
G	chr1	+	195339940	A	00000367409	1389	ASPM	1687	2	T/I	89
C	chr4	+	5806442	A	00000382674	3383	EVC	448	2	T/K	78
A	chr1	+	103126725	G	00000358392	779	COBA1	1546	1	S/P	74
C	chr8	+	6466449	T	00000344683	43689	MCPH1	760	2	A/V	64
G	chr17	+	17637255	C	00000353383	11188	RAI1	89	2	G/A	60
T	chr7	+	42054746	C	00000395925	5465	GLI3	182	1	T/A	58
C	chr4	+	141708517	T	00000262999	3753	UCP1	63	1	A/T	58
A	chr2	+	215584410	T	00000272895	33372	ABCAC	776	1	S/T	58
A	chr21	+	46676180	G	00000359568	33592	PCNT	2791	2	Q/R	43

T	chr17	+	1320267	C	00000359786	42226	MYO1C	825	2	Q/R	43
A	chr21	+	46646015	G	00000359568	33592	PCNT	1638	1	I/V	29
G	chr8	+	31118821	T	00000298139	6082	WRN	1073	3	L/F	22

Supplemental Online Material 12

Neandertal segmental duplications and copy number variations

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

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

METHODS: The whole-genome shotgun sequence detection (WSSD) method identifies regions >10 kb in length with a significant excess of read depth within 5 Kbp overlapping windows (*S60*). WSSD analysis was performed using 79 million "Illuminized" (36 bps) reads from a combination of three Neandertal individuals (sequences were retrieved from the alignments from SOM S.3, thus bacterial contamination was eliminated), following correction methods specific to next-generation sequencing data (such as biases in the GC distribution) as previously described (*S61*). This sequence library showed a good correlation with a training set of BAC clones with known copy number in humans (Figure S17). 20.6 million reads were mapped to a repeatmasked version of the human genome using *mrFAST* (microread fast alignment search tool) within an edit distance of 2, an algorithm that tracks all read map locations allowing read-depth to be accurately correlated with copy number in duplicated regions (*S61*).

We first compared a known human segmental duplication (SD) map (NA18507) (*S62*) (>10 kb) to Neandertal SDs (Table S33, Figure S18 and S19) and we found that 77/87 Mb (88.9%) of the Neandertal duplications were shared (compared to only 57 Mb (65.5 %) that are shared with the published chimpanzee genome (*S63*)). Only 1.3 Mb of the shared duplications between Neandertal and chimpanzee were single copy in human. We next compared annotated duplication maps for four human genomes (JDW (*S64*), NA18507 (*S62*), YH (*S65*)) and the Celera WGS database (*S60, 66*). From this comparison, we detected 111 potential Neandertal-specific sites (average size=22,321 bps and total length 1,862 Kbps) that did not overlap with the previous dataset (Figure S20). In the absence of an external validation of Neandertal duplications, we calculated the single-nucleotide diversity (differences supported by at least two reads) in all these regions (Figure S21). We permuted the distribution of sizes of the regions queried to a set of control regions (regions in which no human or non-human primate SDs have been detected). Potential Neandertal-specific SDs with a nucleotide diversity beyond 3 standard deviations of the control regions were considered for further analysis (N=90; 1,480 Mbps). A

visual inspection of the data revealed that almost all of them correspond to putative human segmental duplications that were below our threshold of detection (see Figure S22 an example of a false positive Neandertal duplication). Only three regions remained as potential Neandertal-specific duplications (with nucleotide diversity higher than control regions, highlighted in red in Figure S21, see Figure S23), (*hg17: chr20:59041000-59055370, chr4:7273000-7286673, chr6:100123000-100138118*), but none of them overlap with a known gene.

Genome Resampling Comparison

Two limitations of the genome comparisons are the low sequence coverage (Neandertal 1X) and the heterogeneity of the sample (i.e. Neandertal sequence data are derived from a pool of three individuals). To correct for this, we resampled sequence from three human genomes generated using the Illumina sequencing platform. We selected three individuals from different geographic origins and randomly selected sequence to match the coverage and diversity of the Neandertal dataset (1.3 X coverage obtained equally from HGDP samples French (HDGP00521), Han (HGDP00778), and Yoruba (HGDP00927)) (Table S34 and Figure S24). The duplication map of the human resampling was found to be slightly smaller than the Neandertal duplication map (88,869 Kbps; N=1,129 versus 94,419 Kbps; N=1,194) (Figure S25) but copy number estimations were comparable to previously published genomes (Figure S26). We repeated the experiment using a second resampled human genome dataset at low coverage to eliminate potential bias in the resampling process. The second subset gave similar results. Both resampled sets showed excellent correlation (Figure S27) and none of them showed any bias in terms of GC content (SOM S3) in any category that would explain our results (Figure S28).

Gene Analysis

We estimated the copy number of each unique autosomal RefSeq gene (N=17,601) in the Neandertal genome and compared it to the copy number estimate from the resampled human genome. We saw an excellent correlation in copy number (R^2=0.91) between human and Neandertal (Figure S29). We searched for genes with potential copy number differences between humans and Neandertal (Table S35, S36). Interestingly, 29/43 (67%) of the most differentiated genes (more than five copies) had higher copies in Neandertal than in humans. Most of these corresponded to genes mapping to core duplicons (*S67*)—genes of considerable variation in copy number among extant human populations (*S61*). The global analysis of genes with more than two copy number differences (N=295) did not show any bias in the distribution having 52%

(153/295) higher copy in humans and 48% (142/295) in Neandertal but increased variance is expected in the copy number estimation because of the 1X coverage of the Neandertal genome and in the human resampling.

17q21.31 Haplotype Analysis: We find no evidence of the 17q21.31 H2 haplotype or the associated segmental duplication prevalent in Europe today, suggesting that the Neandertal carries the more common H1 haplotype (*S58, 68*) (Figure S30). We remapped the Neandertal reads to human reference genome using mrFAST and tracked the underlying sequence variation in the mapping data. We then compared the mapping information with 381 SNPs that are informative of the H2 haplotype in the 17q21.31 region (*S58*). We observed that the sequence reads mapping to the 17q21.31 region supported only 2 possible SNPs (rs4074462, and rs1467969) diagnostic of the H2 allele, whereas 130 SNPs were consistent with the H1 haplotype. These findings do not support the proposed introduction of the H2 haplotype into human populations from Neandertals (*S58, 69*).

FIGURES

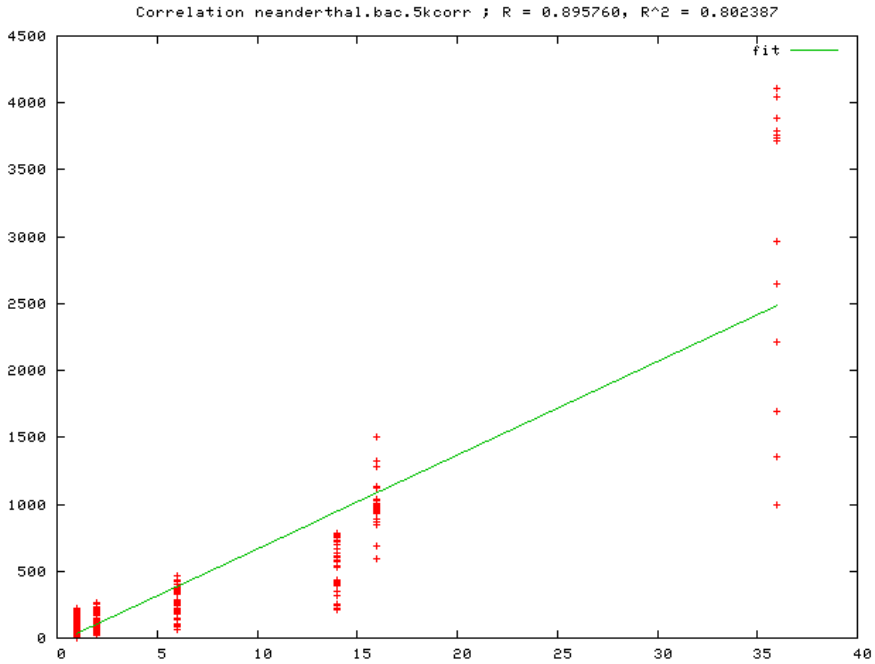


Figure. S17: Correlation of Neandertal read-depth with BAC sequences of known human copy number (R^2=0.80).

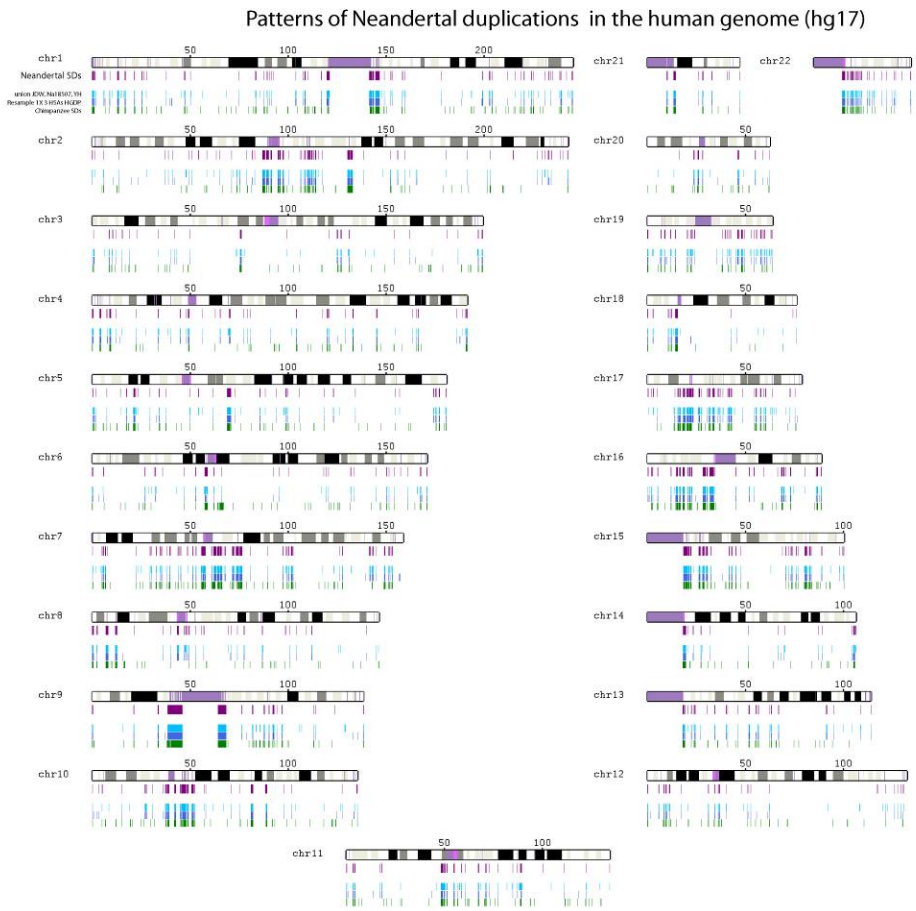


Figure S18: Genomic distribution of Neandertal SDs (>10 kb) in comparison to other humans and chimpanzees.

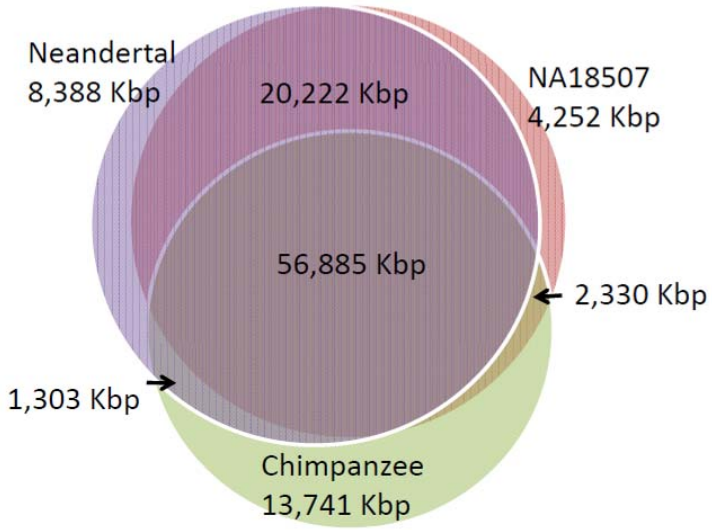


Figure S19: Comparison of Neandertal (3 individuals, 1X coverage) with the known duplication of a human (NA18507) and chimpanzee (Clint).

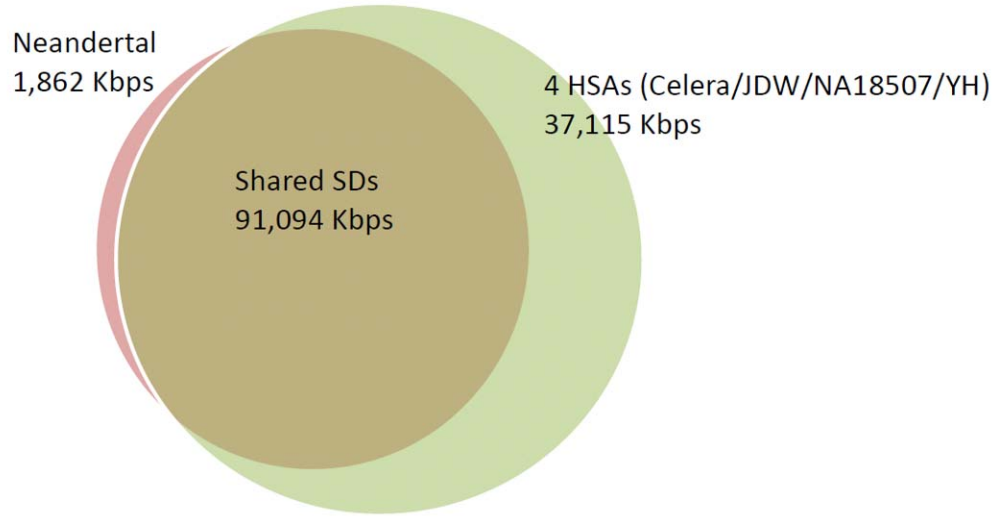


Figure S20: Venn diagram with the comparison of Neandertal (>10 kb) with a union of four humans (Celera WGS, JDW, YH and NA18507 genomes).

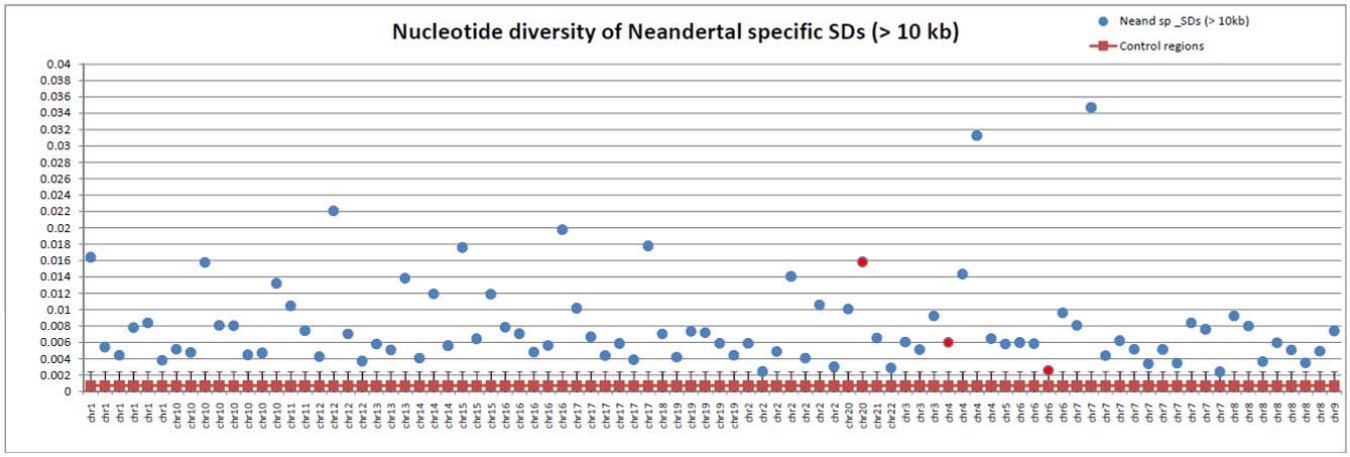


Figure S21: Excess Single Nucleotide Diversity for Predicted Neandertal-Specific Duplications. We identified regions showing excess read-depth and computed single-nucleotide diversity for each Neandertal-specific duplication comparing them to control regions (regions in which no known duplication in any human or primate has been previously detected). Error bars in the control regions correspond to the 3 standard deviations from the permuted sample set. Notice the general trend of higher nucleotide diversity consistent with true duplicated sequences. While not classified as human-specific duplications, we note that many of these regions also show evidence of increased read-depth in one or more human genomes. We have highlighted in red the three unique sites that after visual inspection where we have found no evidence of duplication in human. This small subset likely corresponds to true Neandertal-specific SDs.

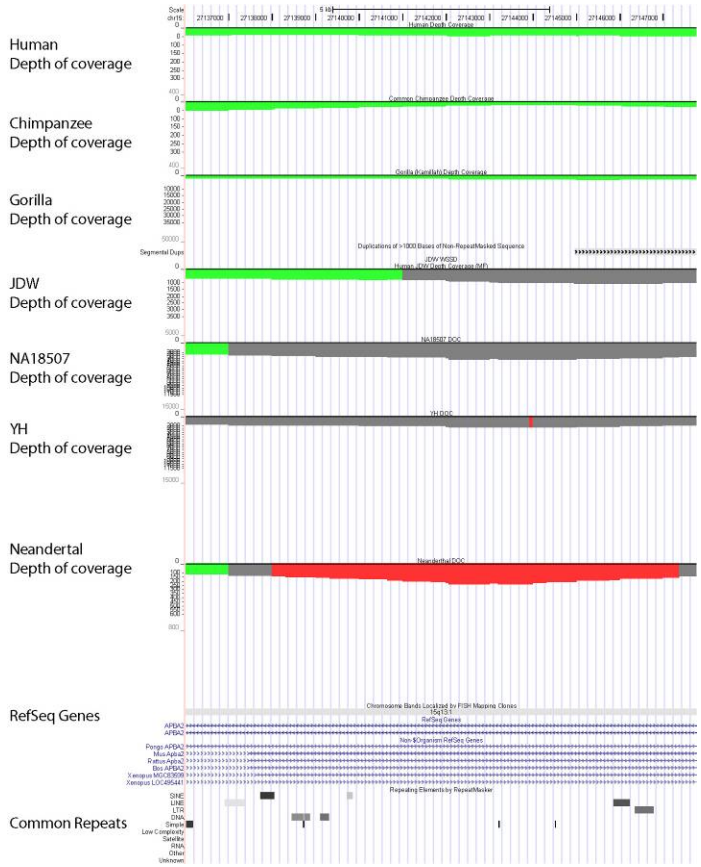
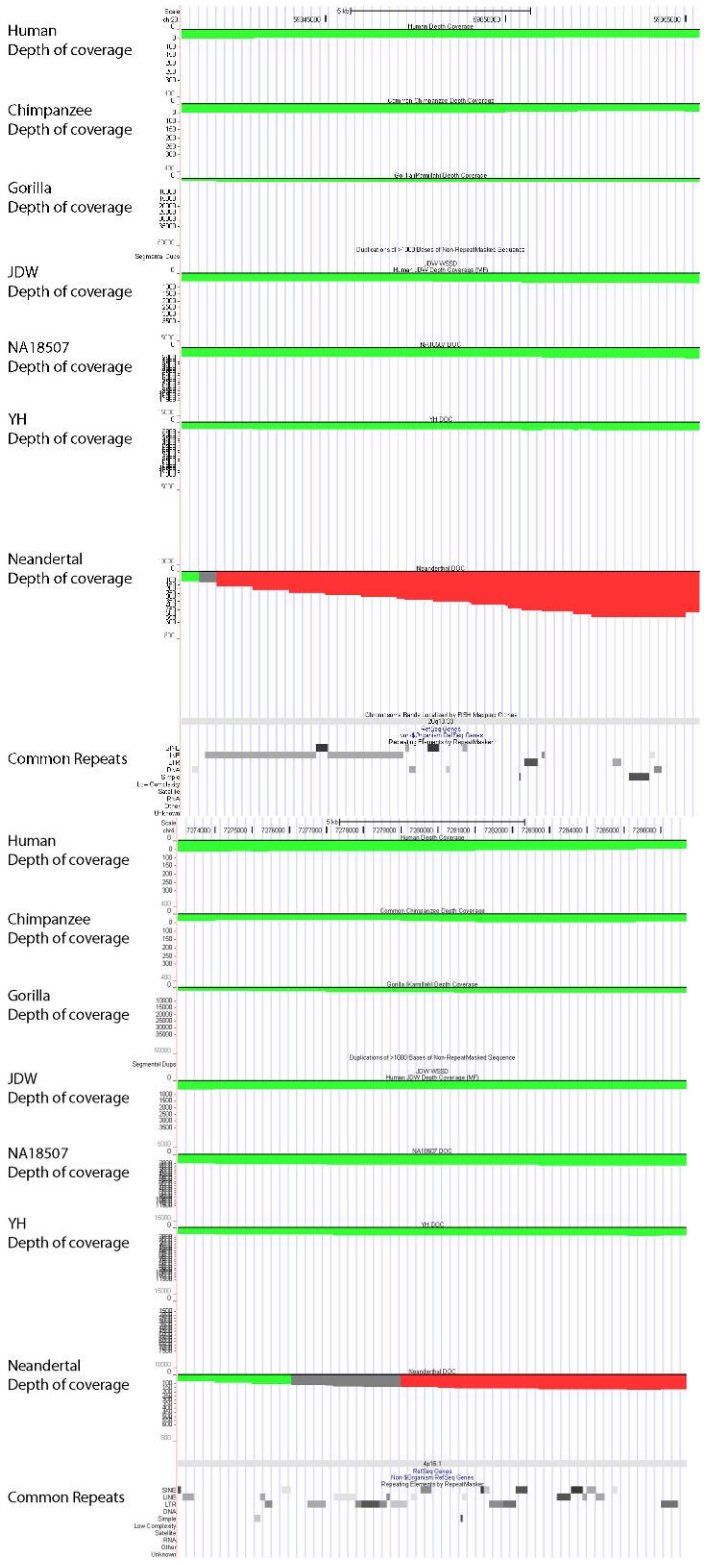


Figure S22: Visual inspection of putative neandertal-specific SD identifies a potentially false negative in other humans. This site was predicted as a Neandertal-specific SD (red denotes an excess read depth > average+3std, criterion to select SDs), but grey regions (intervals with > average+2std read depth) reveals that they might be potential real duplications in other humans. We reason that in this case, the three human genomes (JDW, NA18507 and YH) likely carry a corresponding small duplication that was not called because of size and read-depth thresholds applied in previous human analyses therefore this region is not likely a Neandertal-specific SD. "Human Depth of coverage" track was generated from Celera WGS data using Sanger-based sequencing.


Human
Depth of coverage
Chimpanzee
Depth of coverage
Gorilla
Depth of coverage
JDW
Depth of coverage
NA18507
Depth of coverage
YH
Depth of coverage
Neandertal
Depth of coverage
Common Repeats
Human
Depth of coverage
Chimpanzee
Depth of coverage
Gorilla
Depth of coverage
JDW
Depth of coverage
NA18507
Depth of coverage
YH
Depth of coverage
Neandertal
Depth of coverage
Common Repeats

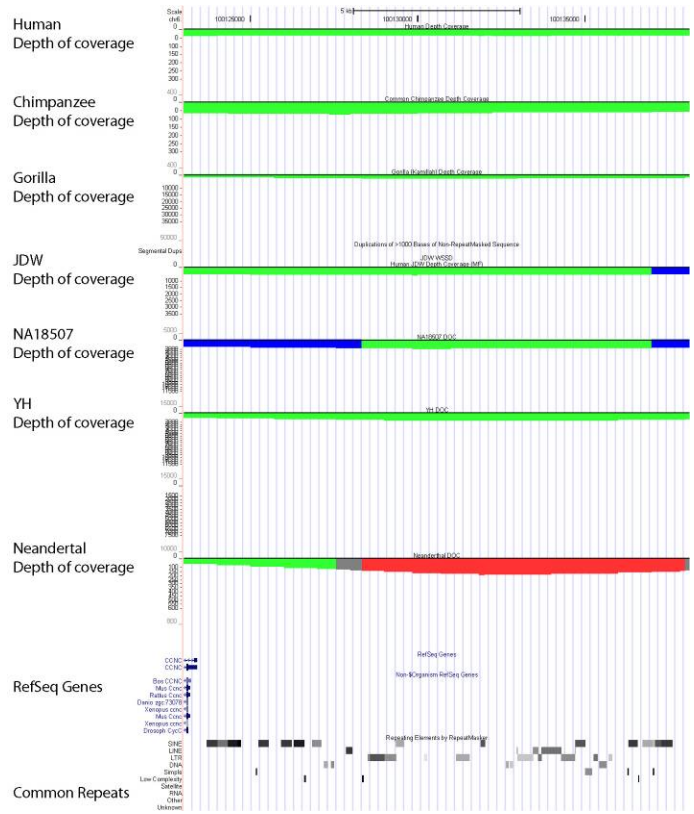


Figure S23: Neandertal-specific SDs. Screenshots of the three loci (total 40 Kbps) containing Neandertal-specific SDs. Red denotes regions with excess of depth-of-coverage (in this case only Neandertal). Notice that none of the humans or primates show any evidence of duplications based on read-depth.

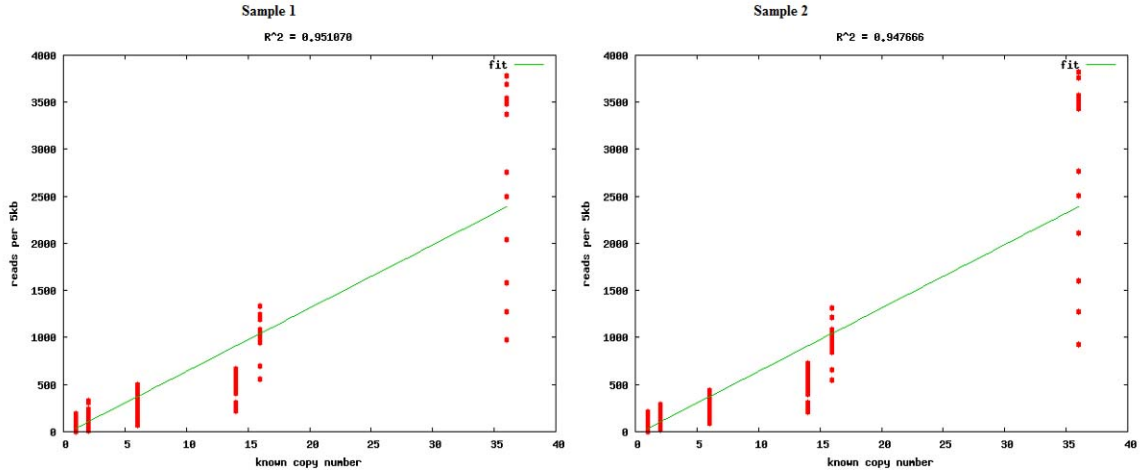


Figure S24: Correlation read depth calculated from the resampling of three humans and comparing predicted copy based on read-depth to experimentally determined copy number for known segmental duplications (Rsq=0.81). Correlation of copy numbers predicted from the resampling of three humans with previously detected copy numbers in JDW (Rsq=0.94).

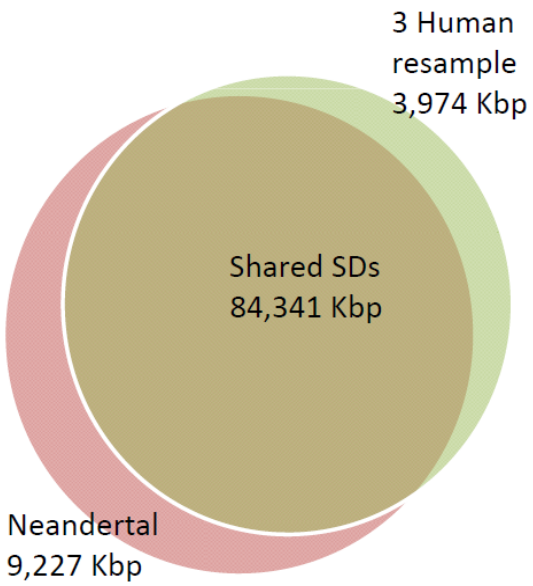


Figure S25: Comparison of Neandertal with a resampled set of three human genomes (1X coverage) (> 10 Kbp).

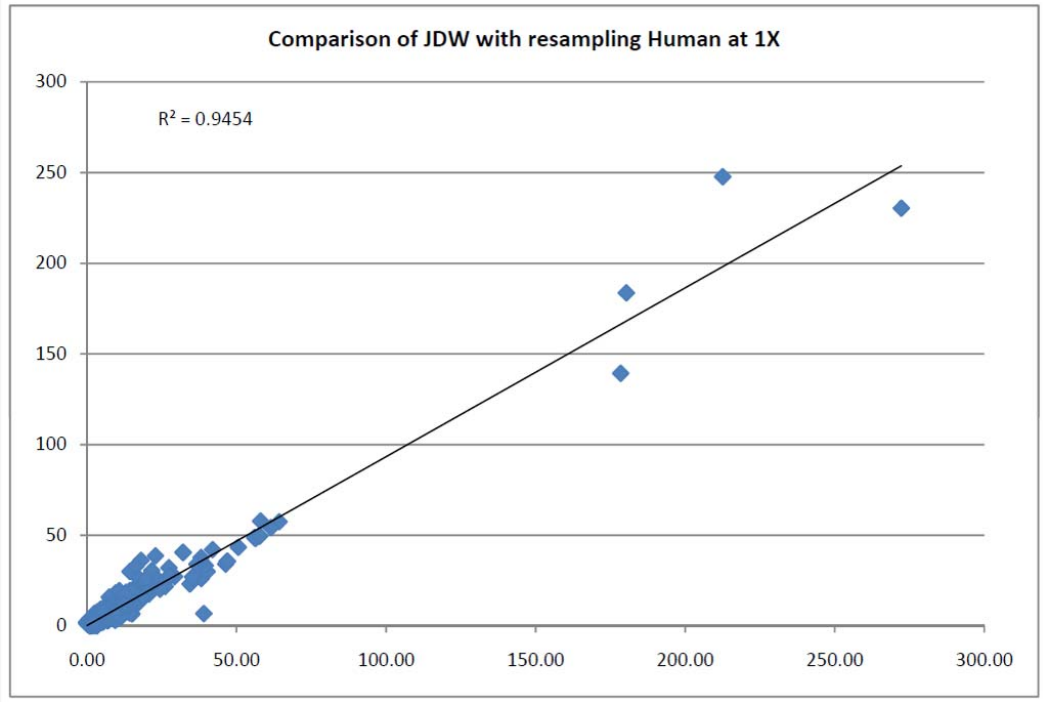


Figure S26: Correlation of copy numbers predicted from the resampling of three humans with previously detected copy numbers in JDW (Rsq=0.94).

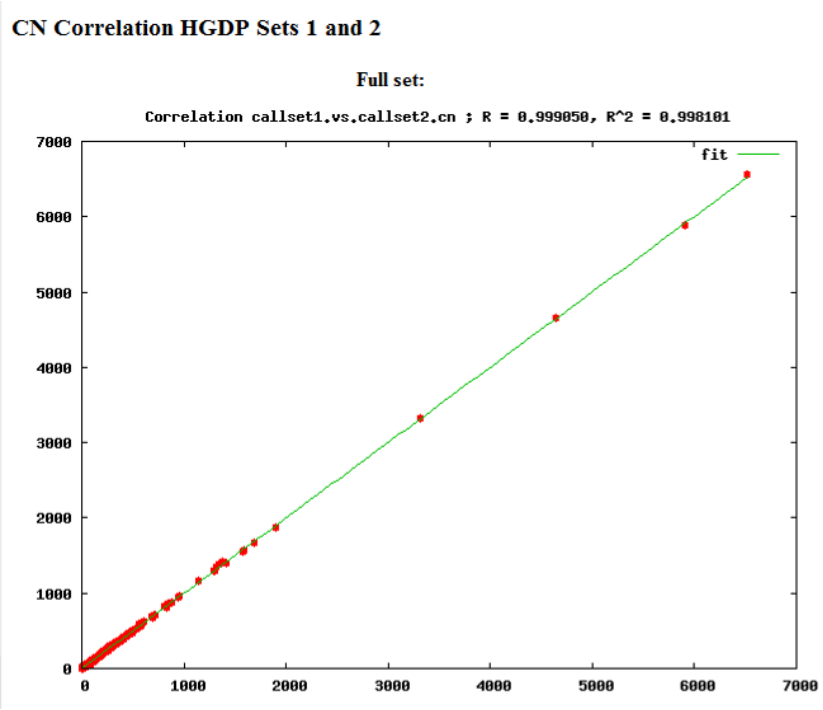


Figure S27: Correlation of copy number (and depth of coverage) between two resampled human genome datasets (independent samples of 1X sequence coverage from a pool of human genomes: (HGDP French (HDGP00521), Han (HGDP00778), and Yoruba (HGDP00927)).

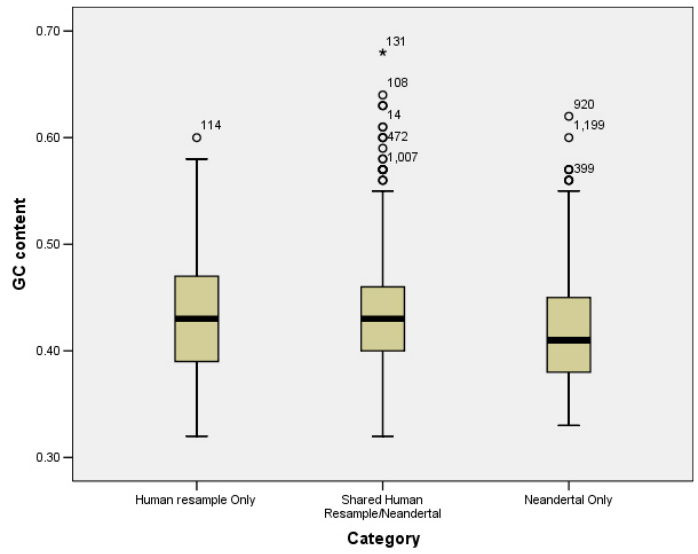


Figure S28: GC content of the predicted duplications stratified by category. Data shows no bias in the GC composition because of the statistical corrections applied before duplication detection.

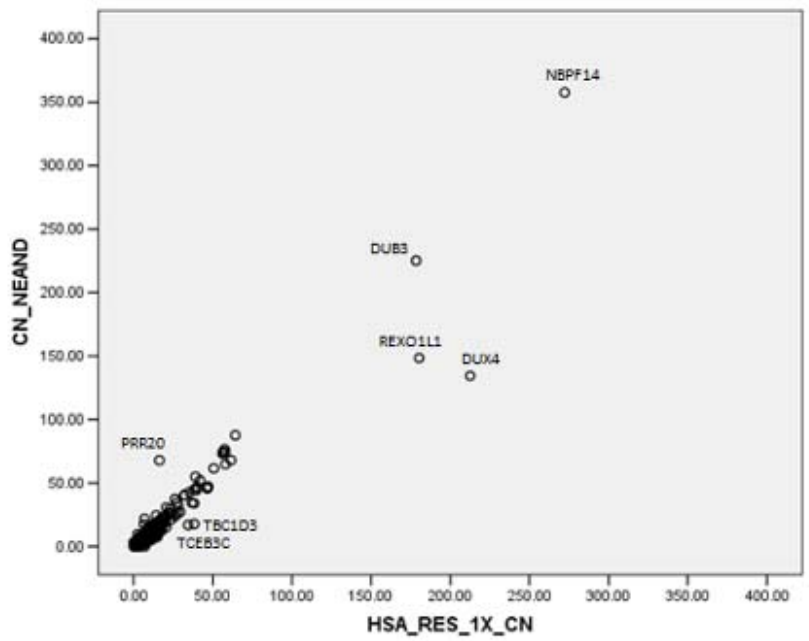


Figure S29: Correlation of predicted copy numbers between three humans at low coverage (1X) and Neandertal genomes (R2=0.91). Genes showing the greatest differences in copy-number are labeled. .

