Supplementary Materials for

A High-Coverage Genome Sequence from an Archaic Denisovan Individual

Matthias Meyer,* Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, Flora Jay, Kay Prüfer, Cesare de Filippo, Peter H. Sudmant, Can Alkan, Qiaomei Fu, Ron Do, Nadin Rohland, Arti Tandon, Michael Siebauer, Richard E. Green, Katarzyna Bryc, Adrian W. Briggs, Udo Stenzel, Jesse Dabney, Jay Shendure, Jacob Kitzman, Michael F. Hammer, Michael V. Shunkov, Anatoli P. Derevianko, Nick Patterson, Aida M. Andrés, Evan E. Eichler, Montgomery Slatkin, David Reich,* Janet Kelso, Svante Pääbo*

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

Published 30 August 2012 on *Science* Express
DOI: 10.1126/science.1224344

This PDF file includes

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/science.1224344/DC1)

1224344s2.xls (Tables S44, S48, S49, S50, S54 and S55)

Materials and methods

Table of Contents

Note 1: Sampling and extraction of side fractions

Marie-Theres Gansauge, Qiaomei Fu and Matthias Meyer*

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

In 2010, two extracts (E236 and E245), each with a volume of 100 µl, were prepared from 40mg of the Denisovan phalanx (*2*) (see Figure S1 for an overview of bone material usage). For initial screening of the bone, sequencing libraries were prepared without prior excision of deoxyuracils from the template strands. When analysis of captured mtDNA fragments revealed that the sample came from a previously unknown hominin (*4*) and that the DNA fragments contained patterns of nucleotide misincorporations typical of ancient DNA (*9*), fractions of the extracts were treated with uracil-DNA-glycosylase (UDG) and endonuclease VIII (*29*) to produce two libraries (SL3003 and SL3004) that were used to generate the nuclear DNA sequences that were analyzed and published earlier (*2*).

We routinely freeze all relevant side fractions arising during DNA extraction. The silica-based DNA extraction process used (*30*) involves three steps: (i) sample lysis in an EDTA/proteinase K buffer, (ii) DNA binding to a silica suspension in the presence of chaotropic salts (iii) ethanol wash and elution in low-salt buffer. From each extraction, the following side fractions are kept: (i) any undissolved bone powder; (ii) binding supernatant; (iii) the silica pellet.

The bone powder from the phalanx was completely dissolved. We used the binding supernatant and the silica pellet of E245 in an attempt to recover additional molecules. In a falcon tube, 30 µl of new silica suspension were added to 5 ml of 'left-over' binding supernatant. After rotating for three hours in the dark, the silica was pelleted by centrifugation and washed twice in washing buffer. DNA was eluted in 28 µl TE buffer and transferred to the old silica pellet for re-elution, thereby combining the molecules recovered from both side fractions (ii and iii) of E245.

Note 2: Library preparation and sequencing

Marie-Theres Gansauge, Qiaomei Fu and Matthias Meyer*

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

Method summary

A schematic overview of the single-stranded library preparation method as well as the two double-stranded methods previously used for library preparation from ancient DNA (*31, 32*) is provided in Figure S2. In more detail, the single-stranded method involves the following steps:

(i) Input DNA is treated with a mixture of uracil-DNA-glycosylase (UDG) and endonuclease VIII, which removes uracils from DNA strands and converts resulting abasic sites into single-nucleotide gaps. This treatment is optional.

(ii) To this reaction, a heat-labile phosphatase is added, which removes both 5'- and 3'-phosphates from DNA ends if present.

(iii) The double-stranded molecules are heat-denatured and again treated with phosphatase to complete dephosphorylation.

(iv) After heat-inactivation of the phosphatase, adaptor oligonucleotides of ten bases, which carry 5'-phosphates and 3'-biotin-linker arms, are ligated in the presence of polyethylene glycol (PEG) to the 3'-ends of the input molecules using CircLigase II^{TM}, an enzyme that preferentially circularizes single-stranded DNA (*33*), but when presented with donor and acceptor molecules with only one ligatable end, efficiently achieves end-to-end ligation (*34*).

(v) Ligation products as well as excess adaptors are subsequently bound to streptavidin-coated magnetic beads.

(vi) A 5'-tailed primer, which carries a priming site for later amplification, is hybridized to the adaptor oligonucleotide, and the original template strand is copied using Bst polymerase (large fragment).

(vii) *T4* DNA polymerase is used to remove 3'-overhangs generated during the primer extension reaction.

(viii) *T4* DNA ligase is used to ligate the 5'-phosphorylated end of a double stranded adaptor to the 3'-end of the newly copied strand. The 3'-end of the adaptor carries a dideoxy modification to prevent adaptor self-ligation. Between all enzymatic reactions, beads are intensely washed, also at elevated temperature, to remove extension primers hybridized to excess adaptor oligonucleotides.

(ix) Finally, the original template molecules and the library molecules, i.e. copies with short adaptor sequences attached to each end, are released by heat-destruction of the beads.

Libraries were prepared from extracts E236 and E245 as well as from the molecules recovered from the side fractions of extraction E245 using the single-stranded method (Table S1). In addition, water controls were carried through the library preparation process. The number of molecules in each library was estimated by qPCR (*35, 36*). According to these estimates, the sample extracts produced at least one order of magnitude more molecules than the water controls, indicating that less than 10% of library molecules are derived from artifacts. Full-length adaptor sequences, carrying sample-specific sequence tags (indexes), were then added to both ends of the library molecules by amplification with 5'-tailed primers (*11*). For a more efficient use of sequencing capacity, these amplified libraries were separated on polyacrylamide gels, and the fractions of library molecules with inserts larger than ~40 bp were excised. Sequence data was produced from amplified libraries both with and without size-fractionation, using a protocol for double-index sequencing described elsewhere (*35*). A special sequencing primer was used for the first read, because the P5 adaptor sequence was truncated by five bases compared to the design described previously (*11*) to avoid interactions between the highly self-complementary P5 and P7 adaptors during single-strand library preparation.

In addition to sequencing the libraries prepared with the new single-stranded method, deeper sequencing was performed for libraries prepared in the previous study with the double-stranded method (SL3003 and SL3004) (*2*).

Method details

Sequences of oligonucleotides used in this work are listed in Table S2.

Deoxyuracil removal, dephosphorylation, heat denaturation and single-stranded adaptor ligation

Libraries were prepared in an ancient DNA clean room from a maximum volume of 28.5 µl DNA extract as follows. In a 0.5 ml safe-lock tube (Eppendorf, Hamburg, Germany), DNA extract was supplemented with water – if necessary – to reach a total volume of 28.5 µl. After adding 8 µl CircLigase II 10x reaction buffer (Epicentre, Madison, USA), 4 µl 50 mM $MnCl_2$ (Epicentre) and 0.5 µl USER enzyme mix (New England Biolabs, Ipswich, USA), the reaction was mixed and incubated at 37°C for 1 h in a thermal cycler with a heated lid. Then, 1 µl (1 U) FastAP (Fermentas, Burlington, Canada) was added and the reaction was incubated for 5 min at 37°C and 2 min at 95°C. The reaction tube was directly placed into an ice-water bath and another 1 µl FastAP was added. The reaction was incubated for 10 min at 37°C, 10 min at 75°C and 2 min at 95 °C and the tube was chilled in an ice-water bath. Then, 32 µl 50% PEG-4000 (Sigma-Aldrich, St. Loius, USA) and 1 µl 10 µM adaptor oligo CL78-2 were added and the reagents

were mixed by intense vortexing. After adding 4 µl CircLigase II (Epicentre), the reaction was intensely mixed again, incubated for 1 h at 60°C in a thermal cycler, and cooled down to 4°C. Before freezing overnight, 4 µl 1% Tween-20 were added to the reaction to avoid binding of DNA to the tube walls.

Immobilization of ligation products on streptavidin beads

Next day, 20 µl of MyOne C1 beads (Life Technologies, Carlsbad, USA) were washed twice with 1xBWT+SDS (1 M NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.05% Tween-20, 0.5% SDS). This and all subsequent wash steps were performed using a magnetic rack for separating beads from supernatant. Beads were resuspended in 250 µl 1xBWT+SDS and transferred to a 1.5ml-siliconized tube (Sigma-Aldrich). The reaction mix from the previous day was thawed, heated to 95°C for 2 min, chilled in an ice-water bath and added to the beads. The tube was rotated for 20 min at room temperature. The supernatant was removed and the beads were washed once with 200 µl 0.1xBWT+SDS (100 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.05% Tween-20, 0.5% SDS) and once with 200 µl 0.1xBWT (100 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.05% Tween).

Primer annealing and extension

The supernatant was removed and the beads were resuspended in 47 µl of a reaction mixture containing 40.5 µl water, 5 µl 10x Thermopol buffer (New England Biolabs), 0.5 µl 25 mM each dNTP (Fermentas) and 1 µl 100 µM CL9 (extension primer). The tube was incubated for 2 min at 65°C in a thermal cycler, immediately chilled in an ice-water bath and transferred to another thermal cycler pre-cooled to 15°C. While placed on the cycler, the tube was opened and 3 µl (24 U) *Bst* DNA polymerase, large fragment, (New England Biolabs) were added. The reaction was incubated first ramping from 15°C to 37°C at a speed of 1°C/minute and then holding at 37°C for 5 min. Beads were kept in suspension by gently mixing every three minutes. The supernatant was discarded, and the beads were washed with 200 µl 0.1xBWT+SDS. Beads were resuspended in 50 µl stringency-wash (0.1x SSC buffer (Sigma-Aldrich), 0.1% SDS) and incubated at 30°C for 3 min in a thermal cycler and then washed with 200 µl 0.1xBWT.

Removal of 3'-overhangs

Beads were resuspended in 99 µl of a reaction mix containing 86.1 µl water, 10 µl 10x Tango buffer (Fermentas), 2.5 µl 1% Tween-20 and 0.4 µl 25 mM each dNTP. After adding 1 µl (5 U) *T4* DNA polymerase (Fermentas), the reaction was incubated for 15 min at 25°C in a thermal cycler. Beads were kept suspended by gently mixing every three minutes. Beads were washed with 0.1xBWT+SDS, stringency wash and 0.1xBWT as described above.

Blunt-end ligation of second adaptor and library elution

A 100 µM solution of double-stranded DNA adaptor was generated by hybridizing two oligonucleotides (CL53 and CL73) as follows: In a PCR reaction tube, 20 µl 500 µM CL53, 20 µl 500 µM CL73, 9.5 µl TE (10 mM Tris-HCl pH 8.0) and 0.5 µl 5 M NaCl were combined. This mixture was incubated for 10 seconds at 95°C in a thermal cycler and cooled to 14°C at a speed of 0.1°C/s. Final concentration of 100 µM was reached by dilution with 50 µl TE.

Beads were resuspended in 98 µl of a reaction mix containing 73.5 µl water, 10 µl 10x *T4* DNA ligase buffer (Fermentas), 10 µl 50% PEG-4000 (Fermentas), 2.5 µl 1% Tween-20 and 2 µl 100 µM adaptor CL53/73. After thorough mixing, 2 µl (10 U) *T4* DNA ligase (Fermentas) were added, and the reaction was incubated for 1 h at 25°C in a thermal cycler. Beads were kept suspended by gently mixing every three minutes. Beads were washed with 0.1xBWT+SDS, stringency wash and 0.1xBWT as described above, resuspended in 20 µl EBT buffer (10 mM Tris-HCl pH 8.0, 0.05% Tween-20) and transferred to single-cap PCR tubes. After incubation for 10 min at 95°C in a thermal cycler with heated lid, the supernatant, containing the single-stranded library, was collected in a fresh tube.

Library amplification

A 40-fold dilution with EBT was generated from 1 µl of each library. Copy number of the library was determined by qPCR in replicates, using 1 µl of the dilution as template for each measurement. Based on amplification plots, an optimal PCR cycle number was determined, which was used to amplify the remaining 19 µl of library avoiding PCR plateau and hence the formation of heteroduplexes. Amplification reactions were set up in the clean room, using AccuPrime *Pfx* DNA polymerase (Life Technologies) with reaction parameters described elsewhere (*37*). For each library (or fraction of library), a unique combination of indexed primers was used. Cycling and subsequent work was performed in a post-PCR laboratory. PCR products were purified using the MinElute PCR purification kit (Qiagen, Hilden, Germany) and eluted in 25 µl TE. Amplification success was verified on a DNA-1000 chip using Agilent's Bioanalyzer 2100. From the amplified library, 4 µl were taken as template for a second round of amplification in a 100 µl PCR reaction using the universal primer pair IS5 and IS6 (*35*) and Herculase II Fusion DNA polymerase (Agilent) under the conditions described elsewhere (*37*).

PCR plateau was again avoided. DNA was again purified using the MinElute PCR purification kit and eluted in 20 µl TE. Library concentration was determined using the Bioanalyzer 2100. The amplified library was either directly used for sequencing or size-fractionated on an acrylamide gel.

Size fractionation

5 µl of amplified library as well as markers of appropriate size were loaded onto a pre-cast 10% acrylamide gel (Criterion 15% TBE gel, BioRad, Hercules, USA). DNA was separated for 2.5 h at 200 volts. After breaking the chamber, the gel was stained for 10 min with SybrSafe (Life Technologies). A gel slice containing library molecules with inserts larger than ~40 bp was excised and transferred to a 0.5-ml tube, into which a whole had been poked with a hot needle. The 0.5-ml tube was then placed on top of a 2-ml tube and centrifuged for 2 min at maximum speed in a table-top centrifuge to fragment the gel. After adding 350 µl diffusion buffer (100 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA), elution was carried out over night at room temperature. Next day, the supernatant was taken off and purified using the MinElute PCR purification kit. Recovery of library molecules was verified by qPCR, the eluate was amplified using primers IS5 and IS6 as described above, and amplification success determined using the Bioanalyzer 2100.

Sequencing

Using Illumina's Genome Analyzer IIx platform, 76-bp paired-end reads as well as two index reads were generated for the new libraries following the methodology described elsewhere (*11*), but replacing the sequencing primer of the first read by CL72. Otherwise, the manufacturer's instructions for multiplex sequencing on the Genome Analyzer IIx platform (FC-104-50xx/v5 sequencing chemistry and PE-203-4001/v4 and PE-203-5001/v5 cluster generation kits) were followed with only one exception; an indexed control φX 174 library was spiked into each lane, yielding a fraction of 1% control reads in each lane.

Deeper sequencing of the libraries prepared in the previous study (SL3003 and SL3004) (*2*) was performed with 101-bp paired-end reads and a single index read using FC-104-40xx/v4 sequencing chemistry and PE-203-4001/v4 cluster generation kits. Spike-in of a φX 174 control library yielded about 2-3% control reads in each lane.

Note 3: Comparison of the single-stranded and double-stranded library preparation methods

Matthias Meyer*

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

Sequence yield

The amount of sequence that is generated from a library is limited both by library complexity, i.e. its content of unique molecules, and sequencing depth. In order to test whether it is possible to exhaustively sequence all unique molecules in a library by oversampling, we sequenced library SL3004, which was prepared with the double stranded protocol previously (*2*), to an average of 12.2 sequences per molecule. No complete saturation was reached (Figure S3), i.e. even at this sequencing depth new molecules are discovered. This is probably the result of biases in amplification and sequencing, which would confound library complexity estimates obtained from fitting a simple exponential model. The following comparisons of sequence yield therefore rely on the assumption that these biases are similar among experiments.

A library prepared from extract E245 with the single-stranded method (B1108) produced about six times more sequence per microliter of extract although it was much less exhaustively sequenced than SL3004, which had been prepared from the same extract with the double-stranded method (Table S3). Library B1133, which was prepared from extract E236 using the single-stranded method, produced 23 times more sequence per microliter of extract than library SL3003, although the latter had been sequenced more exhaustively. As the first number in particular underestimates the gain in recovery (due to differences in sequencing depth), we conclude that the single-stranded method increased the recovery of library molecules by approximately one order of magnitude.

Fragment size distribution and percentage of endogenous DNA

Irrespective of the library preparation method, the Denisovan phalanx shows an unusually low level of microbial contamination compared to other samples from non-permafrost environments. In addition, the fraction of sequences that can be mapped to the human genome increases with fragment length, indicating that microbial DNA is more fragmented than endogenous DNA in this sample (Figure S4, dashed red line).

Since both library methods were applied to the same DNA extracts, sequencing results are directly comparable. The fragment size distribution obtained with the single-stranded method is wider and flatter, indicating an improved recovery of long molecules in particular. In addition, the single-stranded method recovered molecules <30 bp, which are almost completely lost with

the double-stranded method (Figure S4, solid lines). As an undesired side-effect it also produced a substantial background of library molecules with very short (~10 bp) artifactual inserts, originating from damaged or incompletely synthesized adaptor oligonucleotides. For the bulk of sequencing, we removed these artifacts together with short-insert library molecules by gel-size-fractionation.

Patterns of miscoding DNA damage and DNA fragmentation

Using untrimmed Denisovan sequence alignments to the human reference genome (filtered for map quality >= 30), we determined the average frequency of each type of substitution at each alignment position (Figure S5). In a CpG context the majority of cytosines are methylated in vertebrate genomes, and, when deaminated, leave thymine instead of uracil. Thus, despite treatment with uracil-DNA-glycosylase and endonuclease VIII, deaminated cytosines are in the majority of cases not removed. With the single-stranded library preparation method, an elevation of CpG->TpG substitutions at both fragment ends indicates the presence of 5' and 3' single-stranded DNA overhangs carrying 5-methyl-cytosines. With the double-stranded method the same pattern is found at 5' ends, but the reverse complement pattern, an excess of CpG->CpA substitutions, is seen at 3' ends. These substitutions are known to be artifacts of blunt-end repair (*9*), during which 3' overhangs (carrying deaminated cytosines) are removed and 5' overhangs are filled (causing G->A substitutions on the opposite strand).

With the single-stranded method, deoxyuracil excision is performed as the first step of the protocol. However, even in non-CpG context deoxyuracils prevail at 5' terminal nucleotides as well as the 3' terminal and penultimate nucleotides. This observation is in line with previous experimental work, which showed that 3'-terminal nucleotides and, in the absence of terminal phosphate groups, 5'-terminal and 3'-penultimate nucleotides are not efficiently excised by *E.coli* UDG (*38*). Since deoxyuracil removal is performed simultaneously with blunt-end repair in the double-stranded method, damage removal is nearly fully efficient. In the interior of molecules, deoxyuracil was effectively removed with both library preparation methods, and C->T substitutions remain elevated only in CpG context (see Figure S5). Note that CpG sites evolve so fast in mammalian genomes that G->A substitutions, which reflect *bona fide* substitutions in the genome at CpG sites, occur at ~1.2%. When this is subtracted from the ~2.2% C->T substitutions, the fraction of deamination-induced substitutions in CpG-context can be gauged to ~1% .

In addition to deamination, patterns of DNA fragmentation can be inferred from the reference base composition around molecule break points. At 5'-ends, patterns obtained from sequences generated with the two methods closely resemble each other and patterns described previously (*9, 39, 40*) (see Figure S6). Most notably, an increased frequency of guanine at the position immediately preceding 5' ends indicates frequent strand breakage after guanine. Increased cytosine frequency at the same position is an artifact of deoxyuracil excision, which creates strand breaks around deaminated cytosines (*29*). As described above, at 3'-ends, the double-

stranded method merely generates a reverse complement pattern. In contrast, the single-stranded method does not require manipulation of 3' ends by a polymerase and therefore preserves actual DNA fragmentation patterns. The frequency of guanine is strongly elevated at the position immediately succeeding 3' ends, suggesting frequent strand breakage before guanine. The evidence of strongly increased strand breakage 5' and 3' to guanines indicates the existence of a molecular mechanism, possibly depurination, which eliminates guanine from ancient DNA strands while generating single-nucleotide gaps. Further interpretation of fragmentation patterns is complicated by the additional activities of endonuclease VIII, which creates gaps upon removal of abasic sites and several damaged pyrimidines (*41*), some of which may be present in ancient DNA. Single-stranded library preparation should represent a powerful tool for further elucidating the mechanism underlying ancient DNA decay in future work.