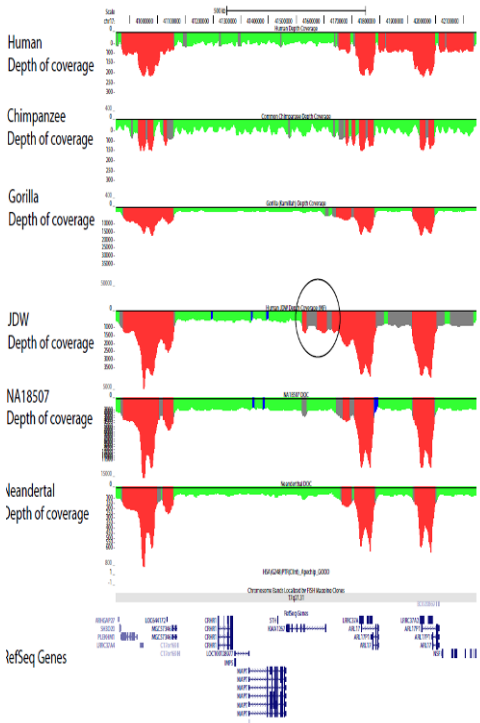


Figure S30: MAPT locus. Notice the extension of the duplication (black circle) denotes an H2 haplotype specific duplication.

TABLES

Table S33: Summary statistics of Neandertal Duplications.

Chr	Length	#N
chr1	7,687,906	115
chr2	8,841,275	94
chr3	1,837,167	36
chr4	3,502,073	56
chr5	4,169,350	45
chr6	1,990,803	46
chr7	9,399,493	128
chr8	2,966,042	46
chr9	10,719,274	69
chr10	5,987,718	65
chr11	3,059,084	43
chr12	1,552,231	37
chr13	1,724,565	40
chr14	1,970,750	26
chr15	6,767,207	66
chr16	7,418,118	60
chr17	4,958,599	90
chr18	1,519,719	11
chr19	2,469,874	58
chr20	1,100,364	13
chr21	1,722,537	14
chr22	3,055,087	36
Grand Total	**94,419,236**	**1194**

Table S34: Summary statistics of the resampling of three humans at 1X coverage.

	# reads (36 bp each)	Mapped to repeatmasked hg17
French	23,000,000	6,980,173
Han	23,000,000	7,224,219
Yoruba	23,000,000	7,370,157
Sum	**69,000,000**	**21,574,549**

Table S35: Genes with high copy number difference between Neandertal and human (resampled at 1X). Only genes with absolute copy number >5 reported.

GeneName	RefSeq ID	CN Neand	CN HSA*	Δ Copy number (HSA-NEAND)	MAX(Δ Copy number 4 HSAs)**	MIN(Δ Copy number 4 HSAs)**	Δ Copy number Chimpanzee
NBPF14	NM_015383	357.58	272.24	-85.35	55.09	3.65	304.19
DUX4	NM_033178	134.44	212.53	78.10	151.14	28.38	44.45
PRR20	NM_198441	67.85	16.27	-51.58	17.15	1.15	9.35
DUB3	NM_201402	225.16	178.41	-46.76	64.29	17.57	252.90
REXO1L1	NM_172239	148.45	180.34	31.90	53.74	4.98	12.42
NBPF16	NM_001102663	87.63	64.20	-23.42	6.51	0.39	22.73
TBC1D3	NM_032258	17.70	38.21	20.51	13.19	1.74	5.86
LOC200030	NM_183372	76.54	57.50	-19.04	5.35	0.23	17.10
NBPF11	NM_001101663	74.59	56.40	-18.19	6.30	0.24	15.91
TCEB3C	NM_145653	16.91	34.41	17.51	15.79	1.27	42.11
NBPF10	NM_001039703	73.17	56.05	-17.12	3.04	0.23	0.39
NBPF20	NM_001037675	74.04	57.74	-16.29	6.32	1.50	16.06
FAM90A7	NM_001136572	54.93	39.07	-15.87	44.20	6.64	8.25
HSPA1B	NM_005346	21.95	6.86	-15.08	2.03	0.64	15.67
LOC339047	NM_178541	37.48	26.14	-11.33	5.72	0.41	5.02
NBPF1	NM_017940	61.51	50.57	-10.94	4.61	0.27	2.44
HSPA1A	NM_005345	17.40	6.47	-10.94	2.29	0.56	14.00
FRG1	NM_004477	24.53	14.33	-10.21	7.14	0.96	2.19
NBPF7	NM_001047980	30.65	20.50	-10.15	6.30	0.72	5.55
PPIAL4A	NM_178230	51.99	41.96	-10.03	8.10	0.20	19.72
FOXD4L5	NM_001126334	40.46	32.10	-8.35	5.55	0.33	15.07
FOXD4L4	NM_199244	40.46	32.10	-8.35	5.55	0.33	15.07
LOC100132247	NM_001135865	42.68	35.26	-7.43	3.55	0.16	5.85
FOXD4	NM_207305	7.60	15.00	7.41	5.86	0.17	8.29
LOC729617	NM_001101668	34.72	27.38	-7.33	9.23	0.06	11.36
CLPS	NM_001832	9.67	2.45	-7.23	0.42	0.10	7.25
HIST1H2BB	NM_021062	0.00	6.96	6.96	3.19	0.20	0.92
LOC100132832	NM_001129851	29.79	22.98	-6.81	2.79	0.12	7.62
LOC650293	NM_001040071	64.79	58.00	-6.79	7.27	0.32	20.71
NPIP	NM_006985	46.90	40.15	-6.75	3.59	0.75	6.94

MGC70863	NM_203302	44.72	38.16	-6.57	6.66	0.78	30.98
NBPF15	NM_173638	67.91	61.46	-6.45	3.47	0.48	14.41
POLR2J	NM_006234	7.28	13.57	6.28	2.56	0.02	0.43
RASA4	NM_001079877	5.34	11.58	6.24	1.70	0.15	0.68
LOC441956	NM_001013729	8.10	14.11	6.02	5.94	0.60	2.22
KIR2DS4	NM_012314	14.06	8.29	-5.77	2.48	0.29	10.18
PCDHB2	NM_018936	6.49	12.13	5.64	8.88	0.42	7.09
PCDHB15	NM_018935	3.55	9.05	5.50	8.62	1.10	9.63
HIST1H2BN	NM_003520	2.65	7.97	5.33	4.28	0.52	6.89
TUBB8	NM_177987	15.33	20.61	5.29	4.89	0.99	4.89
GOLGA6	NM_001038640	12.64	17.80	5.16	9.57	0.48	6.31
PSG9	NM_002784	17.88	12.80	-5.08	5.06	0.22	6.44
FLG	NM_002016	16.19	11.15	-5.04	5.80	0.71	3.73

* Copy number estimated from resampling at low coverage 3 human from HGDP
** Copy number estimated from 4 humans (JCV, JDW, NA18507 and YH) sequenced at high coverage

Table S36: Genes (>10 kb) with the greatest copy number difference between Neandertal and humans (resampled at 1X). Only genes with absolute copy number >5 reported. +only one representative of the gene family reported.

GeneName	**RefSeq ID**	**CN Neand**	**CN HSA***	**Δ Copy number (HSA-NEAND)**	**MAX(Δ Copy number 4 HSAs)****	**MIN(Δ Copy number 4 HSAs)****	**Δ Copy number Chimpanzee**
NBPF16+	NM_001102663	87.63	64.20	-23.42	6.51	0.39	22.73
TBC1D3	NM_032258	17.70	38.21	20.51	13.19	1.74	5.86
LOC200030	NM_183372	76.54	57.50	-19.04	5.35	0.23	17.10
FAM90A7	NM_001136572	54.93	39.07	-15.87	44.20	6.64	8.25
LOC339047	NM_178541	37.48	26.14	-11.33	5.72	0.41	5.02
FRG1	NM_004477	24.53	14.33	-10.21	7.14	0.96	2.19
LOC100132247	NM_001135865	42.68	35.26	-7.43	3.55	0.16	5.85
LOC100132832	NM_001129851	29.79	22.98	-6.81	2.79	0.12	7.62
NPIP	NM_006985	46.90	40.15	-6.75	3.59	0.75	6.94
MGC70863	NM_203302	44.72	38.16	-6.57	6.66	0.78	30.98
RASA4	NM_001079877	5.34	11.58	6.24	1.70	0.15	0.68
KIR2DS4	NM_012314	14.06	8.29	-5.77	2.48	0.29	10.18
GOLGA6	NM_001038640	12.64	17.80	5.16	9.57	0.48	6.31
PSG9	NM_002784	17.88	12.80	-5.08	5.06	0.22	6.44
FLG	NM_002016	16.19	11.15	-5.04	5.80	0.71	3.73

* Copy number estimated from resampling at low coverage 3 human from HGDP
** Copy-number estimated from 4 humans (JCV, JDW, NA18507 and YH) sequenced at high

Supplemental Online Material 13

Selective sweep screen

Richard E. Green* and Michael Lachmann

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

A selective sweep occurs when a genetic variant is increased in frequency due to a fitness advantage relative to other genetic variants at that locus. As a selective sweep occurs, a region of the genome linked to the selected variant, the haplotype on which the selected variant originated, also comes to high frequency or fixation. Present-day human populations still harbor a large fraction of the polymorphisms that were present when the ancestral population of modern humans and Neandertals split. However, in regions of the genome that fixed because of a selective sweep in modern human ancestors that occurred since the modern human/Neandertal population split all such shared polymorphism will have been lost. That is, in these regions the variation present within modern humans will not be shared with Neandertals. We would therefore like to search for regions of the genome where the allelic variation among modern humans is not shared with Neandertals.

We searched for regions of the reference human genome that are especially reduced in Neandertal derived sites at current human polymorphic positions. These regions could arise under at least two plausible evolutionary scenarios: (1) an instance of positive selection in modern human ancestors or (2) continuous or recent purifying or positive selection such that few derived alleles are seen even in current humans. As we are interested in the first case, and especially interested in instances of positive selection that occurred early in modern human evolution, we conditioned our Neandertal expectation for derived alleles on the frequency of human derived alleles. In this way, we enrich for selective sweeps in which the allele frequency spectrum in modern humans has largely recovered since the sweep by accumulation and drift of new mutations. By simulating both neutral sequence evolution and positive selection we demonstrate the power of this method under various selection coefficients and across a range of local recombination rates.

Description of the method

To discover human SNP sites, genome-wide, we randomly sampled the BWA mappings of the HGDP data at each genomic position for the first read passing the map-quality and base quality criteria as described above. This sampling strategy, while discarding potentially useful information about heterozygous sites in these individuals, has the benefit of not biasing towards or against heterozygous sites as happens when attempting to call a genotype in low-coverage data. Chimpanzee aligned sequence was extracted from whole-genome alignment as provided by UCSC (http://hgdownload.cse/ucsc/edu/goldenPath/hg18/vsPanTro2/).

We filtered these data for sites that passed the following criteria:

1. No CpG sites
2. Observed to be bi-allelic among the 5 HGDP and 1 reference human
3. At least one Neandertal observation of one of the two human alleles
4. The chimpanzee base was one of the two human alleles (and used to infer the ancestral allele)

At these sites, we then calculated the fraction of Neandertal alleles that were observed in the derived state as a function of the derived allele frequency in humans (Fig. S31A). As expected, there is a strong correlation between seeing an allele in the high-frequency derived state in humans and finding it in the derived state in Neandertals. For example, at polymorphic positions where 5 of the 6 humans show the derived allele, there is >50% chance of observing the Neandertal base in the derived state. For comparison, we made the same measurement for the Craig Venter genome (*S26*). In this case, the derived allele frequency in the other humans was nearly perfectly predictive of the probability of seeing the Venter base in the derived state (Fig. S31B).

To maximize the predictive power of the human allele states, we further characterized the probability of observing a Neandertal allele under each configuration, *g*, of ancestral and derived alleles in the human data. The observed rate of Neandertal derived alleles at each *g* is shown in panel C of Fig. S31. The human reference allele is

the most predictive of the Neandertal state, indicating increased mapping sensitivity when a Neandertal read matches the human reference sequence. Using the exact configuration of human derived alleles to train the Neandertal expectation circumvents the need to develop a demographic model of the five humans and the reference genome for this experiment. Because of the abundance of polymorphism data, good estimates of the fraction of Neandertal derived alleles can be trained for each *g*. For example, the least abundant configuration (derived allele in the reference and Han, ancestral in the French, and missing data for the 3 other human samples) is observed 658 times and we see Neandertal derived at 46.7% of these sites. The most abundant configurations are observed more than 500,000 times. We use these genome-wide averages to calculate an expected number of Neandertal derived alleles within regions under consideration.

We investigated the interaction between local recombination rate and the human derived allele frequency and its predictive power for Neandertal alleles. We used a recently published high-density recombination map (*S70*) to calculate the local recombination rate within 100kb of each human polymorphic position. As shown in Fig. S32, the frequency spectrum is skewed towards more low frequency derived alleles in regions of low recombination. This may be due to tighter linkage to genomic features under purifying selection in regions of low recombination (*S71*). The salient question for our purposes is whether this phenomenon affects the predictive power of Neandertal derived alleles.

To determine if local recombination rate interferes with the ability of human derived alleles to predict Neandertal derived alleles, we performed the same analysis described above, but this time binning by local recombination rate in humans. As shown in Figure S33 the human derived allele frequency spectrum retains its predictive power for Neandertal derived alleles across the range of recombination rates. In summary, recombination rates skew the human allele frequency spectrum making it less likely to see high frequency (and hence older alleles) in regions of low recombination. However, when high frequency derived alleles are seen, regardless of the local rate of recombination, they are predictive of the Neandertal state. That is, they contain equivalent information about the age of the allele.

Our scan for positive selection relies on identifying regions of the human genome with a dearth of Neandertal derived alleles at human polymorphic positions. To do this, we rely on the observed human derived allele configurations at each polymorphic position, g_i. In regions of the genome that are under strong purifying selection or that have undergone a recent selective sweep, the allele frequency spectrum may be skewed towards lower frequency human alleles (*S72*). In such regions, where the human gene trees are shallow, human variation will less often be old enough to include Neandertals. Thus, we would not expect to see Neandertal derived alleles. Generating an expected number of Neandertal derived alleles based on the genome-wide averages for each configuration of observed human derived alleles, naturally disregards regions of recent selection or strong purifying selection. On the other hand, in regions of older selective sweeps, in which the human allele frequency spectrum has recovered, we will have maximum power to detect sweeps.

For each polymorphic position, we calculated the expected number of Neandertal derived alleles $E(N_{D,s,e}) = \sum_{i=s}^{e} p(N_D \mid g_i)$ within a window of 100 kilobases, given the configuration of human derived alleles at each polymorphic position in the window. While a 100 kilobase window size is arbitrary, we note that our aim is to detect strong selective sweeps and the strength of selection will be proportional to the width of the signal. Given that the size of a swept window is approximated $0.01s/r$ base pairs (*S73*), a 100kb window would correspond to strong selection (s = 0.01 for r = 1E-8). Further, given the age of the split and the transit time of beneficial mutations, $-\log(1/2N_e)/s$ in generations, we are by definition looking for signals of very strong selection (*i.e.*, $s >$ 0.001) to have fixed since the split for N_e=10,000. And while 100 kilobases may preclude detection of sweeps whose signals are significantly narrower than this, it does not preclude the detection of regions with wider window-sizes. We then counted the observed number of Neandertal derived alleles $O(N_{D,s,e})$ in each window for comparison.

The difference between $E(N_{D,s,e})$ and $O(N_{D,s,e})$ was normalized by the variance of $N_{D,s,e}$ to arrive at a measure S, that we analyze to find regions of Neandertal derived

allele depletion. We then identified each region of at least 25kb in which all polymorphic sites were at least 2 standard deviations below the expected value. We then recalculated *S* for each identified region by re-computing $E(N_{D,s,e})$ and $O(N_{D,s,e})$ using for *s* and *e* the start and end polymorphic site of the region, respectively.

This first-pass screen identified 4,235 regions of at least 25 kilobases in which the observed number of Neandertal derived alleles is depleted relative to the expectation. To prioritize these regions for further analysis, we sorted them by *S* and took the top 5% of these, a total of 212 discrete regions spanning a total of 53,103,538 base pairs. This corresponds to an *S* cutoff of -4.3202. We also sorted this list of 212 regions by genetic width. *S* is related to the statistical depletion of Neandertal alleles within a region. It can thus be considered a ranking of confidence that there is a signal for a selective sweep at all. Genetic width, calculated using the recombination map of Myers and co-workers (*S70*), should be proportional to the strength of the selective sweep, if one occurred. This is because a stronger sweep will pass through a population in fewer generations, and thus fewer recombinations, than a slower sweep. The top candidate by genetic width prioritization, containing the gene *THADA* is shown in Figure 4C. The top candidate by *S* contains the gene *AUTS2* as shown if Figure S34. The 212 regions and the genes contained within them, sorted by genetic width are listed in Table S37.

One class of genetic changes that may cause a selective sweep is amino acid substitutions in proteins. We thus asked if derived substitutions that are fixed in present-day humans and change amino acids in proteins occur in any of the 212 regions. There are 139 such substitutions discovered here (see above) and in an accompanying study where non-synonymous substitutions in protein-coding genes on the human lineage were specifically captured and sequenced in a Spanish Neandertal. Four genes (*MRPL53*, *SSH2, ZFYVE26,* and, *C19orf52*) were found to be both in regions of a putative selective sweep and have a fixed amino acid difference between Neandertals and humans, whereas between zero and seven would be expected by chance if these regions were arranged randomly in the genome Table S38. There is, therefore, no significant increase of amino acid substitutions in these regions. Thus, although the list of fixed amino acid substitutions that occurred in modern humans after their divergence from Neandertals is not yet exhaustive, there is no indication that such substitutions underlie a large

proportion of the regions where selective sweeps might have occurred early during modern human evolution.

Another class of genetic changes that may cause a selective sweep is changes in gene expression. We thus asked if genes which show expression differences between humans and chimpanzees in five tissues (*S74*) or genes that changed their timing of expression during human development (*S75*) occur in the 212 regions more often than expected by chance. Again, there is no statistically significant enrichment for these genes within the candidate regions of selective sweeps.

We note that outlier approaches to identify positive selection such as the method developed here suffer from low power and high false positive rates (*S76, 77*), to an extent that depends on many factors that are generally unknown (*S78*). Nevertheless, this approach potentially identifies positive selection in a time period that is crucial for modern human origins and where other approaches have no or very limited power (*S79*). To explore this, we compared the overlap between the 212 regions identified here to the previously identified regions compiled in a recent review of positive selection (*S78*) No significant overlap between regions identified in these previous scans was detected compared to placing them randomly across the genome. The Tajima's D based method of Carlson et al. (*S80*) showed the highest enrichment, 9 of 58 autosomal regions overlapping. Randomly placing these regions 200 times indicates that 7.62 overlapping regions can be expected by chance (5 - 9; 95% confidence interval). We hope that further methodological development in conjunction with sequencing of the Neandertal genome to higher accuracy as well as a comprehensive knowledge of the variation among present-day humans (*S81*) will allow refined versions of the analyses presented here.

Power of selective sweep screen as determined by simulations

To measure the power of these two methods, we simulated various neutral and selective scenarios using a modified version of the "`sweep_der`" program from Thornton and Jensen 2007, (*S82*),

downloaded from http://www.molpopgen.org/software/ThorntonJensen2007/ .

The program was modified to allow us to run selective and neutral scenarios under otherwise identical conditions. We used a simple split between modern humans and

Neandertals that occurred 300ky ago. We simulated a region of the genome of size 300kb, which in the selected case always had a selective event happen in the middle, or for neutrality a region of the same size, without a sweep. We simulated a sample of 12 human chromosomes, and 6 Neandertal chromosomes, from each diploid we sampled one base at each site, and for Neandertal we further sampled from the three Neandertals one single base for each site. The effective population size of humans was fixed to 17,000 to match the level of observed derived alleles in Neandertal at human polymorphic sites, without invoking a model that involves both selection and growth. The conditional probabilities for seeing a Neandertal derived allele at a polymorphic site with a certain frequency among six chromosomes, which is used for calculating the S score, was calculated based on a neutral simulation. We note that these demographic parameters may differ from reality in important ways that are not currently knowable. Therefore, these results must be interpreted with appropriate caution.

Figure S35 shows region width vs. *S* score for all significant regions that were detected under various conditions. Shown are significant regions for a neutral simulation with recombination rates from 1cM/Mb to 0.1 cM/Mb, and for selection simulations with a recombination rate of 1cM/Mb with a selection pressure of s=0.2% to 2%. In each case we simulated around 1% of the total length of the human genome, and for selection, a total of 200 events. Figure S35 can be compared to Figure 4B in the main paper, and one can see that wide regions with high *S* score are highly enriched for selection.

Table S37: Top five percent *S* score regions ranked by genomic width in centimorgans

position (hg18)	*S*	width (cM)	gene(s)
chr2:43265008-43601389	-6.04	0.5726	ZFP36L2;THADA
chr11:95533088-95867597	-4.78	0.5538	JRKL;CCDC82;MAML2
chr10:62343313-62655667	-6.1	0.5167	RHOBTB1
chr21:37580123-37789088	-4.5	0.4977	DYRK1A
chr10:83336607-83714543	-6.13	0.4654	NRG3
chr14:100248177-100417724	-4.84	0.4533	MIR337;MIR665;DLK1;RTL1;MIR431;MIR493; MEG3;MIR770
chr3:157244328-157597592	-6	0.425	KCNAB1
chr11:30601000-30992792	-5.29	0.3951	
chr2:176635412-176978762	-5.86	0.3481	HOXD11;HOXD8;EVX2;MTX2;HOXD1;HOXD10; HOXD13;HOXD4;HOXD12;HOXD9;MIR10B;HOXD3
chr11:71572763-71914957	-5.28	0.3402	CLPB;FOLR1;PHOX2A;FOLR2;INPPL1
chr7:41537742-41838097	-6.62	0.3129	LOC285954;INHBA
chr10:60015775-60262822	-4.66	0.3129	LOC728640;BICC1
chr6:45440283-45705503	-4.74	0.3112	RUNX2;SUPT3H
chr1:149553200-149878507	-5.69	0.3047	SELENBP1;POGZ;MIR554;RFX5;SNX27;CGN;TUFT1;PI4KB;PSMB4
chr7:121763417-122282663	-6.35	0.2855	RNF148;RNF133;CADPS2
chr7:93597127-93823574	-5.49	0.2769	
chr16:62369107-62675247	-5.18	0.2728	
chr14:48931401-49095338	-4.53	0.2582	
chr6:90762790-90903925	-4.43	0.2502	BACH2
chr10:9650088-9786954	-4.56	0.2475	
chr7:18345533-18552832	-4.57	0.2407	HDAC9
chr2:144413853-144892859	-6.57	0.2404	ZEB2;GTDC1
chr5:17957038-18092279	-4.68	0.2313	
chr10:74945897-75146102	-4.56	0.2279	AGAP5;USP54;MYOZ1;SYNPO2L;BMS1P4
chr4:172667677-172874237	-4.5	0.2244	
chr6:49896913-50209330	-5.04	0.2129	DEFB110;DEFB133;DEFB114;CRISP1; DEFB112;DEFB113
chr8:19429699-19541950	-4.39	0.2107	CSGALNACT1
chr4:81396953-81916240	-7.43	0.208	FGF5;C4orf22

chr5:27051867-27187162	-4.34	0.2074	CDH9
chr15:62380190-62739939	-5.33	0.2005	KIAA0101;ZNF609;CSNK1G1;TRIP4
chr1:198167237-198296992	-4.78	0.2002	NR5A2
chr8:92722769-92975785	-4.54	0.196	
chr5:71492503-71683233	-4.78	0.1844	MAP1B;PTCD2;MRPS27
chr5:19644379-20016964	-6.66	0.1842	CDH18
chr18:39554551-39819494	-4.62	0.1836	
chr2:103122470-103338449	-5.54	0.1762	
chr3:71482762-71697708	-5.16	0.1749	MIR1284;FOXP1
chr4:67100391-67305566	-4.65	0.1702	
chr16:60402585-60655562	-4.66	0.1656	CDH8
chr1:97826151-98217969	-6.57	0.1636	DPYD
chr15:81807775-82015388	-5.01	0.1636	SH3GL3
chr18:29811830-30121233	-4.39	0.1609	NOL4
chr2:98551855-98844723	-5.46	0.1605	C2orf64;INPP4A;C2orf55;MGAT4A;UNC50
chr14:104707157-104828815	-4.52	0.1589	BTBD6;NUDT14;BRF1
chr8:34717797-34950744	-6.41	0.1575	
chr1:97002477-97282169	-5.18	0.1568	PTBP2
chr3:183178200-183393223	-5.2	0.1474	
chr12:87533879-87890894	-5.03	0.1451	
chr5:43925639-44140841	-4.98	0.1383	
chr3:178188759-178323090	-4.38	0.1382	TBL1XR1
chr5:109063395-109262205	-5.12	0.1352	MAN2A1;MIR548C
chr12:15297164-15463435	-4.44	0.131	PTPRO
chr6:128152465-128347581	-4.65	0.1306	THEMIS;PTPRK
chr11:45281604-45469462	-4.91	0.1262	
chr8:58590492-58766200	-5.03	0.1257	
chr20:29711520-29944293	-4.41	0.1207	MYLK2;TPX2;BCL2L1;TTLL9;FOXS1;DUSP15
chr13:83030002-83319286	-5.31	0.1191	
chr5:160719697-160959255	-4.69	0.1182	GABRB2
chr9:101374549-101669700	-5.52	0.1175	NR4A3
chr2:148263712-148719018	-7.46	0.1171	ACVR2A;ORC4L
chr5:45086344-45429511	-5.32	0.1153	HCN1

chr11:45702865-45867881	-4.76	0.1129	DKFZp779M0652;MAPK8IP1;SLC35C1;CRY2
chr1:84839222-84950292	-4.59	0.1125	C1orf180;SSX2IP
chr4:45867870-46251900	-5.78	0.1096	GABRA2
chr7:114198584-114416858	-4.68	0.1083	MDFIC
chr2:145309061-145508148	-4.74	0.1083	
chr12:85802694-86029908	-4.53	0.1083	
chr1:63634658-63864734	-4.81	0.1077	ALG6;PGM1;ITGB3BP;EFCAB7;DLEU2L
chr6:84355396-84538906	-4.32	0.1071	SNAP91
chr4:75202676-75367373	-4.87	0.1062	MTHFD2L
chr2:205129991-205360069	-4.49	0.103	PARD3B
chr2:74302337-74568746	-5.91	0.1027	C2orf81;LOC100189589;WDR54;CCDC142;MOGS;INO80B;SLC4A5;WBP1;TTC31;RTKN;MRPL53;DCTN1
chr15:82110672-82287115	-4.56	0.1013	ADAMTSL3
chr8:28829534-29210051	-5.99	0.1009	KIF13B;HMBOX1
chr7:98359455-98551217	-4.39	0.0955	TRRAP;SMURF1
chr17:65441257-65529179	-5.19	0.0943	
chr8:65031265-65286479	-5.11	0.0936	
chr6:132144446-132238150	-4.33	0.0919	ENPP1
chr8:116428760-116638732	-4.69	0.0888	TRPS1
chr20:11371195-11503408	-4.33	0.0883	
chr2:151955667-152143530	-5.43	0.0879	NEB;RIF1
chr2:22515667-22682347	-4.32	0.0846	
chr14:58745905-58892484	-4.85	0.0833	DAAM1
chr11:41612105-41740879	-5.01	0.0816	
chr3:95848272-96105925	-5.21	0.0812	
chr7:48625255-48798384	-4.82	0.0809	ABCA13
chr2:32572353-32771275	-4.83	0.0792	BIRC6;MIR558;TTC27
chr3:157975708-158226727	-4.99	0.0786	LEKR1;PA2G4P4
chr2:15330431-15504097	-5.04	0.0786	NBAS
chr10:103696031-103997252	-4.58	0.0764	GBF1;LDB1;ELOVL3;NOLC1;PITX3;HPS6;C10orf76;PPRC1
chr16:45363049-45702326	-5	0.076	GPT2;NETO2;DNAJA2;C16orf87
chr2:142676611-142845194	-4.51	0.0743	
chr8:49415454-49640151	-4.63	0.0729	
chr11:56383816-56653224	-4.77	0.0722	OR5AK2
chr4:98433976-98748547	-5.13	0.0717	C4orf37

chr8:63894744-64036436	-4.53	0.0689	NKAIN3
chr3:79287548-79620463	-5.57	0.0678	ROBO1
chr1:114917957-115180316	-4.33	0.067	AMPD1;BCAS2;CSDE1;NRAS;SIKE1;DENND2C
chr10:50344537-50485591	-4.49	0.0602	PGBD3;ERCC6
chr12:16776014-16991152	-5.16	0.0599	
chr8:35892186-36187045	-4.8	0.0597	
chr8:66469790-66571309	-4.39	0.0595	
chr6:99092000-99250034	-4.53	0.0594	
chr17:30145601-30423028	-4.75	0.0591	CCT6B;LIG3;ZNF830;RFFL
chr10:106703798-106838410	-4.72	0.0577	SORCS3
chr5:61717832-61996376	-4.4	0.0563	DIMT1L;KIF2A;IPO11;LRRC70
chr8:127160081-127242303	-4.49	0.0559	
chr5:127026941-127270624	-5.33	0.0556	
chr12:64081271-64211458	-4.52	0.0536	MSRB3
chr11:27442372-27677598	-5.03	0.0526	BDNFOS;BDNF;LIN7C;LGR4
chr7:132638934-132840070	-4.38	0.0482	EXOC4
chr14:67346003-67533244	-4.54	0.0472	ZFYVE26;RAD51L1
chr3:13944056-14192304	-5.22	0.0467	TPRXL;XPC;TMEM43;CHCHD4
chr18:32649779-33078352	-6.75	0.0464	BRUNOL4;KIAA1328;C18orf10
chr16:34127293-34595203	-6.69	0.0455	LOC146481;LOC283914;UBE2MP1
chr11:31018381-31627465	-6.89	0.0448	DCDC1;IMMP1L;DNAJC24;ELP4
chr3:48588024-48870224	-4.67	0.0431	UQCRC1;PRKAR2A;SLC26A6;TMEM89;CELSR3;IP6K2;COL7A1;NCKIPSD;SLC25A20;MIR711
chr1:113859346-114182823	-5.28	0.043	PTPN22;RSBN1;PHTF1;MAGI3
chr1:46327105-46490961	-5.3	0.0428	PIK3R3;RAD54L;TSPAN1;POMGNT1;C1orf190
chr13:67521482-67822858	-5.93	0.0423	
chr3:25772581-25951561	-4.59	0.0407	OXSM;NGLY1
chr2:36949298-37111916	-4.53	0.0402	STRN;HEATR5B
chr6:140379904-141018010	-7.63	0.0382	
chr7:126965775-127449571	-6.4	0.0379	GCC1;C7orf54;ARF5;SND1;FSCN3;PAX4
chr3:121457784-121597604	-4.32	0.037	FSTL1;MIR198;LRRC58
chr22:26905509-27244746	-5.28	0.037	TTC28
chr7:40016635-40158534	-4.43	0.0367	C7orf10;CDC2L5;C7orf11

chr5:114144540-114378603	-6.02	0.0363	
chr20:13336947-13608066	-4.4	0.0348	TASP1
chr2:57801690-57969981	-4.5	0.0347	
chr20:33889382-34074304	-4.4	0.0343	SCAND1;PHF20;C20orf152
chr17:27971458-28238024	-5.36	0.0338	MYO1D
chr11:55404491-55652591	-5.16	0.0329	OR10AG1;OR8H3;OR5W2;OR7E5P;OR5F1;OR5AS1;SPRYD5;OR8I2;OR8H2;OR5I1
chr8:53602010-53801129	-4.76	0.0322	RB1CC1;FAM150A
chr12:88771521-89023115	-5.19	0.0318	
chr17:24434840-24619077	-4.81	0.0312	CRYBA1;MYO18A;NUFIP2
chr2:31587161-31824722	-4.9	0.031	SRD5A2
chr10:24947714-25071096	-4.44	0.0305	ARHGAP21
chr1:29123223-29454711	-5.76	0.0302	EPB41;SFRS4;TMEM200B;PTPRU;MECR
chr5:81320640-81673716	-5.53	0.0301	ATG10;ATP6AP1L;RPS23
chr8:49757256-49991278	-4.62	0.0285	EFCAB1
chr1:208202965-208435865	-4.76	0.0283	SYT14
chr2:145851338-146053880	-4.35	0.0282	MIR548Q
chr13:50842278-50980438	-4.54	0.028	INTS6
chr5:86585148-86880429	-5.02	0.0275	CCNH;RASA1
chr2:123291693-123431971	-4.33	0.0272	
chr2:232543676-232888785	-5.85	0.0269	DIS3L2
chr14:82678608-82846353	-4.65	0.026	
chr6:79583236-79877504	-5.01	0.025	PHIP;IRAK1BP1
chr11:59613965-59797766	-5.37	0.025	MS4A2;MS4A6A
chr1:94411925-94611718	-4.66	0.0246	ARHGAP29
chr3:182447686-182606512	-4.57	0.0236	
chr14:70819866-71281353	-6.09	0.0222	SIPA1L1;SNORD56B;LOC145474
chr2:73403433-73647717	-4.85	0.0217	ALMS1
chr3:58816825-59112587	-5.85	0.0216	C3orf67
chr7:43572180-43854900	-5.38	0.0214	STK17A;BLVRA;C7orf44
chr11:108048953-108305665	-4.91	0.0213	DDX10
chr4:124179502-124320006	-4.48	0.0207	SPATA5
chr1:77825306-78062178	-5.47	0.0201	USP33;FAM73A;ZZZ3
chr1:241728677-241961996	-5.03	0.0201	SDCCAG8;AKT3

chr8:71166934-71357997	-4.61	0.0199	NCOA2
chr7:68662946-69274862	-8.7	0.0199	AUTS2
chr14:74496358-74716061	-4.67	0.0199	EIF2B2;FAM164C;TMED10;ACYP1;NEK9;MLH3
chr5:37416335-37685205	-5.65	0.0185	WDR70
chr8:99702821-100083787	-5.86	0.0183	STK3;OSR2
chr9:37775699-37917533	-4.91	0.0179	MCART1;SHB;DCAF10
chr2:61278408-61596041	-5.4	0.0179	XPO1;USP34;SNORA70B
chr4:128839023-129212754	-5.07	0.0171	C4orf29;SLC25A31;LARP1B;INTU;PLK4;HSPA4L;MFSD8
chr19:10792684-11008333	-4.64	0.0171	C19orf52;TMED1;C19orf38;CARM1;SMARCA4;YIPF2;DNM2
chr2:62835302-63143312	-4.5	0.0168	EHBP1;LOC100132215;OTX1
chr3:100977224-101317172	-5.4	0.0166	COL8A1;C3orf26;FILIP1L
chr3:96721658-96906463	-4.4	0.0165	
chr2:155409169-155646459	-5.35	0.0163	KCNJ3
chr9:83984606-84225041	-4.68	0.0161	
chr11:54780414-55198258	-6.23	0.0151	OR4S2;OR4A15;OR4C15;OR4A16;OR4C11;OR4C6;OR4P4;OR4C16;TRIM48
chr13:59438447-59687908	-5.19	0.0147	DIAPH3
chr3:44300268-44638217	-4.87	0.0146	C3orf23;C3orf77;ZNF445;ZNF167;ZNF660
chr11:56765112-56857424	-4.54	0.0143	SSRP1;TNKS1BP1
chr3:161462587-161761512	-4.71	0.0142	MIR15B;MIR16-2;KPNA4;TRIM59;SMC4;SCARNA7;IFT80
chr8:47827581-48223238	-6.14	0.0131	BEYLA
chr20:32937878-33108019	-4.68	0.0125	TRPC4AP;GSS;ACSS2;MIR499;MYH7B
chr2:63781390-64042155	-4.57	0.0122	VPS54;UGP2
chr3:112102198-112387037	-5.42	0.0116	PVRL3
chr17:59899486-60137336	-4.43	0.0115	DDX5;POLG2;SMURF2;CCDC45
chr5:93076209-93576784	-6.15	0.0114	POU5F2;C5orf36;FAM172A
chr3:116166959-116388836	-4.52	0.0112	ZBTB20
chr7:106797749-106964241	-4.63	0.011	GPR22;COG5
chr2:98940428-99199203	-5.21	0.0106	C2orf15;MITD1;MRPL30;TSGA10;LIPT1
chr2:187605993-187813463	-4.39	0.0104	
chr11:72195016-72440608	-4.49	0.0099	FCHSD2;ATG16L2
chr10:74083477-74337153	-4.8	0.0098	OIT3;CCDC109A
chr9:124758063-125065903	-5.68	0.0097	MIR600;STRBP;GPR21;C9orf45;RABGAP1
chr2:135561601-135841635	-5.5	0.0095	ZRANB3;RAB3GAP1

chr20:32451776-32699045	-4.76	0.0093	PIGU;MAP1LC3A;DYNLRB1;MIR644;ITCH
chr15:47463250-47649741	-4.52	0.0082	FGF7;C15orf33
chr3:132117783-132343447	-4.64	0.0079	ATP2C1;NEK11;ASTE1
chr2:72544029-72706415	-4.56	0.0074	EXOC6B
chr17:55626114-55863364	-4.79	0.0073	SCARNA20;C17orf64;USP32
chr3:49072090-49374867	-6.02	0.0056	RHOA;LOC646498;GPX1;USP19;KLHDC8B; C3orf62;LAMB2L;CCDC71;QARS;USP4; CCDC36;LAMB2;QRICH1
chr10:32919972-33181723	-5.01	0.0055	C10orf68
chr17:25248108-25537005	-5.15	0.0048	SSH2;EFCAB5;MIR423;CCDC55
chr13:19432302-19555647	-5.05	0.0047	ZMYM2
chr1:50780760-51008404	-4.42	0.0045	FAF1
chrX:63432354-63720324	-4.64	0.0042	MTMR8
chr1:45799119-45986770	-4.84	0.0039	GPBP1L1;AKR1A1;NASP;IPP;RPS15AP10;TMEM69;CCDC17
chr4:33686539-33924340	-4.39	0.0035	
chr15:69929973-70183174	-4.55	0.0025	MYO9A
chr3:137691478-137912085	-4.74	0.0024	STAG1
chr5:131041314-131291254	-4.65	0.0023	FNIP1
chr6:126709867-127030785	-5.47	0.0018	C6orf173
chr16:34775211-35006524	-5.27	0.0015	
chr8:100187079-100377442	-4.44	0.0001	VPS13B

Table S38: Specific categories of genes are not enriched in regions of putative positive selection. [1]Genes identified in this report and in the accompanying paper, Burbano et al., to have fixed non-synonymous changes in modern humans since Neandertal divergence. [2]Genes with a human-chimpanzee expression difference ($p<0.05$) identified by Khaitovich et al. [3]Genes with an age-related expression trajectory that differs in humans compared to chimpanzees and rhesus.

Gene category	**number of genes**	**observed is SS regions**	**Expected [95% CI]**
Fixed non-synonymous difference[1]	139	4	3.4 [0-7]
Human/chimpanzee expression change[2]	9,370	119	112.3 [99-127]
Human-lineage age-related expression change[3]	442	19	12.9 [6-20]

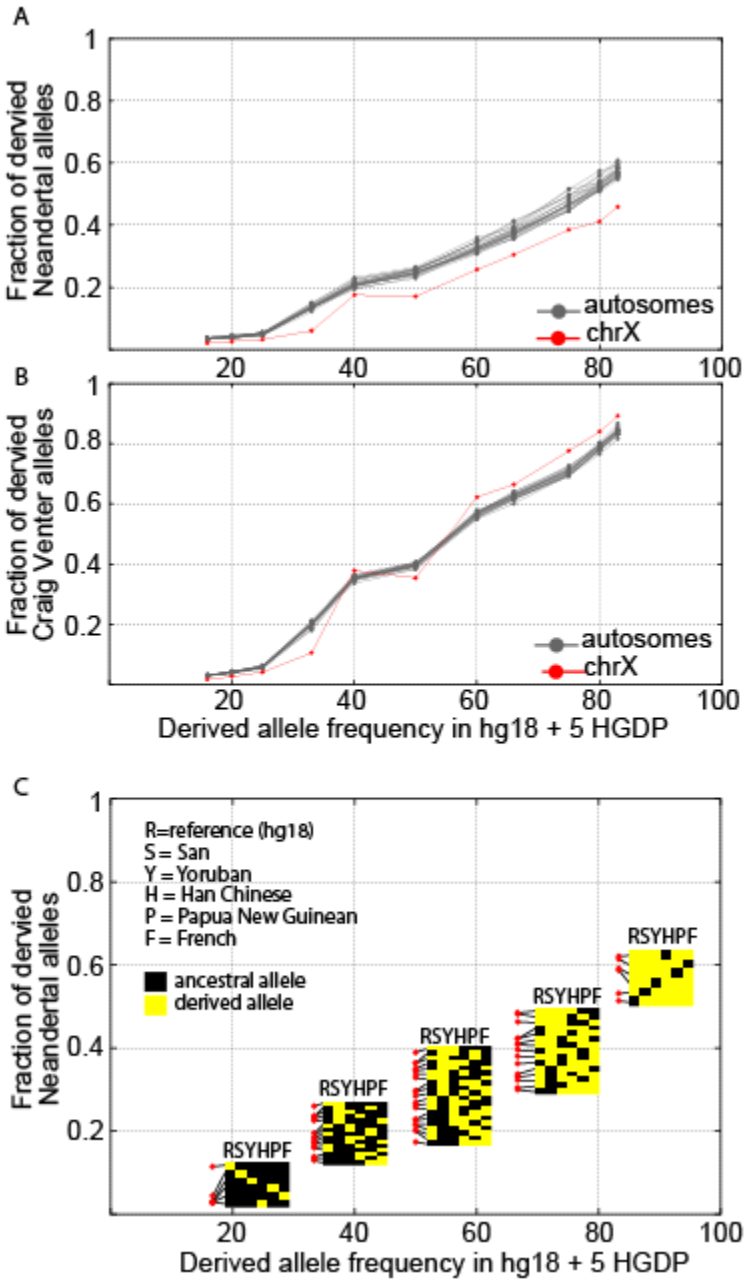


Figure S31: Human derived allele frequency is predictive of Neandertal derived alleles. A. For each human derived allele frequency bin, we calculated the fraction of derived Neandertal alleles. The chrX, with a smaller Ne than the autosomes, is expected (and observed) to share a smaller fraction of derived alleles. B. The same analysis using Craig Venter's genome instead of the Neandertals. As expected, human data are more predictive of Craig Venter derived alleles than Neandertal derived alleles. C. The specific configuration of human derived alleles is informative about the fraction of Neandertal derived alleles.

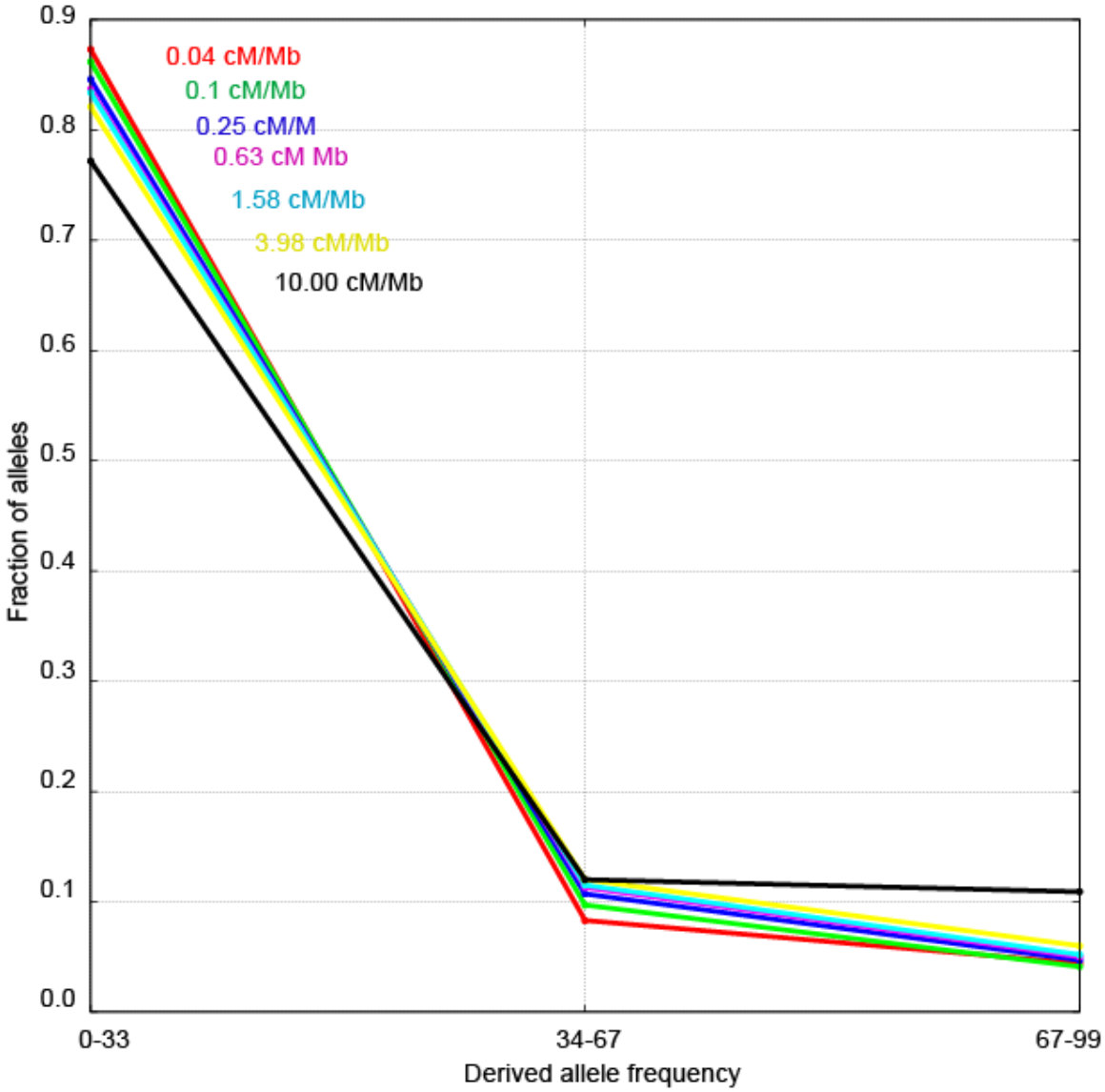


Figure S32: The derived allele frequency spectrum as a function of local recombination rate. Three bins of derived allele frequency were made. At high rates of recombination, there is a higher concentration of alleles in the low-frequency derived bin. As recombination rate increases, more derived alleles are observed to be at higher frequency derived state.

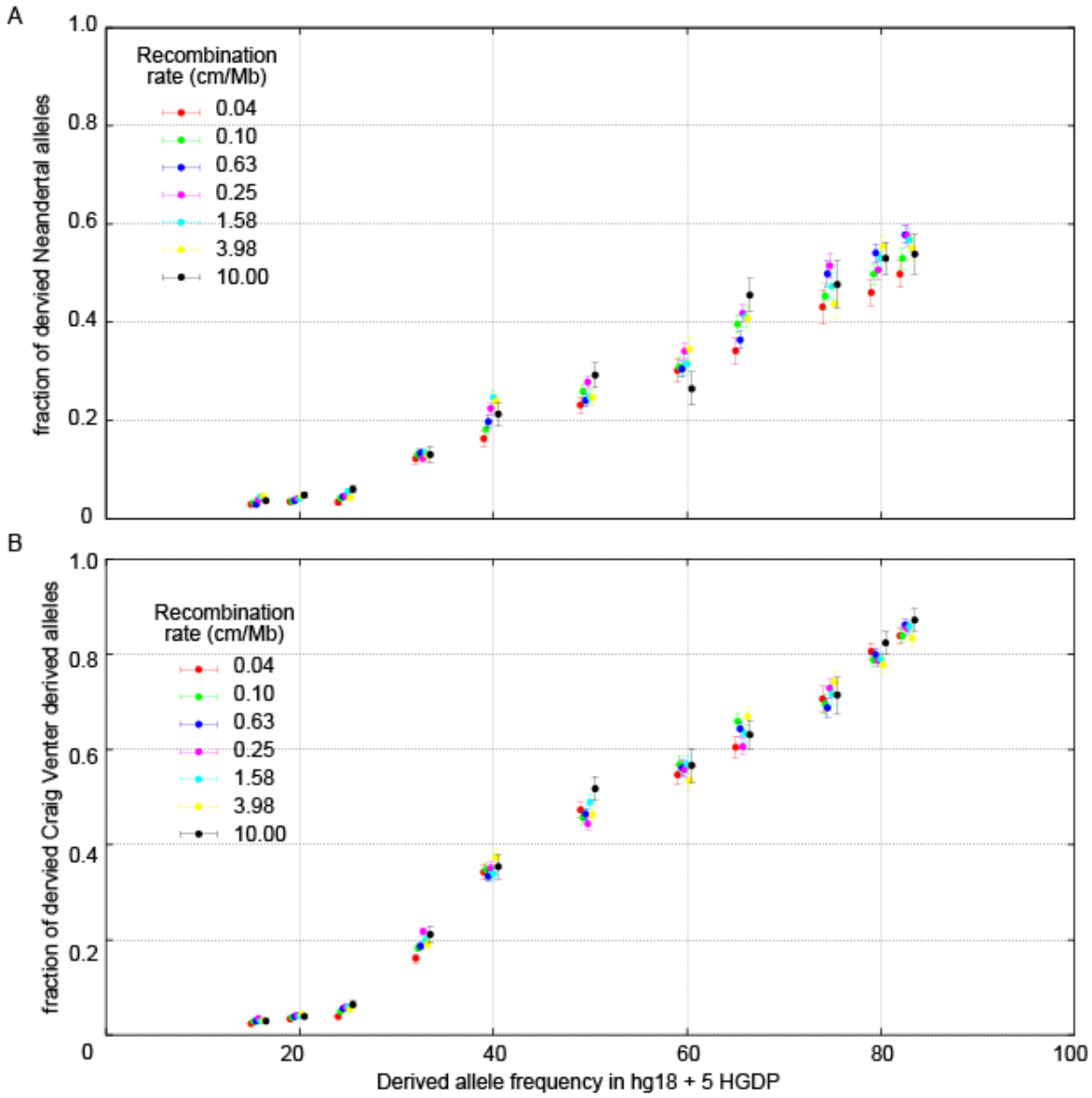


Figure S33: The fraction of Neandertal and Craig Venter derived alleles as a function of local recombination rate and human derived allele frequencies. A. The human derived allele frequency spectrum remains predictive of Neandertal derived alleles across the range of recombination rates. B. The same analysis for Craig Venter alleles shows an equivalent pattern.

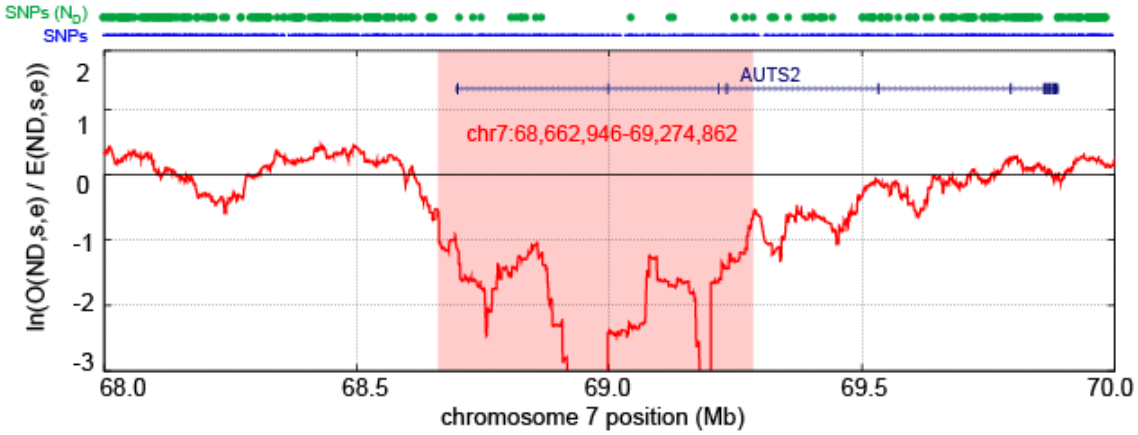


Figure S34: Selective sweep screen region of top *S* score. This region of chromosome 7 contains the gene *AUTS2*. The red line shows the log-ratio of the number of observed Neandertal derived alleles versus the number of expected Neandertal derived alleles, within a 100 kilobase window. Above the panel, in blue is the position of each human polymorphic site. Green indicates polymorphic position where the Neandertal carries derived alleles. The region identified by the selective sweep screen is shown highlighted in pink.

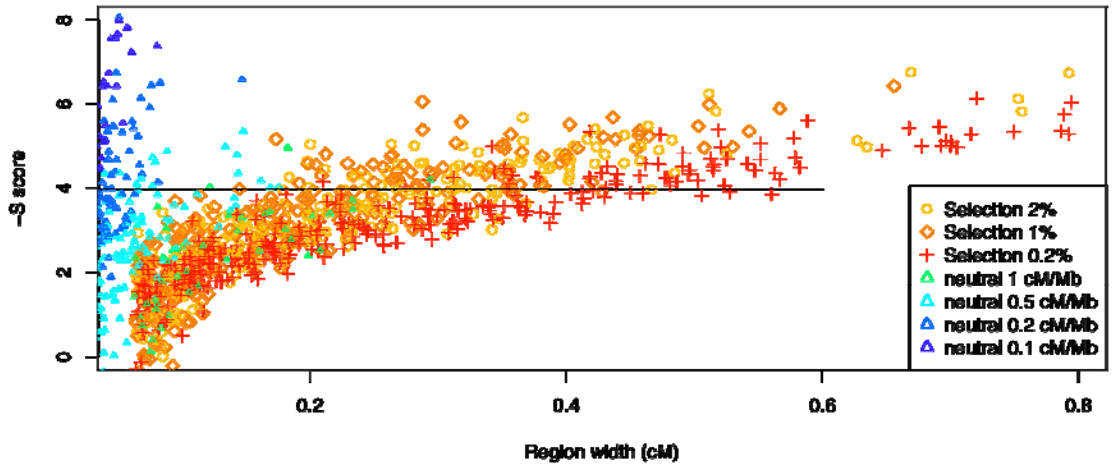


Figure S35: Width of region in centi-Morgan, vs *S* score of region, for significant regions resulting from coalescent simulation under a neutral and selective scenario. The horizontal line corresponds to the 95% quantile of S of the neutral simulation with a recombination rate of 1cM/Mb.

Supplementary Online Materials 14

Date of population divergence between Neandertals and modern humans

Heng Li, James Mullikin and David Reich*

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

In what follows, we estimate that the ancestral populations of Neandertals and modern humans diverged 272,000-435,000 years ago. West Africans are used to represent modern humans in this analysis because in SOM 15-17, we find no evidence for gene flow between Neandertals and the ancestors of West Africans after their initial divergence. This makes interpretation of the population divergence date much simpler.

A panel of 2,192,854 SNPs heterozygous between the two chromosomes of a Yoruba individual

We discovered 2,192,854 SNPs as differences between the two chromosomes of a single Yoruba individual (NA18507) who had been sequenced to 40-fold coverage using paired end 35-41 base pair reads via the Illumina technology (*S62*). We mapped all reads from this individual to the human genome reference sequence (Build 36) using the BWA software with default parameters (*S33*). We called SNPs using SAMtools (*S20*) using the default configuration. The software for calling SNPs automatically removes loci in repetitive regions and with unusually high or low coverage, thus minimizing false-positive SNPs due to repeats, structural variation or mismapping. We further restricted to sites where we were able to obtain both a chimpanzee and orangutan allele based on Ensembl's EPO alignment (*S16*). SNPs aligning to paralogous regions were also removed from our analysis.

Table S39 Percent of SNPs discovered in two Yoruba for which another population X has the derived allele

	F_{chimp} % derived alleles determined by comparison to chimpanzee		$F_{chimp+orang}$ % derived alleles determined by comparison to chimp & orangutan		$F_{corrected-1}$ % derived alleles corrected for recurrent mutation	
	Transversions	Transitions	Transversions	Transitions	Transversions	Transitions
Yoruba	31.1%	31.8%	30.8%	31.1%	30.5%	30.4%
Han	30.3%	31.0%	30.0%	30.4%	29.7%	29.8%
French	30.4%	31.2%	30.0%	30.5%	29.6%	29.9%
Papuan	30.1%	30.7%	29.7%	30.1%	29.3%	29.5%
San	26.9%	27.6%	26.5%	26.8%	26.2%	26.0%
Vi33.16	19.2%	22.3%	18.7%	21.2%	18.2%	20.2% *
Vi33.25	19.1%	21.6%	18.5%	20.3%	17.9%	19.1% *
Vi33.26	19.1%	22.0%	18.4%	20.7%	17.7%	19.5% *

* There is a discrepancy in the estimate of $F_{corrected-1}$ for transitions and transversions for the three Neandertal bones, which we hypothesize reflects the fact that ancient DNA degradation primarily affects transition sites. We restrict to transversions for our analyses below.

Proportion of SNPs for which a third sample from population X carries the derived allele

We measured the proportion of SNPs discovered in two Yoruba chromosomes at which a random sample from another population X carries the derived allele. Here, X is one of the 5 humans we sequenced (Yoruba, Han, French, Papuan and San) or one of the Vindija Neandertal bones (Vi33.16, Vi33.25 and Vi33.26). We randomly drew one read to represent each sample at sites with multiple coverage. The proportion of sites at which population X carries the derived allele is highest for Yoruba, declines for the three non-African populations, declines further for San, and is lowest for the Neandertal bones (Table S39). This decline is expected for populations with increasing divergence from Yoruba, since the more ancient the divergence, the greater the probability that the SNP occurred in present-day humans after divergence.

The proportion of sites that are estimated to be derived varies depending on which group of SNPs we analyze. Taking X=Neandertal as an example, 21.6% of alleles at transitions and 19.1% at transversions match chimpanzee, and 20.7% at transitions and 18.5% at transversions match both chimpanzee and orangutan. We hypothesize that the differences in these rates are due to recurrent mutation, which can cause mislabeling of the derived allele and is most effectively excluded when both the chimpanzee and orangutan

are required to match, and when we restrict to transversions with their low mutation rate. Recurrent mutation on the hominin lineage since the divergence from chimpanzees, however, cannot be detected by using the orangutan sequence. We therefore developed a computational correction for recurrent mutation.

Correcting the estimated proportion of derived alleles for the effect of recurrent mutation

To estimate the proportion of sites where population X carries the derived allele, accounting for the process of recurrent mutation, we make several simplifying assumptions:

1. We assume that exactly one mutation has occurred since the common genetic ancestor of the two Yoruba samples and the sample from population X (at time t^H; Figure S36).
2. We assume that no more than two mutations have occurred since humans and chimpanzees diverged (at time t^{HC}; Figure S36).
3. We assume that no mutations have occurred on the orangutan lineage at the site.

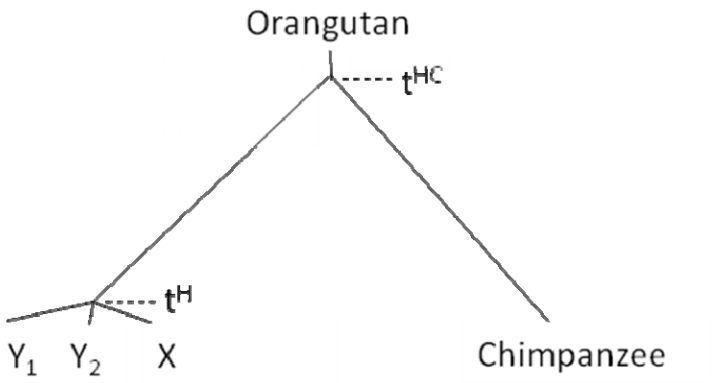


Figure S36: Genealogical tree relating two present-day Yoruba, a third sample X (Neandertal or another present-day human), chimpanzee and orangutan. (We show a trifurcation in this figure, but in fact several topologies are possible relating Y_1, Y_2 and X.) We define the time since the most recent common genetic ancestor of all the humans as t^H, and the time since the most recent common genetic ancestor of humans and chimpanzees as t^{HC}.

We now define the following quantities, which we can estimate empirically from the data:

F_{chimp}	Proportion of Yoruba SNPs where X does not match chimpanzee.
$F_{chimp+orang}$	Proportion of Yoruba SNPs where X does not match chimpanzee and orangutan (we filter out sites where chimpanzee and orangutan are discrepant).
$F_{corrected-1}$	Proportion of Yoruba SNPs where X has the new mutation, correcting for recurrent mutation.
P	Proportion of SNPs affected by recurrent mutation on the lineage leading from chimpanzee to the ancestor of Neandertals and present-day humans (total branch length = $2t_{HC}$-t_H)
Q	Proportion of recurrent mutation SNPs that are expected to have occurred on the lineage leading to present-days humans, equal to (t_{HC}-t_H)/(2t_{HC}-t_H).

With these definitions, we can write down a system of two equations corresponding to the expected values of F_{chimp} and $F_{chimp+orang}$, which takes into account the fact that a proportion of the sites that are analyzed in the data are inevitably due to recurrent mutation:

$$E[F_{chimp}] = (1-P)(F_{corrected-1}) + (P)(1-F_{corrected-1}) \quad \text{(S14.1)}$$

$$E[F_{chimp+orang}] = \frac{(1-P)(F_{corrected-1}) + (P)(1-F_{corrected-1})Q}{(1-P)+(P)Q} \quad \text{(S14.2)}$$

Here, Q is the proportion of recurrent mutations that cannot be detected by the inclusion of orangutan. Although we cannot identify these mutations directly as they occurred on the lineage leading to modern humans, we can estimate how often they occur. Specifically, Q is expected to equal the historical opportunity for mutations on the human side of the tree in Figure S36 (proportional to t_{HC}-t_H), divided by the opportunity on the chimpanzee side (proportional to t_{HC}): Q = (t_{HC}-t_H)/(2t_{HC}-t_H) = (1-t_H/t_{HC})/(2-t_H/t_{HC}). From the divergence estimates from SOM 10, t_H/t_{HC} is empirically about 0.1 for present-day humans, and hence we estimate that Q = (1-0.1)/(2-0.1) = 47%. We also explored the effect of varying t_H/t_{HC} over a wide range that more than spans the values in Neandertals and present-day humans, and found that when we varied it

between 0-0.2, Q varied between 44-50%. This degree of uncertainty in Q has a negligible effect on our correction for recurrent mutation, and thus our inferences are insensitive to the exact value of Q.

We solved the system of equations to estimate $F_{corrected\text{-}1}$ and P. Encouragingly, $F_{corrected\text{-}1}$ estimates are indistinguishable for transitions or transversions for the 5 present-day humans (Table S39), and we correctly estimate a much higher recurrent mutation rate for transitions (P = 3.2%) than transversions (P = 1.7%). For Neandertal, by contrast, the transition- and transversion-based estimates are very different (Table S39). This is likely to reflect the high rate of DNA degradation that is known to occur at transitions in Neandertals (C→T and G→A mutations). In what follows, we therefore restrict our analysis to transversions. Averaging across three Neandertal bones, we estimate that $F_{corrected\text{-}1}$ = 17.9%.

False-positive SNPs in Yoruba individual NA18507 have a negligible effect on our inferences
The Yoruba individual in whom we discovered SNPs, NA18507, was sequenced to approximately 40-fold coverage on the Illumina/Solexa Genome Analyzer. While this in principle provides the basis for confident identification of the great majority of SNPs, we were nevertheless concerned that false-positive SNPs might bias our inferences. At a site that is called as a SNP in NA18507 but that is in fact not heterozygous in that individual, the new sample from a population X would have a very high probability of carrying the ancestral allele (since about 99% of sites in the genome are concordant between humans and chimpanzees). Thus, the presence of false-positive SNPs in the Yoruba in our data set would result in overestimation of the proportion of ancestral alleles and an underestimation of the proportion of derived alleles.

To estimate a false-positive SNP rate from the sequencing of individual NA18507, we took advantage of the fact that a whole-genome SNP array was also run on this individual (*S62*). Comparing sequencing calls to calls from SNP array-based genotyping, Bentley et al. estimated a false-positive rate of R = 0.5%. Defining the further-corrected proportion of derived alleles as $F_{corrected\text{-}2}$, we obtain:

$$E[F_{corrected-1}] = (1-R)(F_{corrected-2}) \quad \text{(S14.3)}$$

Solving for $F_{corrected\text{-}2}$, we obtain a minor upward correction of the estimated proportion of derived alleles by 0.09-0.15% depending on the population (Table S40).

Error from our Illumina sequencing is expected to have a negligible effect on our inferences
We also considered the rate at which sequencing error in the read from population X will cause it to be misread as the alternate allele at the SNP at that nucleotide. We can estimate the substitution-specific transversion error rate based on the numbers derived in SOM 10. Suppose that we have n^{all} nucleotides aligned for three samples: chimpanzees, the human reference sequence, and sample X. We now define allele B as the one carried by chimpanzee (which can be any of A, C, G or T). Let $n^{H}_{B\to D}$ denote the number of sites where the reference human sequence is unique in having the alternate allele D, and $n^{X}_{B\to D}$ the number of sites where X is unique in having D. Under the simplifying assumptions that the chimpanzee and human reference sequences have no errors, and that the same number of true substitutions have occurred in the two hominin samples since their divergence (that is, a molecular clock applies), we can estimate the $B\to D$ specific error rate as the increased number of substitutions on the X lineage compared with the human reference sequence lineage since they diverged:

$$Error\ rate(B \to D) = \frac{n^{X}_{B\to D} - n^{H}_{B\to D}}{n^{all}} \quad \text{(S14.4)}$$

We computed the substitution-specific error rate for each of the 8 possible transversions classes (A→C, A→T, C→A, C→G, G→C, G→T, T→A, and T→G) and averaged them to provide a single transversion-specific error rate $ERR_{transversions}$ for each sample X (Table S40). The error rate $ERR_{transversions}$ in the present-day humans (0.19%) is substantially higher than in the Neandertal (0.04%). This is not surprising given that each nucleotide in the present-day human samples was read only once, whereas in Neandertal, many were read multiple times due to bi-directional sequencing of each clone and the fact that some of the clones were

sequenced multiple times (SOM 2). To infer the proportion of Yoruba SNPs where population X has the derive allele, correcting for errors in the Illumina reads from population X, we can now write the corrected proportion $F_{corrected\text{-}3}$ as:

$$E[F_{corrected-3}] = (1 - ERR_{transversion})(F_{corrected-2}) + (ERR_{transversion})(1 - F_{corrected-2}) \quad \text{(S14.5)}$$

Solving for $F_{corrected\text{-}3}$, we obtain a slight downward correction of the proportion of derived alleles, which varies depending on the sample from 0.03-0.15%.

We conclude that sequencing error in NA18507, or in the read from population X, has a negligible effect on the inferred proportion of derived allele in each population ($F_{corrected\text{-}1}$, $F_{corrected\text{-}2}$, and $F_{corrected\text{-}3}$ are similar in Table S40). Nevertheless, to be maximally careful, we use the $F_{corrected\text{-}3}$ estimates in what follows.

Table S40: Proportion of transversion SNPs where population X is derived for all estimates

	$ERR_{transversion}$ Transversion substitution error rate estimated for this sample	$F_{chimp,transversion}$ % derived alleles at transversions using chimp to determine the ancestral allele	$F_{chimp+orang,transversion}$ % derived alleles at transversions using chimpanzee and orangutan	$F_{corrected\text{-}1}$ % derived alleles at transversions corrected for recurrent mutation	$F_{corrected\text{-}2}$ % derived alleles at transversions corrected for recurrent mutation and 0.5% false-positive SNPs	$F_{corrected\text{-}3}$ % derived alleles at transversions corrected for recurrent mutation, false-positive SNPs, and empirically estimated rate of allele calling error in pop. X
Yoruba	0.17%	31.1%	30.8%	30.5%	30.7%	30.6%
Han	0.16%	30.3%	30.0%	29.7%	29.8%	29.8%
French	0.10%	30.4%	30.0%	29.6%	29.7%	29.7%
Papuan	0.36%	30.1%	29.7%	29.3%	29.4%	29.3%
San	0.16%	26.9%	26.5%	26.2%	26.3%	26.3%
Neandertal pool	0.04%	19.1%	18.5%	17.9%	18.0%	18.0%

Translating from *F* to a population divergence date

The proportion of sites at which a random sample from population X carries the derived allele at a SNP discovered between two Yoruba chromosomes contains information about the date of population divergence of X and Yoruba. Importantly, complexities in the demographic history of population X after its final divergence from Yoruba do not affect the estimate of the final divergence time using this approach. In particular, changes in population size in X do not affect the expected proportion as long as population X did not continue to exchange genes with other populations on the lineage leading to present-day Yoruba.

To understand why complexities in the demographic history of population X since its split from Yoruba do not affect inferences, denote an allele's frequency at the time of t_X of divergence from Yoruba as p_i. In this case, the probability of a single randomly chosen allele from population X being derived today is also p_i. Although subsequent demographic events may dramatically increase or decrease its frequency, they will do so with equal probability so that the expected value is unchanged, a phenomenon that we confirmed by simulations. Thus, we only need to accurately model the demographic history of Yoruba in order to estimate this population's divergence time from a range of other populations.

To build a calibration curve relating the observed proportion of derived alleles in a population X to the divergence time of that population from Yoruba, we carried out coalescent simulations of Yoruba demography using four models that have been fitted to different aspects of data from present-day humans.

- Schaffner et al. *Genome Research* 2005 (*S83*). This model is based on fitting sequence divergence, linkage disequilibrium, and allele frequency differentiation patterns across populations to data from three present-day human populations (West Africans, European Americans, and East Asians). We used the *cosi* simulation software provided with that paper to make inferences about the proportion of derived alleles in population X assuming different times of population divergence from Yoruba.

- Wall et al. *Mol. Biol. Evol.* 2009 (*S30*). This model was developed to search for a signal of ancient substructure in Europeans (possibly due to gene flow from Neandertals). We used the command line calls to the Hudson's *ms* coalescent simulator (*S29*) given in the supplementary materials for that

paper. For population X, we used a constant-sized population and varied the population divergence time. We rescaled the mutation rate to give the same average time since the most recent common ancestors as in the two simulations above.

- Keinan et al. *Nature Genetics* 2007 (*S84*): This model was fitted to allele frequency spectrum data from the Yoruba population, using SNPs discovered in two chromosomes of West African ancestry, and then genotyped in a much larger number of Yoruba samples. The simulation involves a single expansion from a population size of N=7,197 to N=12,855 at a time 5,903 generations ago. We assumed that population X diverged instantaneously from the Yoruba, and was constant in size at 12,855 individuals since that time (as described above, the demography of population X since its divergence from Yoruba does not affect results). Simulations were carried out using Hudson's *ms*.

- Pairwise Sequential Markovian Coalescent model (PSMC): This model has been developed by author Heng Li to infer the distribution of the times since the most recent common ancestor of two haplotypes from a single individual, genome-wide. The model was fit to sequence data from Yoruba individual NA18507, the very individual used to identify the SNPs analyzed here. This distribution was converted into an inference about the changes in Yoruba effective population size over 60 piecewise constant intervals of time. The inferences based on this model, as shown in Figure S37, are consistent with the above three models and support their robustness.

Neandertal-Yoruba population divergence time is estimated to be 272,000-435,000 years ago

To estimate the Neandertal-Yoruba population divergence time τ^{XY}, we simulated models 1-4 for a range of population divergence times to generate the expected proportion of SNPs where population X has the derived allele under this model. We chose to express this proportion as a function of the ratio τ^{XY}/t^{YY}, where t^{YY} is the time since the genetic ancestor of two Yoruba. The value of studying this ratio is that it is not affected by assumptions about mutation rate and generation time, which is valuable as these vary somewhat across the four models above.

We can now convert to an estimate of τ^{XY}/t^{HC}, where t^{HC} is human-chimpanzee average genetic divergence. To do this, we multiply τ^{XY}/t^{YY} by the ratio t^{YY}/t^{HC} = 8.77%, which we measured empirically from an alignment of two West African sequences and a chimpanzee in SOM 10. Figure S37 plots the expected value of τ^{XY}/t^{HC} corresponding to an observed value of $F_{corrected\text{-}3}$, for each of models 1-4. For the empirically observed $F_{corrected\text{-}3}$ = 18.0% for Neandertal, this leads to τ^{XY}/t^{HC} = 0.0488-0.0526. Importantly, the mutation rate and the generation time do not enter into this analysis. By expressing the simulation results as a ratio τ^{XY}/t^{YY}, uncertain quantities like the mutation rate and generation time cancel and do not affect the inference.

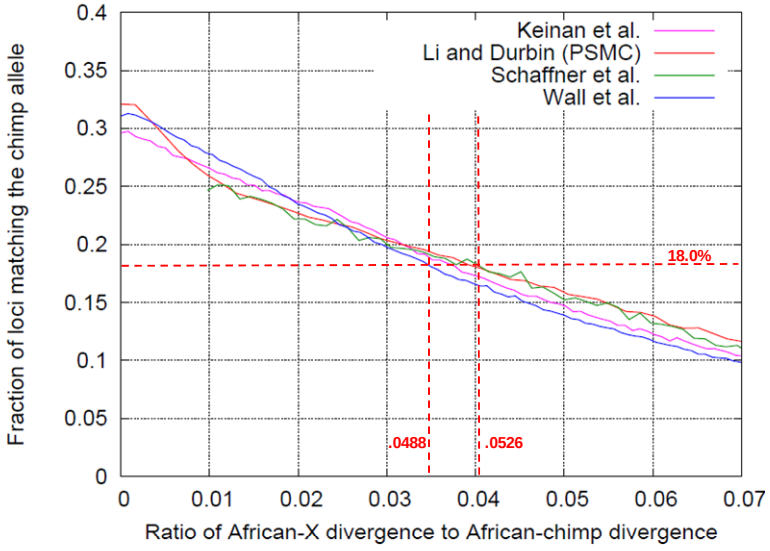


Figure S37: Computer simulations of the proportion of SNPs for which a population X has the derived allele, for SNPs discovered in two Yoruba chromosomes and simulated models of Yoruba demographic history that have been fitted to various aspects of Yoruba data. The four previously reported models all predict that a proportion of derived alleles $F_{corrected\text{-}3}$ = 18.0% for Neandertals corresponds to a divergence time of 0.0488 (t_{HC}) - 0.0526(t_{HC}).