Base composition

We have previously reported a GC-bias in sequences generated from libraries prepared with the double-stranded method (*1, 42*), including the draft sequence of the Denisovan genome (*2*). This bias decreases with fragment size (see Figure S7). With the single-stranded library preparation method, we find a similar decrease in GC-content with fragment size but the GC-content is below genome average throughout the range of fragment sizes. Thus, the high-coverage genome sequence described here is AT-biased. In addition, A and T, as well as G and C, are not equally represented. Most notably, there is an overrepresentation of A in short fragments. It is unclear whether similar biases also exist in libraries prepared with the double-stranded method, because strand information is absent there. Further research will be needed to disentangle biases in extraction and library preparation from actual patterns of DNA degradation in this and other ancient samples.

Note 4: Processing and mapping raw sequence data from Denisova

Martin Kircher

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

Base calling and raw sequence processing

Denisovan sequencing runs generated from the new libraries (B1107, B1108, B1109, B1110, and B1133) were analyzed starting using the BCL and CIF intensity files from the Illumina Genome Analyzer RTA 1.9 software. Sequencing runs of the previously prepared SL3003 and SL3004 were analyzed starting from the QSEQ sequence files and CIF intensity files from the Illumina Genome Analyzer RTA 1.6 software. Raw reads showing the control index ('TTGCCGC') in the first index read were aligned to the φX 174 reference sequence to obtain a training data set for the Ibis base caller (*43*), which was then used to recall bases and quality scores of each run from the CIF files.

The so-obtained raw reads were filtered for the presence of the correct library index sequences, allowing for one substitution and/or the skipping the first base of the index (*35*). For the double-indexed data generated from the new libraries, a minimum base quality score of 10 was required in both index reads. Index sequences for each library are listed in Table S4. The remaining sequence reads were merged (and adapters were removed) by searching for an >= 11nt overlap between the forward and the reverse reads (*44*). For bases in the overlapping sequence part, a consensus sequence was obtained by determining consensus quality scores and calling the base with the highest consensus quality. Reads with more than 5 bases with base quality scores below a base quality score of 15 were rejected.

Mapping

Only merged sequences were used for mapping with BWA (*12*) 0.5.8a to three reference sequences: the human genome (GRCh37/1000 Genomes release), the revised Cambridge Reference Sequence (rCRS) of the human mitochondrial genome (NC_012920.1) and the chimpanzee genome (CGSC 2.1/pantro2) using parameters (-l 16500 -n 0.01 -o 2) that deactivate seeding, allow more substitutions and up to two gaps (instead of 1). Prior to alignment, the first and last two bases were trimmed from the reads to reduce the effects of remaining ancient DNA damage (see Note 3). For the analyses presented in Note 3, sequences prior to trimming were aligned to the human genome. Using BWA's samse command, alignments were converted to the SAM format, and then via samtools (*45*) 0.1.18 to coordinate-sorted BAM format. BAM files were filtered by removing non-aligned reads as well as reads shorter than 35 bp. Furthermore,

BAM NM/MD fields were recalculated using samtools calmd, and reads with an edit distance of more than 20% of the sequence length were removed. This step was included to correct for non-A,C,G,T bases in the reference genomes being replaced by random bases when generating the BWA alignment index. For each library, reads which map to the same outer reference coordinates were replaced by a consensus sequence to collapse duplicate reads (*44*). Table S5 summarizes the number of filter-passed reads mapped to the human genome as well as the number of reads remaining after duplicate removal.

Local realignment for resolving insertions and deletion

After duplicate removal, the BAM files for all libraries were combined. For the alignments to the human and chimpanzee genomes, the Genome Analysis Tool Kit (GATK) (*13*) v1.3-14-g348f2db was used to identify genomic regions with many differences to the corresponding reference genome (RealignerTargetCreator). The GATK IndelRealigner was used to realign sequences in the identified genomic regions. After local realignment, BAM NM/MD fields were again calculated using samtools calmd, and reads with an edit distance of more than 20% of the sequence length were removed.

Note 5: Sequencing and processing of 11 present-day human genomes

Swapan Mallick, Nadin Rohland, Heng Li, Arti Tandon, Jacob Kitzman, Michael F. Hammer, Jay Shendure and David Reich*

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

Sample preparation

We generated deep genome sequences from 11 diverse individuals. We chose 10 of the individuals to be from the CEPH-Human Genome Diversity Panel cell lines (*46*) (of whom 8 overlapped with individuals we previously sequenced to 1-6× coverage (*1, 2*)). The 11th individual was a Dinka from Sudan, from a mouthwash sample.

To optimize the dataset for population genetics, we minimized differences across samples:

(1) We sequenced all samples together to minimize biases due to instrument variability. We generated four barcoded libraries for each sample, pooled the 44 = 4×11, and then sequenced the pool and decoded samples based on the barcodes. Because all samples were sequenced in a pool, sequencing errors which can vary from lane to lane and over time should affect all the samples equally, and thus should not bias inferences about population relationships.

(2) We chose all samples to be of the same gender, specifically males since for San, we only had access to males. This minimized artifacts due to different rates of mismapping of sex chromosome reads in males and females.

For each of the 11 samples, we started with 3μg of DNA. We then split the sample into four and prepared individually barcoded, whole genome shotgun sequencing libraries. The sample preparation was previously described in detail in Rohland and Reich 2012 (*47*). Briefly, we sheared each sample into short fragments using a Covaris E210 instrument. We then performed a dual fragment size selection with SPRI beads to select molecules with mean lengths of around 300bp and to minimize the fraction of molecules <200bp or >400bp. After blunting, we ligated 4 different barcoded, partially double-stranded adapters to each sample. The barcodes were designed to differ by at least 2 bases from each other (their sequences are in Table S6). We carried out a fill-in reaction and performed enrichment PCR to complete the adapter sequences.

Sequencing of pooled libraries and data access

The 44 libraries were combined together and then sequenced in one of the following two pools:

- Pool 1: We initially pooled the libraries by normalizing them based on their estimated read counts from qPCR. We sequenced the resulting pool on two flow cells (SRR359311-2) of an Illumina HiSeq2000 instrument at the University of Washington (UW).
- Pool 2: To increase the evenness of coverage across libraries, we repooled the libraries according to the read distribution from Pool 1. We sequenced 4 flow cells (SRR359307-10) at UW, 2 flow cells (SRR359305-6) at Harvard Medical School (HMS), and 4 flow cells (SRR446824-5 and SRR446835-6) at Beijing Genome Institute (BGI).

A total of 6,816,086,594 paired reads (the total read count is twice this number) passed standard Illumina quality thresholds. These are shown by barcode in Table S6. We deposited these data into the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/sra?term=SRX103808), as experiment "SRP009145", sample "SRS269343", and runs SRX103805-130812.

Mapping to the human and chimpanzee genomes

We identified the sample associated with each read pair based on its barcode (the first 6 bases of the first read, given in Table S6). We allowed up to one mismatch to the barcode for the reads we mapped to human, and no mismatches for the reads we mapped to chimpanzee. Allowing one mismatch increases the read count by 1.6%.

We mapped the barcode-trimmed read pairs to the human (*hg19/GRCh37* which we extended by adding the Epstein Barr virus) and chimpanzee (*panTro2*) reference sequences using BWA (*12*) using the command line "bwa aln -q15", which removes the low-quality ends of reads. If a read pair maps to multiple equally good positions, BWA randomly maps the pair to one and assigns a mapping quality zero to the pair. We used Picard to identify potential PCR duplicates. Table S7 summarizes the number of reads that remain after each filter.

We created BAM files based on the mappings to both human and chimpanzee; these files are available on request from the authors. The BAM files were the basis of all the population genetic analyses of the data from these 11 individuals, and were also used to produce the Extended Variant Call Files (VCFs) described in Note 6.

Note 6: Extended Variant Call Format files

Martin Kircher*, Swapan Mallick, Cesare de Filippo, Aida M. Andres and Janet Kelso

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

We used the UnifiedGenotyper from GATK (*13*) (v1.3-14-g348f2db) to produce genotype calls for single nucleotide variants (SNVs) and insertions and deletions (InDels) over all sites separately for each individual (parameters: --output_mode EMIT_ALL_SITES --genotype_likelihoods_model BOTH). Genotypes were called using alignments against both the human and chimpanzee genomes for the Denisovan (Note 4) and the 11 human individuals (Note 5). Genotypes were called using alignments to only the human reference for the six high-coverage 1000 Genomes trio individuals and the Vindija Neandertal. These genotype calls were deliberately left unfiltered but heavily annotated to explore various filtering options as described below. These files provide the basic input for most analyses. We provide per-chromosome Variant Call Format (VCF) files in block-gzip compressed form with a tabix *(http://samtools.sourceforge.net/tabix.shtml)* index file, which allows fast access to a certain genomic position/region. To match the 1000 Genomes trio data format, calls for mappings to the non-chromosomal or 'random' contigs are stored in one file (nonchrom).

Iterating GATK variant calls

The UnifiedGenotyper of GATK allows for only one alternative, i.e. non-reference, allele to be called per site. Thus, at positions of heterozygosity in which both alleles are non-reference alleles, only one allele can be described. To correct this issue and to reduce the potential reference bias that may be introduced by considering the reference sequence during genotype calling, positions where at least one non-reference allele was called for a certain individual were recalled as follows: (1) The reference genome was modified by replacing the reference allele with the alternative allele called in this individual. (2) GATK was run again using the modified reference genome with the same parameters as in the first iteration of genotype calling.

Table S8 shows the concordance between the first and the second genotype calls from Denisovan sequences aligned to either the chimpanzee or the human reference. The bias introduced by the reference genomes is apparent when looking at heterozygotes (0/1) ascertained in the first iteration of which ~29% and ~10% (in chimpanzee and human, respectively) become homozygous for the alternative allele (1/1) in the second iteration. However, this strong reference genome bias is mainly associated with low-quality genotypes (Table S8). When only heterozygous positions with genotype quality scores (GQ) >= 30 are considered, ~0.004% (chimp) and ~0.005% (human) of positions are called as homozygotes (1/1) in the second iteration. These percentages are further reduced when choosing a higher cutoff (GQ >= 40).

For sites that were homozygous for one alternative allele in the first iteration (1/1) and heterozygous for two alternative alleles in the second iteration (1/2), and sites that were heterozygous for one alternative and the reference allele in the first iteration (0/1) and heterozygous for two alternatives alleles in the second iteration (1/2), there was only a 2-fold reduction in the number of tri-allelic sites (heterozygotes 1/2 in Denisova) when using GQ filters for both human and chimpanzee alignments (Table S8). This indicates, that some of these sites are most likely truly heterozygous for two alternative alleles in Denisova. We therefore exchanged the genotypes of the first iteration to the genotypes of the second iteration for such sites regardless of genotype quality. More specifically, this applied to sites that were homozygous for one alternative allele in the first iteration and heterozygous for two alternative alleles in the second iteration and likewise for sites that were heterozygous for one alternative and the reference allele in the first iteration and heterozygote for two alternatives alleles in the second iteration Other sites with discordant genotype calls were kept as determined in the first iteration but tagged 'LowQual' in the FILTER field of the VCF file.

Bi-allelic sites which are different from the reference were reported as 'AF=0.50,0.50' and 'AC=1,1' in the INFO field and 'GT 1/2' in the SAMPLE field. Thus, in all cases the REFERENCE field still refers to the corresponding reference sequence. While recalling SNPs, some positions were identified showing a single nucleotide difference for one allele and an InDel for the other. In this case, if the InDel differed in length from an InDel identified in the first iteration, the original one was replaced by the new call. InDels identified only in the second iteration were not considered.

Extending GATK VCF files by additional annotation

To support downstream analysis, we supplemented the GATK VCF files by the following information inferred from the Ensembl Compara EPO 6 primate whole genome alignments (*48, 49*) (Ensembl release 64) and the integrated variant files of the 20101123 release of the 1000 Genomes project (*26*) (only available for the human reference alignments):

- Inferred ancestor bases for the human-chimpanzee, human/chimpanzee-orang, human/chimpanzee-gorilla ancestors as well as the rhesus macaque base were added to the INFO field.
- A string providing a one-letter code of the species aligned in the corresponding EPO block was added to the INFO field. The letter encoding is 'H' (Human), 'P' (Pantro), 'G' (Gorilla), 'O' (Orangutan), 'M' (Macaque) and 'C' (Callithrix). 'HH' for example indicates a block where two human homologous regions were aligned, but no outgroups are available. "HPGOMC" is a perfect block with 1:1 assignment of all 6 species available in the EPO alignments.
- A flag ('CpG') was added to the INFO field if a position is within a CpG dinucleotide context in the corresponding reference genome or the human-chimpanzee ancestor.

- A Repeat masking flag ('RM') was added to the INFO column, which is present if the reference sequence is lower-case, i.e. soft-masked, when obtained from the EPO alignment.
- dbSNP IDs and alternative alleles provided in the 20110521 release of the 1000 Genomes project (*26*) were added. The dbSNP ID was inserted into the ID field, while alternative alleles observed in the 1000 Genomes release (even if absent in the respective sample) were added to the list of alleles in the ALT field.
- Average allele frequencies as well as population frequencies from the 1000 Genomes 20101123 intermediate release were added to the INFO field where available.

We also added information from other sources. Note that these values and flags are only available for genomic blocks with a clear orthology assignment of human to chimpanzee in the EPO 6 primate alignments, because they are only reported after coordinate projection between the human and chimpanzee reference sequences.

- Background selection scores

  B-statistic, or B scores (*50, 51*), which incorporate information about the recombination rate and local density of conserved segments, were added to the 'bSC' field. B scores for hg18/NCBI36 were downloaded from *http://www.phrap.org/software_dir/mcvicker_dir/bkgd.tar.gz* and transferred to hg19/GRCh37 coordinates using the UCSC liftOver tool (*52*). Regions which failed to lift successfully, because they are deleted in the new assembly, were dropped. Regions which are 'split' or 'partially deleted' in the new assembly were broken down into smaller sections and the liftOver step was repeated. Some regions with different B values in the hg18 assembly overlapped in the new assembly. In these regions, the average of the two B-statistic values was used.

- Mammal conservation scores excluding human ('mSC' field)

  The 35 Eutherian mammal EPO alignments (*48, 49*) were downloaded from Ensembl release 64. The alignments were used to build chromosomal alignments from the alignment blocks and stitched to fit chimpanzee genome (pantro2/CHIMP2.1) coordinates. Positions with apparent chimpanzee deletions were removed to maintain coordinates, and the human sequence was removed from the alignment. PhastCons (*53*) was run on these alignments. PhastCons uses a two-state phylo-HMM with a state for conserved regions and a state for non-conserved regions. To avoid parameter tuning, parameters of the phylogenetic model used for generating the PhastCons conservation scores in the UCSC browser (*52*) were used. The conservation score given at each site is the posterior probability of the corresponding alignment column being generated by the conserved state. Since the tree used by UCSC did not contain all species of the EPO dataset, an alternate tree was obtained from NCBI taxonomy. PhastCons was run using

the parameters --msa-format FASTA $*.fa --target-coverage 0.3 --expected-length 45 --rho 0.31 --not-informative panTro2_ref.

- Primate conservation scores excluding human ('pSC' field)

  The 6-primate EPO alignments (*48, 49*) were downloaded from Ensembl release 59 and conservation scores were computed following the approach described for the 35 Eutherian mammal alignments above.

- Duke mappability scores of 20mers ('Map20' field)

  The 20mer mappability score reflects the uniqueness of 20-mer sequences in the genome with the score being assigned to the first base of each 20-mer. Scores are normalized to range between 0 and 1, with 1 representing a completely unique sequence and 0 representing a sequence that occurs more than 4 times in the genome (excluding 'random' chromosomes and alternative haplotypes). A score of 0.5 indicates the sequence occurs exactly twice, likewise 0.33 indicates three, and 0.25 four occurrences. The Duke Uniqueness tracks were generated for the ENCODE project (*54*) and downloaded from UCSC *(http://hgdownload.cse.ucsc.edu//gbdb/hg19/bbi/wgEncodeDukeMapabilityUniqueness20bp.bigWig)*

- Segmental duplications

  A flag ('UR') was added to the INFO field whether the position is in a control region for segmental duplication analyses, i.e. is not considered a candidate for structural variation, as defined by the Eichler lab.

- Systematic errors

  A flag ('SysErr') was added to the INFO field if a position was found to be prone to systematic errors. We define positions prone to systematic error as positions where the sequence context leads to elevated error rates in Illumina sequencing. One well established approach for the identification of such errors is the presence of an (erroneous) allele on only one DNA strand, i.e. strand bias (SB). Calling the UnifiedGenotyper with the 'sl' option, strand bias is quantified in the SB field of the output VCF, with higher positive values in the SB value indicating more evidence of strand bias. Since the vast majority of Denisovan paired-end reads were merged during primary processing, the signal of SB may be weakened or even vanish completely in this data set. To nevertheless assess problematic positions, we assumed that systematic errors are largely constant with the same sequencing platform. Therefore, we used the high-coverage trio data from 1000 Genomes Project (*26*) to identify systematic error-prone positions based on strand bias.

  In detail, BWA-aligned Illumina sequences from the Yoruba (NA19238 and NA19239) and European parents (NA12891 and NA12892) were considered. Aligned reads were filtered for mapping quality (MQ >= 30; likelihood of a read being incorrectly placed in a genomic region expressed in PHRED-scale (-10•log10)) and the presence of alignment flags indicating either aligned single-reads or properly aligned paired-end reads required.

Genotype calling was performed with the GATK UnifiedGenotyper excluding InDel calls and using all individuals together. A total of 144,421 heterozygote calls with SB >= 10 and a total coverage of at least 32x in all four samples combined were identified as sites of potential systematic error.

Lastly, the number of A, C, G, and T bases observed for each strand at each position as well as the number of reads starting an insertion/deletion at each position were added to the sample fields of the VCF files. These numbers were obtained using samtools mpileup (*45*). The exact identifiers of the corresponding fields are available from the VCF header lines. During processing, the AD value was removed from the sample fields in order to obtain an equal number of entries for variant and non-variant sites.

Combined VCF files

A set of combined VCF files containing all individuals were generated to simplify analyses involving multiple individuals. When merging different VCF files, multiple alternative alleles might be present. For this purpose alleles were re-numbered in the combined files, following VCF conventions (0-reference, 1-first alternative, 2-second alternative, and so forth). However, sample GQ and PL values reflect the original GATK three state model of homozygote reference, heterozygote and homozygote alternative genotype calls. Only due to the performed recall may the reference values refer to another alternative allele.

Since the first fields of each VCF line describe per-site features, we tried to integrate as many values as possible from all individuals there. However, for some values it was not possible using the information available from all individual VCF files. Thus, for example the RMS MQ field is taken from the Denisovan VCF file directly. This is documented in the corresponding VCF header lines. Individual-specific MQ and M0 values were maintained as sample specific features. The value 'LowQual' in the FLAG field was set if any of the samples had it, thus filtering on LowQual is probably overly conservative.

Note 7: Estimates of present-day human contamination in Denisova

Michael Siebauer, Qiaomei Fu, Matthias Meyer and Martin Kircher*

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

Mitochondrial contamination

Mitochondrial contamination estimates were generated following an approach described in more detail elsewhere (*14*). Briefly, all sequences were trimmed to remove damage-derived errors, aligned with BWA, and filtered for PCR duplicates exactly as described in Note 4, using the revised Cambridge Reference Sequence (rCRS, NC_012920.1) as the sole reference sequence for mapping. All aligned sequences were converted into FastQ format and re-aligned against the Denisovan mtDNA genome (NC_013993.1) using MIA (*14*) without further filtering (parameter -H 1, no iteration). Sequences overlapping one of 278 'diagnostic' positions, where the Denisovan sequence (FN673705) differs from >99% of the sequences from 311 present-day humans, were classified either as 'Denisovan' or 'human' according to which variant they matched. The mitochondrial contamination estimate for the complete data set is 0.35% (C.I. = 0.33 – 0.36), and variation among individual libraries is very low (Table S9).

Autosomal contamination

Neandertals and Denisovans fall within the variation observed for human nuclear sequences. Thus, only few fixed differences can be identified. However, using sites in the human genome that acquired a high frequency derived state after the divergence from the human-chimpanzee ancestor, we can obtain a maximum likelihood co-estimate (MLE) of sequence error, contamination as well as two population parameters (which are correlated with divergence and heterozygosity) for a sample from a divergent hominin lineage. This estimate assumes that any person contaminating the sample would add human derived alleles for which the ancient population sampled is at least partially ancestral. In this set-up, we can measure present-day human contamination as the excess of derived alleles. More precisely, contamination can be inferred from low frequency allele counts at homozygote positions, which can be caused by either contamination or sequence errors, as well an overrepresentation of the derived alleles at heterozygous sites, where contamination again interacts with error, especially in the case of low coverage data. The two effects, sequencing error and contamination, can be separated using genomic sites of different coverage.

To perform this estimate, we need high-frequency derived sites and sequencing data of a bone where these sites are covered with varying depth. To select for high-frequency derived alleles,

we used sites with at least 99% population frequency in the 1000 Genome intermediate release 20110521. We present a maximum likelihood estimator for the inference of two population parameters (depending on the divergence and heterozygosity of the Denisovan individual), a per-base sequencing error estimate, as well as contamination from present-day humans, which is similar to the one used in the Denisova draft genome paper (*2*), Supplemental Information 3. In our notation, we refer to a human-like allele as "derived" (index d) and a human-chimpanzee-ancestor like allele as "ancestral" (index a).

Let $\Omega = \{c, p_{ad}, p_{dd}, \varepsilon, f\}$ 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 the Denisovan individual with probability $1-c$.
- $p_{ad} \rightarrow$ probability of the Denisovan individual being heterozygous, given that humans and the human-chimpanzee ancestor differ at this site.
- $p_{dd} \rightarrow$ probability of the Denisovan individual being homozygous for the human allele, given that humans and the human-chimpanzees differ at this site.
- $\varepsilon \rightarrow$ probability of an error. We observe the human derived allele when the truth is ancestral (or vice versa) with probability ε.
- $f \rightarrow$ probability of a contaminating allele being human-like (derived). When examining only sites of fixed differences between humans and chimpanzees, $f = 1$.

We write the probability of the observed numbers of derived alleles (n_d) as the product of the probabilities of the L individual sites, conditional on the number of reads (n) at each site:

$$\mathrm{lik}(\Omega) = \Pr(n_{1,d}, \ldots, n_{L,d} \mid n_1, \ldots, n_L, \Omega) = \prod_i \Pr(n_{i,d} \mid n_i, \Omega) \qquad \text{(E7.1)}$$

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

$$\Pr(n_{,d} \mid n, \Omega) = \sum_{t=0}^{2} \Pr(t \mid p_{ad}, p_{dd}) \Pr(n_d \mid t, n, c, \varepsilon, f) \qquad \text{(E7.2)}$$

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

$$\Pr(t \mid p_{ad}, p_{dd}) = \begin{cases} 1 - p_{ad} - p_{dd} & t = 0 \\ p_{ad} & t = 1 \\ p_{dd} & t = 2 \end{cases} \qquad \text{(E7.3)}$$

The second term inside the sum, the probability of the observed number of derived alleles follows a binomial distribution:

$$\Pr(n_d \mid t,n,c,\varepsilon) = \binom{n}{n_d} q_t^{n_d} (1-q_t)^{n-n_d} \qquad \text{(E7.4)}$$

$$\begin{aligned} q_2 &= cf(1-\varepsilon) + c(1-f)\varepsilon + (1-c)(1-\varepsilon) \\ q_1 &= cf(1-\varepsilon) + c(1-f)\varepsilon + (1-c)(1-\varepsilon)/2 + (1-c)\varepsilon/2 \\ q_0 &= cf(1-\varepsilon) + c(1-f)\varepsilon + (1-c)\varepsilon \end{aligned} \qquad \text{(E7.5)}$$

The overall likelihood of the data given the parameters can be calculated from (E7.1), by substituting (E7.2), (E7.3), (E7.4) and (E7.5) in turn. Finally we estimate our parameters of interest (c and ε) by maximizing the likelihood of the data for parameters $\{c, p_{ad}, p_{dd}, \varepsilon, f\}$. We reduce the number of dimensions by assuming that $f = 1$; eventhough we are only requiring high frequency derived alleles (>= 0.99). The optimization was performed in the statistical software package R (*55*) using the bbmle package and the bound estimator "L-BFGS-B" of the mle2 method. The profile of the fit was used to obtain confidence intervals for the estimate. The R code is available in Figure S8.

When applying this estimate to the Denisova data, we considered only sites where a chimpanzee, gorilla and orang utan ancestor sequence is available (in a tree of the common phylogeny and with no second human or chimpanzee sequence present) and the inferred ancestor base identical. For this purpose, we used the Variant Call Format files (Note 6) and used the base count observations for derived and ancestral alleles directly from this file. Only sites with Duke 20mer mappability of 1.0 and covered by less than 100 reads were considered. We obtain a contamination estimate of 0.224% (95% CI: 0.217% - 0.232%) and per-read error (including sequencing, mapping and other effects) is estimated with 0.147% (95% CI: 0.147%-0.148%), see Table S10 for details.

We checked how well the model fits the actual data (Figure S9). The number of predicted and observed sites is in good agreement close to homozygous ancestral and derived as well as clear heterozygous states (close to 50:50 allele counts). These sites represent the vast majority of the data used for estimating the parameters of the model. However, the model deviates in intermediate states, which are not predicted by a binomial sampling of alleles. Binomial sampling is also the model applied for genotype calling with GATK and we currently lack a better model that matches the actual read sampling process observed from sequencing data. As a result, confidence intervals obtained from the model fit could be underestimates.

Y-chromosomal contamination

We have previously presented an estimate of male contamination in a female sample by counting aligned reads in Y unique regions (*2*). For this purpose, we inferred Y unique regions from the human reference (NCBI36) and assumed unbiased sampling of genomic regions. With the new release of the human genome (GRCh37), we again identified Y unique regions by two approaches that are motivated by the previous strategy.

First, we generated all possible 30-mers from the Y chromosomal sequence and aligned these back to the complete genome with up to 3 mismatches (but no gaps) using bowtie (*56*). We required that 30-mers align only to a single place in the human genome, the position from which the 30-mer was extracted, and kept only consecutive 30-mer regions of at least 500 bp. Using that approach we identify 144 Y-unique regions, which we further filtered for regions that are covered by reads in the four female individuals of the 1000 genomes trio data (*26*). By rigorously removing all regions that show any alignments in these females, a total of 28 out of 144 regions were removed, keeping 74 kb out of 93 kb Y-unique bases. In the remaining regions, we find 11 reads in the Denisova data while we would expect 18,403 (= #ALIGNED_READS x TOTAL_REGION_SIZE / GENOMESIZE x 0.5 = 1418957698 x 74444 / (2.87×10^9) x 0.5) when evenly sampling molecules from a male individual. Thus, we obtain a Y-contamination estimate of 0.06% (= 11/18,403) with a 95% confidence interval of [0.03,0.1%] from a binomial distribution.

In a second approach, we slightly weakened the previous criteria and used available mapability tracks of 40-mers and 75-mers as well as repeat masking information (UCSC table browser tables wgEncodeCrgMapabilityAlign40mer, wgEncodeCrgMapability Align75mer and rmsk). Non-repeat masked Y regions of a maximum 40-mer mapability and a minimum length of 300 bp were intersected with non-repeat masked 75-mer maximum mapabillity regions of at least 800bp, resulting in a total of 499 regions (~437 kb). After filtering out regions covered by reads in at least one of the four 1000 Genomes trio females, we kept 373 regions (~303 kb) in which we identified 52 contaminating reads (while 74,837 would be expected for a male). This corresponds to a Y contamination estimate of 0.07% [0.05%,0.09%].

Note 8: Assessing genome sequence quality

Martin Kircher, Michael Siebauer and Matthias Meyer*

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

Comparison of genomic coverage in the uniquely mappable genome

To exclude difficult-to-align parts of the genome, including repetitive and duplicated sequence, coverage statistics were computed for the 'uniquely mappable' regions of the genome. The 'uniquely mappable' genome is defined here as the sum of all positions, where 20-mer sequences can be placed uniquely in the genome according to the Duke Uniqueness 20 bp scores (see Note 6), which amounts to ~69% of the genome. We note that this filter is very conservative and that a larger fraction of the known genome (~85%) may be accessible to sequence determination by short-read sequencing technologies if less stringent criteria are applied (*26*). Figure S10 shows the distribution of coverage for the Denisovan genome as compared to modern human genomes. These distributions were used to determine upper and lower coverage cutoffs for several subsequent analyses, which exclude the 2.5% of sites with highest and lowest coverage (see Table S12). Figure S11 shows that the coverage distribution of the Denisovan genome remains largely unaltered even if the mappability filter is not applied.

Since we found a mild bias towards AT-rich sequences in Denisova (see Note 3), we examined the uniquely mappable parts of the genome in a 100 bp sliding window and determined the average coverage of Denisova and the present-day human genomes as a function of GC-content of the human reference genome. Figure S12 shows that coverage is positively correlated with AT-content in the Denisovan genome. Coverage declines rapidly when GC-content is above 60% (e.g. from 23x at 60% down to 15x at 70% GC). However, regions of very high GC-content are rare in the human genome. Furthermore, almost identical GC-dependencies are seen in the coverage plots of the eleven present-day human genomes, indicating that single-strand library preparation from ancient DNA did not introduce more bias than the library preparation methods used for modern DNA. In fact, it is unclear whether these biases are introduced during sample preparation, sequencing or mapping. In conclusion, the distribution of genomic coverage for the Denisovan genome is indistinguishable from those seen for the genomes of the eleven present-day humans generated in this study.

Comparison of per-base error rates

To estimate sequencing errors in individual sequence reads, we used positions in the human reference genome where only very little divergence is expected among the genomes of Denisova and present-day humans. We selected these positions from the Ensembl 35-way GERP elements annotated for human (ftp://ftp.ensembl.org/pub/release-64/bed/35way_gerp_elements.Homo_sapiens.bed.gz),

and filtered them for a primate conservation score of >= 0.98 using the extended VCF files described in Note 6. To minimize the possible influence of alignment error we also required a mapability score of one in the Duke 20mer uniqueness scoring (see Note 6), obtaining in total ~5.6 million positions.

When comparing genotype calls to the reference, we found that 0.03 – 0.05% of genotypes differ (Table S11), and this in a pattern that is consistent with sequence divergence from the reference genome; i.e. Denisova shows the largest number of differences, followed by Africans and then the other present-day humans. Thus, despite extensive filtering, sequence divergence contributes to the estimates of per-base sequence error rate which are obtained by counting differences between individual sequence reads and the human reference genome. We therefore subtracted genotype divergence from the per-base error rate to obtain 'divergence-corrected' per-base sequence error rates. These rates are highest (1.2 – 1.7%) in the thousand genomes trios and much smaller in the eleven present-day human genomes (0.17 – 0.19%) (Table S11). At 0.13%, the error rate estimated for Denisova is lower than that obtained for any of the present-day humans, and is comparable to the co-estimate of error and autosomal contamination obtained in the maximum-likelihood approach described in Note 7 (0.15%).

Note 9: Estimates of sequence divergence of Denisovans and present-day humans

Martin Kircher*, Montgomery Slatkin, Matthias Meyer

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

Strategy for estimating sequence divergence and branch lengths from genotype data

To estimate the sequence divergence between all possible pairs of genomes generated in this study as well as the human reference genome, we follow a strategy similar to what was described previously (*1, 2*). However, we adjusted the approach to use an inferred human-chimpanzee ancestor for the outgroup allele, and genotype data rather than individual sequence reads. Briefly, for each divergence estimate, we select one genome as 'reference' and a second as 'sample' and identify positions where one or both of them are derived, i.e. different from the inferred human/chimpanzee ancestor state. At heterozygous positions, we randomly sample one allele. We then count the number of positions where sample and reference both show the derived state ('C'; derived alleles in common) as well as the number of positions with reference-specific ('R') and sample-specific ('S') derived alleles (Figure S13).

We calculate divergence as R/(C+R), which is the reference-specific fraction of the branch leading from the reference to the human-chimpanzee common ancestor. When selecting high-quality sequences as reference, we assume a negligible error rate and that the differences identified between reference and outgroup are therefore due to divergence only.

Differences specific to the sample ('S') are not used in this divergence calculation, but allow for determining the relative length of the sample-specific branch compared to the reference-specific branch (S/R). S/R ratios of ~5.3 and ~1.6 were previously determined for the Neandertal (*1*) and Denisovan (*2*) draft genome sequences, respectively. This indicates an excess of error in these genomes due to low-coverage sequencing and ancient DNA damage.

We estimate sequence divergence and branch length ratios using the outlined method. To explore the stability of these estimates, we include the 1000 genomes (1000G) trios in the analysis, which are of lower quality compared to Denisova and the eleven present-day humans sequenced in this study. Further, we also compute direct pair-wise measures of sequence differences, an analysis that is possible with the high-quality of the Denisovan genome.

Data filtering

Using genotype calls from the combined VCF files described in Note 6, we extracted information and applied filters as follows:

(1) We restrict analysis to regions of maximal mappability (Map20 = 1.0).
(2) To establish orthology, a 'HPGO' substring is required in the 'TS' field, which indicates that three ape outgroups are present and excludes regions with paralogs. Further, we explicitly require that no more than one chimpanzee and one human sequence are present in the 'TS' field. Due to the large block size in the Ensembl EPO alignments, these filters seem not sensitive to the more local effects of incomplete lineage sorting between these apes, and more than 90% of EPO blocks remain after filtering.
(3) We use all three inferred ancestor alleles (the human-chimpanzee ancestor, the human-gorilla ancestor and the human-orangutan ancestor) and require these to be identical with the following exception: If the human-chimpanzee ancestor allele differs from the two other ancestor alleles, we consult the Denisovan genotype call. If in this case Denisova shows the allele present in the human-gorilla and human-orangutan ancestors, this allele is used. Otherwise the position discarded. This procedure eliminates a human-reference bias in the inference of the human-chimpanzee ancestor allele, which would otherwise lead to miscounting positions with differently segregating alleles in the ancestor population (i.e. incomplete lineage sorting), parallel mutations or shared errors.
(4) We use the ancestor sequences of neighboring positions to define whether a position is in a CpG dinucleotide. Due to mammalian 5'-cytosine methylation in CpG dinucleotides, CpGs show a very different mutation mechanism and rate (*57, 58*). In addition, post-mortem deamination of 5-methylcytosines generates thymines in the Denisovan genome, and these are not removed by the enzyme treatment described in Notes 2 and 3. We therefore excluded CpG dinucleotides from the relative divergence calculation.

Before computing all possible pair-wise divergences, we stratified divergence to the human reference genome (GRCh37) by coverage to evaluate the stability of our estimates (see Figure S14). We first note that divergence estimates for all samples become unstable at the extreme tails of the coverage distribution, probably due to alignment problems in repetitive or duplicated segments of the genome. However, divergence estimates of the eleven present-day humans remain stable over a broad range of coverage surrounding the mean. Divergence of the 1000G trios reduces with increasing coverage without reaching a stable plateau, suggesting that differences in sequence quality may be the major effect in this analysis. For Denisova a plateau is reached, albeit shorter than for the eleven present-day humans. Since the sequencing error rate is low in Denisova, this increased sensitivity to coverage in the divergence calculation can probably be attributed to an alignment bias caused by shorter reads – the only parameter for which Denisova falls in-between the 1000G trios and the eleven present-day humans. In subsequent analysis, we eliminate positions that are within the upper and lower 2.5% tails of the coverage distribution in all samples (see Note 8 for coverage cutoffs).

Autosomal divergence and branch shortening

Using either the eleven present-day humans or the 1000G trios as reference, we calculate the Denisova-human sequence divergence as a percentage of the human-chimpanzee divergence. The divergences are 12.4% [12.2%-12.5%] and 12.2% [12.2%-12.4%] respectively (Table S13). Divergence between the eleven human genomes ranges from 6.3% to 9.7% and between the 1000G trios it ranges from 4.9% to 8.7% (Table S14). The lower range in the 1000G trios is due to the sampling of different populations and the inclusion of the daughters in each trio, which show an expected 24-25% reduction of divergence when compared to each of their parents. Excluding the trio daughters (NA12878 and NA19240), all human-human divergence values range from 6.3% to 9.7% and are thus lower than the divergence seen to Denisova.

Divergence is slightly lower for the human reference genome sequence (GRCh37) when it is used in pair-wise comparisons, not only in pair-wise comparisons to Denisova (12.0%), which might be expected due to the presence of a Neandertal component in the Eurasian parts of the human reference genome (*1*), but in pair-wise comparisons to all humans (e.g. all Africans show lower divergence from GRCh37 than from any other present-day human; see Table S14). These differences in divergence estimates are probably the result of two confounding factors: (i) Sequences generated from whole genome shotgun sequencing and mapping assemblies are of lower quality than the finished human reference genome. Thus, if genotypes from one of the shotgun genomes are used as the 'reference' in the divergence calculation, sequence errors will increase the inferred length of the reference-specific branch (R) and hence divergence. (ii) Alignment bias to GRCh37 may lead to a preferential loss of non-reference genome alleles. This would not only reduce the length of the 'sample'-specific branch (S), but causes an over-representation of alleles in common with the 'reference' (C) compared to 'reference'-specific alleles (R), thereby reducing divergence. Since these two counter-acting biases are difficult to disentangle, we prefer to limit divergence calculations to comparisons among genomes of similar quality (i.e. excluding GRCh37).

When comparing the Denisovan branch-specific counts (S) to the counts of the eleven present-day humans and the 1000G trios (Table S13), we see that the Denisovan branch is on average 9.4% [9.2% to 10.1%] and 8.5% [7.9% to 9.4%] shorter, respectively. Such an extreme difference in branch length is not observed for any pair-wise human comparison (Table S15). When comparing within the eleven humans, the average branch length difference is 0.2% [-1.6% to 1.6%], and 0.6% [-0.7% to 2.8%] for the lower quality trio data. When comparing the eleven humans and the 1000G trios, the trios show slightly shorter branches (0-3%), in line with stronger alignment bias in these genomes. When comparing to GRCh37 all human samples show branches that are 4-7% longer, while Denisova still shows a 6% shorter branch. The observed shift in branch length ratios are most likely explained by the confounding effects of alignment bias and lower quality shotgun genomes.

These results suggest that branch-shortening is not simply an artifact. Instead, it appears to reflect that, as expected for an ancient fossil sample, the Denisovan genome lacks several tens of thousands of years of molecular evolution. Based on the differences in branch length to the common ancestor of human and chimpanzee (1.13% to 1.27%; see Table S13), we estimate that the observed branch shortening corresponds to 73,614 – 82,421 (average 75,443) years assuming

a human-chimpanzee divergence time of 6.5 million years. This range limits comparisons to the 11 present-day human genomes generated in this study, which are most comparable in quality to the Denisovan genome. We note, however, that a sufficiently accurate molecular dating of this fossil is confounded by differences in genome quality and alignment bias in all samples both ancient and modern.

X chromosomal divergence

When using the present-day humans as reference in the divergence calculation, we infer an average X-chromosomal sequence divergence between Denisova and the other humans of 12.2% [11.8%-12.5%] (Table S16). We also observe branch shortening for the Denisova on chromosome X (Table S17), however due to the lower amount of data we see a considerably larger variation for this estimate. Within humans (excluding the trio daughters), divergence ranges from 4.5% to 9.3% on the X chromosome. Thus, human-human average divergence on the X chromosome reduces from 8.4% to 7.2% (~86%), while Denisovan divergence is not different from the numbers reported for the autosomes. The applied relative divergence is a ratio in which the individual values depend on branch lengths, the effective population sizes and the difference in male and female mutation rates. Particularly, the absolute branch lengths and the effective population sizes on the R and C branches are currently not well known. We will therefore revisit the X-autosomal ratio for pair-wise differences below, for which better estimates exist.