Population divergence time of Neandertals and Yoruba in years

To translate the inference of τ^{XY}/t^{HC} into an absolute time, we assumed 15-20 Mya for human-orangutan speciation time, obtained by assessment of the fossil record relevant to human-orangutan divergence (*S36*). Assuming that human-orangutan autosomal genetic divergence time is not more than 2 Mya greater than speciation time, this translates to 15-22 Mya range for human-genetic autosomal geneic divergence time, which in turn translates to t^{HC} = 5.6-8.3 Mya for human-chimpanzee autosomal genetic divergence based on the inference that human-chimpanzee genetic divergence time is 2.66-times less than human-orangutan genetic divergence time (*S36*).

Using the lower bound of $t^{HC}{}_{years}$ = 5.6 Mya and multiplying by the lower bound of $\tau^{NY}{}_{years}/t^{HC}{}_{years}$ = 0.0488 from our modeling in Figure S37, we obtain a lower bound of 272,000 years for Neandertal-Yoruba population divergence time. Using the upper bound of $t^{HC}{}_{years}$ = 8.3 Mya and multiplying by the upper bound of $\tau^{NY}{}_{years}/t^{HC}{}_{years}$ = 0.0526, we obtain an upper bound of 435,000 Mya for Neandertal-Yoruba population divergence time (Table S41).

Table S41: Yoruba - Neandertal population divergence times vs. genetic divergence times

	Category	Lower bound	Upper bound
Fossil calibration	(A) Human-orangutan speciation time from fossils used for calibration (using a range of speciation times consistent with the fossil record)	15 Mya	20 Mya
Autosomal inferences	(B) Human-chimpanzee genetic divergence time (calibrated) (assuming human-orangutan genetic divergence is at most 2 Mya greater than speciation time, and human-chimpanzee autosomal genetic divergence time is 2.66-times less than human-orangutan divergence time, from (*S36*))	5.6 Mya	8.3 Mya
	(C) Neandertal-Yoruba population divergence time as a fraction of human-chimpanzee genetic divergence from modeling (we quote the full range from the modeling of Figure S37)	0.0488-0.0526	0.0488-0.0526
	(D) Neandertal-Yoruba population divergence time in years (obtained by multiplying B×C)	272,000-293,000	403,000-435,000
	(E) Neandertal-Yoruba genetic divergence time in years (obtained by using the estimate of 13.1% = (Neandertal-African)/(Human-chimpanzee) genetic divergence time from SOM 10)	734,000	1,087,000
	(F) Ancestral diversity of Neandertals and modern humans (average time since the most recent common genetic ancestor in the ancestral population, obtained by subtracting E-D)	441,000-462,000	652,000-684,000

Comparisons of genetic and population divergence times for Yoruba and Neandertal

It is illuminating to compare our estimate of Neandertal-Yoruba population divergence time to Neandertal-Yoruba genetic divergence time.

Autosomes: Assuming times of human-chimpanzee genetic divergence of 5.6-8.3 Mya, and the ratio of (Neandertal-West African genetic divergence)/(Human-chimpanzee genetic divergence) = 13.1% from SOM 10, we obtain an estimate of 734,000-1,087,000 for the genetic divergence of Neandertals and present-day humans. Subtracting the population divergence from the genetic divergence, we estimate 441,000-684,000 years for the average time since the most recent common ancestor (t_{MRCA}) in the ancestral population.

Mitochondrial DNA: We were curious about how our estimates of population divergence times from the autosomes reconcile with mtDNA genetic divergence estimates, when the same fossil calibration times are used. To convert the estimate of autosomal genetic divergence time of humans and chimpanzees of 5.6-8.3 Mya to mtDNA genetic divergence, we used the estimates of parameters in the ancestral population of humans and chimpanzees from Burgess and Yang (*S85*), in which the ratio of the average time to the most recent common autosomal genetic ancestor (t_{MRCA}) in the ancestral population of humans and chimpanzees to human-chimpanzee speciation time was inferred to be 4.1:3.9. We conservatively considered a range of 0-67% for mtDNA to autosomal t_{MRCA} as potentially consistent with our data (the population genetic

expectation is 25%). This led to an estimate of human-chimpanzee genetic divergence in mtDNA between 48-83% of the autosomes; that is, (3.9)/(3.9+4.1) to (3.9+4.1×0.67)/(3.9+4.1).

(i) Assuming 5.6 Mya for human-chimpanzee autosomal divergence, this inference corresponds to 2.7-4.6 Mya for mtDNA divergence, which we truncated to 4.5-4.6 Mya based on the consideration that the fossil record of australopithecines makes a divergence of <4.5 Mya unlikely.

(ii) Assuming 8.3 Mya for human-chimpanzee divergence, this inference corresponds to 4.5-6.9 Mya.

We used these calibration times to estimate plausible times for human-Neandertal mtDNA genetic divergence time that are consistent with our data. A recent study of the (Neandertal-modern human)/(Human-chimpanzee) genetic divergence time ratio in mtDNA has estimated a 95% confidence interval of 5.5-9.4% (*S3*). Multiplying this by the 4.5-4.6 Mya and 4.5-6.9 Mya ranges above, we obtain the inference that Neandertal-mtDNA genetic divergence occurred 250,000-650,000 years ago. By subtraction these numbers, we infer that the mtDNA t_{MRCA} within the ancestral population is 0-210,000 years ago. This is a broad interval, but is entirely plausible.

Estimate of the population divergence time of Yoruba from other present-day humans

We also applied the same analysis to the divergence of other present-day humans we sequenced. We infer that the San diverged from Yoruba between 67,000-164,000 years ago, while all three non-African populations diverged from Yoruba <73,000 years ago, taking into account the full uncertainty of 5.6-8.3 Mya for human-chimpanzee genetic divergence time (Table S42). We caution that our inferences are only as precise as the accuracy of the models we explored. The models produce estimates for the divergence from the three non-African populations that are very imprecise; the difference between the largest and smallest estimate is at least 76% of the largest estimate, as shown in Table S42. Encouragingly, however our estimates for the Neandertal divergence are fairly similar across models (varying by only 7%). Thus, for Neandertal, almost all the uncertainty is due to calibrations against the fossil record.

Table S42: Estimate of population divergence time between Yoruba and each other population X

	Range of t^{XY}/t^{HC} across models	**Range of t^{XY}/t^{HC} across models** as a fraction of the largest estimate $\left(\max\left(t^{XY}/t^{HC}\right)-\min\left(t^{XY}/t^{HC}\right)\right)/\max\left(t^{XY}/t^{HC}\right)$	**Estimated value of t^{XY} in years** (taking into account model-based estimated as well as fossil uncertainty in t_{HC} of 5.6-8.3 Mya)
Han*	0.0008 - 0.0073	89%	5,000 - 60,000
French*	0.0012 - 0.0076	85%	7,000 - 63,000
Papuan*	0.0022 - 0.0088	76%	12,000 - 73,000
San	0.0120 - 0.0198	39%	67,000 - 164,000
Neandertal pool	0.0486 - 0.0524	7%	272,000 - 435,000

* The inferred Neandertal-to-non-African human gene flow from SOM 15-17 means that our results for the Han, French, and Papuan actually reflect an average of the divergence time of Yoruba from the modern human and Neandertal-related ancestors of non-Africans. However, as show in SOM 18-19, the Neandertal mixture proportion is only likely to be a couple of percent, and thus results in only a negligible bias to the inference.

Supplementary Online Materials 15

Neandertals are more closely related to non-Africans than to Africans

Nick Patterson, Nancy Hansen, James Mullikin and David Reich*

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

To test whether Neandertals share more alleles with some present-day human populations than with others, we compared the Neandertal sequence that we generated to sequence from present-day human samples of diverse ancestry. Specifically, we discovered single nucleotide polymorphisms (SNPs) by comparing exactly two chromosomes from different individuals (H_1 and H_2). We then assessed whether a test individual (H_3, e.g. Neandertal) tended to match either H_1 or H_2 more often at sites where H_3 has the derived allele relative to chimpanzee. Under the null hypothesis that H_3 belongs to an outgroup population, it should match H_1 and H_2 equally often. In contrast, if gene flow has occurred, H_3 may match one more than the other. For these analyses, we used Neandertal data generated using both the Illumina GA_{II} and Roche 454 technologies, and present-day human data generated using both the ABI3730 and Illumina GA_{II} technologies.

Processing of the ABI3730 data from 8 present-day humans

We first analyzed data from 8 present-day humans (4 West Africans, 2 East Asians, and 2 European Americans) that had been generated using an ABI3730 sequencer using a protocol that produced pairs of long reads (average of 752 bp) with an approximate spacing of 40 kb. A range of 1.9-3.3 million sequencing reads were generated from these samples (*S86*).

We next applied a series of read and nucleotide filters to the data, producing 0.67-1.42 Gb of nucleotides of sufficient quality to call SNPs for each sample (Table S43):

1. We restricted to reads from fosmid clones that mapped to the PanTro2 chimpanzee sequence. For all the analyses that follow, reads were mapped to chimpanzee to avoid a bias toward matching the samples used to generate the human reference sequence.
2. We used SSAHA SNP to identify single nucleotide polymorphisms (SNPs), applying the criterion that each SNP had a quality score of at least Q≥40 and 5 flanking bases on either side had quality scores of Q≥15 (*S87*).
3. For each present-day human ($i = 1 \ldots 8$) we identified a cutoff X_i such that 99.5% of nucleotides in the chimpanzee reference sequence had read coverage of less than X_i. We removed nucleotides with coverage greater than this to filter out potential sites of copy number variation and poor mapping.

A summary of the data we analyzed from these 8 present-day humans is presented in Table S43.

Table S43: ABI3730 sequencing data from 8 present-day humans

Sample ID	Our label	Population	ABI 3730 Reads*	Number of nucleotides analyzed to discover SNPs
NA18517	YRI	Yoruba	2,076,237	744,994,084
NA18507		Yoruba	3,331,676	1,410,858,692
NA19240		Yoruba	2,118,546	910,314,367
NA19129		Yoruba	2,053,392	672,021,814
NA18956	ASN	Japan	2,076,828	895,269,089
NA18555		China	1,966,644	825,605,562
NA12878	CEU	CEPH	2,168,656	870,647,385
NA12156		CEPH	2,021,844	1,027,157,061

* Numbers reproduced from Kidd et al. *Nature* 2008.

Processing of the Illumina GA_{II} data from 5 present-day humans

For each of the 5 males from the CEPH-Human Genome Diversity Project panel (French, Han, Papuan, San and Yoruba group), we only used data that passed the following filters:

1. We only used paired-end reads where both ends of the clone were available and mapped to the PanTro2 chimpanzee sequence with opposite orientation and the expected spacing.
2. We required that there were no insertion/deletion polymorphisms when aligned to chimpanzee.

3. We required that all reads be exactly 76 bp in length.
4. We removed bases with quality scores $Q>40$, since at sites with higher quality scores we found that divergence was significantly elevated relative to chimpanzee. We could not discern a cause for this, but we note that it affected a small fraction of bases, and had a negligible effect on our analyses.
5. We restricted to sites from the 5 present-day humans where we had coverage from at least one Neandertal read with a base of good quality (next section). "Duplicate reads" with the same endpoints on the reference were removed, and low complexity read pairs (entropy < 1.0) were also removed.
6. For the Papuan sample we found that base 34 of each read had significantly higher divergence with chimpanzee and other humans than other positions in the read. This is likely due to sequence error, and we dropped these bases from the analysis. This only makes a small difference to the results.

Processing of the Illumina GA_{II} data from Neandertal

For the comparison of the Neandertal Illumina data to the Illumina data from 5 present-day humans, we applied several additional filters to the Neandertal data:

1. We required that all reads map to the PanTro2 chimpanzee genome sequence.
2. We required there to be no insertion/deletion polymorphisms when aligned to chimpanzee.
3. We restricted to autosomal reads.
4. We required that no more than 10% of bases in the reads disagree with chimpanzee.
5. We only accepted nucleotides with a map quality (MAPQ) of at least 43 as assessed by the ANFO software, which is specialized for mapping ancient DNA reads (*S13*).
6. We did not analyze nucleotides within 5bp of either end of the Neandertal read, because it is known that sequence quality is lower in these regions (*S28*).

Calibration curves used to make allele calls in the Illumina GA_{II} data

An important feature of Illumina GA_{II} data is that each sample has its own subtle biases in terms of which sequencing errors are most common. These biases vary according to which nucleotide type and strand is being examined. Thus, a potential source of false-positives could be that samples H_1 and H_3 have a correlated error process, which could make them artifactually match at a higher rate than H_2 and H_3, even though in fact H_1 and H_2 are a clade and there has been no history of gene flow to or from H_3.

To minimize these biases, we built calibration curves in a base-, sample-, and strand-specific manner, and discarded half the data (the lower quality half) for each combination of these three categories. The goal was to ensure that samples with a bias toward lower quality scores at particular classes of alleles did not have a higher rate of data dropout. Briefly, for each base $a \in \{A,C,G,T\}$, each strand of the chimpanzee reference sequence to which a read mapped $s \in \{\text{forward, reverse}\}$, and each present-day human sample $x \in \{French, Han, Papuan, San, Yoruba\}$, we computed a quality score threshold $T(a, s, x)$ such that the probability of acceptance > 1/2. For bases with quality scores that exactly overlapped the 50^{th} percentile, we randomly chose nucleotides for inclusion in our analysis, such that the probability of acceptance was exactly 1/2.

Statistical details of the test for gene flow

We developed a statistical test to assess whether two samples H_1 and H_2 are consistent with descending from a common ancestral population, which diverged at an earlier time from the ancestors of a third sample H_3. In the basic version of this test, for each position in the genome where we have coverage from three samples H_1, H_2 and H_3, we chose one allele from each sample. For the ordered set $\{H_1, H_2, H_3, Chimpanzee\}$, we denote the chimpanzee allele as "*A*", and restrict our analysis to biallelic sites at which H_1 and H_2 differ and the alternative allele *"B"* is seen in H_3. At these sites, we have observed two copies of both alleles, making it unlikely that the base substitutions we are analyzing have arisen due to sequencing error.

We call the two possible patterns of SNPs "ABBA" or "BABA". If H_1 and H_2 descend from a common ancestral population that diverged at an earlier time from the ancestral population of H_3, then we expect ABBA and BABA to occur with exactly equal frequencies. Alternatively, if H_1 and H_2 do not form a clade *vis-à-vis* H_3, they can occur at different rates. We note that differing mutation rates in the H_1 and H_2 populations since their divergence are not expected to contribute to a false-positive deviation from 0%. The

reason for this is that by restricting to sites where H_3 is derived relative to chimpanzee, we are almost certainly restricting to mutations that arose prior to divergence of the ancestral populations of H_1, H_2 and H_3. Figure S38 illustrates why a test for an equal rate of ABBA and BABA sites provides a formal test of the null hypothesis that populations H_1 and H_2 form a clade in a phylogenetic tree relative to H_3.

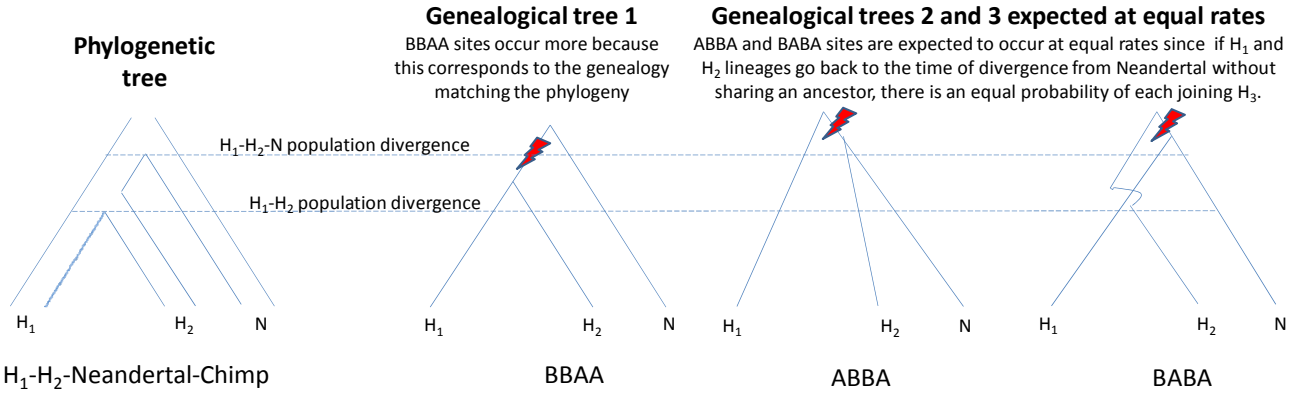


Figure S38: Comparing ABBA and BABA counts provides a test for whether Neandertals are an outgroup to modern humans. We test the null hypothesis that H_1 and H_2 form a clade, with Neandertals more distantly related. If we look at a single sample from each population, there are three possible trees relating the three hominins and chimpanzees, as explored in more detail in SOM 19. (Tree 1) The most common is a tree in which H_1 and H_2 are most closely related, matching the phylogeny. Mutations ancestral to H_1-H_2 divergence (red mark) will leave a BBAA pattern. (Trees 2 and 3) If the H_1 and H_2 lineages go back to divergence from Neandertal without coalescing, there is an equal probability of each joining N, resulting in an equal rate of ABBA and BABA sites. Differences in the demographic history of the H_1 and H_2 populations have no affect on the prediction of an equal rate of ABBA and BABA sites (despite the profound demographic differences between some human populations, e.g. non-Africans than Africans, that have resulted in substantial differences across populations like differences in the averaged derived allele frequency). By picking a single sample from each population, the lineages are guaranteed to go back to the common ancestral population without coming together. Thus demography after the H_1-H_2 split does not affect the relative rates of ABBA and BABA sites.

We developed a formal test for a difference between the rates of ABBA and BABA sites. To do this, we computed a statistic corresponding to the difference in the count of ABBA and BABA sites across the n base pairs in the autosomes for which we have alignment of all four samples, normalized by the total number of observations. In this statistic, $C_{ABBA}(i)$ and $C_{BABA}(i)$ are indicator variables; they can be 0 or 1 depending on whether an ABBA or BABA pattern is seen at base i.

$$D(H_1, H_2, H_3, chimpanzee) = \frac{\sum_{i=1}^{n}\left[C_{ABBA}(i) - C_{BABA}(i)\right]}{\sum_{i=1}^{n}\left[C_{ABBA}(i) + C_{BABA}(i)\right]} \qquad \text{(S15.1)}$$

To evaluate the D statistic in the Illumina data at sites with multiple read coverage in a sample, we picked one read at random to represent the sample. To evaluate D in the ABI3730 data, we pooled reads at sites with multiple read or sample coverage, allowing us to measure D with higher precision. To summarize the data from each population, we estimated the derived allele frequency $\hat{p}_{ij}$ of SNP i in population j. For a single sample, $\hat{p}_{ij}$ was estimated by examining all reads mapping to the nucleotide; if all reads match chimpanzee, $\hat{p}_{ij}$=0%; if only some do, $\hat{p}_{ij}$=50%; and if all are derived, $\hat{p}_{ij}$=100%. For a pool of m samples (e.g. 4 West Africans, 2 East Asians, 2 European Americans, or 3 Neandertal bones), we weighted the individual estimates by the expected number of that individual's two chromosomes that were sampled: w_{ik}=0 with no coverage, and w_{ik} = 2-0.5^{r-1} with $r\geq 1$ read covered. Then, $\hat{p}_{ij} = \left(\sum_{k=1}^{m} \hat{q}_{ik} w_{ik}\right) \Big/ \left(\sum_{k=1}^{m} w_{ik}\right)$, where $\hat{q}_{ik}$ are the frequency estimates per sample. We can now write a generalized equation that allows us to combine data sites with multiple coverage in a population, which allows us to measure the D-statistic more precisely.

$$D(H_1, H_2, H_3, chimpanzee) = \frac{\sum_{i=1}^{n}(1-\hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3} - \hat{p}_{i1}(1-\hat{p}_{i2})\hat{p}_{i3}}{\sum_{i=1}^{n}(1-\hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3} + \hat{p}_{i1}(1-p_{i2})\hat{p}_{i3}} \qquad \text{(S15.2)}$$

Mathematically, Equation S15.2 reduces to Equation S15.1 for random sampling of one allele from each individual at sites that are multiply covered.

Block jackknife to compute standard errors

To compute a standard error on the *D*-statistic as well as other statistics in this study, we used a Weighted Block Jackknife (*S88*). Briefly, we divided the genome into a number *M* of contiguous stretches (non-overlapping) that are chosen to be many Mb in size (chosen to be larger than the extent of linkage disequilibrium). By computing the variance of the statistic over the entire genome *M* times leaving each block of the genome in turn, and then multiplying by *M* and taking the square root, we can obtain an approximately normally distributed standard error using the theory of the jackknife. For the genome-wide analysis of the ABI3730 data, we divided the genome into *M*=100 contiguous blocks with an equal number of divergent sites, and for chromosome-specific analyses, use *M*=20 contiguous blocks. For the analysis of the Illumina data, we divided the genome into *M*=604 contiguous 5 Mb blocks, and weighted their contribution to the jackknife to account for the variable number of SNPs (*S89*). We found in practice that the standard error of our jackknife was stable as long as blocks were set to be at least 2 Mb (Table S44), a threshold that was met for all the computations that follow.

Table S44: Standard errors for the block jackknife analysis are reliable as long as blocks are ≥2 Mb

	D(YRI,CEU,Neandertal,Chimp)			***D(YRI,ASN,Neandertal,Chimp)***			***D(ASN,CEU,Neandertal,Chimp)***		
Blocks	**D**	**std.err.**	**Z**	**D**	**std.err.**	**Z**	**D**	**std.err.**	**Z**
50	4.57%	0.40%	11.4	4.81%	0.39%	12.5	-0.35%	0.48%	-0.7
100	4.57%	0.36%	12.7	4.81%	0.39%	12.2	-0.35%	0.46%	-0.8
200	4.57%	0.34%	13.3	4.81%	0.38%	12.6	-0.35%	0.50%	-0.7
400	4.57%	0.34%	13.6	4.81%	0.35%	13.6	-0.36%	0.49%	-0.7
800	4.57%	0.33%	13.8	4.81%	0.35%	13.7	-0.36%	0.48%	-0.7
1600	4.57%	0.33%	13.8	4.81%	0.35%	13.7	-0.36%	0.48%	-0.7
3200	4.57%	0.31%	14.8	4.81%	0.32%	15.0	-0.36%	0.44%	-0.8
6400	4.57%	0.31%	14.8	4.81%	0.32%	15.0	-0.36%	0.44%	-0.8

Note: Results for different jackknife block sizes on the ABI3730 present-day human data, pooling 4 West Africans, 2 European Americans, 2 East Asians, and 3 Neandertal bones. Standard errors are stable as long as the number of blocks is ≤1600 (span of ≥2 Mb).

Neandertals are more closely related to non-Africans than to Africans

We computed $D(H_1,H_2,Neandertal, Chimpanzee)$ for all 28 pairs of present-day humans in the ABI3730 data (Table 4). Neandertal matches the 4 non-Africans significantly more than the 4 Africans in all 16 comparisons ($3.6 < Z < 9.9$ standard deviations), but all comparisons of pairs of African samples or of pairs of non-African samples are non-significantly deviated from 0 after correcting for 12 hypotheses tested ($|Z| << 2.7$). Pooling samples from each ancestry, the African-to-non-Africans difference increases in significance ($Z = 12.2$-12.7), and we continue to observe no European-East Asian difference ($Z = -0.8$).

Table S45: Consistency of D(YRI,CEU,Neandertal,Chimpanzee) for various subsets of data

Category 1 vs. Category 2	**Category 1**	**Category 2**	**2-sided P-value for heterogeneity**
454 vs. Illumina Neandertal*	4.86 ± 0.51%	4.50 ± 0.38%	0.37
Transversions vs. Transitions*	4.54 ± 0.49%	4.58 ± 0.38%	0.92
Non-CpG vs. CpG sites *	4.63 ± 0.37%	4.39 ± 0.58%	0.69
≤50 bp vs. ≥50 bp Neandertal read	3.80 ± 0.38%	4.51 ± 0.40%	0.021
>5bp vs. ≤5 bp from end of read	4.50 ± 0.38%	3.25 ± 0.42%	0.0010
Vi33.16 vs. Vi33.25 bone	4.48 ± 0.45%	3.82 ± 0.38%	0.11
Vi33.16 vs. Vi33.26 bone	4.48 ± 0.45%	3.95 ± 0.42%	0.19
Vi33.25 vs. Vi33.26 bone	3.82 ± 0.38%	3.95 ± 0.42%	0.76

Note: All analyses are restricted to SNPs discovered in the ABI3730 fosmid data, pooling the data from 4 YRI to represent West Africans and 2 CEU samples to represent Europeans. The P-values for heterogeneity are obtained by a block jackknife of the Category 1 - Category 2 difference across the genome, and reporting the number of standard deviations from 0.
* The first 3 rows combine Illumina and 454 Neandertal data; the later rows are all Illumina.

We repeated the analysis in various subsets of the data. We found no statistically significant difference between results on the Illumina and 454 Neandertal data, transitions and transversions, hypermutable CpG dinucleotides and all other sites, and the three Vindija bones (Vi33.16, Vi33.25 or Vi33.26). We did find that

shorter Neandertal reads have significantly smaller *D*-statistics than longer (3.80% for <50bp vs. 4.51% for ≥50bp; P=0.021), which may reflect a higher rate of read mapping errors in shorter reads. We also found that nucleotides within 5bp of the ends of Neandertal reads produce smaller *D*-statistics than nucleotides in the interior (3.25% vs. 4.50%; P=0.0010), probably reflecting the high error rate at the ends of reads (*S28*). In all other analyses reported below and in the main manuscript, we filter sites 5bp from the end of each Neandertal read.

We computed *D*-statistics for all possible nucleotide substitution classes to explore the consistency of our results. While there is significant heterogeneity across substitution types, when we pool the data into 6 classes that are symmetric by strand (e.g. pooling C→A and G→T), there is no longer substantial evidence of heterogeneity (Table S46). We also computed *D* statistics by chromosome, and found that while the signal is widespread across chromosomes, there is also significant heterogeneity across chromosomes (Table S47). We hypothesize that this reflects variability in the sequencing error process across chromosomes and samples that is not captured by the jackknife analysis. There is no evidence of particular loci that are contributing to most of the signal; when we analyzed the most outlying chromosome (chromosome 12: *D(YRI,ASN,Neandertal,Chimpanzee*) = 13.0 ± 2.5%), there is no single locus driving the signal.

Table S46: Signal of gene flow is widespread across substitution classes

Mutation	**Excess matching to CEU at CEU – YRI sites**			**Excess matching to ASN at ASN – YRI sites**			**Excess matching to CEU at CEU – ASN sites**		
	% match	Std. err.	Normal Z score	% match	Std. err.	Normal Z score	% match	Std. err.	Normal Z score
A→C	3.4%	1.2%	2.8	5.6%	1.2%	4.6	-3.3%	1.8%	-1.8
A→G	5.6%	0.6%	9	7.0%	0.7%	9.8	0.0%	0.8%	0
A→T	3.5%	1.3%	2.8	3.9%	1.4%	2.8	-1.1%	2.1%	-0.6
C→A	5.7%	1.2%	4.9	4.0%	1.3%	3	-5.6%	1.6%	-3.6
C→G	7.4%	1.2%	6.2	2.7%	1.3%	2.1	4.9%	1.6%	3
C→T	3.2%	0.6%	5.6	3.2%	0.6%	5.1	0.9%	0.9%	1.1
G→A	5.4%	0.6%	9.3	5.4%	0.6%	9	-1.3%	0.9%	-1.3
G→C	3.3%	1.2%	2.8	5.7%	1.2%	4.6	1.4%	1.8%	0.8
G→T	2.0%	1.2%	1.6	4.6%	1.3%	3.6	0.1%	1.8%	0
T→A	5.5%	1.4%	4	5.8%	1.6%	3.7	-0.1%	1.9%	0
T→C	4.3%	0.7%	6.3	4.2%	0.7%	6	-1.3%	0.9%	-1.4
T→G	5.5%	1.3%	4.4	5.3%	1.4%	3.9	1.4%	2.0%	0.7
Heterogeneity?			*P = 0.0072*			*P = 0.017*			*P = 0.0014*
A→C or T→G	4.5%	0.9%	5	5.4%	0.9%	6	-0.9%	1.4%	-0.6
A→G or T→C	4.9%	0.5%	9.8	5.6%	0.6%	9.7	-0.7%	0.7%	-1
A→T or T→A	4.5%	1.0%	4.4	4.9%	1.0%	4.7	-0.6%	1.4%	-0.4
C→A or G→T	3.8%	0.8%	4.7	4.3%	1.0%	4.5	-2.8%	1.2%	-2.3
C→G or G→C	5.3%	0.9%	6	4.2%	0.9%	4.7	3.1%	1.3%	2.5
C→T or G→A	4.3%	0.4%	9.9	4.3%	0.5%	9.2	-0.2%	0.7%	-0.2
Heterogeneity?			*P = 0.51*			*P = 0.78*			*P = 0.032*

Note: We calculate P-values for heterogeneity using χ^2 tests with 11 and 5 degrees of freedom.

European contamination cannot explain the skew of the *D*-statistics

We considered the possibility that the skew we detect could be due to European contamination in the Neandertal extract, which might reflect contamination from laboratory personnel or archaeologists (*S34, 90*). While there are multiple lines of evidence against contamination based on mtDNA, Y chromosome, and autosomes (Table 1 and SOM 5-7), and the consistency of *D* across bones (Table S45), we nevertheless considered this possibility in further detail.

Under the hypothesis that all the skew in the *D* statistics is due to European contamination, we expect that Neandertals will be more closely to related to Europeans than to East Asians, just as they are more closely related to Europeans that to West Africans. However, in our data, *D(YRI,CEU,Neandertal,Chimpanzee)* =

4.57 ± 0.36%, whereas *D(ASN,CEU,Neandertal,Chimpanzee)* = -0.35 ± 0.46%, a skew that is (non-significantly) in the opposite direction from what is expected from contamination. We conclude, there is no evidence at all for Neandertals being more closely related to Europeans than to East Asians, and hence there is no evidence for European contamination in these data.

Replication of the finding that Neandertal is closer to non-Africans in Illumina data

We next calculated the *D* statistics for all 10 pairs of present-day human samples sequenced using the Illumina technology. Neandertal is more closely related to non-Africans than to Africans in all 6 pairwise comparisons of African and non-African populations ($7.1 < Z < 10.8$ standard deviations), producing consistent values of the *D* statistics when compared with the ABI3730 data (Table 4). A particularly important result is that we find no statistically significant difference in the relationship of the three non-African populations to Neandertal ($|Z| \ll 1.7$), and no statistically significant difference in the relationship of San and Yoruba to Neandertal ($|Z| = 0.3$). These results suggest that the skews in the *D* statistics reflects an event that occurred before non-African populations separated from each other, and either (a) after the separation of non-Africans from all Africans, or (b) before the separation of Yoruba and San.

Table S47: Signal of gene flow affects many chromosomes although there is heterogeneity

Chro-mosome	Excess matching to CEU at CEU-YRI sites			Excess matching to ASN at ASN-YRI sites			Excess matching to CEU+ASN at CEU+ASN-YRI sites			Excess matching to CEU at CEU-ASN sites		
	D	Std err	Z score	D	Std err	Z score	D	Std err	Z score	D	Std err	Z
1	4.3%	1.2%	3.6	6.4%	1.2%	5.2	6.1%	1.1%	5.8	-2.3%	1.4%	-1.6
2	5.7%	1.3%	4.6	5.2%	1.4%	3.7	5.3%	1.0%	5.1	0.3%	1.7%	0.2
3	3.7%	1.4%	2.7	6.0%	1.6%	3.8	5.2%	1.2%	4.3	-3.6%	2.0%	-1.8
4	3.2%	1.3%	2.4	2.3%	1.8%	1.3	3.3%	1.3%	2.5	-5.3%	2.3%	-2.4
5	5.9%	1.6%	3.6	5.2%	2.3%	2.3	5.3%	1.3%	4.1	-0.7%	2.4%	-0.3
6	5.6%	1.7%	3.3	6.7%	1.4%	4.7	5.9%	1.2%	4.9	2.6%	1.9%	1.4
7	5.1%	1.2%	4.2	2.5%	1.4%	1.8	3.9%	1.2%	3.2	0.8%	1.7%	0.5
8	3.9%	1.7%	2.2	2.7%	2.1%	1.3	3.7%	1.6%	2.3	-1.2%	2.8%	-0.4
9	5.0%	1.3%	3.9	7.8%	1.5%	5.4	6.9%	1.1%	6.3	-1.5%	2.0%	-0.8
10	3.1%	1.7%	1.8	6.5%	1.8%	3.6	4.5%	1.6%	2.9	-0.9%	2.1%	-0.4
11	4.9%	1.5%	3.2	6.1%	1.5%	4.1	4.8%	1.2%	3.9	-2.7%	2.1%	-1.3
12	8.4%	2.2%	3.8	13.0%	2.5%	5.3	10.3%	1.9%	5.4	1.9%	2.1%	0.9
13	5.5%	1.9%	2.8	7.6%	2.4%	3.2	7.1%	1.7%	4.3	3.3%	1.8%	1.8
14	7.5%	1.4%	5.4	2.4%	1.7%	1.4	5.0%	1.3%	3.8	1.4%	2.3%	0.6
15	4.0%	1.8%	2.3	1.3%	1.7%	0.7	2.1%	1.5%	1.5	3.1%	2.1%	1.5
16	5.0%	1.4%	3.7	3.4%	1.2%	2.8	5.0%	1.0%	5.1	1.8%	2.4%	0.8
17	-0.9%	2.4%	-0.4	2.9%	1.5%	2	1.0%	1.7%	0.6	-8.1%	2.3%	-3.6
18	5.1%	1.6%	3.2	5.4%	1.6%	3.4	4.9%	1.1%	4.7	2.5%	2.9%	0.9
19	7.1%	1.5%	4.7	1.9%	1.8%	1.0	4.8%	1.2%	4.1	5.2%	2.6%	2.0
20	4.4%	2.3%	2	1.0%	3.1%	0.3	3.0%	2.2%	1.4	7.1%	2.9%	2.4
21	0.4%	2.9%	0.1	-1.4%	2.3%	-0.6	0.3%	2.4%	0.1	1.3%	3.4%	0.4
22	0.5%	2.3%	0.2	0.6%	3.0%	0.2	2.5%	2.3%	1.1	0.6%	4.1%	0.2
1-22	4.57%	.30%	12.7	4.81%	. 39%	12.2	4.82%	. 32%	14.9	-	.46%	-0.8
X	3.6%	2.6%	1.4	11.1%	2.8%	4.0	5.9%	1.6%	3.7	-2.4%	2.9%	-0.8
Heterogeneity?			$P = 1.0\times10^{-4}$			$P = 4.9\times10^{-8}$			$P = 2.9\times10^{-8}$			$P = 1.3\times10^{-4}$

Note: For each chromosome-specific jackknife analysis we use 20 blocks (rather than 100 as in the genome-wide analyses) to ensure that each block contains at least several Mb. P-values for heterogeneity across chromosomes are from a χ^2 test with 22 degrees of freedom.

The significant skews in the *D*-statistics cannot be due to modern-human-to-Neandertal gene flow

The analyses above are based on computing the statistic $D(H_1,H_2,H_3,Chimpanzee)$ using H_3 = Neandertal. However, additional information can be obtained by using other present-day humans in place of Neandertal. Table S48 reports $D(H_1,H_2,H_3,Chimpanzee)$ statistics for all 60 informative combinations of H_1, H_2 and H_3 in the Illumina present-day human data. (There are actually fewer than 60 independent *D* statistics here, as some of them are linear combinations of the others.)