We note that the larger difference in divergence on X and autosomes that was previously observed in SI 2 Reich et al. (*2*) (reduction of the divergence on the X chromosome compared to the autosomes in Neandertal from 12.1% to 10.1% and in Denisova from 11.9% to 9.6%), was likely caused by a quality difference between the chimpanzee autosomes and the X chromosome. The previous paper used the chimpanzee allele rather than an ancestral allele inferred from multiple reference genomes as outgroup. When polarizing changes between the human and chimpanzee reference sequence by the inferred ancestor allele, we see an excess of 3% on the chimpanzee autosomes while we see an excess of 33% for the X chromosome. If we correct for this effect, the numbers reported by Reich et al. increase to 12.3%/11.8% for Neandertal and 12.1%/11.2% for Denisova.

Pair-wise differences

As evidenced by the observation of branch shortening for the Denisova and equal branch lengths for the human samples, data quality is sufficiently high that we no longer need to polarize changes by an outgroup. We therefore determined pair-wise differences between the genotypes determined for all individuals, applying only filters (1) and (2) (see *Data Filtering* above). Table S18 and S19 give the numbers of transitions and transversions for all pair-wise comparisons for autosomes and the X chromosome. We note that both for ancestor divergence and for pair-wise differences, the lowest Denisova-human divergence is observed to the Papuan individual (HGDP00542). This is in agreement with the unique signal of Denisovan admixture detected in that population (*2, 6*). This effect is however small.

On the autosomes, the Denisovan individual differs on average at 1017 transitions and 508 transversions per million base pairs from a human sample, while the eleven humans differ on average by 717 transitions and 379 transversions from each other. On the X chromosome, the Denisovan individual differs on average in 759 transitions and 395 transversions per million base pairs from a human genome, while the eleven present-day humans differ on average by 479 transitions and 269 transversions. The ratio of the average Denisova-human to the average human-human pair-wise differences on the autosomes (65%) is in good agreement with the human-human ancestor divergence being at about 69% of the Denisova ancestor divergence, as determined with the method outlined above.

The number of differences between human and Denisova on the X chromosome is 76.2% of the number determined for the autosomes. This reduction is even more pronounced in present-day human comparisons (68.9%). This observation is consistent with what is known about male-biased mutation in chimpanzees and humans when differences in average coalescence times of X and autosomes in the ancestral population are taken in account. For autosomes, the average genomic divergence is $2(T+N)\mu_A$ and for the X, it is $2(T+3N/4)\mu_X$ where μ_A is the autosomal mutation rate, μ_X is the mutation rate for the X and N is the size of the ancestral population. From standard theory (*59*), $\mu_A = (\mu_m + \mu_f)/2$ and $\mu_X = (2\mu_f + \mu_m)/3$, where μ_m and μ_f are the mutation rates in males and females. Therefore, the ratio of genomic divergences of the X to the autosomes will be

$$R_{XA} = \frac{2(2+\alpha)}{3(1+\alpha)} \frac{T + 3N/4}{T+N}$$

where $\alpha = \mu_m / \mu_f$. Scally et al. (*60*) estimated α to be 2.3, somewhat lower than previous estimates, which were 3 or larger (*61*). From Tables S18 and S19, R_{XA} is about 75.1% for the eleven humans compared to the Denisovan, and about 72.3% for the San (HGDP01029) compared to other humans. These values are roughly consistent with the theoretical expectation. For example, for humans and Denisovans, T≈16,000 generations (=400,000 years). If α=2.3 and N=10,000, R_{XA}=75.3%. For San compared to other human populations, T≈4,000 generations, if α=2.3 and N=10,000, R_{XA}=71.3%. These calculations do not take account of the uncertainties in estimates of T and α and do not allow for population size changes in the ancestral population, which affect the X and autosomes differently (*62*).

Note 10: Date of population divergence of modern humans from Denisovans and Neandertals

Heng Li, Swapan Mallick and David Reich*

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

Motivation

An important date in human evolution is when the ancestors of modern humans diverged from Denisovans and their sister group the Neandertals. In the paper on the draft sequence of the Neandertal genome, we estimated this date for Neandertals (*1*). Since Denisovans are a sister group of Neandertals (*2*), they should have approximately the same population divergence; however, we never assessed this directly. Furthermore, the inference in the Neandertal genome paper was based on assumptions about mutation rates from early 2010. Since that time, better data have become available, lower mutation rates have been suggested, and the true value of the mutation rate has become less certain. It is important to obtain a new date estimate in light of this.

Here we estimate the population divergence date of Denisovans and Neandertals from Yoruba using the same strategy we described in the Neandertal genome paper. Our strategy is to discover high-confidence polymorphisms within the two chromosomes of a single present-day Yoruba individual, and then to estimate how often a randomly sampled allele from another sample (Denisova or Neandertal) carries the derived allele (the new mutation). The older the divergence, the less often we expect to observe the derived allele (since the older the divergence, the higher the probability that the mutation arose in the Yoruba population since the divergence from the population to which they are being compared).

A strength of this strategy is that the rate of ticking of this molecular clock is unaffected by demographic complexities in the history of the non-Yoruba population we are analyzing. To estimate divergence time, all we need to have is a model of Yoruba history. Fortunately, there are a number of models that have been fit to the data (this is why we focus on Yoruba, rather than one of the other sub-Saharan African populations we sequenced). Another advantage of calibrating to divergence from Yoruba is that unlike non-African populations, they do not have evidence of admixture from Neandertals or Denisovans that could complicate inferences.

Three datasets for estimating derived allele frequencies at sites heterozygous in Yoruba

The first dataset we use was reported in the Neandertal draft genome paper (*1*) (the final column of Table S40 in SOM Text 14). It was generated from a set of SNPs that were identified as polymorphisms in a single deeply sequenced Yoruba individual (NA18507) (*32*). The analysis was restricted to transversions (which have a reduced rate of recurrent mutation), and to autosomal sites with data from both chimpanzee and orangutan. At each site, 1-6× genome sequence data from each of six humans (5 present-day humans and Neandertal) was used to compute the probability that an allele randomly sampled at these sites was derived. To infer the probability of being derived while correcting for mislabeling of the ancestral allele due to recurrent mutation, we used the difference between the derived allele frequency when only

chimpanzee was used to learn the ancestral allele, and when chimpanzee and orangutan were required to agree (*1*).

The second dataset consists of autosomal SNPs discovered as candidate heterozygotes in a Yoruba individual (HGDP00927) and mapped to the chimpanzee (*1*), restricting to sites where we also had aligned sequences from Denisova and Neandertal. Collaborating with Affymetrix, we developed a screening array to validate these candidate heterozygous sites in the same sample in which they were discovered, using an Axiom® SNP array. After this screening, a second Axiom® array containing only validated SNPs was used to genotype 934 unrelated people from 53 populations in the CEPH-Human Genome Diversity panel, a dataset we released on August 12, 2011 (ftp://ftp.cephb.fr/hgdp_supp10/). We also genotyped 5 Dinka on this array. To estimate the derived allele frequency in each test population, we restricted to transversions. We also corrected for recurrent mutation using the same strategy as for data set 1, in this case comparing results when only chimpanzee is used to infer the ancestral allele, to results when both chimpanzee and gorilla are required to agree.

The third dataset consists of autosomal SNPs that we identified as high confidence heterozygous single nucleotide substitutions by sequencing a single Yoruba individual (HGDP00927) and mapping the reads to the human reference genome sequence hg19 resulting in 32× average coverage (Note 5). We used the Genome Analysis Toolkit (GATK) (*13*) to generate genotyping calls as described in Note 6, and restricted to transversion polymorphisms where we had data from both chimpanzee and gorilla allowing us to apply the same correction for recurrent mutation as in data sets 1 and 2. This resulted in a total of 665,220 sites that we could use for our analysis. To infer the probability of observing the derived allele in other populations, we used the fraction of sequencing reads aligned to that site that carried the derived allele (restricting to sites that had sequence coverage within the 95% central interval of that sample's genome-wide distribution). Empirically, we found that directly using the reads gave indistinguishable results from using genotyping calls from GATK. We preferred using the reads because this allowed us to compare to the low coverage Neandertal data.

Table S20 reports the derived allele frequency estimates in each of up to 11 modern human populations, as well as Neandertal and Denisova, based on each of these datasets. Neandertal and Denisova have overlapping probabilities of carrying the derived allele, consistent with their being sister groups (*2*). We date them together in what follows.

Estimating the probability that Denisova and Neandertal carry the derived allele

Table S20 shows that AT→CG substitutions have a higher derived allele frequency on average than CG→AT substitutions, while A↔T and C↔G substitutions have an intermediate frequency. This is consistent with biased gene conversion, which prefers transmitting C/G base pairs over A/T base pairs at heterozygous sites that overlap a gene conversion tract (*63*). This affects the probability of an allele being derived by up to a few percent, and is not taken into account by neutral population genetic models.

To minimize the potential confounding factor of biased gene conversion, we computed the *difference* in the probability of carrying the derived allele between the test population and that in the Yoruba. We note that for the deep sequencing data, we do not have a second Yoruba individual, so we calibrate to the Mandenka assuming that its probability of carrying the derived

allele is -0.1% below the Yoruba based on the empirical observations from the SNP array data. (We do not model errors in this correction factor. However, inspection of the SNP array data indicates that such errors are not likely to be more than ~0.1%, which is small compared to the range of uncertainty of 1.7% in the probability of archaic samples carrying the derived Yoruba allele, so we neglect this source of error in what follows.) In Table S21 we show that differences between the AT→CG, CG→AT and A↔T/C↔G classes become smaller when we focus on the difference rather than the absolute numbers.

Inspection of Table S21 reveals that the fall-off in the probability of carrying the derived allele is slower for the genome sequence data than for SNP array data. We hypothesize that this reflects a bias in the SNP array data, due to the fact that SNPs included on the array were required to have high genotyping completeness over all samples analyzed. Poor SNP array genotyping is correlated with oligonucleotide primers that overlap polymorphisms. Thus the requirement for completeness biases against segments of the genome with a higher polymorphism rate, which in turn biases against segments of the genome with deep gene trees. We therefore focus our inferences on the sequence data, which we do not expect to have this bias to the same degree.

To make our inferences about population divergence, we focused on two columns of Table S21. The first is "All" transversions together, where the Neandertals and Denisovans have a -12.3% to -13.2% lower probability of carrying the derived than the Yoruba. The second is A↔T and C↔G transversions, since these are not subject to biased gene conversion; here, Neandertals and Denisovans have an -11.5% to -12.4% lower probability of carrying the derived allele than the Yoruba. We are not sure which range is more appropriate for our purposes, since the former range corresponds to the type of data that was previously used to fit population genetic models (the models we use for date calibration), while the latter range is not subject to biased gene conversion. To be conservative in what follows, we assume that the true probability of an archaic human carrying the derived allele at a site that is polymorphic in Yoruba is reduced by -11.5% to -13.2%, and explore how this range affects the inferred date of the population split.

Inference of population divergence time as a fraction of human-chimpanzee divergence

To estimate a population divergence time, we follow the same strategy as SOM 14 of the Neandertal paper (*1*). Specifically, we compare our observations to simulations that inferred the probability of an individual carrying the derived allele at a Yoruba heterozygous site, as a function of the time separation between that population and Yoruba, for previously fitted models of Yoruba demographic history. Here, we only report results from 3 of the 4 models used in SOM 14 of the Neandertal paper: Keinan et al. 2007 (*64*), Wall et al. 2009 (*65*) and Li and Durbin 2011 (*22*). We no longer report results from the older modeling study of Schaffner et al. 2005, which was fit to a smaller dataset that was affected by SNP ascertainment bias (*66*).

The simulations allowed us to convert the numbers in Table S21 to dates of divergence of a test population from Yoruba as a fraction of the average time to the most recent common genetic ancestor within Yoruba. As in SOM 14 of the Neandertal paper, we divide this by 11.4 to infer the population divergence as a fraction of human-chimpanzee, since human-chimpanzee genetic divergence has been estimated from genetic data to be 11.4 times that of Yoruba-Yoruba divergence (Table S22).

Inference of absolute divergence time in years

To convert the ratios in Table S22 into absolute estimates of time, we used one of two strategies. We discuss these below, and summarize the results in Table S23.

(a) Fossil calibrations assuming a constant molecular clock: 170-440 kya

In SOM 14 of the Neandertal paper, we argued based on two different calibrations to the fossil record, and taking into account relative genetic divergence times, that human-chimpanzee genetic divergence on the autosomes was 5.6-8.3 million years ago (*1*). By multiplying by the lower and upper bounds by the numbers in the final column of Table S22, we obtain new estimates of population divergence dates based on these estimates of the lower and upper bounds for human-chimpanzee genetic divergence (Table S23). Our updated inference is 170-440 kya for the divergence of Yoruba from Neandertals and Denisovans (vs. 270-440 kya in the Neandertal paper).

(b) Using a suggested lower mutation rate: 410-700 kya

Recent studies based on direct observations of mutation rates in families have suggested that human mutation rates per base pair might be substantially lower than the rates implied by the fossil calibrations used in the Neandertal paper, which correspond to mutation rates of 0.83×10^{-9} to 1.22×10^{-9} per year using 1.31% as the divergence per base pair at aligned regions of the autosomes that we measure from the EPO alignments (*48*). In the gorilla genome paper, the authors proposed that various lines of evidence might be reconciled through a mutation rate as low as 0.5×10^{-9} per year, combined with a hypothesized parallel slow-down in the molecular clock rate in the ancestors of both African great apes and orangutans (*60*). Using this much lower mutation rate and an estimated heterozygosity of 0.00104 per base pair for Yoruba from Note 15, we obtain an estimated range of 410-700 kya.

The greatest uncertainties in these date estimates come from uncertainties about the human mutation rate, for which plausible estimates vary >2-fold. As better estimates of the rate of the molecular clock become available, it will be possible to obtain more accurate genetic inferences about the population split time. At the moment, our estimate contributes little to the interpretation of the fossil record. For example, fossils from Sima de los Huesos (Spain) that show Neandertal-like traits, were previously dated to 600 kya (*67*), but have recently been suggested to be as young as 350 kya (*68*). Both dates fall within the range of plausible divergence times of the ancestral population of Neandertals and Denisovans from modern humans estimated here.

Note 11: D-statistics and interbreeding between archaic and modern humans

Nick Patterson, Swapan Mallick and David Reich*

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

In our paper on the draft Denisovan genome (*2*), we tested whether Denisova is more closely related to some present-day humans than to others. This analysis, which we carried out using the *D*-statistics first developed for the Neandertal genome paper (*1*), documented that Denisova is significantly more closely related to Papuans than to mainland Eurasians.

Here we use deep coverage data from Denisova and 11 present-day humans to extend this work. The advantage of this reanalysis is not just the higher coverage Denisovan genome, but also the greater uniformity of the data. Previously, the data from present-day humans came from individuals who we sequenced on separate lanes and using different instruments and reagents (*2*). Variation in the error process of the sequencer over time and across lanes could cause some samples to appear more closely related to Denisova than to others, if the error process for some samples happened to be correlated to that in Denisova. Our new sample preparation solves this by pooled sequencing. As described in Note 5, we pooled DNA from all samples into a single tube and sequenced the pool. Thus, the error process is the same for all samples, and is not expected to cause some modern humans to appear closer to Denisova than others.

We note that despite the high-coverage Denisova genome sequence, which we leverage heavily in what follows, we are still not able to carry out a formal test for whether the Neandertals and Denisovans hybridized are symmetrically related to present-day Africans. The reason for this is that as described in the first Denisova genome paper (2), the quality of the Neandertal genome is too low to support such an analysis. It may be possible to carry out such a test once a high coverage Neandertal genome of equal quality to the Denisova sequence becomes available.

Data processing

We restricted analyses to reads mapped to chimpanzee (*panTro2*), because chimpanzee is equally distant to all present-day humans, and thus we do not expect there to be a bias toward some present-day humans mapping better to the reference genome than others. We restricted to sites where just two alleles were observed, and used both transitions and transversions because we found empirically that restricting to transversions gave the same results. We analyzed the data in two ways:

(i) Analysis directly from the reads: We filtered out reads and nucleotides according to the procedure we previously developed for *D*-statistic analysis. The details of these filters are in Supplementary Information 6 of the Denisovan draft genome paper (*2*).

(ii) Analysis using GATK genotype calls: We analyzed the genotypes from the combined VCF obtained by alignment to the chimpanzee genome (Note 6). We restricted to autosomal sites

with a mapability score of Map20=1, and coverage for the individuals analyzed in the central 95% of the genome distribution (Note 8). For chromosome X, we excluded pseudo-autosomal regions and the X-Y transposed region.

We computed D-statistics measuring whether Denisova shares more derived alleles with sample H_1 than sample H_2. For each site in the autosomes, we computed three probabilities for the two present-day humans being compared (H_1 and H_2), and Denisova. For the read data these were:

p^i_{H1} = fraction of H_1 reads aligned to site i that are derived relative to chimpanzee

p^i_{H2} = fraction of H_2 reads aligned to site i that are derived relative to chimpanzee

p^i_D = fraction of Denisovan reads aligned to site i that are derived relative to chimpanzee

For the genotypes, we set p^i_j to be 0, 0.5 or 1 depending on sample j's genotype at SNP i.

Denoting the chimpanzee allele as "A" and the non-chimpanzee allele as "B", there are two possible patterns: {H_1-H_2-*Denisova*-*Chimp*} = {B-A-B-A} or {A-B-B-A}. The probability of a "BABA" at site i is $p^i_{H1}(1- p^i_{H2})p^i_D$, and the probability of an "ABBA" is $(1-p^i_{H1})p^i_{H2}p^i_D$. Denoting the expected count of each across all sites as n_{BABA} and n_{ABBA}, our D-statistic is then:

$$D(H_1, H_2, Denisova, Chimp) = \frac{n_{BABA}-n_{ABBA}}{n_{BABA}+n_{ABBA}} = \frac{\sum_{i=1}^{N} p^i_{H1}(1-p^i_{H2})p^i_D-(1-p^i_{H1})p^i_{H2}p^i_D}{\sum_{i=1}^{N} p^i_{H1}(1-p^i_{H2})p^i_D+(1-p^i_{H1})p^i_{H2}p^i_D} \qquad (1)$$

A value consistent with zero indicates that H_1 and H_2 are consistent with descending from a homogeneous ancestral population since the split from Denisova, without subsequent genetic interchange with relatives of Denisova.

We use a Block Jackknife (*1, 2, 69*), dividing the genome into 500 equally sized contiguous blocks on the autosomes, to compute a standard error. We obtain a Z-score by dividing the D-statistic by this standard error. We use a 2-sided Z-test (from a normal distribution) to evaluate whether Denisova is more closely related to H_1 or H_2.

To assess the robustness of the D-statistics to different ways of preparing the data, we computed D-statistics separately using the reads and using the genotypes, and compared the results for all 55 = 10×9/2 possible statistics (Table S24). The statistics are highly correlated: r^2 = 0.985, and the regression slope is 0.9998. The D-statistics in the two datasets also overlap in their confidence intervals. We conclude that it makes no qualitative difference to our analysis whether we compute the D-statistics from the GATK genotypes or from the read data. In what follows, we carry out analyses on the genotype data where possible, because it makes more efficient use of the available information, slightly increasing the precision of our inferences.

Less archaic ancestry in Europeans than in Eastern non-African populations

Replication of more Denisovan ancestry in Papuans than other non-Africans

Table S24 confirms a major finding from the first Denisova genome paper (*2*): that Denisova shares significantly more derived alleles with Papuans than with other non-Africans: D(Papuan, X, Denisova, Chimpanzee) is 6.0-7.6% (Z = 9.8 to 12.5), where X = any of 5 non-African populations other than Papuans.

A novel observation that emerges from our updated D-statistic analyses is a significant difference in the degree of relatedness to Denisova detected in a pair of non-Papuan populations

outside of Africa: *D(French, Dai, Denisova, Chimpanzee)*. In the read data, D = -1.7±0.4% (Z = -4.0) (Table S24), and the signal is also seen in the genotype data although it is weaker (D = -1.2±0.4%; Z = -2.9). This signal is not predicted by our previous finding of Neandertal gene flow into the ancestors of all non-Africans (*1*). A possible explanation for these findings is Skoglund and Jakobsson's recent conclusion, based on analysis of the draft Neandertal and Denisova genomes and SNP array data from present-day humans, that: (a) there is more Denisovan affinity in East than West Eurasians, and (b) there is more Denisovan affinity in southeast Asians (including Dai) than northeast Asians (including Han) (*17*). However, alternative explanations are also possible, and we therefore carried out further work to investigate this signal.

Enhanced *D*-statistics document less archaic ancestry in Europe than eastern non-Africans

To enhance our power to detect archaic gene flows into non-Africans, we restricted to the subset of the genome where a panel of 35 sub-Saharan Africans for which we had sequence data are consistent with all carrying the chimpanzee allele. The effect of this restriction can be seen by studying the expected value of the *D*-statistic (*70*):

$$E[D] = \frac{variants\ arising\ from\ gene\ flow\ (contributing\ to\ BABA)}{variants\ arising\ from\ gene\ flow + variants\ from\ ancestor\ (noise\ contributing\ equally\ to\ BABA\ and\ ABBA)} \quad (2)$$

The requirement that African genotypes are all ancestral retains sites that arose as mutations in Denisova ancestors since the split from the ancestors of modern humans (the "signal" term in the numerator of Equation 2 which only contributes to BABA), while filtering out sites that are due to variation inherited from the ancestral population (the "noise" term in the denominator, which contributes equally to BABA and ABBA). Thus, we have an "Enhanced *D*-statistic" with an improved signal-to-noise ratio, which in theory can improve our power to detect real gene flow.

It is important to explain why the Enhanced *D*-statistic gives valid results. Consider the hypothesized phylogeny shown in Figure S15 (which is not to scale). Here there is a split at X into a lineage leading to archaic hominins, and another leading to modern humans. After a split between ancestral Africans and ancestral Eurasians, there is introgression from a Neandertal lineage, giving an ancestral Eurasian lineage EA. Suppose the derived allele has frequency *ea* at EA. Then conditional on *ea*, the expected derived allele frequencies in both Europe and Asia are *ea*. It follows that *D(Europe,Asia;Neandertal,Chimpanzee)* = 0, and this is still true if we condition on allele frequency in Africa, since Africa is symmetrical in this phylogeny with respect to Asia and Europe. In particular, if we condition on the African samples we analyze all carrying the ancestral allele, then the derived alleles observed in Eurasians will have an increased chance of being introgressed from the archaic lineage (if introgression in fact occurred) and this will 'enhance' the *D*-statistic signal.

Variations on this phylogeny in which there are multiple independent introgressions from the same archaic population (but with the same total archaic flow into Europe and Asia) still leave the expected value of *D* as 0 after conditioning on African alleles. It is relevant here, however, that Europe and Asia are assumed to be symmetrically related to Africa. If this is not true, and for example, there was African flow into Europe but not Asia after the 'Out of Africa' event, this would not only dilute the archaic flow into Europe, but also bias the Enhanced *D*-statistic; only the "Basic" *D*-statistic would be valid in this case.

Computation of the Enhanced D-statistic

To compute the Enhanced *D*-statistic, we used data from 35 sub-Saharan African individuals whose reads we had mapped to chimpanzee, requiring that ≥18 (at least half) had read coverage at each site we analyzed. We also required that ≥99% of reads carried the ancestral allele. The 35 samples were:

- 5 individuals sequenced as described in Note 5: San, Dinka, Yoruba, Mandenka and Mbuti.
- 30 YRI (Yoruba) individuals sequenced as part of the 1000 Genomes Project (*26*). We downloaded all YRI data from the Pilot of the 1000 Genomes Project that contained at least 75 base pair paired-end reads to improve mapping accuracy (only 30 of the YRI samples had reads that satisfied this). We mapped them to *PanTro2* using the same parameters as in Note 5. The samples were NA18486, NA18487, NA18498, NA18516, NA18520, NA18853, NA18856, NA18867, NA18868, NA18871, NA18873, NA18874, NA18908, NA18910, NA18917, NA18923, NA18924, NA18933, NA18934, NA19116, NA19130, NA19172, NA19197, NA19198, NA19213, NA19223, NA19235, NA19236, NA19247 and NA19248.

Table S25 compares the Basic *D*-statistics to the Enhanced *D*-statistics, showing the power of the approach to amplify the significance of signals of gene flow. (This is also shown in Figure 4 of the main text.) For example, statistics involving Papuans are significantly more skewed from zero in terms of absolute magnitude and in terms of statistical significance than is the case for the Basic *D*-statistics. We find that $D_{enhanced}$(Papuan, Pool of 5 non-Africans, Denisova, Chimpanzee)=55.4±1.9% (Z=29.4), while D_{basic}(Papuan, Pool of 5 non-Africans, Denisova, Chimpanzee)=6.2±0.5% (Z=12.5).

The increased power afforded by the Enhanced *D*-statistics leads to two novel insights:

(1) <u>*Less archaic ancestry in Europeans than in eastern non-Africans*</u>

We detect significantly less archaic ancestry in Europeans than in eastern non-Africans. Computing all possible *D*-statistics of the form $D_{enhanced}$*(East, Europe, Denisova, Chimp)*—to enhance our power to detect the skew from zero we were originally detecting using the D_{basic}*(Dai, French, Denisova, Chimp)* statistic—we find that the majority are highly significant at |Z|>4 standard errors from zero (Table S25). Focusing on pools of samples to increase power, $D_{enhanced}$*(3 East, 2 Europe, Denisova, Chimp)* = 11.0 ± 2.1% (Z=5.2). We note that our finding of less archaic ancestry in Europe than in eastern non-Africans is likely to reflect the same patterns Skoglund and Jakobsson reported when they detected less archaic ancestry in Europeans than in eastern non-Africans (*17*). However, our findings do not support Skoglund and Jakobsson's interpretation that the excess archaic ancestry is Denisovan rather than Neandertal in its affinity, as we discuss below.

(2) <u>*No signal of more Denisovan ancestry in Southeast than Northeast Asians*</u>

We used the Enhanced *D*-statistics to test whether there is more affinity to Denisova in southeast Asians (like Dai) than in northeast Asians and their relatives (Han and Karitiana) as was suggested by Skoglund and Jakobsson (*17*). Writing the statistic in the order $D_{enhanced}$*(Southeast Asian, Northeast Asian, Denisova, Chimp)* so that a positive value supports the hypothesis, we find:

D_{basic}*(Dai, Han, Denisova, Chimp)*	= 0.5 ± 0.4%	(Z = 1.2)
D_{basic}*(Dai, Karitiana, Denisova, Chimp)*	= 0.5 ± 0.4%	(Z = 1.2)
$D_{enhanced}$*(Dai, Han, Denisova, Chimp)*	= -2.4 ± 2.6%	(Z = -0.9)
$D_{enhanced}$*(Dai, Karitiana, Denisova, Chimp)*	= 3.8 ± 2.6%	(Z = 1.6)

These results suggest that Skoglund and Jakobsson's finding of more Denisovan affinity in the Dai than in the Han is consistent with a statistical fluctuation. We note that the *D*-statistic that Skoglund and Jakobsson report as evidence of more Denisova affinity in Southeast Asia than in Northeast Asia in Table S4 of their paper, *D(Southeast Asia, Northeast Asia, Denisova, Chimpanzee)* = 0.55 ± 0.23% (Z=2.4), is not significant after correcting for the 28 hypotheses they tested (P=0.46 based on a 2-sided test with a Bonferroni correction). In another paper where we surveyed Denisovan ancestry in many Asian populations (using SNP array data like Skoglund and Jakobsson) we also did not find compelling evidence of significantly more Denisovan ancestry in Southeast than in Northeast Asians (*2, 6*).

The extra archaic ancestry in Eastern non-Africans is consistent with Neandertal gene flow

The *D*-statistic analyses show that Denisova is significantly more closely related to Eastern non-African populations (represented here by Dai, Han and Karitiana) than to European populations (represented here by Sardinian and French). However, the analyses do not indicate what archaic group is responsible for the extra archaic ancestry in Eastern populations, and do not rule out the possibility that it could be Neandertals that are responsible for it (since they are related to Denisova (*2*)).

To evaluate if the extra archaic ancestry in eastern non-African populations is more closely related to Neandertals or to Denisova, we used the *S*-statistics defined in Equation S8.2 of the Denisova draft genome paper (*2*) (these are the numerators of the *D*-statistics). The *S*-statistics measure the absolute excess of BABA over ABBA sites comparing two human genomes (H_1 and H_2) at sites where a third genome that we are comparing to (Neandertal or Denisova) carries the derived allele, and are expected to be directly proportional to the excess of archaic ancestry present in sample H_1 vs. sample H_2:

$$S(H_1, H_2; Neandertal, Chimp) = \sum_{i=1}^{N} [p_{H1}^i (1 - p_{H2}^i) - (1 - p_{H1}^i) p_{H2}^i] p_{Neandertal}^i$$
$$S(H_1, H_2; Denisova, Chimp) = \sum_{i=1}^{N} [p_{H1}^i (1 - p_{H2}^i) - (1 - p_{H1}^i) p_{H2}^i] p_{Denisova}^i \quad (3)$$

To compute the *S* statistics in practice, we merged our dataset with the filtered Neandertal BAM files that we previously used for computing *D*-statistics in the Neandertal draft genome paper (*1*). Since we did not have genotype calls for the Neandertal, we instead used the raw reads to compute the probability of Neandertal carrying the derived allele at each site (we excluded sites with C→T and G→A substitutions because of their known high error rate in the Neandertal genome (*1*)). For symmetry, we filtered Denisovan raw reads in the same way. We then focused on the difference between these two *S*-statistics as a quantity that is informative about whether the introgression is from Denisova or Neandertal:

$$\Delta S(H_1, H_2) = S(H_1, H_2; Denisova, Chimp) - S(H_1, H_2; Neandertal, Chimp) \quad (4)$$

If $\Delta S(H_1, H_2)$ is positive, then the excess archaic ancestry in sample H_1 compared with H_2 is more closely related to Denisova (since it shares more derived alleles with Denisova), and if it is negative it is more closely related to Neandertals. We computed $\Delta S(H_1, H_2)$ empirically for three pairs of present-day human populations (Table S27). For the tests that did not involve a comparison to Africans, we also computed an Enhanced $\Delta S(H_1, H_2)$ statistic, requiring that ≥99%

of reads from 35 sub-Saharan Africans to carry the ancestral allele, thus amplifying signals of true introgression.

$\Delta S(Europe, Africa)$ Z_{Basic} = -13.0

Here Europe=Sardinian+French, and Africa=Yoruba+Dinka+Mandenka+Mbuti. This is known Neandertal introgression, so the negative Z-score is expected.

$\Delta S(East, Europe)$: Z_{Basic} = -1.8 $Z_{Enhanced}$ = -3.5

Here, East = Dai+Han+Karitiana and Europe = Sardinian+French. The negative Z-score shows that the excess archaic material present in eastern non-African populations compared with Europeans is more Neandertal than Denisova related.

$\Delta S(Papuan, East)$: Z_{Basic} = +5.7 $Z_{Enhanced}$ = +5.9

Here East = Dai+Han+Karitiana. This is a known case of Denisovan introgression, and so the positive statistic is expected.

We conclude that the extra archaic material in eastern non-Africans compared with western non-Africans is consistent with being more closely related to Neandertals than to Denisova, and in particular, the difference between the *S*-statistics does not have the positive sign that is characteristic of Denisova gene flow into Papuans.

These findings are inconsistent with the parsimonious model suggested by the Neandertal genome paper (*1*): that all the Neandertal material in non-Africans is due to introgression of Neandertals into the common ancestral population of non-Africans before they diverged, and that there was no subsequent human migration out of Africa. The next-simplest scenarios that could explain these patterns are the following:

(1) *At least two episodes of gene flow*. There were at least two gene flows that were entirely independent: one into European and one into East Eurasian ancestors, along the lines suggested by Currat and Excoffier (*18*).

(2) *An additional wave of gene flow into the ancestors of Eastern populations*. There could have been a common gene flow event into the ancestral population of Europeans and Eastern non-Africans, followed by additional gene flow into Eastern non-African ancestors.

(3) *Dilution of the Neandertal proportion in West Eurasians.* A qualitatively different explanation is a single Neandertal gene flow event into the common ancestors of all non-Africans, followed by dilution of the proportion of Neandertal material due to additional mixing with populations without Neandertal ancestry (perhaps additional flow from Africa). In the next section, we estimate that the proportion of Neandertal ancestry in Europeans is 64-88% of that in eastern non-Africans (95% C.I.), so this would imply that later migrations out of Africa contributed one minus this, or 12-36% to the ancestry of Europeans. This proportion is too large to be explained by the fact that some southern Europeans have up to 5% sub-Saharan African ancestry due to gene flows in the last few thousand years (*71*); thus, if such a history explains the data, it must reflect more ancient African gene flows.

The scenarios above are merely the simplest that are consistent with our data, and the truth may be even more complicated. The fact that we now have evidence of different proportions of

Neandertal ancestry in two groups of Eurasians suggests that as higher resolution ancestry estimates become available and we examine more present-day humans, we may detect additional variation in the proportion of archaic gene flow.

Estimates of Neandertal and Denisovan ancestry in present-day humans

Having documented variability in the proportion of archaic ancestry in different groups of non-Africans, we used our data and new methods to obtain updated mixture proportion estimates.

Statistics for estimating ancestry proportion

We wish to estimate how much more archaic ancestry is present in one present-day human sample H_1 than another H_2. To do this, we use a modification of the *S*-statistic ratio strategy that we developed in the draft Denisova genome paper (*2*).

We first describe our estimate $Nea(H_1,H_2)$ of Neandertal ancestry proportion. Denote the Neandertal proportion in H_1 as f_1, and in H_2 as f_2. We assume that neither H_1 nor H_2 has any Denisovan ancestry. We now estimate f_1-f_2 with the statistic:

$$Nea(H_1,H_2) = \frac{S_{Denisova}(H_1,H_2)}{S_{Denisova}(Neandertal,H_2)} = \frac{\sum_{i=1}^{N}[p_{H_1}^i(1-p_{H_2}^i)-(1-p_{H_1}^i)p_{H_2}^i]p_{Denisova}^i}{\sum_{i=1}^{N}[p_{Neandertal}^i(1-p_{H_2}^i)-(1-p_{Neandertal}^i)p_{H_2}^i]p_{Denisova}^i} \quad (5)$$

$Nea(H_1,H_2)$ provides an estimate of f_1-f_2, as documented pictorially in Figure S16. Specifically, $S_{Denisova}(H_1,H_2)$ measures the excess rate of matching of derived Denisovan alleles to sample H_1 vs. sample H_2. If H_1 and H_2 both have the same proportion of Neandertal ancestry, $S_{Denisova}(H_1,H_2)$ has an expectation of zero. If H_1 is a Neandertal and H_2 is an unadmixed modern human, the statistic has a non-zero expected value (which we arbitrarily denote as *K)*. Thus:

$$\begin{aligned} E[S_{Denisova}(H_1,H_2)] &= E[S_{Denisova}(f_1 Neandertal + (1-f_1)Modern,\ f_2 Neandertal + (1-f_2)Modern)] \\ &= f_1 f_2 0 + f_1(1-f_2)(K) + (1-f_1)f_2(-K) + (1-f_1)(1-f_2)0 = (f_1-f_2)K \end{aligned} \quad (6)$$

By dividing $S_{Denisova}(H_1,H_2)$ by $S_{Denisova}(H_1,Neandertal)$, we measure what fraction of the way sample H_1 is toward having entirely Neandertal ancestry, compared with H_2 as a baseline. *S*ince 1-$f_2 \approx 1$ (Neandertal mixture is small in all modern humans), we can write:

$$E[Nea(H_1,H_2)] = \frac{(f_1-f_2)K}{(1-f_2)K} = \frac{(f_1-f_2)}{(1-f_2)} \cong f_1 - f_2 \quad (7)$$

A similar argument allows us to estimate the excess Denisovan ancestry in sample H_1 versus sample H_2 assuming that they have the same proportions of Neandertal ancestry:

$$Den(H_1,H_2) = \frac{S_{Neandertal}(H_1,H_2)}{S_{Neandertal}(Denisova,H_2)} = \frac{\sum_{i=1}^{N}[p_{H_1}^i(1-p_{H_2}^i)-(1-p_{H_1}^i)p_{H_2}^i]p_{Neandertal}^i}{\sum_{i=1}^{N}[p_{Denisova}^i(1-p_{H_2}^i)-(1-p_{Denisova}^i)p_{H_2}^i]p_{Neandertal}^i} \quad (8)$$

A key feature of these statistics is that they directly estimate ancestry proportion which is a quantity of historical interest. Thus, even if there are differences in effective population size across different parts of the genome, they still work to estimate ancestry proportion (for example, they are equally valid on chromosome X as on the autosomes, even though the effective

population size of chromosome X is expected to be ¾ that of the autosomes). This contrasts with *D*-statistics whose values are affected by population sizes.

Our most interesting analyses are based on four classes of statistics:

(1) $Nea(Europe, Africa)$
This estimates the Neandertal ancestry proportion in Europeans, relative to sub-Saharan Africans with 0% Neandertal ancestry (Yoruba, Dinka, Mandenka or Mbuti).

(2) $Nea(Eastern, Europe)$
This estimates excess Neandertal ancestry in Eastern populations vs. Europeans.

(3) $Den(Papuan, Eastern)$
This estimates excess Denisovan ancestry in Papuans relative to a baseline assumed to be 0% in Eastern populations (Dai, Han and Karitiana).

(4) $Den(Dai, Han)$
This estimates excess Denisovan ancestry in southeast Asians (Dai) relative to a baseline assumed to be 0% in northeast Asians (Han).

Enhancing the power of the ancestry estimates

To improve the precision of our ancestry estimates, we use the same idea that we report for enhancing the power of the *D*-statistics; that is, we restrict to sites that meet certain criteria based on the genotypes in outgroups (sub-Saharan Africans or archaic humans).

As shown in Figure S15, under the null hypothesis that H_1 and H_2 have equal proportions of archaic ancestry, *Nea*(H_1,H_2) and *Den*(H_1,H_2) will have an expected value of zero if we condition on the genotypes in samples symmetrically related to H_1 and H_2. Under the alternative that H_1 and H_2 have different proportions of archaic ancestry, we can restrict to sites at increased likelihood of being due to introgression. This amplifies the rate of archaic-derived sites in the numerator and denominator by the same factor (since in both cases the archaic material is drawn from archaic populations in the same clade), and so the expected value of the ratio is not biased. To compute enhanced ancestry estimates in practice, we restrict to sites meeting two criteria:

(1) *We require Saharan African outgroups to H_1, H_2 to carry the ancestral allele.*

For computing *Nea(West,African)*, we do not perform any enhancement because Africans are directly involved in the computation.

For computing *Nea(East,West)* and *Den(Papuan,East)*, we restrict to sites where >99% of the reads from the 35 sub-Saharan Africans carry the ancestral allele, and where we have representation for at least 18 of these individuals. (These are the same criteria that we use for the Enhanced *D*-statistics above).

(2) *We require one of the two archaic samples to carry the derived allele.*

For computing *Nea*(H_1,H_2), we require that all Denisovan reads carry the derived allele, and for computing *Den*(H_1,H_2) we require that all Neandertal reads carry the derived allele. This requirement enriches for sites that genuinely arose due to gene flow from archaic humans. The choice of which archaic population we use (Denisova for *Nea*(H_1,H_2), and Neandertal for

Den(H_1,H_2)) is motivated by the fact that we need to use an archaic sample for conditioning that is equally an outgroup to the populations in the numerator and the denominator.