Table S48: Symmetry tests for all combinations of 5 HGDP present-day humans and Neandertal

H_1	H_2	H_3	Number ABBA sites	Number BABA sites	D=(ABBA-BABA)/(ABBA+BABA) Value%	Std.err%	Z-score for D≠0	Interpretation (″ indicates same as previous row)
San	Yoruba	Neandertal	99,515	99,788	-0.1	± 0.3	**-0.4**	Neandertal equally close to Africans
French	Han	Neandertal	74,477	73,089	0.9	± 0.5	**1.7**	Neandertal equally close to non-Africans
French	Papuan	Neandertal	70,094	70,093	0	± 0.5	**0.0**	″
Han	Papuan	Neandertal	67,022	68,260	-0.9	± 0.6	**-1.4**	″
French	San	Neandertal	95,347	103,612	-4.2	± 0.5	-9.3	Neandertal gene flow with non-Africans
French	Yoruba	Neandertal	84,025	92,066	-4.6	± 0.4	-10.5	″
Han	San	Neandertal	94,029	103,590	-4.8	± 0.5	-9.9	″
Han	Yoruba	Neandertal	82,575	91,872	-5.3	± 0.5	-10.5	″
Papuan	San	Neandertal	90,059	97,088	-3.8	± 0.5	-7.0	″
Papuan	Yoruba	Neandertal	79,529	86,570	-4.2	± 0.6	-7.5	″
French	Neandertal	Han	74,477	364,200	-66	± 0.4	-148	Han closer to present-day humans than to Neand.
Papuan	Neandertal	Han	68,260	327,968	-66	± 0.5	-138	″
San	Neandertal	Han	103,590	293,958	-48	± 0.5	-105	″
Yoruba	Neandertal	Han	91,872	308,073	-54	± 0.5	-117	″
French	San	Han	86,228	188,799	-37	± 0.5	-76	Han closer to non-Africans than to Africans
French	Yoruba	Han	90,557	157,229	-27	± 0.5	-55	″
Papuan	San	Han	82,405	176,072	-36	± 0.6	-65	″
Papuan	Yoruba	Han	86,436	146,942	-26	± 0.6	-46	″
San	Yoruba	Han	142,189	108,897	13.3	± 0.5	29	Han closer to Yoruba than to San
French	Papuan	Han	101,021	101,445	-0.2	± 0.6	**-0.4**	Han equally close to French and Papuan
Han	Neandertal	French	73,089	364,200	-67	± 0.4	-164	French closer to present-day humans than to Neand.
Papuan	Neandertal	French	70,093	324,631	-65	± 0.4	-153	″
San	Neandertal	French	103,612	303,340	-49	± 0.4	-119	″
Yoruba	Neandertal	French	92,066	319,721	-55	± 0.4	-134	″
Han	San	French	88,271	188,799	-36	± 0.5	-72	French closer to non-Africans than to Africans
Han	Yoruba	French	93,140	157,229	-26	± 0.5	-50	″
Papuan	San	French	87,390	170,225	-32	± 0.5	-60	″
Papuan	Yoruba	French	92,602	141,892	-21	± 0.6	-37	″
San	Yoruba	French	146,387	111,576	13.5	± 0.5	30	French closer to Yoruba than to San
Han	Papuan	French	91,693	101,445	-5	± 0.6	-8.5	French closer to Han than to Papuan
Han	Neandertal	Papuan	67,022	327,968	-66	± 0.5	-138	Papuan closer to present-day humans than to Neand.
French	Neandertal	Papuan	70,094	324,631	-65	± 0.5	-134	″
San	Neandertal	Papuan	97,088	265,775	-47	± 0.5	-92	″
Yoruba	Neandertal	Papuan	86,570	277,960	-53	± 0.5	-103	″
French	San	Papuan	79,253	170,225	-37	± 0.5	-75	Papuan closer to non-Africans than to Africans
French	Yoruba	Papuan	82,610	141,892	-26	± 0.5	-51	″
Han	San	Papuan	75,849	176,072	-40	± 0.5	-79	″
Han	Yoruba	Papuan	78,898	146,942	-30	± 0.5	-55	″
San	Yoruba	Papuan	129,364	99,763	12.9	± 0.5	28	Papuan closer to Yoruba than to San
French	Han	Papuan	101,021	91,693	4.8	± 0.6	8.3	Papuan closer to Han than to French (!)
Han	Neandertal	Yoruba	82,575	308,073	-58	± 0.4	-147	Yoruba closer to present-day humans than to Neand.
French	Neandertal	Yoruba	84,025	319,721	-58	± 0.4	-156	″
Papuan	Neandertal	Yoruba	79,529	277,960	-56	± 0.4	-132	″
San	Neandertal	Yoruba	99,515	304,399	-51	± 0.4	-139	″
French	San	Yoruba	107,418	146,387	-15.4	± 0.4	-35	Yoruba closer to non-Africans than to San
Han	San	Yoruba	106,701	142,189	-14.3	± 0.5	-30	″
Papuan	San	Yoruba	104,863	129,364	-10.5	± 0.5	-22	″
Han	Papuan	Yoruba	78,898	86,436	-4.6	± 0.5	-9.0	Yoruba closer to Han+French than to Papuan (!)
French	Papuan	Yoruba	82,610	92,602	-5.7	± 0.5	-11.4	″
French	Han	Yoruba	90,557	93,140	-1.4	± 0.4	-3.1	Yoruba closer to French than to Han (!)
Han	Neandertal	San	94,029	293,958	-52	± 0.4	-133	San closer to present-day humans than to Neand.
French	Neandertal	San	95,347	303,340	-52	± 0.4	-132	″
Papuan	Neandertal	San	90,059	265,775	-49	± 0.4	-119	″
Yoruba	Neandertal	San	99,788	304,399	-51	± 0.4	-128	″
Han	Yoruba	San	106,701	108,897	-1	± 0.4	-2.6	San closer to Han+French than to Yoruba (!)
French	Yoruba	San	107,418	111,576	-1.9	± 0.4	-5.0	″
Han	Papuan	San	75,849	82,405	-4.1	± 0.5	-8.8	San closer to Han+French than to Yoruba (!)
French	Papuan	San	79,253	87,390	-4.9	± 0.4	-11.1	″
Papuan	Yoruba	San	104,863	99,763	2.5	± 0.4	5.9	San closer to Yoruba than to Papuan (!)
French	Han	San	86,228	88,271	-1.2	± 0.4	-2.8	San closer to French than to Han (!)

Note: Z-scores are calculated by using a weighted block jackknife over 5 Mb blocks (a naïve binomial test for a difference in the rates of ABBA and BABA counts is incorrect because of linkage disequilibrium). We highlight in **bold** |*Z*|-scores less than 2 standard deviations from 0.
! An exclamation point denotes a pattern that is not predicted under current models of human history.

To test whether Table S48 is consistent with entirely by Neandertal to modern human gene flow, we considered the hypothesis that the positive *D(African,non-African,Neandertal,Chimpanzee)* statistic is due to flow from a non-African-related modern human population OOA_1 into Neandertals. In this scenario, Neandertal sequence can be thought of as a mixture of a proportion α of OOA_1 ancestry and (1-α) of "*Outgroup*" ancestry:

$$Neandertal = \alpha OOA_1 + (1-\alpha)Outgroup \tag{S15.3}$$

Let OOA_2 denote a present-day human non-African population, and *AFR* denote a present-day African population. The expected value of $D(OOA_2,AFR,Neandertal,Chimpanzee)$ is then:

$$\begin{aligned} &E[D(OOA_2, AFR, Neandertal, Chimpanzee)] \\ &= \alpha E[D(OOA_2, AFR, OOA_1, Chimpanzee)] + (1-\alpha)E[D(OOA_2, AFR, Outgroup, Chimpanzee)] \\ &= \alpha E[D(OOA_2, AFR, OOA_1, Chimpanzee)] \end{aligned} \tag{S15.4}$$

The second term has an expected value of 0, since all present-day humans form a clade relative to the *Outgroup* component of Neandertal ancestry under our null hypothesis. Equation S15.4 allows us to estimate the mixture proportion α by a ratio that should be the same whether *AFR* = San or *AFR* = Yoruba assuming that the hypothesis of gene flow from a modern non-African related population to Neandertal is correct.

$$\alpha = \frac{E[D(OOA_2, San, Neandertal, Chimpanzee)]}{E[D(OOA_2, San, OOA_1, Chimpanzee)]} = \frac{E[D(OOA_2, Yoruba, Neandertal, Chimpanzee)]}{E[D(OOA_2, Yoruba, OOA_1, Chimpanzee)]} \tag{S15.5}$$

To test this expectation in our data, we defined the statistic:

$$C(OOA_2, OOA_1) \equiv \frac{D(OOA_2, Yoruba, Neandertal, Chimp)}{D(OOA_2, Yoruba, OOA_1, Chimp)} - \frac{D(OOA_2, San, Neandertal, Chimp)}{D(OOA_2, San, OOA_1, Chimp)} \tag{S15.6}$$

From Equation S15.5, we see that the expected value E[C(OOA_2,OOA_1)]=0. We computed C(OOA_2,OOA_1) for all six possible permutations of French, Han and Papuan to represent the non-African samples OOA_1 and OOA_2 (Table S49). This quantity is significantly different from zero ($3.7 < |Z| < 4.5$ standard deviations), rejecting the hypothesis that our results can be explained by gene flow from a modern human population related to non-Africans into Neandertal (or contamination from any present-day non-African population, which would have the same effect on this statistic).

While our finding that $C(OOA_2,OOA_1)$ is inconsistent with 0 rejects the hypothesis that the observed skews in the *D*-statistics can entirely be explained by modern human-to-Neandertal gene flow, it is important to recognize that these results do not preclude the possibility that a proportion of the putative gene flow was in that direction. Additional haplotypes-based analysis, however, shows that there is no evidence at all in our data for gene flow from modern humans into Neandertals (presented in SOM 16).

Table S49: Data are inconsistent with modern human-to-Neandertal gene flow

OOA_2	OOA_1	No. of stand. dev. from 0 of $C(OOA_2, OOA_1)$	$\frac{D(OOA_2, San, OOA_1, Chimp)}{D(OOA_2, Yoruba, OOA_1, Chimp)}$	$\frac{D(OOA_2, San, Neandertal_1, Chimp)}{D(OOA_2, Yoruba, Neandertal_1, Chimp)}$
French	Han	-4.0	1.54 ± 0.02	1.03 ± 0.10
French	Papuan	-4.0	1.53 ± 0.02	1.03 ± 0.10
Han	French	-4.5	1.57 ± 0.03	1.03 ± 0.09
Han	Papuan	-4.0	1.47 ± 0.02	1.03 ± 0.09
Papuan	French	-4.2	1.68 ± 0.03	1.00 ± 0.11
Papuan	Han	-3.7	1.55 ± 0.02	1.00 ± 0.11

Neandertal-to-modern-human gene flow is consistent with the data

To test the null hypothesis that gene flow occurred in the reverse direction, from Neandertals to non-Africans after the non-Africans / African divergence, we denote a non-African population *OOA* as a mixture between a Neandertal-related population N_1 and a population AFR_1 that is in a clade with present-day Africans:

$$OOA = (\alpha)N_1 + (1-\alpha)AFR_1 \quad \text{(S15.7)}$$

It is now of interest to calculate the expected value of $D(OOA,AFR_2,Neandertal,Chimpanzee)$, where AFR_2 is a present-day sub-Saharan African population that is different from AFR_1. This quantity partitions into two terms, one of which has an expected value of 0 since Neandertal is an outgroup to AFR_1 and AFR_2:

$$\begin{aligned} &E[D(OOA, AFR_2, Neandertal, Chimpanzee)] \\ &\quad = \alpha E[D(N_1, AFR_2, Neandertal, Chimpanzee)] + (1-\alpha)E[D(AFR_1, AFR_2, Neandertal, Chimpanzee)] \\ &\quad = \alpha E[D(N_1, AFR_2, Neandertal, Chimpanzee)] \end{aligned} \quad \text{(S15.8)}$$

The expected value of $D(N_1,AFR_2,Neandertal,Chimpanzee)$ is independent of which AFR_2 population we choose under the null hypothesis that present-day African populations do not inherit any Neandertal gene flow. Thus, for all 12 possible choices of populations OOA_i, OOA_k, AFR_j and AFR_l, we can compute $D(OOA_i,AFR_j,Neandertal,Chimpanzee)$-$D(OOA_k,AFR_l,Neandertal,Chimpanzee)$, and assess the number of standard deviations that this quantity is from 0 using the weighted block jackknife. We find that our data are consistent with Neandertal gene flow into the ancestors of non-Africans (Table S50). In particular, $|Z| \ll 2.4$ for all 12 comparisons, which is not significant after correcting for 12 hypotheses tested.

Table S50: Consistency of data with Neandertal-to-modern human gene flow

OOA_1	AFR_1	OOA_2	AFR_2	Number of std. dev. from 0 of $D(OOA_1,AFR_1,Neandertal,C) - D(OOA_2,AFR_2,Neandertal,C)$
French	San	French	Yoruba	0.3
French	San	Han	San	-1.3
French	San	Han	Yoruba	-0.8
French	Yoruba	Han	Yoruba	-1.4
French	San	Papuan	San	1.2
French	San	Papuan	Yoruba	1.0
French	Yoruba	Papuan	Yoruba	1.0
Han	San	Han	Yoruba	0.3
Han	San	Papuan	San	2.4
Han	San	Papuan	Yoruba	2.0
Han	Yoruba	Papuan	Yoruba	2.2
Papuan	San	Papuan	Yoruba	0.0

We conclude that the observed *D* statistics for many pairs of populations are consistent with Neandertal gene flow into the common ancestors of non-Africans, but are inconsistent with gene flow in the reverse direction.

Appendix 1: Sequencing errors are insufficient to explain the significantly non-zero *D* statistics

To test whether differential rates of sequencing error in the Illumina present-day human samples could explain the degree of greater genetic proximity of Neandertals to non-Africans than to Africans that we measure, we focused on the alignment of {French,San,Neandertal,Chimpanzee}, and designated each allele by a different letter (including multiallelic sites). Table S51 shows the counts for 834 million nucleotides that passed our data quality filters. The 'AAAA' entry corresponds to cases where all 4 bases agree. The 'ABBA' entry corresponds to cases where French=Chimpanzee, and San=Neandertal. The BCDA entry, with only 36 counts, corresponds to bases where all four samples disagree.

Our null model in the absence of a history of Neandertal-to-modern human gene flow is that (French, San) are phylogenetically symmetric with respect to (Neandertal, Chimpanzee). If the phylogeny is symmetric, and assuming that mutation rates have been constant on both the French and San lineages since their divergence, the difference in the rates of ABAA and BAAA counts (689,594 and 756,324 respectively) must be due to a difference in the error rate of the sequencing for French and San. Writing lower case letters for

the true pattern in the absence of sequencing error, *abaa* denotes the case where the San carries the allele not observed in the other samples. We then see that:

$$P(ABAA) \approx P(abaa) + P(aaaa)m_2(a \to B) \tag{S15.9}$$

where $m_2(a \to B)$ is the probability that a San base was sequenced in error. We ignore cases such as P(*bbaa*→*ABAA*) , where the true pattern is *bbaa* and the French base was incorrectly sequenced, because the total count of *P(BBAA)* is much less than *P(AAAA)*, and thus contributes negligibly to the expectation.

Table S51: Counts of all possible allele patterns in a French-San-Neandertal-Chimpanzee alignment

French-San-Neandertal-Chimpanzee	**Counts**
AAAA	818,322,920
AABA	5,827,247
ABAA	689,594
ABBA	95,347
ABCA	2,995
BAAA	756,324
BABA	103,612
BACA	3,544
BBAA	303,340
BBBA	8,156,936
BBCA	32,607
BCAA	972
BCBA	6,147
CBBA	6,264
BCDA	36

Our null hypothesis is that *P(abaa)-P(baaa) = 0* (assuming here that the French and San are a clade and that they have had the same mutation rate since their divergence). Thus:

$$\begin{aligned} &P(ABAA) - P(BAAA) \\ &\quad = P(abaa) + P(aaaa)[m_2(a \to B)] - P(baaa) - P(aaaa)[m_1(a \to B)] \\ &\quad \approx P(AAAA)\left[m_2(a \to B) - m_1(a \to B)\right] \end{aligned} \tag{S15.10}$$

where m_1 is the probability that a French base was read in error. We can then use our observed counts in Table S52 to compute the sequencing error rate difference between French and San averaged over all sites contributing to the counts. We find that the estimate is less than 1 in 10,000:

$$m_2(a \to B) - m_1(a \to B)_2 \approx \frac{P(ABAA) - P(BAAA)}{P(AAAA)} = -0.000082 \tag{S15.11}$$

To compute the expected contribution of this error to the *D* statistics, we focus on *ABBA* and *BABA* counts:

$$\begin{aligned} E\left[P(BABA)\right] &\approx P(baba) + p(aaba)[m_1(a \to b)] + p(bbba)\left[m_2(b \to a)\right] + p(baaa)[m_3(a \to b)] \\ E\left[P(ABBA)\right] &\approx P(abba) + p(bbba)[m_1(b \to a)] + p(aaba)\left[m_2(a \to b)\right] + p(abaa)[m_3(a \to b)] \end{aligned} \tag{S15.12}$$

We are interested in E[P(ABBA-BABA)]. Under the null hypothesis that French and San are a clade, *P(baba) = P(abba)* and *P(abaa) = P(baaa)*. Further, allele label does not matter so *m(a→b) =m(b→a)*. Thus:

$$
\begin{aligned}
&E\left[P(ABBA)-P(BABA)\right] \\
&\quad = P(abba)+P(bbba)\left[m_1(b\to a)\right]+P(aaba)\left[m_2(a\to b)\right]+P(abaa)[m_3(a\to b)] \\
&\qquad -P(baba)-P(aaba)[m_1(a\to b)]-P(bbba)\left[m_2(b\to a)\right]-P(baaa)[m_3(a\to b)] \\
&\quad = \left[P(aaba)-P(bbba)\right]\times\left[m_2(a\to b)-m_1(a\to b)\right] \\
&\quad \approx \left[5{,}827{,}247-8{,}156{,}936\right]\times\left[-0.000082\right]=190
\end{aligned}
\qquad \text{(S15.13)}
$$

Thus, the numerator of the statistic *D(French,San,Neandertal,Chimpanzee)* is only expected to be inaccurate by 190 counts due to sequencing error. This translates to a bias in the *D*-statistic (without gene flow) of:

$$
\begin{aligned}
&Expected\left[D(French,San,Neandertal,Chimpanzee)\right] \\
&\quad = E\left[\frac{P(ABBA)-P(BABA)}{P(ABBA)+P(BABA)}\right]\approx\frac{E\left[P(ABBA)-P(BABA)\right]}{E\left[P(ABBA)+P(BABA)\right]}=\frac{190}{95{,}347+103{,}612}=0.00095
\end{aligned}
\qquad \text{(S15.14)}
$$

We note that the observed value of *D(French,San,Neandertal,Chimpanzee)* = -0.0415 from the counts in Table S51. This is many standard deviations away from the skew of 0.00095 expected from sequencing error. Thus, the difference in ABBA and BABA counts for San and French as a representative pair of non-African and African populations cannot be due to sequencing error.

Appendix 2: Comparison of HGDP sequencing and genotyping further supports results

The 5 present-day human samples that we studied were all from the CEPH-HGDP panel, and all were therefore genotyped at ~650,000 SNP sites on an Illumina 660Y array (*S91*). By comparing our allele calls from sequencing to these genotypes, we could assess the quality of our data and search for biases. We note that we only analyzed HGDP genotyping data over sections of the genome that were not inverted between humans and chimpanzees. This avoided some bioinformatic difficulties, and we do not expect our results to be affected by removing the regions of the genome affected by these inversions.

A small bias toward matching the chimpanzee reference in HGDP present-day human sequencing data

We first restricted our analysis to autosomal SNPs that were heterozygous in the SNP array data. If there is no differential rate in sequencing error across samples, it should be equally probable that our Illumina sequencing matches or mismatches chimpanzees at sites that were also genotyped. Table S52 shows that there is a significant but small bias for the allele of the heterozygous pair to match the chimpanzee, probably reflecting bias due to aligning of short reads to the chimpanzee reference genome sequence. However, there is no reason *a priori* that this should cause a skew in the *D*-statistics.

Table S52: Probability of sequencing data matching chimpanzee at heterozygous sites

	Matching chimpanzee	**Mismatching chimpanzee**	**% matching chimpanzee**
French	72,443	69,735	51.0%
Han	65,147	62,228	51.2%
Papuan	49,989	45,420	52.4%
San	63,179	60,474	51.1%
Yoruba	72,487	68,877	51.3%

Error rate in the sequencing data from 5 present-day humans of <1/1,000

To place an upper bound on the error rate in the Illumina present-day human data, we restricted our analysis to homozygous sites in the HGDP genotyping data. Ideally there would be no mismatches between the genotyping and sequencing data at these sites. In fact, we obtain a discrepancy rate of about 1/1,000 per base. Some of this must be due to HGDP genotyping error, and hence this is an upper bound on sequencing error.

Table S53: Probability of sequencing data matching genotyping data at homozygous sites

	Matching	Mismatching	% mismatching
French	327,298	324	0.10%
Han	337,203	344	0.10%
Papuan	344,140	323	0.09%
San	385,090	410	0.11%
Yoruba	339,810	309	0.09%

Study of the *D* statistic skews in the HGDP genotyping data

We directly computed the *D* statistics in the genotyping data, examining all autosomal SNPs from the HGDP in which our randomly chosen Neandertal base differed from chimpanzee. For any pair of present-day human samples H_1, H_2, we then computed the probability that the 4 alleles (h_1, h_2, *Neandertal, Chimpanzee*) from our sequencing data are ABBA vs. BABA, and compared it to our genotyping data (Table S54). Adding the counts for all the (non-African,African) pairs in Table S54, and computing a standard error by a block jackknife, we found that $D = -0.024 \pm 0.005$ ($Z = -4.87$) , which is lower than in our main analysis of the sequencing data, but still highly significant. The absolute values of the *D* statistics should not be taken as unbiased, as the SNPs in HGDP are affected by ascertainment bias. However, we believe that it is unlikely that differences in *D* between genotyping and sequencing would be affected by ascertainment bias.

To perform a formal test for a difference between the genotyping and sequencing results in a way that maximizes the information from our limited amount of data, we computed the deviation of the sequencing data from the 'gold standard' expectation from the genotyping data. Specifically, we computed the deviation X averaging the numerator and denominator across the genome:

$$X = \frac{\sum_i \left(ABBA_i^{\,sequencing} - BABA_i^{\,sequencing} \right) - \sum_i \left(ABBA_i^{\,genotyping} - BABA_i^{\,genotyping} \right)}{\sum_i \left(ABBA_i^{\,genotyping} + BABA_i^{\,genotyping} \right)} \qquad \text{(S15.15)}$$

We carried out a weighted block jackknife in 5 Mb blocks using the ABBA+BABA values from the HGDP genotyping as weights for each block. We find that $X = -0.4 \pm 0.3\%$. A 2-sided test gives a P-value of 0.18 against the null hypothesis that $X = 0$, indicating no evidence of bias.

Table S54: Symmetry Test statistics comparing genotyping and sequencing data

		Expected values HGDP genotyping			Observed values from our sequencing		
S_1	S_2	ABBA	BABA	*D*	ABBA	BABA	*D*
San	Yoruba	8,059	8,227	-1.0%	7,979	8,183	-1.3%
French	Han	6,967	6,987	-0.1%	6,866	7,039	-1.2%
French	Papuan	6,467	6,472	0.0%	6,301	6,567	-2.1%
Han	Papuan	6,007	6,004	0.0%	5,921	5,890	0.3%
French	San	8,477	8,840	-2.1%	8,374	8,890	-3.0%
French	Yoruba	7,861	8,344	-3.0%	7,658	8,405	-4.7%
Han	San	8,201	8,557	-2.1%	8,138	8,489	-2.1%
Han	Yoruba	7,602	8,059	-2.9%	7,516	7,995	-3.1%
Papuan	San	7,658	7,946	-1.8%	7,643	7,876	-1.5%
Papuan	Yoruba	7,121	7,481	-2.5%	7,136	7,494	-2.4%

Supplementary Online Materials 16

Haplotypes of Neandertal ancestry in present-day non-Africans

Heng Li, Markus Hsi-Yang Fritz, Nancy Hansen, James Mullikin, Ewan Birney and David Reich*

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

Overview

In this section, we present an independent line of evidence for Neandertals being unusually closely related to the ancestors of non-Africans. Our idea is to take advantage of the very different prediction about the pattern of genetic divergence in the case of gene flow and non-gene-flow scenarios.

Gene flow: In the case of Neandertal-to-non-African human gene flow, present-day non-African haplotypes that are observed to have an unusually low divergence from Neandertal are also predicted to have an unusually high divergence with most other present-day humans.

No gene flow: If all present-day humans descend from a homogeneous ancestral population that diverged at earlier time from Neandertals, present-day non-African haplotypes with an unusually low divergence between from Neandertal are predicted to tend to have low divergence from other present-day humans. The reason for this is that in the absence of gene flow, low divergence is expected to be due to a stochastically short gene tree, a low mutation rate since the split from chimpanzee, or a high human-chimpanzee time divergence that affects the normalization. All these effects will result in low estimates of divergence for all pairs of samples, which will not be restricted to Neandertals.

In what follows, we capitalize on these qualitatively different predictions to search for an unambiguous signal of Neandertals being genetically closer to non-Africans than to Africans: (a) an excess of haplotypes of low nucleotide divergence to Neandertals in non-Africans compared with Africans, and (b) a concomitant observation that these haplotypes tend to have high divergence from other present-day humans.

An excess of low divergence haplotypes to Neandertals in non-Africans (compared to Africans)

If there has been gene flow between Neandertals and the ancestors of non-Africans, then non-Africans are expected to harbor an excess of low divergence segments compared with Africans.

To identify haploid present-day human sequences of known ancestry in which to implement this test, we took advantage of the fact that the human reference sequence is haploid over substantial scales, because it is comprised of a tiling-path of Bacterial Artificial Chromosomes (BACs) that are of typical size 50-150 kb (*S92*). About two thirds of the human genome reference sequence is comprised of clones from a single individual (RPCI-11), which we have determined is likely to be an African American (*S93*). We have now carried out further analysis to show that this individual has about 50% European and 50% sub-Saharan African ancestry, and that we can identify multi-megabase segments (admixture linkage disequilibrium blocks) where RPCI-11 is confidently inferred to have both their chromosomes being of European or African ancestry (SOM16 Appendix). In such multi-megabase sections of the genome of RPCI-11, there is essentially no ambiguity in clone ancestry assignment. A particular advantage of this procedure is that it provides us with a list of European and African ancestry clones that are ascertained in the same way, in which we are able to carry out a rigorous comparison to Neandertal (Supplemental Data File). After restricting to clones of ≥50 kb for which at least 40% of nucleotides are covered in present-day human, Neandertal and chimpanzee, we had 2,825 RPCI-11 clones of confident West African ancestry and 2,797 RPCI-11 clones of confident European ancestry for our analysis.

To compare our Neandertal data to present-day humans, we pooled data from the three Vindija Neandertal bones. This pooling meant that our Neandertal data was effectively hexaploid instead of haploid, which weakens our power to detect low divergence regions since we are averaging over the divergence between the human reference sequence and six Neandertal haplotypes. However, it was necessary in order to increase the amount of coverage we had from our limited amount of Neandertal sequence data. At sites where we had coverage from multiple Neandertal reads, we randomly sampled one allele.

Within each clone *i*, we tabulated three classes of sites: "n^C_i" Chimpanzee-only, "n^H_i" Human-only, and "n^N_i" Neandertal-only. The ratio $(n^H_i+n^N_i)/(n^H_i+n^C_i)$ estimates the divergence time ratio between human and Neandertal at the locus, but because of the high rate of error and misincorporation affecting Neandertal sequence, we focused on only using substitutions on the human side of the tree, and thus used the statistic $2n^H_i/(n^H_i+n^C_i)$ (as discussed in SOM 10).

We ordered all the clones of each ancestry by this statistic, and plotted them in Figure S39a normalized by the genome average. We observe an excess of European clones of very low divergence to Neandertal compared with African clones. For example, if we count the proportion of clones of each ancestry with a ratio less than 30% of the genome-wide average, there are 23 European vs. 10 African clones, which is a moderately significant excess (P=0.016 by a one-sided Fisher's exact test). However, the significance depends on the threshold we choose, and hence this analysis does not provide clear evidence of gene flow.

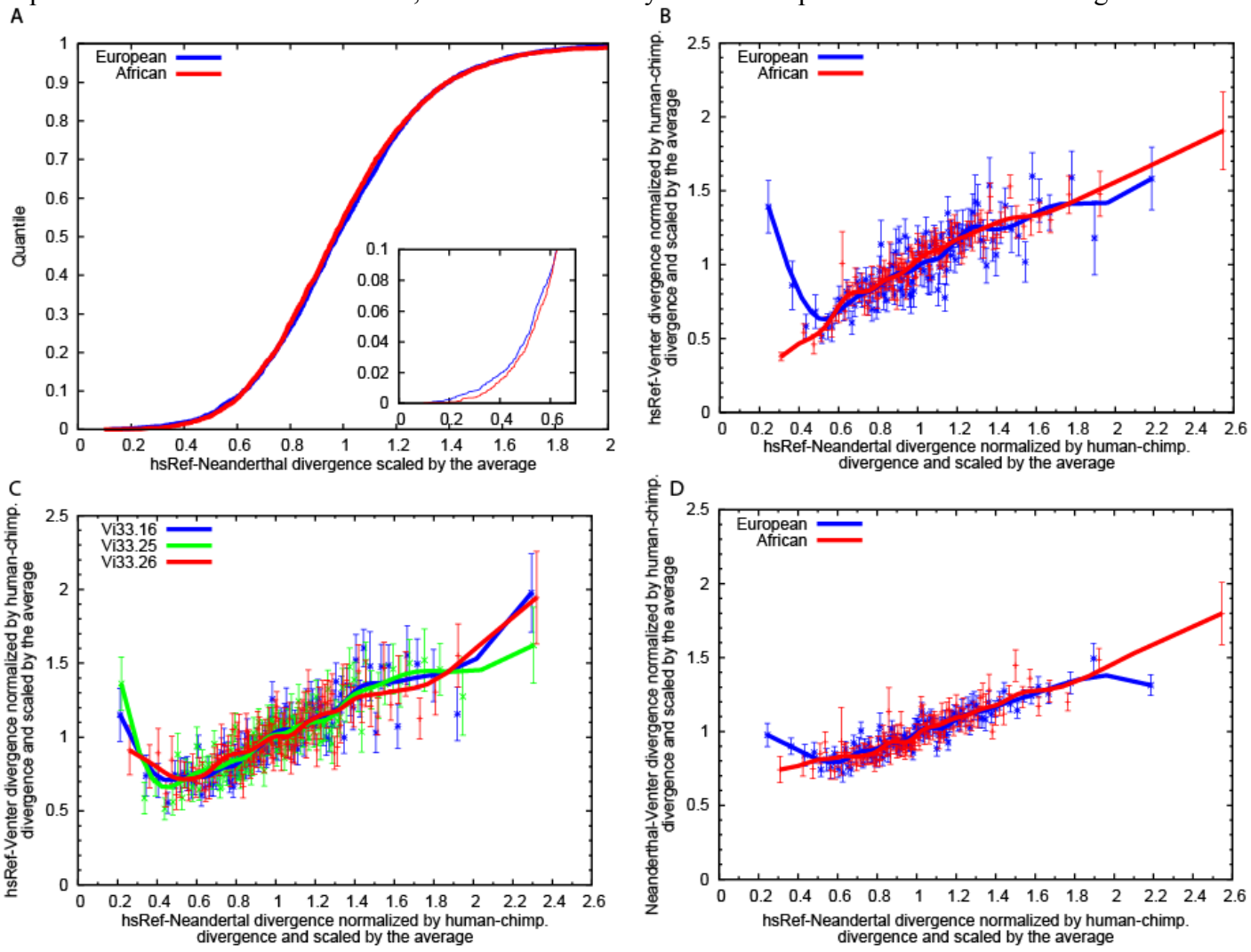


Figure S39: We examined 2,825 segments in the human reference genome that are of African ancestry and 2,797 that are of European ancestry. (A) Rank-ordering by their divergence from Neandertal, we observe a suggestive excess of regions of low divergence between Neandertal and European segments compared with Neandertal and African segments (P=0.016). (b,c) European segments with few differences to Neandertal also tend to have many differences to other present-day humans, a non-monotonic pattern that is consistent across bones and is highly statistically significant ($P<10^{-9}$). The same pattern is not seen in present-day African segments, and hence there is no evidence for Neandertal gene flow into the ancestors of sub-Saharan Africans. (d) An important observation that rules out substantial modern human gene flow into the ancestors of the Neandertals (or contamination) is that at segments of low Neandertal-human reference sequence divergence, we do not tend to observe extraordinarily low divergence of Neandertals and other modern humans.

Theoretical predictions at segments of low Neandertal-human divergence for different demographies
At loci where Neandertal-human reference sequence divergence is very low, there is expected to be a characteristic pattern of divergence to other present-day humans, which will be qualitatively different depending on the demography, allowing us to distinguish among these scenarios (see also Table S55).

(A) *Prediction without gene flow:* At clones showing unusually low divergence between the human reference sequence and Neandertal, the divergence of both the human reference and Neandertal from other humans is expected to be slightly reduced because there is an ascertainment bias toward loci with short hominid gene trees, high human-chimpanzee divergence, or mutation rate slowdown in hominids relative to chimpanzees.

(B) *Prediction of Neandertal-to-modern human gene flow:* In the context of Neandertal-to-modern human gene flow, segments of the human reference sequence that have unusually low divergence from Neandertal are expected to be Neandertal-like. We therefore expect the divergence of the human reference sequence from other humans at these loci to be very high.

(C) *Prediction of modern human-to-Neandertal gene flow:* In the context of modern human-to-Neandertal gene flow, segments of the Neandertal sequence that have very low divergence from present-day humans are expected to be closely related to other present-day humans. We therefore expect very low divergence of the Neandertal from other humans at this locus (not just from the reference sequence).

We validated these expectations by simulating each scenario using Hudson's "ms" software (*S29*).

Table S55: Expected pattern at loci where the human reference and Neandertal have low divergence

	Divergence of human reference genome sequence from other present-day humans	**Divergence of Neandertal from present-day humans other than the reference sequence**
	↓	↓
Scenario A **No gene flow**	Slightly reduced relative to average (low divergence between human reference and Neandertal may reflect a short human tree, high chimp divergence time, or hominin mutation rate slowdown)	Slightly reduced relative to average (low divergence between Neandertal and other modern humans may reflect a short human tree, high chimp divergence time, or hominin mutation rate slowdown)
	↑↑	↑
Scenario B **Neandertal-to-modern human gene flow**	Greatly increased relative to average (the divergence of the human reference from other humans at this locus is expected to be very high: typical of Neandertal-human divergence)	Slight increase relative to expectation (loci with low Neandertal-present-day human divergence will be genuinely due to gene flow, and hence not affected by the same ascertainment bias as Scenario A)
	↓	↓↓
Scenario C **Modern human-to-Neandertal gene flow**	Slightly reduced relative to average (same reason as Scenario A)	Greatly reduced relative to average (since Neandertal is similar to present-day humans, it will have a reduced divergence to other present-day humans)
	↓	↓↓
Scenario D **No gene flow, non-African contamination**	Slightly reduced relative to average (same reason as Scenario A)	Greatly reduced relative to average (since Neandertal has modern human ancestry at this locus, it will have a reduced divergence to other present-day humans)

Further evidence that loci of low European-Neandertal divergence are due to gene flow
To test the predictions of the different gene flow models against real data, we computed the divergence of each clone from not only to the Neandertal data, but also to the diploid genome sequence of Dr. Craig Venter (*S26*). We find a clear signal of the pattern expected due to gene flow from Neandertals into the ancestors of present-day non-Africans, but no evidence of other gene flow.

African segments provide no evidence for gene flow: We rank-ordered "African" segments of the human reference genome sequence by their divergence from Neandertal, grouped them into 100 bins, and calculated the divergence of each bin from the Venter sequence. We found that divergence from other

present-day humans increases monotonically with the divergence from Neandertal (Figure S39b), consistent with no gene flow into African ancestors.

European segments provide powerful evidence for gene flow: We rank-ordered "European" segments of the human reference genome sequence by their divergence from Neandertal, and grouped them into 100 bins. We find that divergence to the Venter sequence shows an initial rise to 1.4-times the genome-wide average, followed by a fall to around 0.65 before rising again. This non-monotonic behavior is only expected for a scenario of Neandertal-to-modern-human gene flow (Table S55), and is significant based on an analysis where we test for an elevation of the lowest bin vs. the genome average ($P<10^{-9}$). The finding is replicated for each individual Neandertal bone (Figure S39c).

A remarkable feature of our results is that simply by restricting to segments of the human reference sequence with unusually low divergence to Neandertal and unusually high divergence to other present-day humans, we can identify segments of the genome that are highly likely to contain Neandertal genetic material. As shown in Figure 5b, if we focus on RPCI-11 segments that have a divergence from Neandertal of <60% of the genome-wide average combined with a divergence from Venter of >150% of the genome average, there are 30 Europeans segments and 2 African segments ($P<10^{-6}$ for enrichment). The posterior probability of Neandertal ancestry at these segments is very high, although it is not 100% certain. It should be possible to further increase the precision of our identification of Neandertal segments in the ancestry of present-day humans by (a) using more Neandertal data once more sequencing data are available; (b) averaging divergence to a larger number of present-day humans; and (c) using a method like a Hidden Markov Model that screens for subsections of non-African sequence where the signal is strongest.