Practical computation of the ancestry estimates

We used the BAM file generated for the draft Denisova genome paper (*2*) to identify sites where (i) all reads from the archaic sample used are derived, and (ii) the substitution is not C→T or G→A (because of the high error rate at such sites in ancient DNA, especially in the Vindija Neandertal).

We next annotated the sites identified in this way using data from the 12 deep genome sequences generated for this paper, as well as the 30 YRI samples from the 1000 Genomes Project mapped to the chimpanzee reference sequence as described above. We restricted analysis to sites with a mappability score of Map20=1, and where the read coverage is within the 95% central interval of the genome-wide distribution, computing the genome-wide distribution separately on chromosome X and the autosomes to control for the lower coverage of chromosome X in males. For chromosome X, we also excluded the pseudoautosomal and X-Y transposed regions.

To estimate the probability of carrying derived alleles (p_{H1}, p_{H2}, $p_{Neandertal}$, and $p_{Denisova}$), we used the reads overlapping each site, rather than the genotypes, since this allowed us to use the Neandertal low coverage data. As discussed above, the *D*-statistics are not affected by whether we compute them using reads or genotypes. We represented the present-day human populations H_1 or H_2 by either individual samples or (in some instances) pools to increase precision. The pools we used were "African" (Yoruba, Dinka, Mbuti), "Europe" (Sardinian, French), "East" (Han, Dai, Karitiana), and "Papuan". For any site, we require at least one sample in each pool to have data passing filters. If multiple samples pass, we average.

A practical challenge that we encountered in estimating *Nea*(H_1,H_2) and *Den*(H_1,H_2) is that by restricting to sites where we have data from all four samples {H_1, H_2, Denisova, and Neandertal} at each site we analyzed, we substantially reduced the size of the dataset. This decreases the power of these statistics in light of the low coverage we have from Neandertal, and so we expect that ancestry estimates such as we generate here will become more precise once a deep Neandertal genome sequence becomes available.

Results

Table S28 and Table S29 present the ancestry estimates we obtain for the autosomes and for individual chromosomes, as well as a Z-score for whether the estimates are different. For the autosomes (chromosome 1-22), we obtain a standard error by dividing the genome into 500 blocks each with an equal amount of data, and computing a Block Jackknife standard error. For individual autosomes, we use 50 equally sized blocks.

(1) We observe that our inferences based on individual genome sequences are consistent with pools of genome sequences from the same regions ("Africa", "Europe", and "East"). The pool-based estimates are more accurate, reflecting the larger amount of data.

(2) Our autosomal ancestry estimates are statistically consistent with those in our previous publication on the draft Denisova genome (*2*). We previously estimated a Neandertal proportion in Eurasians of 2.5 ± 0.6%, and here we estimate 1.0 ± 0.3% in West Eurasians

plus an additional 0.7 ± 0.2% in Eastern non-Africans (a total of 1.7%). We previously estimated Denisovan ancestry in Papuans to be 4.8 ± 0.5%, and here we estimate 3.0 ± 0.8%.

(3) Our Denisovan ancestry estimates in Papuans are larger on the autosomes (3.0 ± 0.8%) than on chromosome X (0.0 ± 0.9%), a significant excess (2.6 standard errors from zero; P=0.01 by a 2-sided test). The signal is observed not just when we compare Papuans to a pool of 3 East Eurasian and related samples, but also when we compare Papuans to individual samples from this region: Han (Z=1.8), Dai (Z=2.4) and Karitiana (Z=3.0). It is also not an artifact of the number of blocks we use in the Block Jackknife, since the standard errors on the chromosome X estimate are similar regardless of the number of blocks we use (±0.97% for 25 blocks, ±0.90% for 50 blocks, ±0.89% for 100 blocks, and ±0.72% for 200 blocks).

Interpretation

These results document more Denisovan ancestry in Papuans on the autosomes than on chromosome X. We considered two scenarios that could explain this finding:

Scenario #1: Sex-biased demographic history.

(#1A) If Denisova males contributed more genes to modern humans than Denisovan females – perhaps because after mating with male Denisovans, pregnant modern humans tended to raise their children among their modern human relatives – the impact on chromosome X variation would be smaller since men carry only one X chromosome for every two in women. Even the extreme scenario of sex-biased gene flow would be expected to produce no less than half the Denisovan ancestry on chromosome X as on the autosomes, and we observe a smaller fraction. However, given our substantial standard errors (3.0 ± 0.8% on the autosomes and 0.0 ± 0.9% on chromosome X), we cannot rule out the possibility that the ancestry proportion on chromosome X is this large.

(#1B) These patterns may reflect sex-biased demography in the expanding modern human population that encountered Denisovans. It is well documented that in many modern human hunter-gatherers, there is more female than male migration among neighboring groups (*72-74*). In a scenario where the expanding modern human population was substructured when it encountered Denisovans, with more female than male migration among neighboring groups, then any introgression would be expected to affect chromosome X less than on the autosomes because of greater rates of within-modern human migration on chromosome X (*19*). Intuitively, the higher rates of female migration has the effect of increasing the effective population size of the expanding population on chromosome X relative to what is expected from the autosomes (since females carry a disproportionate fraction of all copies so chromosome X), so that any introgression from Denisova has a proportionally smaller effect on chromosome X.

Scenario #2: Natural selection or meiotic drive removing Denisovan chromosome X

An alternative scenario that might explain these findings is natural selection or meiotic drive, which removed archaic X chromosomes from the admixed populations after introgression. Archaic populations were diverged from modern humans when the two met. As population divergence increases, hybrid sterility and inviability factors accumulate and tend to be concentrated on chromosome X (the "Large X Effect") (*20*). In light of this, it is plausible that hybrid incompatibility factors might have existed between modern and archaic humans when they met and interbred, so that after the gene flow, natural selection or meiotic drive removed the

genetic material derived from archaic humans quickly enough to affect large proportions of chromosome X via negative selection, thus reducing the impact of the introgression on chromosome X (*75, 76*).

To search for evidence that natural selection affects archaic ancestry proportion as it varies across the genome, which would support Scenario #2, we first considered a potentially analogous observation: the recent finding that the ratio of chromosome X to autosomal genetic diversity in humans is significantly lower close to genes than far away from genes, an effect that can only be ascribed to the effects of natural selection (*77*). To test whether proximity to genes is also affecting archaic ancestry, we stratified our data based on the proximity to genes and conserved non-coding elements, using the B-statistic of McVicker et al. (*50*) which is known to be correlated to genetic diversity both within the autosomes and within chromosome X, and indeed is strongly predictive of the X-to-autosome genetic diversity ratio (*78*). Table S29 shows that there is no correlation of archaic ancestry proportion on the autosomes to proximity to B to within the limits of our resolution, in sharp contrast to what is observed for genetic diversity. Thus, our B-statistic analysis provides no support for Scenario #2.

As a second way of searching for evidence of an effect of natural selection on archaic ancestry proportion as it varies across the genome, we computed individual estimates of ancestry proportion on each chromosome (not just on chromosome X), which we report in Table S30. This analysis shows evidence of some natural selection (providing evidence of at least some influence of Scenario #2) in that it shows that there is significant variability in Denisovan ancestry proportion compared with the genome-wide average not just on chromosome X, but also on individual autosomes that have estimates that are also lower (or higher) than the genome-wide average. For example, on chromosome 11, Denisovan ancestry is estimated to actually be lower in Papuans than in East Eurasians, $Den_{chr11}(Papuan,East)$ = -4.9 ± 1.9%, opposite to the rest of the genome.

The results of Table S30 provide some corroborating evidence for natural selection playing some role in affecting ancestry proportions as it varies across the genome. Chromosome X stands out, however, in consistently having a low estimated archaic ancestry proportion compared with the genome-wide average, whether we estimate *Nea(Europe,Africa), Nea(East,West)* or *Den(Papuan,East)* (Table S30), a pattern that is not seen, for example, on chromosome 11 or on any other autosome. This suggests that reduced evidence of archaic gene flow may be a general feature of chromosome X.

A goal for future work should be to better understand these patterns, and in particular to understand whether scenario #1 or scenario #2, a combination of these two, or another scenario altogether, explains the data.

Note 12: Relationship between Denisova and 11 present-day human genomes

Qiaomei Fu, Udo Stenzel, Martin Kircher, Janet Kelso*

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

We previously studied the relationship between Denisova and present-day humans using *D*-statistics (*2*) (see also Note 11) and the related *S*-statistics (*6*). To further explore the relationship between Denisova and present-day humans, we used the program TreeMix (http://code.google.com/p/treemix/) which estimates population splits and admixtures for a set of populations from their genome-wide allele frequency distributions. Potential migration events are inferred when migration provides a significantly better explanation of the observed data than the maximum-likelihood tree alone (*16*).

We obtained allele distributions from the combined genotype call files described in Note 6. We extracted genotype information for Denisova and the eleven present-day humans for sites on the autosomes that fulfill all of the following criteria:

(1) Sites fall in regions of maximal mappability (Map20 = 1.0).

(2) Sites are in "simple" regions with a clear phylogeny, i.e. the 'HPGO' substring is required in the 'TS' field of the genotype call files to ensure that three ape outgroups are present. Further, we require that no more than one chimpanzee and one human sequence be present in the 'TS' field, thus avoiding regions of human- or chimpanzee- specific duplication.

(3) The coverage of the site is within the coverage ranges for all individuals, as defined in Note 8.

We used the base of the human-chimpanzee ancestor to define the ancestral allele as described in Note 9 and counted the number of ancestral and derived alleles in each of the individuals. We did not require the derived allele to be identical between individuals and considered only sites with at least one derived allele. Following this approach, we ascertained a total of 5,115,249 genomic sites.

Comparing Denisova and eleven present-day humans using TreeMix (Figure S17), we see that the method reconstructs previously known population splits, with African populations separating early from all non-Africans. The San and Mbuti have the deepest population split, and the Dinka is the African population most closely related to all non-African populations. The residual signal in the population covariance matrix indicates a relationship between the Denisovan and Papuan

populations, and when one migration/admixture event is included we see that this relationship is best explained by an admixture from a population related to Denisova into Papuans. In order to quantify the weight and confidence of this migration event, we use a block jackknife (with 5,000 sites in each block, ~3 Mb) to obtain standard errors and P-values. The TreeMix migration event weight for the Denisovan ancestry in the Papuan individual studied is 6.0% ± 0.9% standard errors ($P=2.14\times10^{-11}$), which is more than the estimated admixture proportions from *S*-statistics (Note 11). However, a deviation from the true admixture proportion is expected due to the exclusion of Neandertal in this analysis.

Note 13: Segmental duplication and copy number variation analysis

Peter H. Sudmant*, Can Alkan, Evan E. Eichler

* To whom correspondence should be addressed (psudmant@u.washington.edu)

METHODS

Segmental duplications. We applied the whole-genome shotgun sequence detection (WSSD) method to identify large (>10 kbp) regions of segmental duplication (*79-81*). We rendered the next-generation sequence (NGS) reads of the Denisovan sample into k-mers of 36 bp (n = 1,908,363,786; total 68.7 Gbp), after removal of PCR duplicates using BWA and SAMtools. Next, we mapped them to a repeatmasked reference genome (NCBI GRCh37) using the mrFAST aligner (*80*) with an edit distance of at most two. After applying GC normalization to correct for sequencing biases, we calculated raw copy numbers within 1 kbp non-overlapping windows. We then predicted regions of segmental duplication by using both raw copy numbers and regions of excess read-depth (5 kbp overlapping windows, sliding 1 kbp across the genome). To compare the segmental duplications detected in the Denisovan genome, we downloaded sequence data corresponding to other branches of primate evolution (modern human, Neandertal, chimpanzee, bonobo, gorilla, and orangutan). We repeated the procedure using the underlying whole- genome shotgun datasets from each species since previous analyses had been performed with different versions of the reference genome. Since coverage of most genomes in the 1000 Genomes Project is low, we additionally analyzed the duplication architecture of ten human genomes from the Human Genome Diversity Panel (HGDP) where underlying NGS data had been generated at 20-30X sequence coverage. All segmental duplications are mapped to GRCh37 coordinates and summary statistics of read-depth as well as estimated number of Mbp of duplication of various size thresholds are presented (Table S31).

Quality control analysis. In order to assess our ability to detect duplications and genotype copy number in the Denisovan genome, we compared a set of quality control metrics quantified in the Denisovan genome to those in the set of ten diverse humans sequenced to similar coverage (Table S32). We first compared the correlation between read-depth and 32 ancestral segmental duplications of known copy number among the Denisova and the ten humans. All individuals demonstrated a correlation in excess of 0.9 indicating that read-depth will provide accurate estimates of copy number. We next estimated the copy number of 131,242 3 kbp unmasked windows contained within 4,835 putative diploid contiguous regions >100 kbp. These regions were selected by subtracting known copy number variants (The Database of Genomic Variants, (*82*)), segmental duplications (UCSC Genome Browser), and genomic gaps from the human reference genome and retaining those fragments >100 kbp. Among all 11 genomes, 98.1%-98.9% of all copy number 2 windows were correctly genotyped with 98.9% of all windows in the Denisova being correctly estimated as copy number 2. The slight increase in accuracy in the Denisova is likely a result of its increased coverage and shorter read length. Read-depth assessment is dependent on the number of independent read counts. Finally, we assessed the

fraction of correctly determined copy number 2 windows as a function of the GC content of the window. GC biases introduced during library construction can often skew read-depth copy number analyses. Within the Denisovan genome in addition to the ten HGDP genomes no GC-associated read-depth biases were observed (Figure S18).

Gene copy numer analysis. We used the RefSeq gene annotation table (May 2010, n = 28,565 gene models) from the UCSC Genome Browser. We assigned each gene model the median copy number value of the windows that it overlaps using non-overlapping 1 kbp windows**.** We additionally performed paralog-specific copy number analysis of the genome using the method of singly unique nucleotide (SUN) k-mers (SUNKs) (*81*). Reads were mapped to the reference genome as above with no mismatches (edit distance=0). Reads were then filtered to only those that could be unambiguously assigned a unique location in the genome and then corrected for GC sequencing biases. Copy numbers were estimated by constructing a calibration curve with read-depth of overlapping SUNKs using regions of known copy number.

RESULTS

We annotated segmental duplications from Illumina genome sequence data from seven representative great apes and *Homo* samples: Denisova (this study), Neandertal (*1*), modern human (NA18507 (*32*), chimpanzee (*83*), bonobo (*84*), gorilla (*85*), and orangutan (*86*) using the same methodology and parameters (Table S31) and mapping to GRCbuild37. We compared the Denisova to each of the other six to identify the number of shared and specific duplications (Table S33). We restricted our initial analysis to autosomal duplications (>20 kbp) to maximize sensitivity and control for gender differences. As expected, the Denisovan genome showed more similarity with modern human and Neandertal than the other primate genomes.

The Venn diagram compares segmental duplications detected in Denisova, Neandertal, and NA18507 (Figure S19A). Of the segmental duplications detected in the Denisovan genome, 99.5 Mbp (86.12%) were also found in both Neandertal and NA18507, and an additional 7.25 Mbp (6.2%) of duplications were found in either Neandertal or NA18507. In this comparison we detected 8.8 Mbp of *potentially* Denisova-specific duplications. Note that the Venn diagram is constructed by basepair-level intersections among the three genomes compared, and intersections with less than 50% reciprocal overlap are also included. Therefore, the numbers differ from the analysis presented in Table S33. Since this initial comparison included only one human genome sample, we repeated the analysis using ten additional diverse modern human samples where comparable deep genome sequence data were available (HGDP)**.** Relaxed thresholds were used to call duplications in these ten individuals and the Denisovan individual by chaining together consecutive windows spanning at least 9 kbp of unmasked sequence with copy number >3 (Figure S19B). 123.3 Mbp of duplications were identified in the Denisova, 123.2 Mbp of which (99.92%) overlapped duplications identified in the ten HGDP individuals demonstrating that the duplication architecture of the Denisova is almost identical to that of modern humans. Seven duplications, however, encompassing 236 kbp of sequence appeared to be Denisova-specific.

We manually inspected these seven "Denisova-specific" duplications and found that five of these duplications were in fact present in one of the ten HGDP individuals but fell below our initial length thresholds. These regions, thus, had simply escaped detection as a result of filtering (i.e., chained windows of >9 kbp unmasked sequence). For the remaining two Denisova-specific duplications for which there was no evidence of duplication in the ten reference genomes, we analyzed 146 individuals sequenced as part of the 1000 Genomes Project Pilot I (*26*) and found neither to be present in this large panel of modern humans (Figure S20).

The first duplication with an estimated copy number of three, lies in pericentromeric chromosome 3, spans 31 kbp (chr3:87,859,313- 87,890,719; Figure S21), overlaps no genes, and is unlikely fixed because of its diploid copy number of 3. This duplication shows no excess of single nucleotide divergence potentially consistent with its more recent origin and the observation that it is copy number polymorphic. The second maps to chromosome 4q13.2 (chr4:68,542,678-68,585,941; Figure S21 and S22), spans 43 kbp, and overlaps two genes *UBA6* and *LOC550112*. We estimate four copies of this segment in the diploid genome suggesting a higher prevalence and that it is likely fixed in the Denisovan genome. The duplication includes the promoter and the first six exons of *UBA6* as well as the two exons of *LOC550112*. *UBA6* is a recently discovered ubiquitin activating enzyme involved in protein degradation and signaling. The gene may play a role in germ cell differentiation, spermatogenesis, and male fertility (Hogarth, 2011). *LOC550112* has no characterized function.

Since experimental validation of these Denisovan duplications is not possible, we examined the extent of sequence divergence reasoning that *bona fide duplications will show an excess of single nucleotide variants due to paralogy.* We computed the number of SNVs that are supported by at least two reads with basepair phred quality value >20 using the mrFAST alignments (to repeatmasked genome) described above. The putatative SD on chromosome 3 showed an average of 2.94 SNVs over 1 kbp unmasked windows, while the second Denisova-specific SD on chromosome 4 had 13.22 SNVs in 1 kbp unmasked windows on average. For control unique regions, we found 2.58 SNVs, on average, per 1 kbp unmasked sequence (standard deviation 2.33). Since the chromosome 4 USP6 duplicaiton shows an excess of SNVs (in excess of 5 standard deviations above the average number of SNVs in the control regions), we conclude that the *UBA6* duplication is *bona fide* and likely unique to the Denisovan lineage as compared to extant humans.

Copy number analysis. We performed a copy number analysis of the entire Denisovan genome using previously described methods (*81*) allowing us to compare patterns of copy number between Denisova and contemporary modern-day humans for any region of interest. Most regions of copy number variation in the Denisova were similar to that of the human species. In particular, we analyzed the human-specific chromosome 2 fusion locus as fusion of the ancestral chromosomes 2A and 2B occurred after the divergence humans and chimpanzees and this locus is known to be highly duplicated (Figure S23). The boundaries and copy number of the duplications are identical between Denisova and modern day humans implying that the fusion is ancestral to both lineages.

We searched for any genes that had had been specifically deleted or expanded in the Denisovan genome when compared to ten deeply sequenced diverse human genomes. We performed SUNK

(single unique nucleotide kmer)-based copy number genotyping across 26,229 autosomal RefSeq gene models. SUNKs correspond to single nucleotide or indel differences that uniquely distinguish a specific location in the genome. The SUNK approach has the advantage that it can estimate the paralog-specific copy number of genes by considering reads uniquely assigned to SUNK positions in the genome. Regions with too few SUNKs will not be informative; thus, only loci with >1500 SUNKs (21,328 total) were considered in the analysis.