Novel findings compared with our other tests of gene flow

This analysis provides an independent line of evidence for gene flow compared with SOM 15 and SOM 17, and also produces two insights that are not obtained in those other analyses.

(A) No evidence for non-African-to-Neandertal gene flow. If there had been substantial gene flow from the ancestors of non-Africans into Neandertals (or contamination), then at loci of very low divergence between Neandertals and European segments of the human reference sequence, we would expect to see unusually low divergence to other non-Africans (Table S55). However, we see no such signal in Figure S39d. Thus, a test for gene flow in the reverse direction (into Neandertal) provides no evidence of gene flow.

(B) No evidence for Neandertal-to-African gene flow: Our analysis shows that there is no evidence of gene flow between Neandertals and the ancestors of present-day sub-Saharan Africans. The evidence for this is the fact that the African curves in Figures 39b (sensitive to Neandertal-to-African gene flow) and Figure S39d (sensitive to African-to-Neandertal gene flow) are monotonic, providing no evidence of gene flow.

A smaller amount of gene flow of type (A) or (B) that falls below the limits of our resolution cannot be ruled out by our analysis. However, our results show that the magnitude of such gene flow is likely to be much less than the signal of Neandertal-to-non-African gene flow in Figure S39b,c, which we estimate in SOM 18 accounts for 1.3-2.7% of the genetic material in the ancestry of present-day non-Africans.

Replication of the findings in data from the 1000 Genomes Project

To replicate these findings in an independent present-day human data set, we used data from two mother-father-child trios from the 1000 Genomes Project of European (CEU) and West African (YRI) ancestry. The children from these trios have been sequenced to high coverage using a variety of sequencing technologies, and the parents have been sequenced at lower coverage. We downloaded the data from ftp://share.sph.umich.edu/1000genomes/pilot2/trioAware/highQualityVariants on October 5th, 2009. After excluding sites from analysis that were heterozygous in all three members of a trio making it impossible to infer phase based solely on the data from the trios, we had two phased haploid genomes from the CEU child and two phased haploid genomes from the YRI child.

Analysis of the 1000 Genomes phased data provided similar results to what we found in analyses of the human reference genome sequence. Focusing on YRI haploid data in sliding 100 kb windows across the genome, we found a monotonic relationship between the divergence of the Neandertal to these segments and the YRI-YRI divergence over the same window. In contrast, for CEU, windows of very low divergence to Neandertal tended to have very high CEU-CEU divergence, followed by a non-monotonic behavior as Neandertal-CEU divergence became higher. These patterns are qualitatively similar to Figure S51.

We do not report the details of the 1000 Genomes data analysis here, both because the 1000 Genomes Data are not fully curated, and because we were concerned that the data analysis would be complicated by the removal of the triply heterozygous sites from analysis. Nevertheless, we were encouraged with the qualitative consistency of the results. It is worth pointing out that there is substantially more haploid sequence data available from the 1000 Genomes haploids than from the human reference sequence clones, in principle permitting a more powerful analysis. By analyzing this signal in detail in the 1000 Genomes data, it should be possible to detect more subtle patterns and make more refined inferences about the history of Neandertal-to-modern human gene flow than we did with the BAC clone analysis.

Appendix: Identifying the ancestry of segments of the human genome reference sequence

To compare Neandertal to present-day human haplotypes for the purpose of population genetic analysis, we needed to have long haploid sequences from present-day humans that were of known ancestry. To identify such segments, we took advantage of the fact that the human reference sequence is haploid over scales of tens of kilobases, because it is comprised of a tiling-path of Bacterial Artificial Chromosomes (BACs) or other clone types that are of typical size 50-150 kb (*S92*). We do not know of any other substantial source of high quality human haploid sequences of the requisite size.

Determining the ancestries of the libraries in the human genome reference sequence using HAPMIX

It is crucial to know the 'ancestry' of a clone to use it in a meaningful population genetic analysis. In what follows, we define 'ancestry' as the geographic region in which a clone's ancestor lived 1,000 years ago, inferred based on its genetic proximity to other individuals from that region today. This definition allows us to classify clones from Chinese Americans as "East Asian," from European Americans as "European", and from African Americans as either "West African" or "European".

To identify the ancestries of the libraries comprising most of the human genome reference sequence, we used a list of 26,558 clones tiling the great majority of the genome, most of which we were able to assign to a library of origin. Restricting to the autosomes, we identified 21,156 clones that seemed to fall into 9 libraries based on the naming scheme: CTA (n=199), CTB (n=356), CTC (n=452), CTD (n=1,426), RPCI-1 (n=740), RPCI-3 (n=456), RPCI-4 (n=716), RPCI-5 (n=802) and RPCI-11 (n=16,009). (In a subsequent re-examination, we identified additional clones that we likely could have classified into libraries, including 953 from RPCI-11, 632 from RPCI-1, and 490 from another library RPCI-13.) The median span of the 21,156 clones we analyzed was 112 kb, and 80% are >50kb in size. About 2/3 came from a single library, RPCI-11.

To make an inference about the ancestry of each clone, we used the HAPMIX software, which makes inferences about the probable ancestry of any haploid (or diploid) segment of DNA based on modeling the observed alleles over that segment as mixtures of two panels of samples that are assumed to represent the ancestral populations (*S35*). HAPMIX then calculates the posterior probability that the clones's haplotype is drawn from each of these proposed ancestral populations. We ran HAPMIX using HapMap YRI and CEU, and HapMap CHB+JPT=ASN and CEU, as proposed ancestral populations. For all runs, we assumed a prior probability of a 50-50% mixture of ancestry from the two populations, and no possibility of recent recombination within the clone.

We ran HAPMIX on clones for which there were a substantial number of SNPs genotyped in HapMap. For many clones, HAPMIX provided a confident assignment of one ancestry over another, although there were

also clones for which the ancestry assignment was less confident, reflecting various levels of haplotype frequency differentiation across the genome (results by clone are given in Supplemental Data File). Examining the ensemble of clones from each of the nine libraries, we find that they fall into three clusters (Figure S40a), which we hypothesize correspond to three different ancestries:

1. RPCI-11 is an African American: RPCI-11, the individual who contributed most of the human genome reference sequence, is consistent with having African American ancestry, with 42% of the clones of confident West African ancestry and 42% of the clones of confident European ancestry, and the ancestry of the remaining clones less confidently inferred. The finding of likely African American ancestry for RPCI-11 was previously reported in a study of the ancestry of RPCI-11 clones spanning the Duffy blood group locus (*S93*), and here we confirm this finding, and also expand the inference to the whole genome.

2. CTD is an East Asian: The majority of clones from CTD, the second largest library in its contribution to the human genome sequence, is likely an East Asian. In a HAPMIX analysis with CEU (European) – CHB+JPT (East Asian) as the proposed ancestral populations, the majority of clones are of confident East Asian origin, and there is no secondary mode of confident European ancestry, as might be expected from a Latino or South Asian individual.

3. The remaining 7 libraries are European: The remaining libraries (CTA, CTB, CTC, RPCI-1, RPCI-3, RPCI-4 and RPCI-5) are inferred to be of European ancestry, since they all have consistent distributions of inferred clone ancestries, with the majority of clones of confident European ancestry in both our HAPMIX analyses and no secondary modes.

Based on these analyses summarized in Figure S40, we classified all the clones from CTD as being of East Asian ancestry, and all the clones from CTA, CTB, CTC, RPCI-1, RPCI-3, RPCI-4 and RPCI-5 as being of European ancestry ('Confident Ancestry (>99%)' column in Supplemental Data File).

Figure S40: HAPMIX inferences of ancestry for ≥50kb clones in the reference sequence. (a) With CEU and YRI as the ancestral populations, we find three patterns, suggesting three ancestries. (b) With CEU and CHB+JPT as the proposed ancestral populations, CTD and the other 7 non-RPCI libraries remain distinct. (c) A histogram of ancestry inferences for RPCI-11 clone shows that it has large numbers of clones with >90% confident African (42%) and European (42%) ancestry, consistent with African American origin, whereas there is no peak of clones of likely African origin in the other libraries. (d) The CEU-CHB+JPT analysis shows that all libraries have their mode at >90% confidence of either one or the other ancestry. CTD seems likely to be East Asian, and the other seven seem likely to be European. We do not see libraries with two peaks, as might be expected for a Latino or South Asian individual.

We can assign West African or European ancestry to 41% of clones in RPCI-11 with >99% confidence
RPCI-11, the library that comprises about 2/3 of the human genome reference sequence, is of crucial value for our analyses, since it is the only library that harbors substantial amounts of African sequence (Figure S40). However, the assignment of ancestry to the clones from this individual required more analysis than for the 8 other libraries, since this individual harbors substantial numbers of clones of likely European as well as African ancestry, and hence all clones cannot simply be assigned to one ancestry from their library membership. We considered the strategy of restricting analysis of RPCI-11 to the clones from RPCI-11 where we were able to obtain a >95% confident ancestry inference from HAPMIX. However, we were concerned that this strategy would result in biased inferences, as restricting only to clones where we could confidently infer ancestry based on genetic data from the clone itself would restrict to haplotypes with high frequency differentiation across populations. It seemed plausible to us that these haplotypes might not be representative of the average pattern of genetic differentiation to other humans.

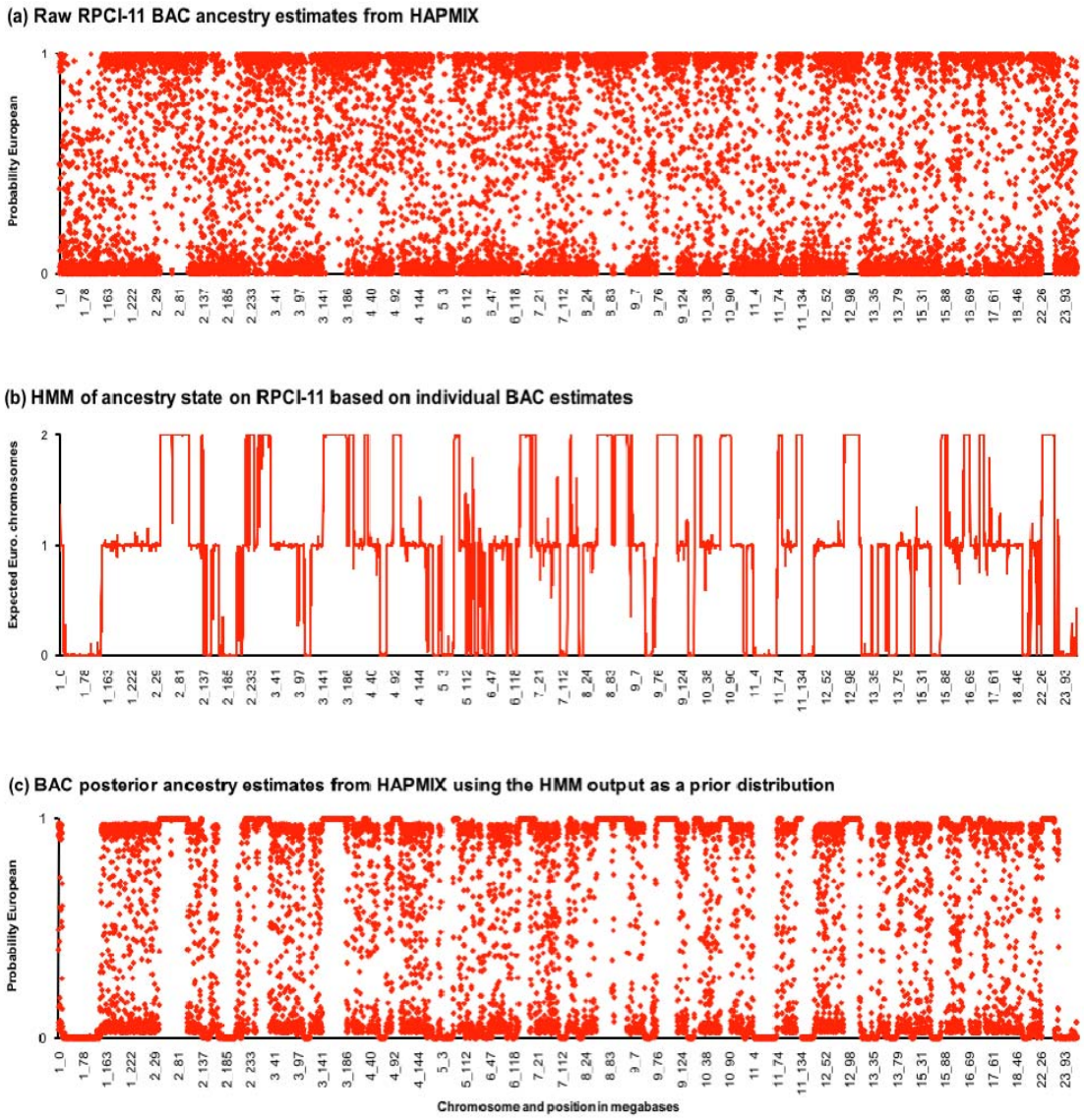


Figure S41: Local ancestry estimates from HAPMIX for RPCI-11, the African American sample that contributes about 2/3 or the human reference sequence. (a) HAPMIX estimates by individual clone. (b) We used a Hidden Markov Model (HMM) to integrate information over many clones to assign loci of 0, 1 or 2 European origin chromosomes. (c) The product of the HMM inference and the local clone ancestry estimate. We only use clones with >99% confidence of either West African or European ancestry, which excludes loci where African Americans are inferred to have both African and European ancestry.

To obtain ancestry inferences of clones from RPCI-11 that were not biased by the genetic patterns present in the clone itself, we took advantage of the fact that African Americans are known to harbor long contiguous segments of European or African ancestry that are typically tens of megabases, about two orders of magnitude longer than the individual clones (*S94*). Figure S41 confirms that this pattern is present in RPCI-

11 by plotting the positions of individual clones against their posterior estimate of probability of European ancestry from HAPMIX. We find three classes of multi-megabase loci:

1. Loci where the great majority of clones are of confident West African ancestry, reflecting a locus where RPCI-11 likely has entirely African ancestry from both parents.
2. Loci where the great majority of clones are of confident European ancestry, reflecting a locus where RPCI-11 likely has entirely European ancestry inherited from both parents.
3. Loci containing substantial proportions of clones of confident West African as well as substantial proportions of clones of confident European ancestry. This is what is expected if RPCI-11 inherited one chromosome each of West African and European ancestry. The fact that there are few clones with intermediate ancestry in these segments reflects the fact that the clones are haploid (sampling only one of the two ancestral chromosomes) and so are usually all of one ancestry or the other.

We wrote a Hidden Markov Model (HMM) that uses the probability of European ancestry estimates from HAPMIX as input, and then provides a posterior estimate of local ancestry by integrating information from many neighboring clones, assuming that nearby loci have the same ancestry state. Briefly, the HMM considers three possible states, in which the true ancestries are either 100% West African, 50% European and 50% West African, or 100% European. The prior probabilities of these states are ¼, ½, and ¼ respectively. Transitions between states are modeled as following an exponential process in which either chromosome had an expected length of 20 Mb before switching ancestries. The 'observed' data used in the HMM are the HAPMIX inferences of probability of European ancestry for each clone (P_i). To be conservative (to prevent us from being overconfident about the estimates), we 'flattened' these ancestry estimates by 5%, using the equation $F_i = 5\% \times (0.5) + 95\% \times (P_i)$. This had the effect that even if HAPMIX was confident of ancestry at a clone, we treated the inference as at most 97.5% confident, to account for the possibility that HAPMIX was erroneously inferring ancestry due to not having sampled the relevant haplotype from the ancestral CEU and YRI population panels. Details of the HMM are available from DR on request.

Figure S41b shows the output of the HMM, indicating many regions of confident European or African ancestry. The HMM identifies 21% of the human genome reference sequence as being from RPCI-11 clones that are of >99% confident African ancestry, and 20% as being of >99% confidence European ancestry. We only compared Neandertal to RPCI-11 data from the segments that are inferred to be confidently of all African or all European ancestry, because at these loci we expect that the particular haplotype structure within the clone itself will not substantially bias the ancestry inference. Figure S41c shows the product of the HMM posterior estimate of each ancestry and the ancestry estimate from HAPMIX using only the data from the clone itself. In analyses where we compared segments of the genome that were entirely of African or entirely of European ancestry, we restricted to clones from RPCI-11 that were >99% confident of all of one ancestry after this multiplication, allowing us to exclude the few clones with a highly discrepant HAPMIX and HMM ancestry inference, and loci that are heterozygous for African and European ancestry.

Supplementary Online Materials 17

Non-African haplotypes match Neandertal at an unusual rate

Weiwei Zhai, David Reich and Rasmus Nielsen*

* To whom correspondence should be addressed (rasmus_nielsen@berkley.edu)

In what follows we describe a scan for "candidate Neandertal gene flow regions" in present-day non-Africans: loci at which there is a much deeper coalescence time in chromosomes of non-African than sub-Saharan African ancestry. The strength of this analysis is that the identification of candidate regions is entirely based on studying present-day human variation, using an algorithm that is blinded to data from our Neandertal genome sequencing. We then "lift the curtain" on the Neandertal sequence and compare the alleles that are candidates for gene flow to Neandertal. If Neandertal matches the alleles at a rate that is greater than 50%, the only possible explanation is gene flow from Neandertal.

Methods

Our screen for candidate Neandertal gene flow regions relies on searching for 50 kilobase (kb) loci in which the inferred Time since the Most Recent Common Ancestor (TMRCA) is much older in a non-African (OOA, or out-of-Africa) than in a West African popualtion. We can then test whether the deeply diverged haplotype that is unique to non-Africans matches Neandertal at an unexpected rate, a phenomenon that is only expected to occur in the case of Neandertal gene flow.

Candidate regions

We analyzed "Perlegen Class A SNPs" that were discovered with a uniform ascertainment scheme genome-wide (*S95*), and were then genotyped in 24 European Americans (CEU), 24 East Asians (ASN), and 23 African Americans (AFR). We used the Perlegen data set rather than HapMap, because the SNP ascertainment does not have regional variation like HapMap. We note that a potential concern about using Perlegen data is that the AFR samples consist of African Americans, who have a small percentage of European ancestry making it more difficult to discover regions where the TMRCA in non-Africans is much higher than in Africans. While the presence of European lineages in African Americans may weaken statistical power to detect regions within higher non-African than African TMRCA, it should not cause false-positives, as we verified by simulations (below).

In each 50kb sliding window in the genome, we estimated the TMRCA for the African (AFR) and out-of-Africa (OOA) population using a tree-based approach that estimates the average number of SNPs since the root. Specifically, we used a UPGMA algorithm (*S96*) to construct a tree for the OOA and AFR samples separately assuming no within-locus recombination. This provides an estimate of the time to the root for the out-of-Africa (OOA = CEU+ASN, CEU or ASN) and African (AFR) populations ($\hat{T}_{OOA}$ and $\hat{T}_{AFR}$ respectively), scaled by the number of mutations. A statistic, $S_T = \hat{T}_{OOA} / \hat{T}_{AFR}$ was then calculated for each 50kb locus sliding along the genome with a step size of 10 kb. Regions in which 6 consecutive windows were all among the highest 0.5% of S_T values in the genome were chosen as candidate regions. The largest region in which all windows were among the 0.5% most extreme in the genome was chosen. We repeated this analysis using Europeans (CEU), East Asian (ASN), or Europeans and East Asian (CEU+ASN) as OOA. We also performed a reciprocal analysis using $\hat{T}_{AFR} / \hat{T}_{OOA}$ to identify candidate gene flow regions into Africans. A potential pitfall for this analysis is that it assumes no within-locus recombination to build the UPGMA trees. However, if recombination has occurred, it is expected to have a conservative effect on our analysis, because it will make it more difficult to detect loci that stand out genome-wide.

Selection of tag SNPs and comparison to Neandertal data

To facilitate comparison with Neandertal and other present-day humans, we identified tag SNPs for each candidate Neandertal gene flow region. First, we estimated a joint tree, again using UPGMA, combining the data from the OOA and AFR samples in the same window. We then identified mutations on the root lineages using parsimony. These root lineages separate a joint (OOA, AFR) "cosmopolitan" clade from a divergent clade containing only OOA individuals in most instances. For the few cases in which AFR individuals appeared on both sides of the root, no tag SNPs were chosen.

The tag SNPs fall into four groups, depending on whether the Neandertal is derived or ancestral relative to chimpanzee, and whether Neandertal matches the deeply divergent clade. We use the term "Derived Match" (DM) to indicate that the allele in the Neandertal is derived relative to chimpanzee and matches the clade that is deeply diverged and only present in the OOA population. "Derived Non-match" (DN) denotes tag SNPs at which the Neandertal is derived but does not match the deep lineage in the OOA population (that is, the Neandertal carries the cosmopolitan allele). Ancestral Match (AM) and Ancestral Non-match (AN) are defined similarly. If the source of the deep haplotypes in non-Africans is gene flow from Neandertals, we expect to observe a majority of Derived Match (DM) and Ancestral Match (AM) tag SNPs at a locus. DM SNPs are particularly informative because a subset of OOA individuals share a derived allele with Neandertals that is not found in Africans. A substantial rate of such SNPs is diagnostic of Neandertal gene flow, as such SNPs are expected to be very rare in the absence of Neandertal gene flow (see simulations below).

Statistical test for contamination

To test whether the excess of matching over non-matching tag SNPs that we detect in candidate Neandertal gene flow regions can be explained by contamination, we take advantage of the fact that at most of the tag SNPs, the allele tagging the clade restricted to non-Africans is much rarer than the allele tagging the clade that is cosmopolitan (the allele frequency is 13% on average for the 13 regions in Table 5). Thus, even in the worst case of 100% contamination by non-African DNA, we do not expect that the OOA-specific allele will be observed in the Neandertal extract at a rate higher than its empirically observed frequency in the OOA sample. To be conservative, we tested this worst case scenario of 100% contamination, which could in principle occur if contamination rates varied across the genome in a locus-specific way.

To implement this idea, for each candidate gene flow region we chose the tag SNP with the lowest frequency in the OOA sample, as this was a tag SNP in which we expected a minimal rate of matching to the OOA-specific clade in the confounding case of contamination. Let the frequency of the allele found in the exclusive OOA clade at SNP i be p_i, i=1,2,..,k. If p_i is an accurate estimate of the allele frequency in SNP i, and if the candidate regions we find are all contaminants from the OOA group, the probability of an allele of type "matching" being observed is p_i (pooling across AM and DM sites). Now, let the indicator function x_i take on the value 1 if the SNP is of type "matching", and 0 if it is of type "non-matching". We define a test statistic $t = \sum_{i=1}^{k} x_i$ (the observed number of matches between the OOA private allele and the Neandertal). The observed value is t_{obs}, and we evaluate its distribution under the null hypothesis of 100% OOA contamination as:

$$\Pr(t \geq t_{obs}) = \sum_{x_1=0}^{1}\sum_{x_2=0}^{2}\cdots\sum_{x_k=0}^{1}\left(I_{\left(\sum_i I_i \geq t_{obs}\right)} \prod_{i=1}^{k} p_i^{x_i}(1-p_i)^{1-x_i} \right) \qquad \text{(S17.1)}$$

The quantity in Equation S17.1 is the probability of observing as many or more matches between the exclusive OOA allele and the Neandertal as is observed in the real data under the null hypothesis of 100% contamination. Equation S17.1 is evaluated using a dynamic programming algorithm.

Simulation framework

We performed simulations using a previously described family of demographic models (*S30*), with and without Neandertal gene flow. Because genome-wide simulations with recombination are extremely slow, we simulated data sets of whole genomes in disjoint 100kb segments under the Wall et al. 2009 model using Richard Hudson's *ms* coalescent simulator (*S29*). The model was modified to take into account the fact that the African Americans represented in the Perlegen data have inherited both African and European ancestry. Specifically, we simulated 20% of their genetic ancestry to be non-African. We also explored modified versions of the model with a modern human-Neandertal divergence time of 480,000 years and a generation time of 20 years, in addition to the default of 320,000 years (Figure S42).

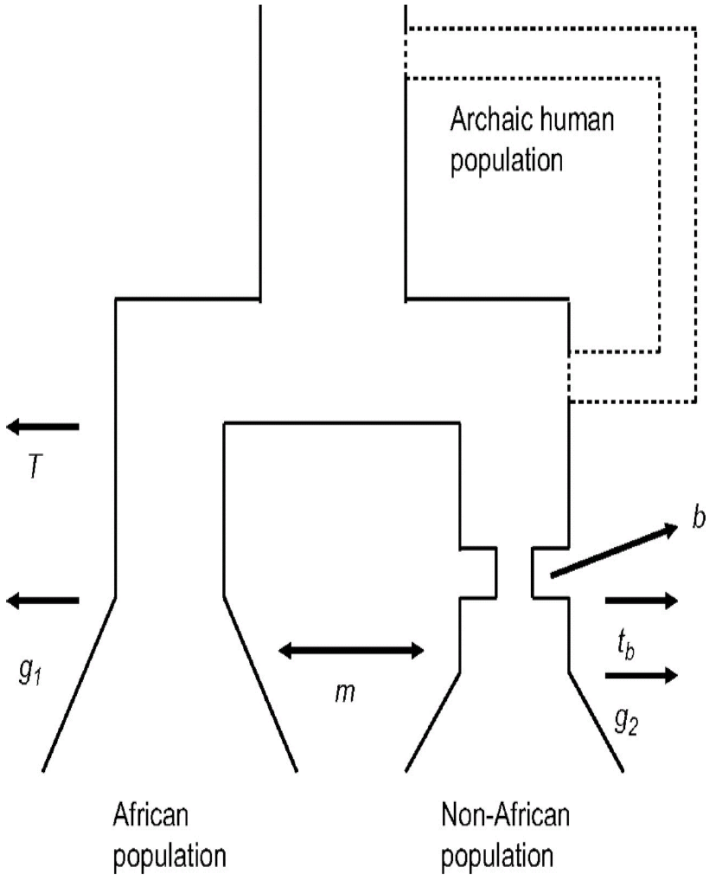


Figure S42: We used the ms software to simulate scenarios with and without Neandertal gene flow and to explore the expected behavior of the statistics in scenarios without gene flow and with gene flow. In the simulations without Neandertal gene flow, time is measured in units of $4N$ generations, with N = 1,000,000 for current Africans. Additional parameters are: (g_1) Growth in Africans starting 0.00035×4×1,000,000=1,400 generations ago from 10,000 to 1,000,000 with a growth rate = 13157.6. (g_2) No growth in Non-Africans, and current size 10,000. (m) Migration between Africans and non-Africans is assumed to be constant at $4Nm$ = 280 since the separation of the two populations. (b, t_b) As variations on the demography, we also simulate a bottleneck 0.00036×4×1,000,000=1,440 generation ago in Europeans, which reduced the population size from 10,000 to 50 for 40 generations, before recovery to its former size. (T) The African / non-African divergence time is simulated to be 0.0006×4×1,000,000=2,400 generations. For the Neandertal gene flow models, we add the following additional parameters: (1) Neandertal gene flow into modern humans 0.0005×4×1,000,000=2,000 generations ago. At that time point we then simulate an instantaneous mixing of populations leading to a group of 86% modern human and 14% Neandertal ancestry. (2) The Neandertal population size is 10,000. (3) Modern humans diverged from Neandertal 0.004×4×1,000,000=16,000 generations ago. We make no claim that these parameters are accurate, but show in this note that they can produce patterns that qualitatively match some features of our data (figure adapted from Wall 2009).

Tailoring the simulations to match the ascertainment of our data

To simulate a pattern of variation similar to that in the Perlegen SNP array data, we used a rejection algorithm to fit the two-dimensional (OOA and AFR) simulated Site Frequency Spectrum (SFS) to the observed SFS, an approach that was previously applied by Voight and colleagues in a scan to detect outlying regions of the genome (*S97*). This correction for ascertainment bias is imperfect: while it captures the genome-wide joint OOA-AFR frequency distribution perfectly, it may not model locus-specific effects particularly well. However, we used this procedure as the best available approach to match our simulations to a uniformly ascertained SNP variation data set.

Results

An excess of candidate Neandertal gene flow regions compared with expectations from simulation

We carried out simulations in the absence of gene flow and compared them to the real data as shown in Figure S43. While our simulated distributions show some sensitivity to recombination rate assumptions, we were not able to generate a tail of high S_T values as extreme as we observed, encouraging us to go further and compare our candidate gene flow regions to the Neandertal sequence.

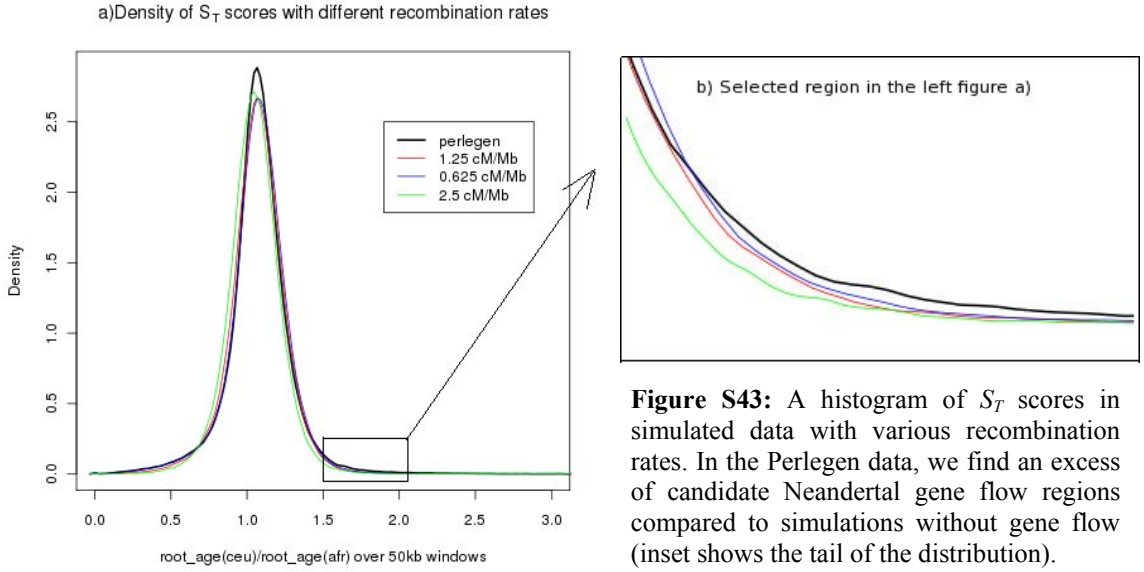


Figure S43: A histogram of S_T scores in simulated data with various recombination rates. In the Perlegen data, we find an excess of candidate Neandertal gene flow regions compared to simulations without gene flow (inset shows the tail of the distribution).

Candidate gene flow alleles match Neandertal at a much higher rate than is expected without gene flow

We identified tag SNPs at the candidate gene flow regions and assessed the rate of matching to Neandertal for both the simulated and the real data. We note that in the simulations without Neandertal gene flow, there were few or no segments with as extreme a value of S_T as in the real data in most genome-wide simulations. We therefore had to generate many genomes' worth of data to produce a sufficient number of candidate Neandertal gene flow regions to compare to the real data.

Table S56: Empirically observed tag SNP patterns in candidate Neandertal gene flow regions

Data type	Focal population[a]	Paired population[b]	No. segments identified	Tag SNP site pattern[c] AN	DN	AM	DM
Simulated data	Without Neandertal gene flow			33%	27%	34%	5%
	With Neandertal gene flow			37%	9%	23%	31%
Real data	OOA=CEU+ASN	African Americans	13	25 (15%)	8 (5%)	57 (34%)	76 (46%)
	OOA=CEU	African Americans	13	20 (14%)	8 (6%)	44 (31%)	68 (49%)
	OOA=ASN	African Americans	11	32 (24%)	9 (7%)	42 (31%)	53 (39%)
	African American	CEU+ASN	25	66 (43%)	26 (17%)	38 (25%)	25 (16%)

a: focal population is the population where we scanned for deep lineages as compared to the paired population.
b: paired population is the population that is used to pair with the focal population to calculate an S_T score.
c: tag SNP patterns, AN(Ancestral Non-match), DM(Derived Match), AM(Ancestral Match), DN(Derived Non-match).

Table S56 shows that in simulations without gene flow, "DM" sites where Neandertal matches the deep OOA-specific clade are expected to be greatly outnumbered by "DN" sites where Neandertal matches the cosmopolitan clade. However, the reverse pattern is expected in the presence of gene flow.

Table S57: Candidate Neandertal gene flow regions obtained by using OOA=CEU and OOA=ASN

Chro-mo-some	Start of candidate region in Build 36	End of candidate region in Build 36	Span (bp)	S_T (est. ratio of OOA/AFR tree depth)	Mean freq. of tag SNPs in OOA clade	Site Patterns#				Qual-itative assess-ment*
						COS-type		OOA-type		
						AN	DN	AM	DM	
Thirteen candidate regions found by using 24 European Americans (CEU) to represent non-Africans										
1	168110000	168220000	110000	2.93	10.4%	1	0	5	10	OOA
1	223760000	223920000	160000	2.78	8.3%	0	0	1	4	OOA
4	171180000	171280000	100000	1.85	10.4%	0	0	1	2	OOA
6	66160000	66260000	100000	5.66	41.7%	0	0	6	6	OOA
9	32940000	33040000	100000	2.83	8.3%	0	0	7	14	OOA
9	114270000	114390000	120000	2.87	10.4%	0	0	3	5	OOA
10	4820000	4920000	100000	3.49	6.3%	0	0	9	5	OOA
10	38000000	38160000	160000	2.61	16.7%	2	0	5	9	OOA
10	69630000	69740000	110000	4.19	20.8%	0	1	2	2	OOA
15	45250000	45350000	100000	2.53	20.8%	1	0	5	6	OOA
17	35500000	35600000	100000	2.93	20.8%	1	0	0	3	OOA
20	20030000	20130000	100000	6.31	91.7%	10	5	0	0	COS
22	30690000	30820000	130000	3.46	8.3%	5	2	0	2	COS
Eleven candidate regions found by using 24 East Asians (ASN) to represent non-Africans										
1	168110000	168210000	100000	2.82	20.8%	1	0	5	10	OOA
1	196790000	196890000	100000	2.25	89.6%	6	2	0	0	COS
1	223760000	223910000	150000	2.76	4.2%	0	0	1	4	OOA
5	28950000	29070000	120000	3.76	6.3%	6	0	16	16	OOA
6	66160000	66260000	100000	5.66	14.6%	0	0	6	6	OOA
6	93290000	93390000	100000	2.31	27.1%	0	0	0	7	OOA
6	116310000	116410000	100000	2.03	79.2%	9	1	0	0	COS
9	114270000	114370000	100000	3.21	10.4%	0	0	3	3	OOA
10	4820000	4920000	100000	3.49	12.5%	0	0	9	5	OOA
10	69630000	69740000	110000	4.19	18.8%	0	1	2	2	OOA
20	20030000	20150000	120000	3.69	37.5%	10	5	0	0	COS
Mean site frequencies when simulated without Neandertal gene flow						33%	27%	34%	5%	-
Mean site frequencies when simulated with Neandertal gene flow						37%	9%	23%	31%	-

Notes: We identified candidate regions of Neandertal gene flow from Neandertal by using either 24 CEU or 24 ASN to represent the OOA population, and 23 African Americans to represent the AFR population. (Table 5 in the main text presents the equivalent analysis for 48 CEU+ASN for the OOA population.) It is possible for there to be both OOA-type and COS-type tag SNPs in a candidate region, reflecting the fact that mutations do not always occur in exactly the same branch of the gene tree and the same genealogy may not apply over the whole region.

* To make a qualitative assessment of the regions in terms of which clade the Neandertal matches, we classified all tag SNPs according to whether the Neandertal matches the cosmopolitan clade (COS) or out-of-Africa specific clade (OOA), and whether the allele is ancestral or derived in Neandertal (we do not list the sites where the matching is ambiguous). If the proportion matching the OOA-specific clade is much more 50%, we classify it as an OOA region, and otherwise COS.

† We present data for loci where we were had Neandertal coverage and found a OOA-unique clade at the root of the tree.

Site patterns: "Match" denotes sites where the Neandertal matches the OOA-specific haplotype, and "Non-Match" where it matches the cosmopolitan type: AN(Ancestral Non-match), DM(Derived Match), AM(Ancestral Match) and DN(Derived Non-match).

Examining our real data, we find that at a majority of tag SNPs, the Neandertal matches the OOA-specific clade. Table 5 gives the results when CEU+ASN are combined to represent the OOA population, and Table S57 gives the results for CEU and OOA separately. When CEU+ASN are combined to represent the OOA

population (Table 5), there are 76 DM tag SNPs where Neandertal is derived and matches the OOA private allele, and only 8 DN tag SNPs where Neandertal is derived and matches the cosmopolitan allele. Out of the 12 regions where we had tag SNPs, all showed a pattern of either all, or almost all, tag SNPs matching the Neandertal, or no tag SNPs matching the Neandertal. We find that 10 segments have an almost complete match to Neandertal and only 2 show the opposite pattern (these latter two have no evidence of Neandertal gene flow, and are likely to be false-positives).

Contamination cannot explain these observations

To evaluate whether contamination could account for the high rate of matching of the OOA-specific clade to Neandertal, we focused on tag SNP where the OOA-specific clade is rarest in each of the 13 candidate Neandertal gene flow regions in Table 5. We then asked whether the rate of observation of the OOA-specific allele at these tag SNPs is significantly elevated compared with the worst case scenario of 100% contamination at these loci (see the Methods above). By this analysis, we reject the hypothesis that our results can be explained by contamination from Europeans (P=0.0025.)

Simulations show that Neandertal-to-modern human gene flow can produce patterns like we observe

To test if skews in the ratio of (DM and AM SNPs) to (AN and DN SNPs) as extreme as we observe can occur in the absence of Neandertal gene flow, and to further test whether they are expected in some scenarios of Neandertal gene flow, we performed coalescent computer simulations of gene flow. The distribution of patterns observed in these simulations is presented in Figure S44.

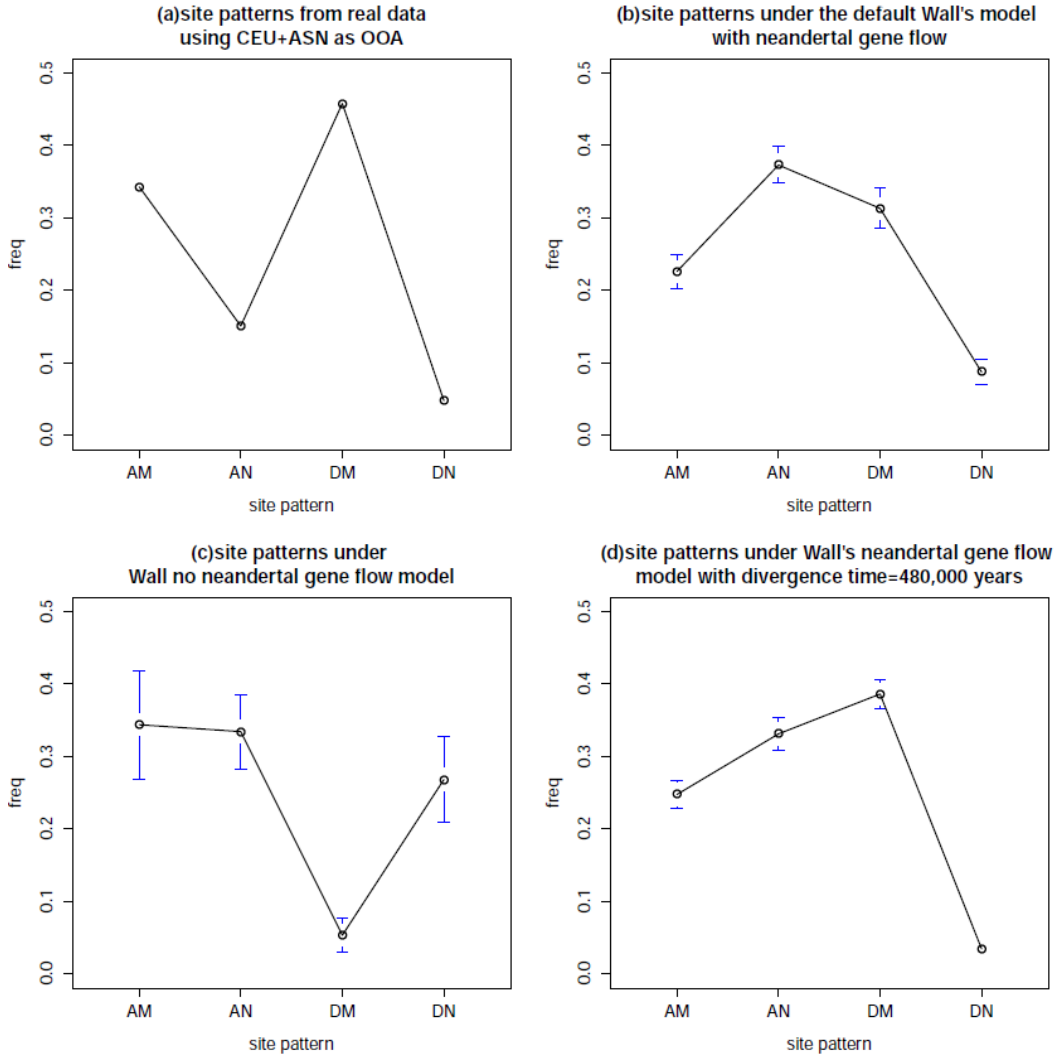


Figure S44: The distribution of the four tag SNP patterns in simulated data under the Wall et al. demographic model. **a.** The observed distribution in the Perlegen+Neandertal data using CEU+ASN as the OOA group. **b.** Neandertal gene flow model (default settings). **c** No Neandertal gene flow model. **d** Modified Neandertal gene flow model assuming Neandertals and present-day humans diverged 480,000 years ago.

A key observation is that the pattern of shared derived alleles between Neandertals and the OOA group (DM) rarely occurs in simulations without Neandertal gene flow. Simulating under the Wall et al. model with Neandertal gene flow, however, we observe the distributions in Figure S44b and S44d. Sites of type Derived Match (DM) now frequently occur, whereas the Derived Non-match (DN) pattern is rare, as in the observed data (Figure S44a). We also repeated the simulations with a lower rate of Neandertal gene flow (4%). The results are qualitatively similar.

Caveat about simulations: A wide range of gene flow scenarios are consistent with the data
Although we have shown that there are demographic scenarios that involve gene flow that are consistent with the data, the specific demographic parameters that we have used for simulating Neandertal gene flow may not correspond to the true history. Our simulations only explore a fraction of parameter space, and there are likely to be many other parameter combinations involving Neandertal gene flow that are as plausible based on the data we have presented. The focus of the simulations is simply to show that there are Neandertal gene flow scenarios that can produce high rates of AM and DM sites (compared to AN and DN sites), as we observe. Likewise, we find no demographic models without Neandertal gene flow that can explain the large proportion of DM sites. The observation of a large proportion of DM sites, appears to be a unique feature of Neandertal gene flow models.

Natural selection in present-day humans since Neandertal divergence would not produce false-positives
The inferences of gene flow that we have made are based on the extreme tail of a genome-wide distribution (Figure S45), which raises the possibility that our ascertainment procedure may have enriched for loci under natural selection in modern humans. Examining these loci more closely, we find that due to ascertainment for a high value of the ratio $S_T = \hat{T}_{OOA} / \hat{T}_{AFR}$, they are characterized both by a high average value of $\hat{T}_{OOA}$ (10.69 for CEU vs. the genome average of 6.69) and a low average value of $\hat{T}_{AFR}$ (2.89 at these loci vs. the genome average of 6.28). The low value of $\hat{T}_{AFR}$ could be due to stochastic variation, but could also be hallmark of a locus that has experienced a complete selective sweep in the history of West Africans since their divergence from Neandertal.

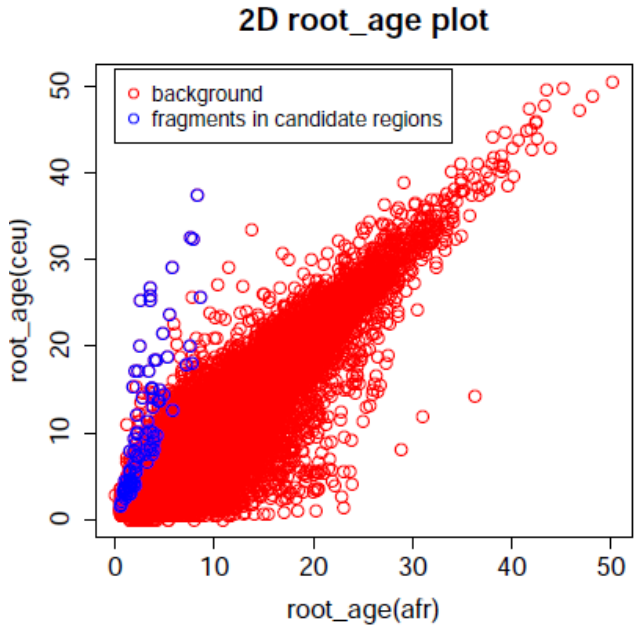


Figure S45: Characteristics of the candidate Neandertal gene flow regions in terms of TMRCA in the AFR and OOA samples. We present the number of mutations from the root of the UPGMA trees in AFR and CEU, for loci with S_T in the top 0.5% of the genome, with the x and y axes scaled in units of number of mutations.

Although it is plausible that the regions we examined are enriched in signals for selection, we do not see how a selective sweep in West Africa could generate a rate of matching of the OOA-specific clade to Neandertal that is significantly above 50%. We emphasize that selective sweeps at a locus in modern humans since

divergence from Neandertal is in fact expected to increase power to detect gene flow at a locus, instead of the opposite. The reason is that at loci with low $\hat{T}_{AFR}$ (either due to a selective sweep or a stochastically short tree), the average coalescence time in the nuclear human genome is expected to be lower than average human/Neandertal genetic divergence time. By contrast, in regions in which the coalescence time in present-day humans is typical, it is difficult to distinguish human-Neandertal genetic identity due to shared ancestral lineages, from a signature of Neandertal to human flow.

No signal of gene flow from Neandertal into Africans

We performed the reciprocal analysis in a sample of West African ancestry, using $\hat{T}_{AFR}/\hat{T}_{OOA}$ as a test statistic based and using CEU+ASN as the OOA population. As seen in Figure S46 and Table S56, the results are radically different. A caveat, however, is that this analysis does not have much power to detect Neandertal gene flow. The power to detect regions of gene flow is lower in Africa because $\hat{T}_{AFR}/\hat{T}_{OOA}$ in general tend to be >1, reflecting higher diversity in African than non-African populations.

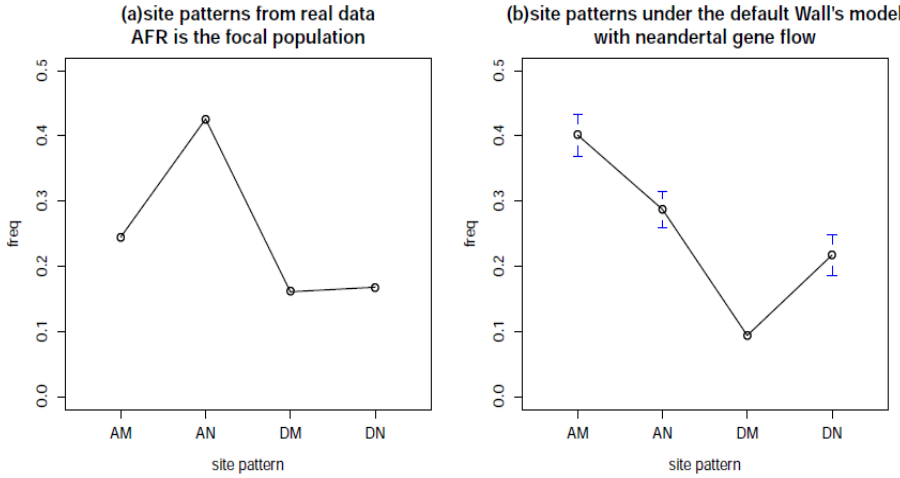


Figure S46: No. signal of Neandertal gene flow from Neandertal into West Africans. **a.** Observed Perlegen+ Neandertal data when AFR is the focal populations (searching for deeply diverged lineages in AFR that are not present in OOA). **b.** Expected pattern from the default Wall demographic model with Neandertal flow into modern humans. We observe no signal even with gene flow, showing that we have little power to detect the signal using AFR as the focal population.

No evidence for these results being an artifact of structural variation

To examine if the regions we identify are enriched for structural variants, we examined the overlap between known structural variants and target regions by comparing to the Database of Genomic Variants (DGV) (http://projects.tcag.ca/variation/). We used randomization to examine if the regions are enriched by randomly choosing loci in the genome of the same length as our candidate regions (Figure S47a). The observed number of hits to entries in the DGV is close to the median of the simulated distribution, suggesting that our target regions are not enriched for structural variants.

Higher power to detect candidate gene flow regions at loci of low recombination

In contrast to the absence of an association with structural variants, the candidate regions are strongly associated with low recombination rate (Figure S47b) (using rate estimates from (*S70*)). The mean recombination rate in the genome is approximately 1.26 cM/Mb, whereas the mean recombination rate in the candidate segments is 0.56 cM/Mb. Random re-sampling of 13 regions with the same lengths as in the real candidate regions indicates that the observed reduction in mean recombination rate is unlikely to occur by chance ($P<0.01$). This association between regions of low recombination and candidate regions is in fact expected under the scenario of Neandertal-to-modern human gene flow, as haplotypes that have arisen through gene flow are expected to be maintained for a longer period in regions of low recombination than in regions of high recombination.

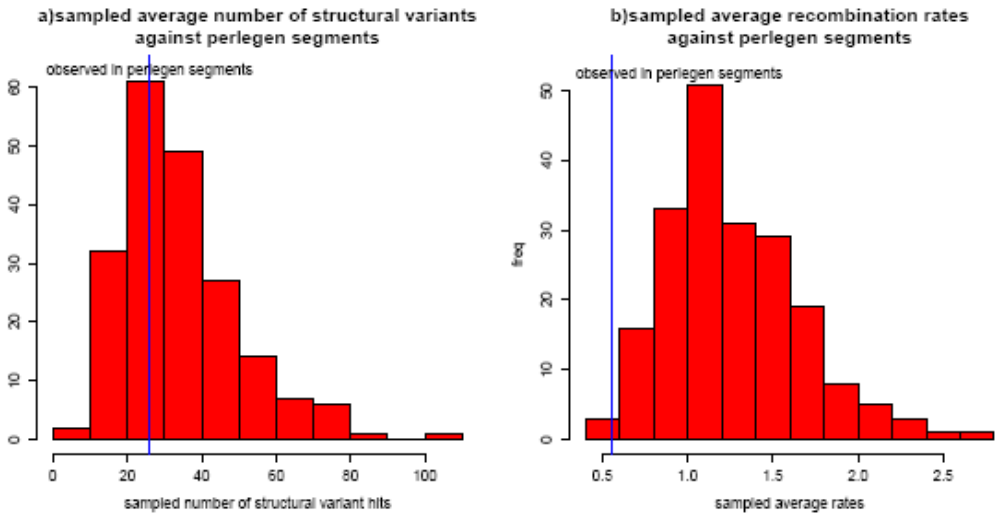


Figure S47. Screens for enrichment of the candidate Neandertal gene flow regions shows **a.** no enrichment for structural variation, but **b.** strong enrichment for regions of low recombination.

Conclusions

The deeply diverged haplotypes that are unique to non-Africans that we identified through our scan show a rate of matching to Neandertal that is well above 50%, a phenomenon that we were not able to explain by any scenario in the absence of gene flow. Moreover, we were not able to identify any simulation scenario that did not involve gene flow from Neandertal that predicts a rate of private derived alleles in OOA populations matching Neandertals that is as high as we observe. By contrast, when we simulate scenario with gene flow, we obtained a qualitative match to these observations.

Supplementary Online Materials 18

Model-free estimate of the proportion of Neandertal ancestry in present-day non-Africans

Nick Patterson and David Reich*

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

Overview:

In this section, we extend ideas from SOM 15 to show that the proportion of Neandertal ancestry in present-day non-Africans is between 1.3-2.7%. Specifically, we define *f* precisely as the proportion of the genome in present-day non-Africans that traces its genealogy through the Neandertal side of the phylogenetic tree, just after the Neandertal-modern human population divergence. Our estimate is robust to assumptions about effective ancestral population sizes and how they have changed over time, as well as assumptions about population divergence times. We only make two assumptions:

Assumption 1: Neandertals formed a clade. The Neandertals who we sequenced and the Neandertals who contributed ancestry to present-day humans are assumed to form a clade relative to present-day sub-Saharan Africans and chimpanzee. In other words, they are assumed to descend from a homogeneous ancestral population that diverged at an earlier time from the ancestors of present-day sub-Saharan Africans.

Assumption 2: Neandertals were homogeneous. We assume for our basic analyses that the Neandertals we sequenced and the Neandertals who contributed ancestry to present-day non-Africans can effectively be treated as a homogeneous population. In fact, however, we are able to obtain meaningful bounds on *f* even allowing for the fact that Neandertals could have been inhomogeneous.

An estimator for the proportion of Neandertal alleles in present-day non-Africans

We extend ideas from SOM 15 to infer *f* in present-day non-Africans. To estimate the proportion of Neandertal ancestry in the genomes of present-day non-Africans, we define the following population names:

C	= Chimpanzee.
AFR	= Yoruba or San
OOA	= French, Papuan, or Han
OOA_A	= the modern human ancestral population of present-day non-Africans
N_A	= the Neandertal population that contributed genes to present-day non-Africans
N_1	= A Neandertal bone that we sequenced
N_2	= A different Neandertal bone that we sequenced

For the ordered set of populations $\{X,AFR,N,C\}$, we denote the chimpanzee allele as "*A*", and restrict our analysis to biallelic sites at which *X* and *AFR* differ and the alternative allele *"B"* is seen in *N*. Thus, the two possible patterns of SNPs are "ABBA" or "BABA". We further define the indicator variables $C_{ABBA}(i)$ and $C_{BABA}(i)$, which can be 0 or 1 depending on whether an ABBA or BABA pattern is seen at base *i*.:

$$S(X, AFR, N_1, C) = \sum_{i=1}^{n} \left[C_{ABBA}(i) - C_{BABA}(i) \right] \quad \text{(S18.2)}$$

We note that the difference S(X,AFR,N_1,C) is just the numerator of the *D*-statistic in SOM 15. It measures the relative rate of matching of the samples *X* and *AFR* to the Neandertal N_1 that we sequenced.

We now model OOA as a linear mixture of ancestral non-Africans and ancestral Neandertal. Thus:

$$OOA = (1-f)OOA_A + (f)N_A, \quad \text{(S18.1)}$$

Combining Equation S18.2 and Equation S18.3, we obtain:

$$E[S(OOA, AFR, N_1, C)] = (1-f)E[S(OOA_A, AFR, N_1, C)] + (f)E[S(N_A, AFR, N_1, C)] \\ = (f)E[S(N_A, AFR, N_1, C)] \quad \text{(S18.3)}$$

The first term E[$S(OOA_A, AFR, N_1, C)$] = 0, since in our model, the unmixed modern human ancestral population of non-Africans OOA_A and present-day Africans *AFR* form a clade relative to Neandertal.

We now algebraically manipulate Equation S18.3 to obtain a ratio that is expected to equal the gene flow proportion:

$$f = \frac{E[S(OOA, AFR, N_1, C)]}{E[S(N_A, AFR, N_1, C)]}. \quad \text{(S18.4)}$$

Intuitive interpretation of Equation S18.4

To understand Equation S18.4 intuitively, it is helpful to consider the meaning of both the numerator and denominator of Equation S18.4:

Numerator: The numerator measures how much closer the Neandertal bone we sequenced is to the *OOA* than to the *AFR* population (it is exactly the same numerator as the *D*-statistic in Equation 1 of the main text). It is significantly skewed from 0, due to gene flow from Neandertals in the ancestors of non-Africans (SOM 15).

Denominator: The denominator is the difference in the rate of matching of the Neandertal we sequenced to N_A and *AFR*, restricting to sites where the Neandertal we sequenced is derived. This gives this skew that would be expected of the *OOA* population was entirely of Neandertal ancestry.

Thus, the ratio in S18.4 is measuring what percentage of the way the *OOA* population is toward having the phylogenetic relationships to Africans that the Neandertal does. It is clear why this is the mixture proportion.

Practical estimation of *f*

Although Equation S18.4 provides a way to estimate the mixture proportion, in practice we cannot estimate it directly. The reason for this is that we do not have access to a sample from N_A, the Neandertal population that actually contributed genetic material to the ancestors of non-Africans. We therefore use a second Neandertal bone N_2 as a surrogate for N_A in Equation S18.4, and empirically explore whether our inferences are consistent when we choose Neandertal bones over a wide geographic and temporal distribution.

We estimated *f* using the same Illumina present-day human and Neandertal data set described in SOM 15, restricting our analysis to autosomal nucleotides that satisfied the following conditions:

1. We required the availability of the chimpanzee reference sequence PanTro2.
2. We required at least one present-day human from outside Africa (French, Han or Papuan).
3. We required at least one sub-Saharan Africans (Yoruba or San).
4. We required two Neandertal bones (Vi33.16, Vi33.25, Vi33.26, Mezmaiskaya, Feldhofer, El Sidron).
5. We used a quality threshold for Neandertal data of ≥32 rather than ≥40, to maximize the data.
6. We only analyzed nucleotides where we had coverage from exactly two Neandertal reads passing our quality filters. We further required that these reads came from different bones. We excluded nucleotides with ≥3-fold coverage to minimize artifacts from structural variation

We write (ooa_i, afr_i, $n1_i$, $n2_i$, c_i) as the alleles we sampled at nucleotide *i* in *OOA*, *AFR*, the two Neandertal bones *n1* and *n2*, and the chimpanzee. We note that both $S(OOA,AFR,N_1,C)/S(N_2,AFR,N_1,C)$ and $S(OOA,AFR,N_2,C)/S(N_1,AFR,N_2,C)$ are both valid estimators of the gene flow proportion according to Equation S18.4, and hence we average these two quantities over all *m* aligned nucleotides:

$$\hat{f} = \frac{\sum_{i=1}^{m}\left(S(ooa_i, afr_i, n1_i, c_i) + S(ooa_i, afr_i, n2_i, c_i)\right)}{\sum_{i=1}^{m}\left(S(n1_i, afr_i, n2_i, c_i) + S(n2_i, afr_i, n1_i, c_i)\right)}. \quad \text{(S18.5)}$$

We compute a standard error by a weighted block jackknife (SOM 15).

In practice, we estimated $\hat{f}$ using any of (French, Han, Papuan) for *OOA*, and any of (Yoruba, San) for *AFR* (Table S58). We chose pairs of Neandertal bones for analysis in two ways:

a. "Full analysis". We required that *n1* and *n2* came from two different bones, which in practice meant that they almost always came from two different Vindija bones.
b. "Restricted analysis". We required that one bone we analyzed was not from Vindija. The Mezmaiskaya infant contributed 93% of the non-Vindija data and hence the vast majority of the comparisons are of Vindija and Mezmaiskaya bones.

Combining the "full" and "restricted" analyses is useful for constraining our estimate of the mixture proportion *f*. As described above, since we cannot directly sample from the Neandertal population that contributed genetic material to non-Africans N_A, we have to replace it with another Neandertal. Our "full" analysis is comparing Neandertals that are more closely related than Vindija-N_A, since the comparisons are mostly Vindija-Vindija. Our "restricted" analysis is likely to be comparing Neandertals that are distantly related, as the Vindija and Mezmaiskaya fossils are from geographically dispersed locations (Croatia and the Caucasus) and the two fossils have very different dates (about 40,000 years vs. 60,000-70,000 years). Thus, the "full" and "restricted" span a wide range of relatedness among Neandertals, which hopefully bracket the true level of *N* and N_A relatedness, and thus hopefully provide estimates of *f* that bracket the truth.

An encouraging aspect of our results (see below) is that we obtained very similar estimates of *f* for both types of comparison (Vindija-Vindija and Vindija-Mezmaiskaya). Thus, the quantitative value of the denominator of Equation S18.4 seems to be similar whatever pair of Neandertals we use, suggesting that the lack of availability of a sample from N_A may not be too much of a problem.

Results

We obtained consistent estimates of $\hat{f}$ for each choice of *OOA* and *AFR*, for both the "full analysis" and the "restricted analysis" (Table S58). Combining across all pairs of *OOA* and *AFR* samples to increase precision, the estimates of the proportion of Neandertal genetic material in the ancestry of non-Africans was $\hat{f}$ = 1.7 ± 0.2% for the "full" and $\hat{f}$ = 1.1 ± 0.8% for the "restricted" analysis.

The "full analysis" combined result ($\hat{f}$ = 1.7 ± 0.2%) provides a conservative minimum of *f* > 1.3% on the proportion of Neandertal genetic material in the ancestry of non-Africans (2 standard deviations below 1.7%). To understand why this is a minimum, we note that the denominator in the estimator of $\hat{f}$ in Equation S18.4 is $S(N_A,AFR,N,C)$: the difference in the rate of matching of the Neandertal *N* to N_A vs. *AFR*. Since two Vindija samples are almost certain to be more closely related than a Vindija sample *N* and the Neandertal who contributed ancestry to modern humans N_A, we are estimating a too-high a rate of matching between N_A and *N*, resulting in an over-inflation in $S(N_A,AFR,N,C)$, and a systematic underestimate of $\hat{f}$.

The "restricted analysis" result ($\hat{f}$ = 1.1 ± 0.8%) provides a likely upper bound of *f*<2.7% on the proportion of Neandertal genetic material in the ancestry of non-Africans (2 standard deviations above 1.1%). Since *N* and N_A are now being represented primarily by Vindija and Mezmaiskaya—which based on geographic and

temporal considerations are potentially as distantly related as Vindija and the Neandertal we sequenced—we may be systematically underestimating $S(N_A,AFR,N,C)$ and thus overestimating f by using Mezmaiskaya to represent N_A, the Neandertal population that actually contributed genetic material to non-Africans.

Table S58: Model-free estimate of the Neandertal ancestry proportion in present-day non-Africans

Full analysis *(primarily Vindija-Vindija comparisons)*				***Restricted analysis*** *(primarily Vindija-Mezmaiskaya comparisons)*			
OOA	*AFR*	$\hat{f}$	std.err. ($\hat{f}$)	*OOA*	*AFR*	$\hat{f}$	std.err. ($\hat{f}$)
French	San	1.7%	0.3%	French	San	0.4%	1.3%
French	Yoruba	1.7%	0.3%	French	Yoruba	1.3%	1.3%
Han	San	2.0%	0.3%	Han	San	1.0%	1.3%
Han	Yoruba	2.0%	0.3%	Han	Yoruba	2.5%	1.3%
Papuan	San	1.4%	0.3%	Papuan	San	-0.4%	1.3%
Papuan	Yoruba	1.5%	0.3%	Papuan	Yoruba	1.4%	1.3%
	combined	1.7%	0.2%		*combined*	1.1%	0.8%

Note: The estimator of $\hat{f}$ is computed as in Equation S18.5.

We confidently infer $f > 1.3\%$, even in the presence of possible confounders

We considered four possible confounders that could bias our estimate of f, but found that none of them were consistent with a mixture proportion of less than 1.3%

1. The estimate of $1.3\% < f < 2.7\%$ is robust to the presence of sequencing error: The numerator of our estimator of $\hat{f}$ in Equation S18.4 is accurately estimated even in the presence of sequencing error, as shown in SOM 15. The denominator $S(N_A,AFR,N,C)$, is likely to be even more accurately estimated, since it is of even larger magnitude than the numerator, and hence less affected by error.
2. The estimate $1.3 < f < 2.7\%$ is robust to the presence of contamination: Although substantial levels of contamination in the Neandertal sequence could affect our inferences, our many analyses of mtDNA, Y chromosome and autosomal data suggest that contamination rates are <0.5% in this data set (SOM 5-7). We also showed in SOM 15 that whatever contamination rates do exist are unlikely to explain the major skews we observe in the *D*-statistics (the *D*-statistics are very similar to the statistics here).
3. f is conservatively estimated be at least 1.3% even if Assumption 1 is violated: If the Neandertal who contributed genetic material to non-Africans N_A had some modern human ancestry, a given mixture proportion would have less of an effect in skewing the numerator of Equation S18.4 than if it had entirely Neandertal ancestry. Thus, this effect would cause us to systematically underestimate f.
4. f is conservatively estimated be at least 1.3% even if Assumption 2 is violated: Even if Neandertals are not homogeneous, our "full analysis" provides a minimum for f. The reason is that as discussed above, N and N_A in Equation S18.4 are supposed to represent the Vindija Neandertal we sequenced and the Neandertals who contributed ancestry to modern non-Africans. If we use two Vindija bones as surrogates for N and N_A, we will be guaranteed to be using two bones that are more closely related than the truth. This will overestimate the denominator of Equation S18.4, and underestimate f.

We conclude that $f > 1.3\%$ from the "full analysis" is a likely minimum estimate on the proportion of Neandertal ancestry in present-day non-Africans, and that $1.3\% < f < 2.7\%$ is a realistic range.

Supplementary Note S19

Fraction of non-African genomes likely to be of Neandertal ancestry

Eric Y. Durand, Philip L. F. Johnson, Anna-Sapfo Malaspinas and Montgomery Slatkin*

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

In this section, we develop a mathematical model that allows us to quantify the amount of gene flow from Neandertals into the ancestors of present-day non-Africans. We analyze the simplest possible model that is consistent with what is known about the history of humans and Neandertals and is consistent with the genetic data, namely, a model in which the gene flow occurred after the divergence of the population ancestral to non-Africans from the population ancestral to sub-Saharan Africans but before the divergence into Han, Papuans and French (Figure S48). A potential time during which this gene flow could have occurred is between 50,000–80,000 years ago, when the archaeological evidence shows that Neandertals and modern humans were both present in the Middle East. This time period is also consistent with the constraints presented in Supplementary Note 15, which show that gene flow occurred before the divergence of the non-African Papuan, Han and French populations. We caution that this is only the simplest model consistent with our observations.

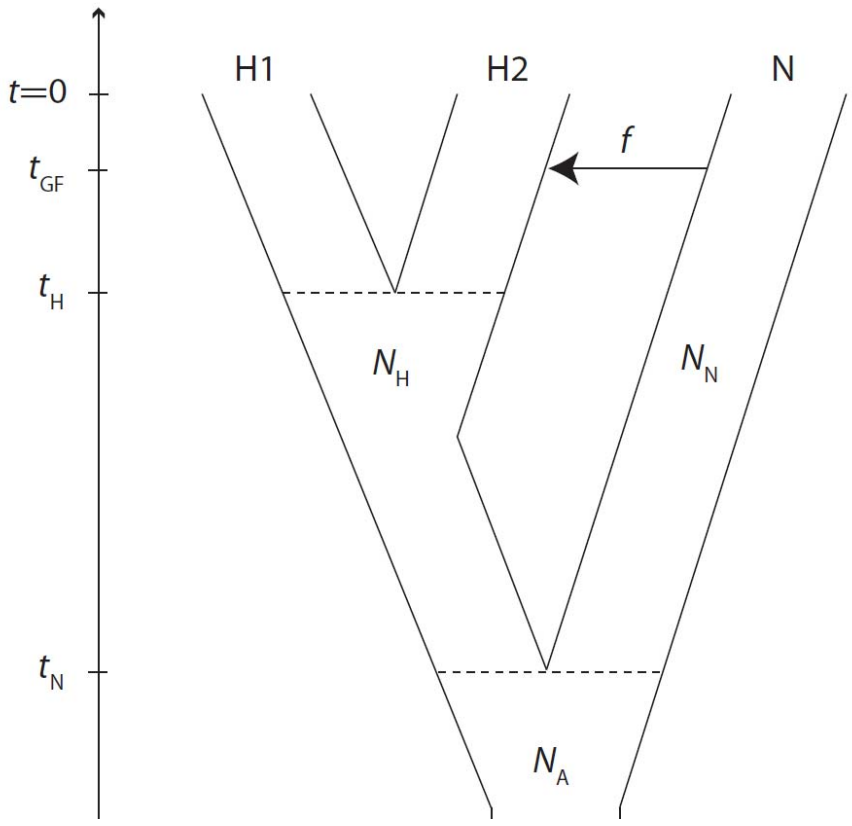


Figure S48: Model of Neandertal-modern human ancestry that we fit to the data. We assume Neandertals and modern West Africans (H_1) split at time t_N, that the ancestors of West Africans and modern non-Africans split at time t_H, and that a single pulse of gene flow contributed a proportion f of Neandertal ancestry to modern non-Africans (H_2) at a time t_{GF} in the past. All population sizes are assumed constant.

In what follows, we begin by estimating the fraction f of Neandertal ancestry in the genome of present-day non-Africans that is necessary in order to produce the observed value of the test statistic *D(African, non-African, Neandertal, Chimp)*. We further constrain the parameters by using our observations of the proportion of sites where Neandertal and Europeans carry the derived allele at SNPs discovered in two Yoruba, and measurements of Neandertal-European and Neandertal-West African divergence. Defining t_N as the number of generations ago when the ancestors of the Neandertal we sequenced and sub-Saharan Africans last exchanged genes, we can then define f as the proportion of alleles in non-Africans whose ancestors were in Neandertal more recently than t_N.

Analytical expectation for the value of the *D* statistic for a given gene flow proportion *f*

We assume that we have one nucleotide site sampled from modern human population 1 (H_1), one from modern human population 2 (H_2), one from a Neandertal (N), and from a chimpanzee (C). By assumption, the chimpanzee carries the ancestral nucleotide (denoted A), and we restrict to biallelic polymorphisms where Neandertal carries the derived nucleotide (denoted B). There are two cases of interest: H_1 has the ancestral nucleotide and H_2 has the derived (denoted ABBA) or the reverse (denoted BABA).

We found that ABBA is significantly more frequent than BABA for H_1=modern sub-Saharan Africans and H_2=modern non-Africans. We measure the excess by the Symmetry Test statistic *D(West African, non-African, Neandertal, Chimpanzee)* = [Pr(ABBA)-Pr(BABA)]/[Pr(ABBA)+Pr(BABA)], which has an observed value of 4-5% empirically (Table 4; SOM 15). We use a value of 4.6% in the model fitting that follows.

A genealogical argument for the expected frequencies of ABBA and BABA sites

We calculated the probabilities of ABBA and BABA under the assumption that only a single nucleotide substitution occurred on the gene genealogy representing the ancestry of the four samples (H_1, H_2, N and C). Figure S49 presents the three possible topologies of the gene genealogy, assuming that C is always the outgroup. For each of these topologies, there are 4 branches on which a substitution can create a derived allele polymorphic within H_1, H_2 and N (denoted by an arrow in Figure S49). The pattern BABA (H_1 and N have the derived nucleotide) is consistent only with Tree IIc. The length of the internal branch is the historical opportunity for a mutation to occur to produce the BABA pattern, averaged across all bases of the genome in which tree topology II applies. Similarly, the pattern ABBA (H_2 and N have the derived nucleotide) is consistent only with tree IIIc, and the length of that internal branch is the time during which mutation can occur to create ABBA, averaged over the bases in the genome over which topology III applies. The probability that a randomly chosen site will have either pattern is the probability of the appropriate topology multiplied by the expected branch length multiplied by the mutation rate, μ. Thus, our problem is to compute the probabilities of the tree topologies and average lengths of the relevant branches under the assumptions of our model. Other trees do not matter because mutations cannot create either the ABBA or BABA pattern.

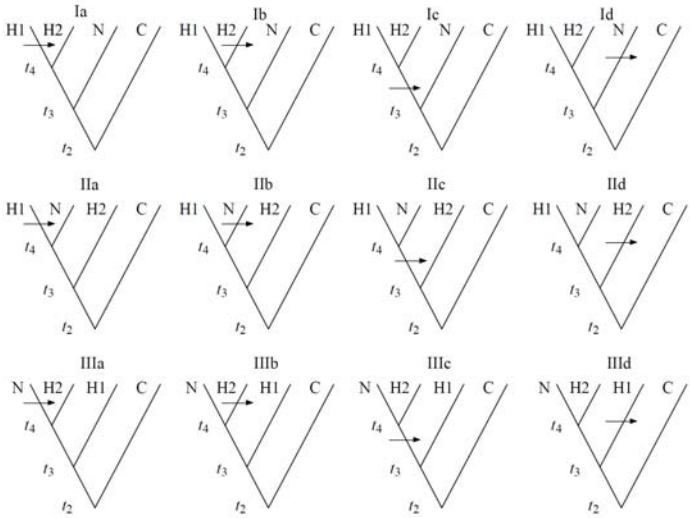


Figure S49: There are three possible topologies of the gene genealogies, I, II, and III, that can relate samples H_1, H_2 and N and the outgroup C. Assuming that any polymorphisms reflect a single historical mutation, ABBA sites can occur only on a Type III gene genealogy due to a mutation that occurs on the branch ancestral to samples H_2 and N (IIIc), and BABA sites can occur only on a Type III gene genealogy due to a mutation on the branch ancestral to samples H_1 and N (IIc). The probability of occurrence of these mutations is expected to be proportional to the lengths of the relevant internal branches of IIc and IIIc, averaged across the genome. We can calculate the expected values of these quantities analytically under the demographic model of Figure S48

Model of gene flow from Neandertals to H_2

In the model, we assume that there was a single episode of gene flow at time t_{GF} in the past (t=0 being the present) from population N to H_2 after the separation of populations H_1 and H_2 (Figure S48). With probability f, H_2 was derived from a Neandertal lineage and, with probability 1-f, H_2 was derived from an H_1-related modern human group. We also considered a slightly more complicated model that assumes a period of ongoing one-way gene flow from N to H_2. We found that this model leads to the same estimate of the mixture proportion. Hence, we just present the simpler model in what follows.