Five genes were predicted to be duplicated in the Denisova (more than 2 copies) (Figure S24) with a copy number greater than what was observed in all ten humans analyzed (Table S34). These included *CYP2D6*—a P450 enzyme involved in synthesis of cholesterol and steroids that was duplicated to four copies, and *KGFLP2*—a keratinocyte growth factor-like protein that was duplicated to three copies. We note that *UBA6* was not flagged in this analysis though *LOC550112* was, as the *UBA6* duplication was incomplete and encompassed only a small portion of the genomic to the gene. We additionally identified 19 genes potentially deleted in the Denisova (predicted copy numbers <2) (Figure S24) yet did not show evidence of deletion in the ten human genomes analyzed. Three of these deletions were homozygous including *GOLGA8*, a core-duplicon that has undergone rapid expansion throughout great ape evolution. Six genes were either diploid or duplicated in the Denisovan genome yet had undergone duplication to even higher copy in all of the ten humans analyzed. These include *TPTE*—a transmembrane phosphatase that is highly expressed in the testis, and the *HERC2P3* pseudogene. *HERC2* variants have been associated with skin, hair, and eye pigmentation (*87, 88*). We finally assayed these gene polymorphisms in 146 individuals sequenced by the 1000 genomes project Pilot 1. All of the gene polymorphisms identified with the exception of the previously mentioned *LOC550112, LOC550112* and *ZNF595* gene duplications were present as polymorphisms in modern humans (Table S34).

Human specific expansions. We finally sought to identify those regions which had undergone copy number expansions in the human lineage following the divergence of modern humans and the Denisova. We thus identified 37 regions which were at increased copy number in the 10 humans analyzed compared to the Denisovan individual. We next assessed the copy number of these 37 regions across 146 individuals sequenced as part of Pilot 1 of the 1000 genomes project identifying 10 of these regions to have expanded in all of the humans analyzed (Table S35). Notably, one of the expansions identified overlapped the 3-prime end of the ROCK1 gene and the segmental duplication associated with the human specific pericentric inversion of chromosome 18 *(89, 90)*. Duplicative transposition of 19kb of sequence in ancestral hominids from the q-arm to the p-arm in resulted in two segmental duplications in inverted orientation. This duplication is not present in any other non-human primates. The inverted orientation of these segmental duplications thus facilitated a peri-centric inversion accounting for the cytological difference between human chromosome 18 and the homologous chimpanzee chromosome 17 (Figure S25a). In contrast to humans, we find the Denisova to harbor only a partial duplication of the ROCK1 chr18 inversion locus (Figure S25b). Proximal to the partial ROCK1 duplication in the Denisova is a homozygous deletion suggesting that following a complete duplication of the 19kb ROCK1 duplicon, a partial deletion of the sub-telomeric copy of the duplication occurred in the Denisovan lineage. It is striking that this deletion is homozygous, suggesting it may be fixed in the Denisovan population, however, we are unable to resolve whether the deletion occurred before or after the pericentric inversion or if the pericentric inversion is indeed present in the Denisova.

Note 14: Chromosome Two Fusion Site

Kay Prüfer*

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

Modern humans differ from all other great apes in their number of autosomal chromosomes. This difference is caused by a fusion of two separate chromosomes (termed 2a and 2b in non-human great apes) into chromosome 2 (*91*). The fusion left a region with telomeric repeats that meet in forward and reverse direction in the interior of chromosome 2, in congruence with a head-to-head fusion of the original chromosomes (*25*). Here, we use the Denisova data to scan for reads that cover the site where the forward and reverse telomeric repeat sequence meet. A total of 12 unique fragments cover this position. In contrast, no alignments are found when testing over 20x genome coverage Illumina shotgun data from 16 chimpanzees and 3 bonobos. The chromosome two fusion is thus found in Denisovans and the fusion event must predate the split of Denisova and modern human.

Alignments to the Chromosome 2 Fusion Site

We scanned the alignments of all Denisovan reads to the human genome (see Note 4) for reads covering the chromosome 2 fusion site (hg19, chr2:114360250-114360750). We identified one read with a trimmed length of 57 bps that shows both the forward and reverse telomere repeat motif and maps uniquely to the region (mapping quality = 37). Due to the palindromic nature of the region forward and reverse alignments of this read both yield an alignment with three mismatches and one gap (see Fig. S26).

Realignment of Denisova Data

In order to identify more reads covering the fusion site, we used the identified read to construct two 500 basepair long target sequences that include the Denisova-specific substitutions observed in forward and reverse direction. We realign all untrimmed Denisovan reads to these constructed fusion sequences using BWA with more permissive alignment criteria (parameters: -o 3 -n 0.001 -l 16500). We identified12 unique fragments (i.e. reads with alignments in different orientation and different start and end coordinates) showing the forward and reverse telomeric repeat sequences (see Fig. S27).

When we realign these 12 unique fragments to the entire human genome (hg19) with the same relaxed parameters we find that fragments align either uniquely (5/12; mapping quality = 37) or have no alignment (7/12).

Testing Chimpanzee and Bonobo Sequences

Using identical alignment parameters and the constructed Denisovan target sequences, we test for the presence of similar sequences in Illumina sequencing data from 16 chimpanzees and 3 bonobos summing to a total of over 20x genome coverage combining all individuals (*84*). We find no read showing forward and reverse telomeric repeat sequences.

Note 15: Estimating heterozygosity in Denisova and 11 present-day human genomes

Cesare de Filippo, Kay Prüfer, Richard E. Green, Michael Siebauer, Katarzyna Bryc, Janet Kelso, Aida M. Andrés, Martin Kircher and Matthias Meyer*

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

Synopsis

We describe three different approaches for estimating heterozygosity in the Denisova and 11 present-day humans. Two of these approaches use allele counts from individual sequence reads while the third uses an aggregate genotype call for each genome position. In the first approach we count the number of reference and non-reference bases at sites with greater than 20-fold coverage to generate relative estimates of heterozygosity. In the second approach we use mlRho (*92*), a maximum likelihood method for estimating heterozygosity from shotgun sequencing data. In the third approach we count high-confidence genotype calls to estimate heterozygosity. Importantly, a genome mapability filter is crucial to produce consistent estimates in each method. With this filter, our heterozygosity estimates are very similar for the three approaches and are highly robust to further filtering strategies. We consistently detect less than half as many heterozygous positions in Denisova than in Karitiana, the individual showing the lowest heterozygosity in our panel of 11 present-day humans. Table S36 summarizes the heterozygosity estimates obtained using the three methods.

[1] Measuring heterozygosity by comparing reference vs. non-reference base frequencies

Due to the small evolutionary distance between Denisova and present-day humans, the vast majority of positions in the Denisovan genome are either homozygous for the human reference allele, homozygous for a non-reference allele, or heterozygous reference/non-reference (*i.e* positions heterozygous for two non-reference alleles are rare (compare Note 6)). Based on this notion, we explored how well genotypes can be differentiated in Denisova and the 11 present-day human genomes by analyzing the frequency spectra of DNA sequences carrying the human reference base or a non-reference base at sites where Denisova and all eleven humans show at least 20-fold coverage. To make the base frequency spectra comparable among samples and sites, we randomly sampled 20 sequences if a position was covered more than 20-fold in an individual. Analyses were performed using VCF files as input (Note 6). Insertions and deletions to the human reference genome sequence were disregarded in this and all other analyses.

We first explored the effect of filtering by selecting genomic positions based on (i) mapability (Duke Uniqueness 20 bp track, score = 1), (ii) map quality (MQ) of reads covering a position (MQ>=30, representing the quadratic mean of map qualities of individual reads), or (iii) upper

coverage cutoffs (eliminating the 2.5% of positions with highest coverage from each sample, a lower cutoff is implicit in this analysis; for upper cutoffs see Note 8, Table S8.2). To ensure full reciprocal overlap of sites for all individuals, we excluded positions that did not pass filters in all individuals.

The base frequency spectra (Figure S28) show central peaks of putative heterozygous sites. If no filter is applied this peak is less symmetric and not clearly centered around 10 reference alleles, indicating that filtering is required to exclude difficult-to-align parts of the genome. While all filters were effective in cleaning the base frequency spectra, we prefer a mapability filter, which is fully independent of the actual sequencing data and restricts to the regions of the genome that are *a priori* amenable to high-confidence read mapping. Since the base frequency spectrum of each individual (provided in Figure 5A and Figure S29 in different scales) is counted using the same set of positions (i.e., the same regions of the genome), the proportion of sites where exactly ten reference alleles are seen in a given genome should be proportional to the heterozygosity of that genome. Thus we can calculate the relative rate of heterozygosity between Denisova and a present-day individual as the ratio of counts of ten reference alleles for Denisova and the count for the present-day individual. In this manner, we estimate the relative heterozygosity of Denisova compared to each of the 11 present-day humans (Table S37). Calculating these ratios under different ways of data filtering also demonstrates that these relative heterozygosity estimates are largely robust to the choice of filters applied.

[2] Measuring heterozygosity using mlRho

We used mlRho (*92*), a maximum likelihood method that co-estimates the population mutation rate (θ) and the sequencing error rate (ε) from deep sequencing shotgun data of a single individual. Under the infinite sites model and for small values of θ, the estimated population mutation rate is a good estimator for heterozygosity. For our analysis, we used the human mapping of the Denisova data and the 11 humans (see Note 4). All alignments were filtered for mapping quality (MQ >= 30) and base quality (BQ >= 30). Table S38 shows the estimated θ and ε for each individual. As before, we tested an additional filter for mapability according the Duke University uniqueness track of 20-mers downloaded from the UCSC genome browser (*93*) to further reduce the contributions of potential mismappings. This filter lowered heterozygosity estimates for all genomes but in particular for Denisova, presumably because mapping error has a larger impact when the true heterozygosity is lower.

[3] Measuring heterozygosity by counting reliable genotypes

In order to estimate heterozygosity from GATK genotype calls in Denisova and the 11 present-day humans as obtained in Note 6, we aimed to identify reliably genotyped positions by applying the following filters to the data, which remove:

[1] positions with extremely high or low coverage (upper and lower 2.5% of the coverage distribution for each sample; see Note 8, Table S8.2);

[2] positions surrounding insertions/deletions (+- 5 bp of the insertion/deletion);

[3] positions identified as prone to systematic error in Illumina sequencing (see Note 6);

[4] positions identified by RepeatMasking (see Note 6);

[5] positions with a 20-mer mapability score < 1 (see Note 6);

[6] positions with genotype quality (GQ as phred-score) < 40, which are genotypes with a probability of being miscalled lower than one in ten thousand;

[7] positions that did not pass filters in any other genome (intersection).

Filters [1] and [7] were used in all analyses, but we explored filters [2]-[6] to assess the effect of filtering on absolute and relative measures of heterozygosity (see Tables S39 and S40). Heterozygosity was estimated by dividing the number of heterozygous genotypes by the total number of identified genotypes per individual genome. The results shown in Table S40 indicate that Denisova has a reduced heterozygosity compared to any of the present-day humans analyzed here, irrespective of the filters applied. In fact, the relative ratios of heterozygosity remain nearly constant once a mapability filter is used. Additional filters reduce absolute heterozygosity, presumably by removing erroneous genotype calls, but do not substantially change the relative ratios of heterozygosity. The most severe drop in absolute heterozygosity is seen with a filter on genotype quality. This filter is in principle problematic, because genotype quality distributions can differ among samples and between homozygous and heterozygous genotype calls. We therefore report heterozygosity disregarding the genotype quality filter, even though it does not change relative estimates of heterozygosity.

The estimates of heterozygosity determined by genotype counting are very similar to the ones obtained by other measures (see Table S36). Thus, we do not detect biases that would affect downstream comparisons among samples using filtered genotype calls.

Note 16: The Denisovan genome lacks a signal of recent inbreeding

Flora Jay* and Montgomery Slatkin

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

To determine whether the low lower level of heterozygosity seen in the Denisovan genome (Note 15) indicates a smaller long-term effective population size or that the Denisovan individual's parents were closely related, we compared the distributions of lengths of runs of homozygosity (ROH) in the Denisovan genome with the lengths of ROH in the 11 present-day humans. Recent inbreeding creates unusually long ROH (*94, 95*).

We identified reliable genotypes by applying all filters described in Note 15 and used only positions for which the genotypes of all individuals are known. For each individual, we noted all positions of this kind that are heterozygous. We then down-sampled the heterozygous positions in the 11 present-day humans to the same density as in Denisova by randomly omitting sites. The down-sampling was done to prevent the higher levels of heterozygosity in the 11 present-day humans from affecting the results. For each individual, the distance between adjacent heterozygous sites was the length of the ROH. In doing this, we ignored tracks that overlap the centromeres. We focused on the distribution of ROH longer than 50 kb.

Figure S30 shows (a) the number of ROH found for several length categories and (b) the sum of these ROH lengths for each category. In each category, the first (red) bar corresponds to Denisova. There is no excess of long tracks in Denisova. Instead, the values are in the range of other individuals. Note that an excess of ROH longer than 400 kb is seen for the Karitiana individual but not for Denisova. These results are consistent with recent inbreeding in the family history of the Karitiana (*95*) but not the Denisovan individual. Figure S31 shows the empirical cumulative distribution of the ROH lengths greater than 50 kb in size. It confirms that the distribution of Denisovan ROH length (red line) is in the range of other individual distributions.

In addition, we analyzed the data without down-sampling to the same number of heterozygous sites. Figure S32 shows the number of ROH and the sum of ROH lengths for these data. Both the number and the sum of lengths of short ROH (<200kb) are greater for Denisova than for the 11 present-day humans. This higher frequency of shorter ROH in the Denisovan genome is consistent with a smaller effective population size (*95*).

We note that the lengths of ROH we found are much smaller than have been found in studies such as that of Kirin et al. (*95*) which used SNP surveys instead of sequence data. ROH will be shorter in sequence data because new mutations occur in genomic regions that are identical-by-descent (IBD). These new mutations will prevent IBD regions from also being identical-by-state (i.e. homozygous). New mutations are much less likely to have the same effect in SNP surveys because the mutations would have to create new alleles in the SNPs examined.

Note 17: Inferred population size changes in the history of Denisovans

Heng Li*, Nick Patterson and David Reich

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

Motivation

Li and Durbin recently reported the Pairwise Sequential Markovian Coalescent (PSMC), an algorithm that can use high coverage genome sequencing data from a single individual to infer the demographic history of the population from which the individual derives (*22*). The PSMC uses a Hidden Markov Model (HMM) to infer the distribution of coalescent times between the individual's two chromosomes across all loci. This can be interpreted in terms of population size change over time, since population size is inversely proportional to the rate of coalescence.

We applied the PSMC to 12 individuals who we sequenced and mapped to the human reference genome *hg19/GRCh37* (Note 4 and Note 5). To generate a diploid consensus genome for each individual, we used SAMtools (*45, 96*) and filtered out the following sites: [1] read depth >2-times or <1/3 of the average shown in Table S41; [2] sites where the root-mean-square mapping quality is below 10; [3] sites within 5bp of a predicted short insertion or deletion; [4] sites where the estimated consensus quality is <30; or [5] sites where at least 18 of 35 overlapping 35-mers from the human reference sequence can be mapped elsewhere with zero or one mismatch. The last filter makes all input data behave as if they are 35bp single-end reads so that all samples have the same mapping quality (we do this since the Denisova data is different from the present-day humans in that the Denisovan fragments are shorter and unpaired).

Results

We inferred population size changes over time for each sample by studying the distribution of the time since the most recent common ancestor using the PSMC (*22*). Since the PSMC splits continuous time into discrete intervals, it needs to set a maximum coalescent depth. In the original PSMC, this was set arbitrarily, and for analyses of present-day humans we found that this simplification of the algorithm did not affect results. However, we found empirically that the choice of the maximum time affected the inference about population size changes in the history of Denisovans, and thus we updated the PSMC algorithm to estimate the maximum time depth from the data. This improves the likelihood, and has a visible effect on the right-end tail of the curve for Denisova (corresponding to deep time depths), while having no important effect on the inferred history of the 11 present-day humans.

Figure 5B of the main paper shows the PSMC inference of how population size changed over time for each of the 12 individuals. The main *x*-axis scale does not give time in units of years, as this requires an assumption about mutation rate per year which is currently uncertain by more than a factor of two (see also Note 10). Instead, we present time in units of the pairwise sequence

divergence at segments of the genome that coalesce at that time. If we assume a mutation rate of 10^{-9} per year, a pairwise sequence divergence 10^{-4} translates to 50kya ($=10^{-4}/10^{-9}/2$). If we assume a mutation rate of half this value as recently suggested (*23, 60*), a divergence of 10^{-4} translates to 100kya. We do not show times of $<10^{-5}$ because the PSMC loses all power for more recent times (*22*). The Denisovan bone is assumed to date to a divergence of 10^{-4} (50-100kya).

We have two striking findings. The first is that modern humans have had a larger effective size than Denisovans since the two diverged. Consistent with the findings of the original PSMC paper (*22*), we infer that non-Africans share similar demographic histories prior to the Last Glacial Maximum 12.5-25kya, and that all modern humans share similar demographic histories prior to the first appearance of anatomically modern humans in the fossil record 100-200kya. By contrast, Denisovans are inferred to have had a consistently smaller effective population size since the split, suggesting that the much lower heterozygosity we detect in Denisovans (Note 15) cannot be due to a single severe bottleneck, and must instead reflect a persistently small population size, or multiple bottlenecks.

The second striking inference from the PSMC is that the estimated effective population sizes for modern humans and Denisovans appear to converge at around a divergence per base pair of 7.5×10^{-4} years in Figure 5B, corresponding to 375-750kya. A naïve interpretation is that this is an estimate of the divergence time of Denisovans from present-day humans. Encouragingly this is consistent with the divergence time estimated in Note 10 based on a different approach.

Caveat about the date of convergence of the modern human and Denisovan curves
A more detailed inspection of Figure 5B reveals reason for caution with regard to interpreting the convergence of the Denisovan and present-day human PSMC curves in terms of a date of divergence. In particular, the PSMC infers that there was a period of time corresponding to a divergence per base pair of 0.75×10^{-3} to 2×10^{-3} during which the Denisovan population was larger in size than modern human ancestors, before finally converging at $>2\times10^{-3}$ on the *x*-axis scale.

To explore whether the inference of a greater population size in Denisovans than in modern humans during this period could be an artifact, we performed a coalescent simulation in which the population was constant in size at N_e = 16,667 until 400kya, after which it crashed to Ne=1,667 and remained at that size until the present (the simulations are carried out for a mutation rate of $\mu=1.5\times10^{-8}$/bp/generation and 25 years/generation, and so we can directly report population sizes and times.) Figure S33 shows that when we run PSMC on data simulated under this scenario, we observe two phenomena: (i) the PSMC estimates the date of the population size crash to be too old (around 500kya), and (ii) the PSMC overestimates the population size in the period before the crash. Specifically, it artifactually smoothes out the sudden change and dates the convergence of curves to be older than the truth. This looks similar to the inference on real data in Figure 5B, and may explain the larger Denisovan population size compared to present-day humans inferred in Figure 5B around a divergence time of 10^{-3}.

In the future, it may be possible to obtain more robust inferences about population divergence times from PSMC-like analyses by analyzing data from more than one individual simultaneously. However, this will require new methodological development.

Note 18: Less effective selection in Denisovans than modern humans due to a smaller population size

Ron Do, Swapan Mallick, Joshua G. Schraiber, Cesare DeFilippo, Montgomery Slatkin, and David Reich*

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

Motivation

In genes, the ratio of non-synonymous-to-synonymous substitutions per site provides a measure of how effective natural selection has been at removing deleterious genetic variation from a population. In large populations, natural selection is expected to be more effective and non-synonymous-to-synonymous ratios are expected to be lower. Conversely, in small populations, selection is expected to be less effective and non-synonymous-to-synonymous ratios are expected to be higher. Thus, the ratio offers a way to learn about past population sizes that is complementary to inferences based on the site frequency spectrum or the distribution of the time since the most recent common ancestor.

We note that in contrast to negative selection, positive selection is expected to produce an increase in the ratio of non-synonymous to synonymous sites. However, several studies have suggested that purifying selection is likely to be more important than positive selection in affecting the ratio of non-synonymous-to-synonymous substitutions in humans (*97, 98*). This is our assumption in the analyses that follow. To further increase the probability that the non-synonymous we are analyzing are under negative selection, in the analyses reported in the main text we focus on the subset of non-synonymous sites that are predicted to be most deleterious (in the "probably damaging" or "possibly damaging" categories) by the PolyPhen 2 functional annotation software (*99*).