Our model in Figure S48 also requires several other demographic parameters. We define the divergence time of the two modern human lineages as $t_H > t_{GF}$, and the modern human-Neandertal divergence time as $t_N > t_H$. The effective population size of the Neandertal lineage is N_N; the effective size of the modern human lineage ancestral to H_1-H_2 divergence is N_H; and the effective size of the population ancestral to Neandertals and modern humans is N_A. All population sizes are assumed to be constant over time and the populations are assumed to be unstructured (i. e. they are randomly mating).

Expected values of ABBA and BABA under the model

There are three genealogical scenarios under the model of Figure S48 that can create either ABBA or BABA sites. The first two produce ABBA and BABA sites with equal probability and the third produces only ABBA sites, contributing to the excess of ABBA over BABA sites that we observe empirically.

1. *The H_2 lineage traces its ancestry through the modern human side of the phylogeny (probability 1-f), and between t_H and t_N, the H_1 and H_2 lineages do not coalesce (probability $\left(1-1/(2N_H)\right)^{t_N-t_H}$).*

   In this case, the H_1 and H_2 lineages trace their ancestry all the way back to the modern human-Neandertal ancestral population without coalescing, so that there are three lineages in the ancestral population (H_1, H_2 and N) that coalesce. The expected time between the first and second coalescent events (which can produce ABBA or BABA sites) is $2N_A$. With probability 1/3 each, the first coalescence is between H_1 and N or H_2 and N, producing the two sites of interest. Thus:

$$\Pr_1(\text{ABBA}) = \Pr_1(\text{BABA}) = \left(1-\frac{1}{2N_H}\right)^{t_N-t_H}\frac{2N_A\mu}{3}(1-f). \qquad \text{(S19.1)}$$

2. *The H_2 lineage traces its ancestry through the Neandertal side of the phylogeny (probability f), and between t_{GF} and t_N the two Neandertal lineages do not coalesce (probability $\left(1-1/(2N_N)\right)^{t_N-t_{GF}}$).*

   In this case, the H_1 and H_2 lineages trace their ancestry all the way back to the modern human-Neandertal ancestral population without coalescing, so that there are three lineages in the ancestral population (H_1, H_2 and N) that coalesce. The expected time between the first and second coalescent events (which can produce ABBA or BABA sites) is $2N_A$. With probability 1/3 each, the first coalescence is between H_1 and N or H_2 and N, producing the two sites of interest. Thus:

$$\Pr_2(\text{ABBA}) = \Pr_2(\text{BABA}) = \frac{2N_A\mu f}{3}\left(1-\frac{1}{2N_N}\right)^{t_N-t_{GF}}. \qquad \text{(S19.2)}$$

3. *The H_2 lineage traces its ancestry through the Neandertal side of the phylogeny (probability f), and between t_{GF} and t_N the two Neandertal lineages coalesce.* This history creates gene genealogies only of type III and results only in ABBA sites (never BABA sites in the absence of recurrent mutation in the same genealogy). The probability that there is a coalescence before t_N is $\left[1-\left(1-1/(2N_N)\right)^{t_N-t_{GF}}\right]$. Once the coalescence occurs, the ancestral lineage cannot coalesce with H_1 before t_N. After t_N, the average coalesce time is $2N_A$. Therefore, the expected length of the internal branch is

$t_N + 2N_A - (t_{GF} + \bar{t})$,where $\bar{t}$ is the expected coalescence time in the Neandertal lineage, given that the coalescence occurs before t_N. A little analysis shows that:

$$\bar{t} = 2N_N - \frac{(t_N - t_{GF})\left(1 - \frac{1}{2N_N}\right)^{t_N - t_{GF}}}{1 - \left(1 - \frac{1}{2N_N}\right)^{t_N - t_{GF}}}. \quad \text{(S19.3)}$$

Thus:

$$\Pr_3(\text{ABBA}) = \mu f\left(1 - \left(1 - \frac{1}{2N_N}\right)^{t_N - t_{GF}}\right)\left[\frac{t_N - t_{GF}}{\left(1 - \left(1 - \frac{1}{2N_N}\right)^{t_N - t_{GF}}\right)} + 2(N_A - N_N)\right]. \quad \text{(S19.4)}$$

$$= \mu f\left[t_N - t_{GF} + 2(N_A - N_N)\left(1 - \left(1 - \frac{1}{2N_N}\right)^{t_N - t_{GF}}\right)\right]$$

The overall probability of ABBA and BABA is obtained by adding the three contributions. The mutation rate parameter μ cancels, and the overall expectation is

$$E(D) = \frac{\Pr(\text{ABBA}) - \Pr(\text{BABA})}{\Pr(\text{ABBA}) + \Pr(\text{BABA})}$$

$$= \frac{3f\left[t_N - t_{GF} + 2(N_A - N_N)\left(1 - \left(1 - \frac{1}{2N_N}\right)^{t_N - t_{GF}}\right)\right]}{3f\left[t_N - t_{GF} + 2(N_A - N_N)\left(1 - \left(1 - \frac{1}{2N_N}\right)^{t_N - t_{GF}}\right)\right] + 4N_A(1-f)\left(1 - \frac{1}{2N_H}\right)^{t_N - t_H} + 4N_A f\left(1 - \frac{1}{2N_N}\right)^{t_N - t_{GF}}} \quad \text{(S19.5)}$$

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

We carried out simulations using the *ms* software (*S29*) that matched the results of Equation S19.5.

Exploration of parameter space

Figure S50 shows the dependence of $E(D)$ on f when t_H=3,000, t_{GF}=2,500, t_N=10,000, and N_H=N_N=N_A=10,000, where all times are measured in generations. A value of f=0.0603 corresponds to D=0.046, which is the average observed in our real data.

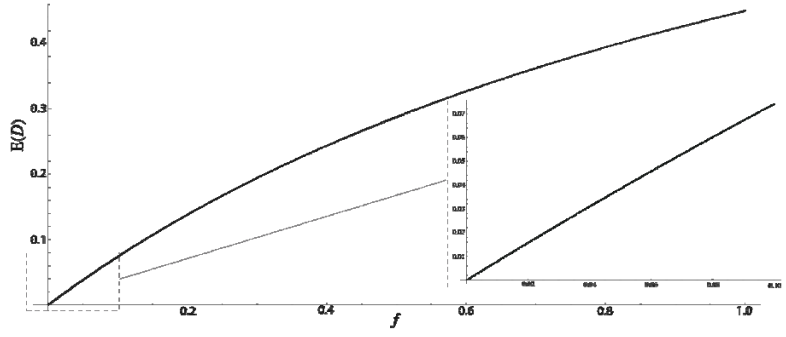


Figure S50: Plot of mixture proportion f against the expected value of $D(H_1,H_2,N,C)$, assuming t_{GF}=2,500 generations (70,000 years assuming 28 years per generation), t_H=3,000 generations (84,000 years), and t_N=10,000 generations (280,000 years).

It is important to recognize that inferences about the gene flow proportion f are sensitive to the choices of the other parameter values in the model.

Sensitivity to times of demographic events: The inference of f is relatively insensitive to choices of t_H and t_{GF}, but that it is highly sensitive to changes in t_N. For example, if t_N=20,000 generations instead of 10,000, f=0.0157. However, other features of the data described in Note S14 and below constrain t_N to be relatively close to 10,000.

Sensitivity to ancestral population sizes: We found that the inference of f is also dependent on assumptions about population size, and especially on N_N. For example, for a 20% decrease in population size, the effect on the inference of f is 8% for N_H, 4% for N_A, and 17% for N_N. We note that N_N is also the only parameter that we have no estimate of from studies of population genetic patterns in present-day humans, and thus uncertainty in this parameter is particularly crucial issue for our inferences. We conclude that by itself, the observed value of D can only broadly constrain f. However, by placing additional constraints on the model using other features of our data, we can tighten the constraints.

Expected fraction of SNPs discovered between two Yoruba chromosomes where Neandertal is derived

The next statistic that we used to constrain the model was the proportion R_N of sites where Neandertal carries the derived allele at SNPs discovered as differences between two Yoruba chromosomes. We present analytical expectations for this proportion in the analyses that follow, and a more extensive exploration of this statistic for more complex demographic models is presented in SOM 14.

We assume that two chromosomes are sampled from one human population, H_1 (e.g. Yoruba West Africans). We then restrict our analysis to sites at which one of them carries and ancestral nucleotide and the other carries the derived allele. In other words, a SNP in H_1 is ascertained by comparing two chromosomes from Yoruba. The problem is to compute the proportion of sites where a Neandertal chromosome carries the derived nucleotide. This model is the same as was simulated in SOM 14 in which the Yoruba population was assumed to be of constant size (cf. Figure S37). The set of trees in Figure S48 illustrate the events of interest, but with the modification that $H_{1.1}$ and $H_{1.2}$ replace H_1 and H_2 (representing the two chromosomes sampled from H_1). If we write the aligned bases at a given nucleotide in the order $H_{1.1}$, $H_{1.2}$, X, C, we are interested in the expected value of the ratio:

$$E(R_N) = \frac{P(ABBA) + P(BABA)}{P(ABBA) + P(BABA) + P(ABAA) + P(BAAA)}. \quad \text{(S19.6)}$$

To allow additional flexibility in the model, we allow the effective population size of H_1, denoted by N_Y, to differ from N_H after the divergence of H_1 and H_2 from each other. The denominator is equal to the probability that the two nucleotides sampled in H_1 differ. If μ is small, the denominator is equal to:

$$4\mu\left(N_Y + \left(1 - \frac{1}{2N_Y}\right)^{t_H} \left(N_H - N_Y\right) + \left(1 - \frac{1}{2N_H}\right)^{t_N - t_H} \left(N_A - N_H\right)\right). \quad \text{(S19.7)}$$

We now derive the numerator of $E(R_N)$. The only scenario that contributes to ABBA and BABA is the one in which $H_{1.1}$ or $H_{1.2}$ first coalesce with N (topologies II or III). The contribution to ABBA and BABA is then proportional to the length of the internal branch:

$$\Pr_N(\text{ABBA}) = \Pr_N(\text{BABA}) = \left(1 - \frac{1}{2N_Y}\right)^{t_H} \left(1 - \frac{1}{2N_H}\right)^{t_N - t_H} \frac{2N_A\mu}{3}. \quad \text{(S19.8)}$$

The expected proportion of alleles where Neandertal is derived is then twice the right hand side of Equation S19.8 divided by the expression in Equation S19.7. Therefore,

$$E(R_{\mathrm{N}}) = \frac{N_{\mathrm{A}}\left(1-\frac{1}{2N_{\mathrm{Y}}}\right)^{t_{\mathrm{H}}}\left(1-\frac{1}{2N_{\mathrm{H}}}\right)^{t_{\mathrm{N}}-t_{\mathrm{H}}}}{3\left(N_{\mathrm{Y}}+\left(1-\frac{1}{2N_{\mathrm{Y}}}\right)^{t_{\mathrm{H}}}\left(N_{\mathrm{H}}-N_{\mathrm{Y}}\right)+\left(1-\frac{1}{2N_{\mathrm{H}}}\right)^{t_{\mathrm{N}}-t_{\mathrm{H}}}\left(N_{\mathrm{A}}-N_{\mathrm{H}}\right)\right)}. \quad \text{(S19.9)}$$

Note that this result does not depend on the gene-flow parameters, t_{GF} and f, or the effective population size of Neandertals, N_{N}. With the parameter values used to generate Figure S50, we find that $E(R_{\mathrm{N}})$=20.2%, which is slightly larger than the observed value of R_{N}=18.0%.

Exploration of parameter space
We explored the sensitivity of the expected value of R_N to varying different parameters in the model (this is further explored in simulation-based studies in SOM 14).

We find that E(R_{N}) is highly informative about t_{N} as well as the effective population sizes. For example, if t_{N} is increased from 10,000 to 15,000 generations, $E(R_{\mathrm{N}})$=15.7%. If the population sizes are all increased from 10,000 to 15,000, $E(R_{\mathrm{N}})$=23.9%. We note that previous studies have carefully fit models for the demographic history of Yoruba, thereby placing meaningful constraints on the effective population sizes. Thus, we can use the observed R_{N} in conjunction with these demographic models to place constraints on t_{N}, as we do in SOM 14 to show that the divergence time of Neandertals and present-day humans t_{N} is less than about 12,000 (cf. Figure S37).

It is important to note that our analytical model is simplified compared with the simulation-based models of SOM 14. In particular, population growth in recent Yoruba history—which we did not incorporate in the models we explored our model but which we do explore in SOM 14—does have a modest effect, and is able to reduce $E(R_{\mathrm{N}})$ to the observed value of 18.0% while keeping t_N at around 10,000 generations (cf. Figure S37)

Expected values of three ratios of divergence times
To further constrain our model, we can also take advantage of three ratios of various genetic divergence times that we have measured empirically from the data in this study:

(Neandertal-H_2)/(Chimpanzee-H_2) per base pair divergence ratio
We denote by t_{C} the average genetic divergence time of chimpanzees and modern humans averaged across the genome (it is important to recognize that this is a genetic divergence time, which is guaranteed to be older than human-chimpanzee speciation time). We assume that t_{C} is large enough so that coalescence between H_2 and N is certain to occur more recently than t_{C}. The expected ratio is then:

$$Div(N,H_2) = \frac{(1-f)\left(t_{\mathrm{N}}+2N_{\mathrm{A}}\right)+f\left[t_{\mathrm{GF}}+2N_{\mathrm{N}}+2\left(N_{\mathrm{A}}-N_{\mathrm{N}}\right)\left(1-\frac{1}{2N_{\mathrm{N}}}\right)^{t_{\mathrm{N}}-t_{\mathrm{GF}}}\right]}{t_{\mathrm{C}}}. \quad \text{(S19.10)}$$

(Neandertal-H_1)/(Chimpanzee-H_1) per base pair divergence ratio
Under the recent gene flow model, the H_1 and *Neandertal* lineages are guaranteed to coalescence at least as old as time t_N, with an expected TMRCA of $2N_A$ above that point:

$$Div(N,H_1) = \frac{t_{\mathrm{N}}+2N_A}{t_{\mathrm{C}}} \quad \text{(S19.11)}$$

(H_2-H_1)/(Chimpanzee-H_1) per base pair divergence ratio
Under the recent gene flow model, the ratio is expected to equal:

$$Div(H_2,H_1)=\frac{(1-f)\left(t_H+2N_H+2(N_A-N_H)\left(1-\frac{1}{2N_N}\right)^{t_N-t_H}\right)+f(t_N+2N_A)}{t_C}. \quad \text{(S19.12)}$$

Using the same parameter values as in Figure S50, we obtain $Div(N,H_1)$=0.129, $Div(N,H_2)$=0.127 and $Div(H_1,H_2)$=0.101 if we set t_C=232,143 generations (corresponding to a divergence time of 6.5 million years if the average generation time is 28 years). These divergence ratios are sensitive to changes in all the parameters. In fact, they are almost linear functions of all the parameters except t_C and they are all inversely proportional to t_C. They are especially sensitive to changes in t_N. However, as described in SOM 14, t_N appears to be strongly constrained by R_N.

Combining measurements to constrain the proportion *f* of Neandertal ancestry in non-Africans
We next compared the analytical expectations of all the quantities we measured to the observed values, to constrain the demographic parameters consistent with the data (in particular, the mixture proportion *f*). To take account of the uncertainty in the observed values, we generated combinations of the 9 parameters of our model, N_Y, N_N, N_H, N_A, t_{GF}, t_H, t_N, t_C and f and kept combinations that resulted in expected values in the following ranges:

$$\begin{aligned}
&0.04<D(H_1, H_2, N, C)<0.05\\
&0.17<R_N<0.2\\
&0.129<Div(N, H_1)<0.131\\
&0.123<Div(N, H_2)<0.127\\
&0.080<Div(H_1, H_2)<0.093
\end{aligned} \quad \text{(S19.13)}$$

Some results are shown in Figure S51. They confirm the numerical results presented above that the effective size of Neanderthals is the primary determinant of *f*. Effective sizes of 5,000 or less are consistent with the estimates of *f* obtained in SOM 18.

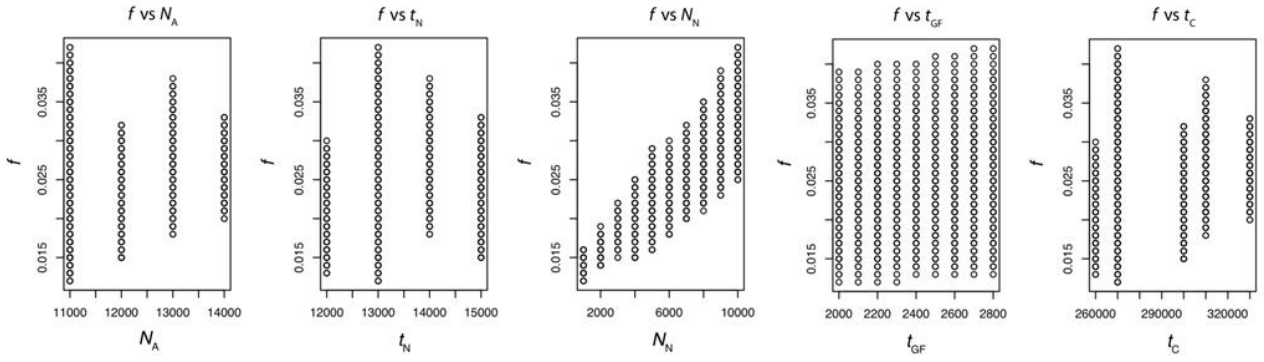


Figure S51: Graphs showing dependence of *f* on other parameters when the 5 observed quantities in Expression (S19.3) are satisfied.

Appendix 1: Effect of more complex demographic models on inferences

We have assumed as simple a demographic model as possible that still allows for gene flow from Neandertals to the ancestors of modern non-African populations.

A potential concern with the procedure that we have used to constrain the mixture proportion *f* is that the apparent precision of our constraints on *f* could be an artifact of having oversimplified the model. However, we note that many of the simplifications are not likely to affect our ability to constrain *f*.

1. Our analyses only analyzed one chromosome from each population. As a result, many demographic details do not matter. In particular, the population sizes and changes in population sizes of both H_1 (Africans) and H_2 (non-Africans) more recently than t_H make no difference to inferences, except in computing R_N (in which a history of population growth in Yoruba affects results; SOM 14).
2. We assume constant effective population sizes N_H, N_N, and N_A. However, if we had allowed them to fluctuate over time, it would not change the generality of our results, as these quantities can be viewed as effective sizes that account for fluctuations in population size.
3. We assumed that the cessation of gene flow between H_1 and H_2 or between the population ancestral to modern humans and the Neandertal population was instantaneous. However, if it as gradual, it would not be problematic for our inferences. A gradual population separation is relatively well captured for our purposes in terms of a change in the effective population size. Only the expected coalescence times of two or three lineages affect our results, and these quantities can be adequately capture by a modification in the effective population size.

We also considered the issue of how gene flow between African and non-African populations after their separation could affect our inferences. In theory, gene flow can have a substantial effect on *D*, because it creates additional opportunities for coalescence between African and Neandertal lineages implying that the value of *f* that corresponds to an observed *D* from our analysis is a minimum. In particular, substantial gene flow between Africans and non-Africans after their initial divergence would require higher influx of genes from Neandertals to non-Africans to produce the same *D*. Empirically, however, there is evidence that such continued gene flow has not affected our inferences. All three groups of non-Africans (X = French, Han and Papuan) show indistinguishable *D*-values in the statistic *D(African,X,Neandertal,Chimpanzee)*. Thus, gene flow after the historical divergence of these populations does not appear to have affected the *D*-statistics.

Our model in Figure S48 assumes for simplicity that the population that contributed genetic material to the ancestors of non-Africans falls in a clade with the Neandertal we sequenced. However, if that population instead was itself already mixed with modern humans, each unit of gene flow would have less of an impact on the *D*-statistics, and we would be underestimating the gene flow proportion. We have no way of discerning whether the population(s) contributing ancestry to non-Africans did in fact fall into a clade with Neandertals. Thus, our analyses should be viewed as providing a minimum on *f*.

Appendix 2: Expectation of $S(N_1, AFR, N_2, C)$ used in SOM 18

The method used above to derive the expectation of the *D* statistic can, with minor modification, be adapted to compute the model-based expectation of the closely related $S(N_1, AFR, N_2, C)$ statistic**,** which is the basis for estimating the admixture fraction, *f*, in Supplement 18.

Assume that N_1 and N_2 are two Neanderthal lineages. The two cases of interest are (i) N_1 and N_2 are two different Neanderthal bones that have been sequenced and (ii) N_1 is the Neanderthal lineage that entered Europeans as a result of admixture and N_2 is a lineage in one of the bones being sequenced.

For both (i) and (ii), let *T* be the most recent time at which N_1 and N_2 had the opportunity to coalesce. In case (i), *T* is the age of the older of the two bones, t_{Fossil}. In case (ii), $T=t_{GF}$, the time of the admixture event. Using the above argument for computing $E(D)$, the only genealogy that results in an excess of ABBA over BABA is the one in which N_1 and N_2 coalesce in the Neanderthal population more recently than it joins with the ancestor of humans at t_N (cf. Figure S49). The contribution to the expectation of $S(N_1, AFR, N_2, C)$ is the

length of the branch connecting the MRCA of the human and Neanderthal lineages to the two Neanderthal lineages. That expected length of that branch is

$$\tau(T) = t_{\mathrm{N}} + 2N_{\mathrm{A}} - T - \bar{t}(T) \tag{S19.14}$$

where $\bar{t}(T)$ is the expected time to coalescence of N_1 and N_2, given that they start at T and coalesce before t_{N}. That time is given by Eq. (S19.3) with T replacing t_{GF}:

$$\bar{t}(T) = 2N_{\mathrm{N}} - \frac{(t_{\mathrm{N}} - T)\left(1 - \frac{1}{2N_{\mathrm{N}}}\right)^{t_{\mathrm{N}} - T}}{1 - \left(1 - \frac{1}{2N_{\mathrm{N}}}\right)^{t_{\mathrm{N}} - T}}. \tag{S19.15}$$

Taking the expected values in Eq. (S18.3) yields

$$E(S(OOA, AFR, N, C) = \mu f \tau(t_{\mathrm{GF}}). \tag{S19.16}$$
$$E(N_1, AFR, N_2, C) = \mu \tau(t_{\mathrm{Fossil}}). \tag{S19.17}$$

Therefore, the expected value of the right hand side Eq. (S18.5) is

$$E(\hat{f}) = f \frac{\tau(t_{\mathrm{GF}})}{\tau(t_{\mathrm{Fossil}})} \tag{S19.18}$$

where t_{Fossil} is the age of the older of the two fossils from which sequence data is available.

The ratio depends on the times, but necessarily $t_{\mathrm{GF}} > t_{\mathrm{Fossil}}$ because the fossils are from after the admixture event. The ratio is close to 1 for reasonable parameter values. For example, $E(\hat{f})/f = 0.974$ for t_{GF}=2,500 generations, t_{Fossil}=1,500 generations, t_{N}=10,000 generations, N_{N}=5,000 and N_{A}=10,000. Thus, the correction to the estimate of the mixture proportion is no more than a few percent.

We conclude that SOM 18's estimate of the mixture proportion of 1.3-2.7% is entirely consistent with our model-based analysis.

References

S1. R. J. Roberts, T. Vincze, J. Posfai, D. Macelis, REBASE--enzymes and genes for DNA restriction and modification *Nucleic acids research* **35**, D269 (2007).

S2. K. Prüfer *et al.*, PatMaN: rapid alignment of short sequences to large databases *Bioinformatics (Oxford, England)* **24**, 1530 (2008).

S3. A. W. Briggs *et al.*, Targeted retrieval and analysis of five Neandertal mtDNA genomes *Science (New York, N.Y* **325**, 318 (2009).

S4. N. Rohland, M. Hofreiter, Comparison and optimization of ancient DNA extraction *Biotechniques* **42**, 343 (2007).

S5. R. E. Green *et al.*, Analysis of one million base pairs of Neanderthal DNA *Nature* **444**, 330 (2006).

S6. R. E. Green *et al.*, A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing *Cell* **134**, 416 (2008).

S7. J. Krause *et al.*, Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae *Nature* **439**, 724 (2006).

S8. M. Meyer, U. Stenzel, M. Hofreiter, Parallel tagged sequencing on the 454 platform *Nat Protoc* **3**, 267 (2008).

S9. J. Krause *et al.*, A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia *Curr Biol* (2009).

S10. M. Margulies *et al.*, Genome sequencing in microfabricated high-density picolitre reactors *Nature* **437**, 376 (2005).

S11. M. Meyer *et al.*, From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing *Nucleic Acids Res* **36**, e5 (2008).

S12. D. R. Bentley *et al.*, Accurate whole human genome sequencing using reversible terminator chemistry *Nature* **456**, 53 (2008).

S13. M. Kircher, U. Stenzel, J. Kelso, Improved base calling for the Illumina Genome Analyzer using machine learning strategies *Genome Biol* **10**, R83 (2009).

S14. Z. Zhang, S. Schwartz, L. Wagner, W. Miller, A greedy algorithm for aligning DNA sequences *J Comput Biol* **7**, 203 (2000).

S15. B. Paten, J. Herrero, K. Beal, E. Birney, Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment *Bioinformatics* **25**, 295 (2009).

S16. B. Paten, J. Herrero, K. Beal, S. Fitzgerald, E. Birney, Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs *Genome Res* **18**, 1814 (2008).

S17. B. Paten *et al.*, Genome-wide nucleotide-level mammalian ancestor reconstruction *Genome Res* **18**, 1829 (2008).

S18. A. Morgulis *et al.*, Database indexing for production MegaBLAST searches *Bioinformatics* **24**, 1757 (2008).

S19. E. W. Sayers *et al.*, Database resources of the National Center for Biotechnology Information *Nucleic Acids Res* **37**, D5 (2009).

S20. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools *Bioinformatics* **25**, 2078 (2009).

S21. B. T. Lahn, D. C. Page, Four evolutionary strata on the human X chromosome *Science* **286**, 964 (1999).

S22. H. Skaletsky *et al.*, The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes *Nature* **423**, 825 (2003).

S23. J. P. Noonan *et al.*, Sequencing and analysis of Neanderthal genomic DNA *Science* **314**, 1113 (2006).

S24. J. Krause *et al.*, The derived FOXP2 variant of modern humans was shared with Neandertals *Curr Biol* **17**, 1908 (2007).

S25. R. W. Schmitz *et al.*, The Neandertal type site revisited: interdisciplinary investigations of skeletal remains from the Neander Valley, Germany *Proc Natl Acad Sci U S A* **99**, 13342 (2002).

S26. S. Levy *et al.*, The diploid genome sequence of an individual human *PLoS Biol* **5**, e254 (2007).

S27. K. A. Frazer *et al.*, A second generation human haplotype map of over 3.1 million SNPs *Nature* **449**, 851 (2007).
S28. A. W. Briggs *et al.*, Patterns of damage in genomic DNA sequences from a Neandertal *Proc Natl Acad Sci U S A* **104**, 14616 (2007).
S29. R. R. Hudson, Generating samples under a Wright-Fisher neutral model of genetic variation *Bioinformatics* **18**, 337 (2002).
S30. J. D. Wall, K. E. Lohmueller, V. Plagnol, Detecting ancient admixture and estimating demographic parameters in multiple human populations *Mol Biol Evol* **26**, 1823 (2009).
S31. C. Lalueza-Fox *et al.*, Neandertal evolutionary genetics: mitochondrial DNA data from the iberian peninsula *Mol Biol Evol* **22**, 1077 (2005).
S32. L. V. Golovanova, J. F. Hoffecker, V. M. Kharitonov, G. P. Romanova, Mezmaiskaya Cave: A Neanderthal Occupation in the Northern Caucasus *Current Anthropology* **40**, 77 (1999).
S33. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform *Bioinformatics* **25**, 1754 (2009).
S34. R. E. Green *et al.*, The Neandertal genome and ancient DNA authenticity *Embo J* **28**, 2494 (2009).
S35. A. L. Price *et al.*, Sensitive detection of chromosomal segments of distinct ancestry in admixed populations *PLoS Genet* **5**, e1000519 (2009).
S36. N. Patterson, D. J. Richter, S. Gnerre, E. S. Lander, D. Reich, Genetic evidence for complex speciation of humans and chimpanzees *Nature* **441**, 1103 (2006).
S37. D. R. Bentley *et al.*, Accurate whole human genome sequencing using reversible terminator chemistry *Nature* **456**, 53 (2008).
S38. M. Hofreiter, V. Jaenicke, D. Serre, A. Haeseler Av, S. Paabo, DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA *Nucleic Acids Res* **29**, 4793 (2001).
S39. R. Grantham, Amino acid difference formula to help explain protein evolution *Science* **185**, 862 (1974).
S40. W. H. Li, C. I. Wu, C. C. Luo, A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes *Mol Biol Evol* **2**, 150 (1985).
S41. A. G. Clark *et al.*, Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios *Science* **302**, 1960 (2003).
S42. E. Oancea *et al.*, TRPM1 forms ion channels associated with melanin content in melanocytes *Sci Signal* **2**, ra21 (2009).
S43. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium *Nat Genet* **25**, 25 (2000).
S44. R. C. Gentleman *et al.*, Bioconductor: open software development for computational biology and bioinformatics *Genome Biol* **5**, R80 (2004).
S45. R Development Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2008), pp.
S46. L. Duret, N. Galtier, Biased gene conversion and the evolution of mammalian genomic landscapes *Annu Rev Genomics Hum Genet* **10**, 285 (2009).
S47. L. Duret, N. Galtier, Comment on "Human-specific gain of function in a developmental enhancer" *Science* **323**, 714; author reply 714 (2009).
S48. N. Galtier, L. Duret, Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution *Trends Genet* **23**, 273 (2007).
S49. J. P. Noonan, Regulatory DNAs and the evolution of human development *Curr Opin Genet Dev* **19**, 557 (2009).
S50. S. Prabhakar *et al.*, Human-specific gain of function in a developmental enhancer *Science* **321**, 1346 (2008).

S51. C. P. Bird *et al.*, Fast-evolving noncoding sequences in the human genome *Genome Biol* **8**, R118 (2007).
S52. E. C. Bush, B. T. Lahn, A genome-wide screen for noncoding elements important in primate evolution *BMC Evol Biol* **8**, 17 (2008).
S53. K. S. Pollard *et al.*, Forces shaping the fastest evolving regions in the human genome *PLoS Genet* **2**, e168 (2006).
S54. K. S. Pollard *et al.*, An RNA gene expressed during cortical development evolved rapidly in humans *Nature* **443**, 167 (2006).
S55. S. Prabhakar, J. P. Noonan, S. Paabo, E. M. Rubin, Accelerated evolution of conserved noncoding sequences in humans *Science* **314**, 786 (2006).
S56. P. D. Evans, N. Mekel-Bobrov, E. J. Vallender, R. R. Hudson, B. T. Lahn, Evidence that the adaptive allele of the brain size gene microcephalin introgressed into Homo sapiens from an archaic Homo lineage *Proc Natl Acad Sci U S A* **103**, 18178 (2006).
S57. S. E. Ptak *et al.*, Linkage disequilibrium extends across putative selected sites in FOXP2 *Mol Biol Evol* **26**, 2181 (2009).
S58. H. Stefansson *et al.*, A common inversion under selection in Europeans *Nat Genet* **37**, 129 (2005).
S59. A. W. Briggs *et al.*, Primer extension capture: targeted sequence retrieval from heavily degraded DNA sources *J Vis Exp*, 1573 (2009).
S60. J. A. Bailey *et al.*, Recent segmental duplications in the human genome *Science* **297**, 1003 (2002).
S61. C. Alkan *et al.*, Personalized copy number and segmental duplication maps using next-generation sequencing *Nat Genet* **41**, 1061 (2009).
S62. D. R. Bentley *et al.*, Accurate whole human genome sequencing using reversible terminator chemistry *Nature* **456**, 53 (2008).
S63. C. S. a. A. Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome *Nature* **437**, 69 (2005).
S64. D. A. Wheeler *et al.*, The complete genome of an individual by massively parallel DNA sequencing *Nature* **452**, 872 (2008).
S65. J. Wang *et al.*, The diploid genome sequence of an Asian individual *Nature* **456**, 60 (2008).
S66. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, Segmental duplications: organization and impact within the current human genome project assembly *Genome Res* **11**, 1005 (2001).
S67. Z. Jiang *et al.*, Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution *Nat Genet* **39**, 1361 (2007).
S68. M. C. Zody *et al.*, Evolutionary toggling of the MAPT 17q21.31 inversion region *Nat Genet* (2008).
S69. J. Hardy *et al.*, Evidence suggesting that Homo neanderthalensis contributed the H2 MAPT haplotype to Homo sapiens *Biochem Soc Trans* **33**, 582 (2005).
S70. S. Myers, L. Bottolo, C. Freeman, G. McVean, P. Donnelly, A fine-scale map of recombination rates and hotspots across the human genome *Science* **310**, 321 (2005).
S71. G. McVicker, D. Gordon, C. Davis, P. Green, Widespread genomic signatures of natural selection in hominid evolution *PLoS Genet* **5**, e1000471 (2009).
S72. J. A. Drake *et al.*, Conserved noncoding sequences are selectively constrained and not mutation cold spots *Nat Genet* **38**, 223 (2006).
S73. N. L. Kaplan, R. R. Hudson, C. H. Langley, The "hitchhiking effect" revisited *Genetics* **123**, 887 (1989).
S74. P. Khaitovich *et al.*, Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees *Science* **309**, 1850 (2005).
S75. M. Somel *et al.*, Transcriptional neoteny in the human brain *Proc Natl Acad Sci U S A* **106**, 5743 (2009).
S76. J. L. Kelley, J. Madeoy, J. C. Calhoun, W. Swanson, J. M. Akey, Genomic signatures of positive selection in humans and the limits of outlier approaches *Genome Res* **16**, 980 (2006).

S77. K. M. Teshima, G. Coop, M. Przeworski, How reliable are empirical genomic scans for selective sweeps? *Genome Res* **16**, 702 (2006).
S78. J. M. Akey, Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res* **19**, 711 (2009).
S79. P. C. Sabeti *et al.*, Positive natural selection in the human lineage *Science* **312**, 1614 (2006).
S80. C. S. Carlson *et al.*, Genomic regions exhibiting positive selection identified from dense genotype data *Genome Res* **15**, 1553 (2005).
S81. J. Kaiser, DNA sequencing. A plan to capture human diversity in 1000 genomes *Science* **319**, 395 (2008).
S82. K. R. Thornton, J. D. Jensen, Controlling the false-positive rate in multilocus genome scans for selection *Genetics* **175**, 737 (2007).
S83. S. F. Schaffner *et al.*, Calibrating a coalescent simulation of human genome sequence variation *Genome Res* **15**, 1576 (2005).
S84. A. Keinan, J. C. Mullikin, N. Patterson, D. Reich, Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans *Nat Genet* **39**, 1251 (2007).
S85. R. Burgess, Z. Yang, Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors *Mol Biol Evol* **25**, 1979 (2008).
S86. J. M. Kidd *et al.*, Mapping and sequencing of structural variation from eight human genomes *Nature* **453**, 56 (2008).
S87. Z. Ning, A. J. Cox, J. C. Mullikin, SSAHA: a fast search method for large DNA databases *Genome Res* **11**, 1725 (2001).
S88. D. Reich, K. Thangaraj, N. Patterson, A. L. Price, L. Singh, Reconstructing Indian population history *Nature* **461**, 489 (2009).
S89. H. Kunsch, The Jackknife and the Bootstrap for General Stationary Observations *The Annals of Statistics* **17**, 1217 (1989).
S90. J. D. Wall, S. K. Kim, Inconsistencies in Neanderthal genomic DNA sequences *PLoS Genet* **3**, 1862 (2007).
S91. J. Z. Li *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation *Science* **319**, 1100 (2008).
S92. E. S. Lander *et al.*, Initial sequencing and analysis of the human genome *Nature* **409**, 860 (2001).
S93. D. Reich *et al.*, Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene *PLoS Genet* **5**, e1000360 (2009).
S94. M. W. Smith *et al.*, A high-density admixture map for disease gene discovery in african americans *Am J Hum Genet* **74**, 1001 (2004).
S95. D. A. Hinds *et al.*, Whole-genome patterns of common DNA variation in three human populations *Science* **307**, 1072 (2005).
S96. R. R. Sokal, C. D. Michener, A statistical method for evaluating systematic relationships *University of Kansas Science Bulletin* **38**, 1409 (1958).
S97. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, A map of recent positive selection in the human genome *PLoS Biol* **4**, e72 (2006).