The ratio of accumulation of non-synonymous to synonymous substitutions is known to vary over evolution. The macaque genome paper (*100*) studied 3-way alignments of humans, chimpanzees, and macaques, and computed the non-synonymous-to-synonymous ratios in lineage-specific substitutions. It found the highest ratio in humans, the second-highest in chimpanzees, and the third-highest in macaques, which the authors hypothesized was consistent with human ancestral populations being smaller than chimpanzee ancestral populations since they separated (and with ape populations being smaller than old world monkey populations since they separated).

Here, we perform the same types of analyses in humans, assessing if the rate of accumulation of non-synonymous and synonymous sites has been different in modern humans than in Denisovans since they diverged.

Data processing

Our primary analyses were performed on the combined VCF files obtained based on mapping to the human reference sequence (Note 6) (we also repeated some analyses on the dataset mapped

to the chimpanzee genome to check robustness). We restricted to sites that pass eight filters: (1) they are on the autosomes; (2) a GATK genotype call was made in the two samples being compared; (3) the sites have a mappability score of Map20=1, (4) the read coverage in the samples being analyzed is within the central 95% of the genome-wide distribution (Note 8); (5) the site is either a synonymous or non-synonymous biallelic substitution; (6) the EPO alignment at that position includes chimpanzee; (7) one of the alleles agrees with chimpanzee; and (8) the genotype quality score is GQ≥30.

We used ANNOVAR (*101*) to classify the biallelic sites as in exons or not, and either synonymous or non-synonymous. This resulted in an average of 15,133 heterozygous sites per individual. We used PolyPhen 2 (*99*) to further annotate the sites based on their predicted phenotypic consequence as assessed from multi-species conservation as well as their predicted effect on protein function. PolyPhen classifies mutations as "benign", "possibly damaging" and "probably damaging". PolyPhen does not use human and chimpanzee to compute its multispecies conservation score, which is a valuable feature of this software as these species are used for choosing sites for our analyses.

Result 1: A higher proportion of non-synonymous sites on the Denisovan than the present-day human lineage

For any pair of samples *A* and *B*, we can compute the expected number of non-synonymous substitutions *qN* and the expected number of synonymous substitutions *qS* on each lineage since the two diverged (we infer the lineage on which the substitution probably occurred by comparing to chimpanzee). A complication is that the samples are diploid, and what we really want to know is the probability that a randomly sampled allele chosen from these two individuals is derived in one sample but not in the other (defined relative to chimpanzee).We defined f_A as the derived allele frequency in sample *A* (the allele not seen in chimpanzee), and f_B as the derived allele frequency in sample *B* (thus, f_A and f_B can be either 0, 0.5 or 1). The expected number of sites in sample *A* since divergence from the other is then $f_A(1-f_B)$, and similarly $f_B(1-f_A)$ for sample *B*.

To compare the *qN/qS* ratios for two samples, we define $Q_{A/B}$ as the ratio of non-synonymous to synonymous sites on the sample *A* lineage compared with the sample *B* lineage since they separated from a common genetic ancestor:

$$Q_{A/B} = \frac{qN_A}{qS_A} / \frac{qN_B}{qS_B} \quad (1)$$

$Q_{A/B} < 1$ suggests that selection has been more effective in population *A* that population *B* since they diverged, while $Q_{A/B}>1$ suggests the reverse.

Figure S34 plots $Q_{A/B}$ for all 66 = 12×11/2 possible pairs of individuals, using the convention that Denisova=A, and randomizing the order for within-modern human population pairs to help visualization. We see that selection has been less effective in Denisovans than in modern humans since divergence ($qN_{Denisova}/qS_{Denisova}$ is always higher than qN_{Modern}/qS_{Modern}), suggesting that population sizes have been historically smaller in Denisovans than in modern humans since the two diverged, an inference that is consistent with our analysis of population size changes over time from pairwise genetic divergence (Note 17). This inference is consistent for the alignments

to the human and to the chimpanzee genome, showing that these findings are not an artifact of modern humans mapping better to the human reference sequence than Denisovans.

We further stratified the analysis into "probably damaging", "possibly damaging", and "benign" sites based on the PolyPhen annotation (Table S42). The average $Q_{Denisova/Modern}$ for the comparison to 11 different modern humans is strongest for "Probably damaging" sites (1.86), weaker for "possibly damaging" sites (1.66), and slightly in the opposite direction for "benign" sites (0.93). In Figure 5 shown in the main text, we therefore focus our analysis on the pool of "probably damaging" and "possibly damaging" sites where the selective constraint is clearly stronger than at synonymous sites. The strengthening of $Q_{Deinisova/Modern}$ with increasing functional constraint is as expected if the qN/qS is reflecting the effectiveness of selection since the two populations diverged, as selection is expected to be more effective when population sizes are larger.

Result 2: A higher *pN/pS* ratio in Denisova than in modern humans

The analyses above draw a single random allele from each population. As a complementary study, we also examined heterozygous sites in the two populations, denoting the count of heterozygous non-synonymous in a population as *pN* and the count of heterozygous synonymous sites in the same population as *pS*. By analogy to $Q_{A/B}$, we define a ratio $P_{A/B}$ measuring the relative rates of heterozygous sites in any pair of two humans, setting *A*=Denisova by convention.

$$P_{A/B} = \frac{pN_A}{pS_A} \Big/ \frac{pN_B}{pS_B} \qquad (2)$$

Figure S34 shows that *pN/pS* ratios are much higher in Denisova than in modern humans, again implying that natural selection has been less effective in Denisova than in modern human history, and again suggesting that effective population sizes have been historically smaller. Table S42 stratifies $P_{Denisova/Modern}$ by Polyphen class, and again finds that the effects strengthen for more selectively constrained classes of sites.

We conclude from both the *dN/dS* and *pN/pS* analyses that population sizes have been historically larger in modern humans than in the history of Denisovans.

Note 19: A complete catalog of features unique to the human genome

Fernando Racimo*, Martin Kircher, Janet Kelso

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

Table of contents

Introduction

The Denisovan genome allows us to identify mutations that rose to fixation or near fixation on the modern human lineage since the split from the last common ancestor with Denisovans. From a comprehensive set of single-nucleotide changes and insertion-deletion events that occurred on the human lineage since the DNA sequence divergence from the common ancestor shared with chimpanzee, we were previously able to determine the Denisovan genotype for 41% of the single nucleotide changes and 22% of the insertions/deletions using the draft Denisova genome (*2*). The high-coverage Denisova genome now allows us to confidently identify Denisovan genotypes for nearly all regions of the human genome accessible by current short-read sequencing technologies.

Using updated whole genome alignments of human (GRCh37), chimpanzee (CHIMP2.1), gorilla (gorGor3.1) and orangutan (PPYG2) as well as human variation data from the 1000 Genomes Project 20110521 release (1000G) (*26*), we first identified sites that are fixed derived in modern humans and ancestral in chimpanzee and at least one other great ape: gorilla or orangutan. This amounts to 13,783,828 single-nucleotide changes (SNCs) and 1,032,499 insertions/deletions (InDels). We also looked at sites with derived alleles that are above 90% global frequency but

are not fixed in humans (693,111 SNCs and 45,857 InDels). We excluded the mitochondrial and Y chromosomes from our analysis, because the 1000 Genomes Project does not include human population data for these chromosomes.

We then selected sites that pass quality filters in Denisova, and where Denisova has at least one ancestral allele. We included sites where Denisova has only one ancestral allele because we are also interested in mutations that arose before the human-Denisova split but only rose to fixation or high frequency after the split. When choosing quality filters for Denisovan genotypes, we aimed to balance between confidence in genotype calls and exhaustive coverage of the genome (see Data and Methods below). For this reason we also include sites that would fail one of the very stringent filters (e.g. positions in a CpG context or close to InDels) but provide flags for recognition. We consider the catalog presented here to provide a complete, high-confidence set of positions that have changed recently on the human lineage.

Electronic access to the catalog

The full catalog of the sites where all modern humans sequenced by 1000G carry a fixed or high-frequency derived allele relative to apes, annotated by the allelic state of the high-coverage Denisovan genome, as well as all tables presented here are available for download at:
http://bioinf.eva.mpg.de/download/HighCoverageDenisovaGenome/

Data and methods

The high-coverage, high-quality Denisovan genome sequence allows us to use standard approaches to variant calling such as the Genome Analysis Toolkit (GATK) (*13, 102*) (see Note 6). We filtered for sites with reliable genotype calls by selecting sites with a PHRED genotype likelihood score ('PL' field) > 40, and a RMS map quality score (the quadratic mean of the map quality scores of individual sequences covering a site) > 30. In addition, to determine coverage cutoffs we calculated the proportion of pair-wise differences between Denisovan genotype calls and the human reference genome sequence as a function of coverage (Figure S35), which provides a composite measure of divergence from the human reference as well as genotype error (including sequencing error and mapping error). Excluding the potential contribution from differences in mutation rates along the genome, divergence should be constant over coverage bins, so elevated divergences at both ends of the coverage distribution indicate excess of genotype errors. To keep genotype error below 0.5%, we excluded sites that are in coverage bins below 14X and above 66X.

Last, we excluded sites deemed to be of low quality due to conflicting genotype calls in a second iteration of GATK or identified as systematic errors (see Note 6). In order to avoid paralogous regions we also required that human and chimpanzee sequences appeared no more than once in the EPO primate alignment blocks ('TS' field, see Note 6). Sites located in CpG regions, sites in repeat masked regions identified by RepeatMasker (http://www.repeatmasker.org) and sites nearby InDels (+/- 5bp) were flagged but not excluded from our analysis.

In order to identify SNCs and InDels of interest, we first searched for sites where the majority of the individuals from the 1000G global sampling of population variation carry the derived allele.

We defined "fixed" changes to be those that either have no recorded alternative allele entries in 1000G or that have an alternative allele with a recorded global frequency equal to 0 (within the resolution of one percent present in the 1000G data). We note that some of these fixed changes have dbSNP entries even though they are recorded to be at 100% frequency in 1000G, so we mark them with an asterisk ("fixed*") in all tables below. "High-frequency" changes are defined as changes with a global derived allele frequency > 90% and < 100%. After applying all filters specified above, we obtained 11,638,419 fixed SNCs, 578,029 high-frequency SNCs, 884,687 fixed InDels and 34,612 high-frequency InDels.

We then identified sites where the Denisovan genome shows either one or two ancestral alleles. In total there were 111,812 fixed and 190,965 high-frequency modern human-derived SNCs as well as 9,499 fixed and 10,966 high-frequency modern human-derived InDels. We used Ensembl's Variant Effect Predictor (VEP) version 2.2 (*103*) and the Ensembl 65 annotation (December 2011) to annotate these changes and predict their effects on protein structure and transcriptional regulation.

SNCs in protein-coding sequences

To identify changes in protein-coding sequences, we excluded sites in transcripts that were non-coding or predicted to undergo nonsense-mediated decay. We refined this set further to use the longest annotated coding sequence for 18,454 CCDS genes (Consensus Coding Sequence project of EBI, NCBI, WTSI, and UCSC - Sep. 7th 2011 release). The CCDS database contains only human genes that are well annotated and have passed several quality control tests. This allows us to identify coding genes with high-confidence, but at the cost of missing genes that may be poorly annotated. Table S43 shows the number of SNCs in each predicted functional category. SNCs that had more than 1 predicted effect (e.g. non-synonymous and 3' UTR) were classified by their most severe predicted effect in the following order: non-synonymous, splice sites, synonymous, 5' UTR, 3' UTR (Figure S36).

Non-synonymous SNCs

We found 260 fixed derived and 393 high-frequency derived non-synonymous SNCs among the CCDS-verified genes. We provide a complete list of the fixed SNCs in Table S44, while the high-frequency SNCs can be found online at:
http://bioinf.eva.mpg.de/download/HighCoverageDenisovaGenome/.

Table S44 also includes the following information:

(i) mammalian conservation scores obtained from a multi-species alignment excluding the human reference (see Note 6).

(ii) primate conservation scores obtained from 6-primate alignments excluding humans (see Note 6)

(iii) Grantham scores (GS), which provide a measure of chemical dissimilarity between amino acids and can hence be used as a proxy for the damaging potential of a SNC.

(iv) predicted effects on protein structure from SIFT (*104*) and PolyPhen (*99*).

We note that the majority of SNCs are predicted to result in benign or tolerated alterations to protein structure.

To prioritize the list of non-synonymous fixed SNCs, we first focused on the 23 positions that had a high primate conservation score (higher than or equal to 0.95) and are therefore strongly conserved for the ancestral state in primates but derived in modern humans. We list these SNCs ranked by the primate conservation score in Table S45, along with the genes they affect and the function or description of their encoded proteins. Eight of these genes (NOVA1, SLITRK1, KATNA1, LUZP1, ARHGAP32, ADSL, HTR2B and CNTNAP2) are associated with brain function or nervous system development.

CCDS genes with multiple non-synonymous SNCs

We identified 29 coding CCDS genes with more than one fixed non-synonymous SNC where Denisova carries the ancestral allele. However, in eight of these (OR2H1, MUC17, TNFRSF10D, MUC6, MUC5B, OR4A16, OR9G1, ERCC5), the Denisovan individual appears to be heterozygous for all SNCs present in the gene, which may indicate that they are the result of duplications or repetitive regions. All but two of these eight genes are either olfactory receptor genes or mucin genes, both of which are gene families known to evolve rapidly and often undergo duplication events (*105, 106*).

We therefore only focus on the remaining 21 genes with more than one SNC where Denisova is homozygous ancestral (Table S46), none of which has more than 3 SNCs. Thirteen of these were previously identified as containing more than one fixed substitution (ANKRD30A, HPS5, ITGB4, RP1L1, SPTA1, SSH2, TTF1, ZNF333) or only one substitution (CASC5, HERC5, OR5K4, SETD2, SPAG5) using the draft Denisovan genome at 1.9X coverage(*2*), while five of them were previously identified as having two substitutions (SPAG17, TTF1) or only one substitution (OR5K4, SPTA1, SSH2) using the draft Neandertal genome (*1*).

Several of the genes with more than one non-synonymous substitution have been highly cited in medical literature. The two substitutions in HPS5 are in sites that are highly conserved in mammals (conservation score = 1). HPS5 is involved in the synthesis and trafficking of intra-cellular vesicles including lysosomes, melanosomes and platelet-dense granules. Mutations in this gene are associated with Hermansky-Pudlack syndrome, a rare recessive disorder that causes albinism (*107*). Another gene with two non-synonymous substitutions, SETD2, is a histidine methyltransferase that is known to be involved in the regulation of gene expression (as a transcription activator). It interacts with Huntingtin (*108*), which when mutated causes Huntington's disease, a neurodegenerative disorder characterized by declines in muscle co-

ordination and cognitive abilities (*109*). HERC5, which has two fixed non-synonymous substitutions, is an E3 ligase involved in regulation of the innate antiviral response to papillomavirus (*110*) and influenza A (*111*).

Gains or losses of STOP codons

We looked for STOP losses and gains that occurred exclusively in present-day humans by checking for VEP-predicted disruptions in STOP codons in the Denisovan genome relative to the human reference genome, and then checking whether the disruption was a derived STOP loss or gain by looking at the ancestral (ape-like) state and the modern human derived state inferred from 1000G data. Among the CCDS-verified genes with fixed and high-frequency changes, two SNCs lead to losses of STOP codons, while two other SNCs lead to gains of STOP codons. The STOP codon losses occur in genes OLFML2B (fixed) and OPRM1 (global human frequency = 97%). OLFML2B is involved in extracellular matrix organization (Gene Ontology Consortium 01 Aug 2011). OPRM1 encodes for the mu opioid receptor, which is the main target of morphine, heroin, methadone and several other opioids (*112*). Variation in OPRM1 contributes to skin pigmentation differences between Europeans and Indigenous Americans (*113*) and has been associated with attachment behavior in infant primates (*114*).

The two STOP codon gains occur in genes FMO2 (global human frequency = 96%) and CASP12 (global human frequency = 96%). FMO2 is a gene involved in NADP metabolism (*115*). It is predominantly expressed in the lungs, but is catalytically inactive in present-day humans (*116*). CASP12 is a cystein protease. Carriers of the read-through variant of CASP12 are found primarily in African populations and are known to be endotoxin hypo-responsive and at risk of sepsis (*117*). There is also evidence that the STOP-gain variant has been subject to recent positive selection (*118*).

We also looked at STOP-gain and STOP-loss SNCs in non-CCDS-verified genes. This category may for example contain genes that were pseudogenized after the split between present-day humans and Denisova, and are therefore not included in the CCDS database. Table S47 shows a combined list of all the Ensembl genes with STOP-gain and STOP-loss changes.

GO analysis of non-synonymous SNCs

We performed two types of Gene Ontology (GO) tests using FUNC (*119*) to find overrepresented biological, functional or structural categories among the present-day human-derived non-synonymous SNCs (fixed and high-frequency) in CCDS-verified genes. We used the UniProt-GOA annotation (retrieved on 06/05/2012) and restricted our analysis to GO terms that contain at least 20 genes.

First, we performed a hypergeometric test, taking all the CCDS-verified genes that have any SNCs on the human lineage since the human-chimpanzee split as background. We then tested for overrepresented categories of genes that acquired non-synonymous SNCs in the modern human lineage after the split from Denisova. The most overrepresented categories for genes with fixed SNCs are listed in Table S48A (raw p-value < 0.01), while the most overrepresented categories for genes with fixed and high-frequency SNCs are listed in Table S48B (raw p-value < 0.01). After accounting for the family-wise error rate (FWER < 0.05) and the false discovery rate (FDR

< 0.05), several categories related to membrane transport and receptor signaling remain significant.

Second, we tested for categories that show an excess of non-synonymous changes on the modern human branch after the split from Denisova given the rate of non-synonymous changes between the human-chimpanzee split and the human-Denisova split. We compared the genes that acquired non-synonymous SNCs in the modern human lineage since the split with Denisova to the genes that acquired non-synonymous SNCs after the split with chimpanzees but before the split with Denisova. We excluded sites where Denisova is heterozygous from both lists. We scored each gene by the number of non-synonymous changes they had in each of the two lists and performed a binomial test on the scored lists. By comparing score distributions, this test controls for the length and nucleotide content of the genes tested. The most overrepresented categories for genes with fixed SNCs are listed in Table S49A (raw p-value < 0.01), while the most overrepresented categories for genes with fixed and high-frequency SNCs are listed in Table S49B (raw p-value < 0.01). We find protein folding, neurogenesis and pigmentation to be among the most overrepresented categories. However, no categories remain significant after accounting for the false discovery rate and the family-wise error rate.

Non-synonymous SNCs in genes with disease associations

To explore whether any of the genes with non-synonymous SNCs are associated with human diseases, we overlapped our list of genes with those present in the database of Mendelian diseases (OMIM, www.omim.org, retrieved through SIFT (*120*) v.4.0.5 on 03/08/2012). Table S50 presents 38 fixed and 77 high-frequency non-synonymous SNCs found in genes with OMIM disease entries.

Among these genes is EVC2 (Ser488Gly), which is known to play a role in skeletal development. Mutations in this gene cause Ellis-Van Creveld syndrome (*121-123*) which, among other symptoms like polydactyly and heart defects, is known to produce taurodontism, a developmental defect of the teeth that involves elongation of the pulp chambers and root reduction or fusion (*124-126*). This trait is also characteristic of Neandertal teeth (*126, 127*) but is not present in the Denisova molar described in Reich et al. 2010 (*2*). One possible hypothesis is that one or more mutations in EVC2, coupled with mutations in other genes, may have caused distinct dental morphologies in different groups of hominins.

We found 4 fixed SNCs in genes involved in diseases associated with skin (HPS5, ERCC5, GGCX, ZMPSTE24) and 6 with eye development (RP1L1, GGCX, FRMD7, ABCA4, VCAN, CRYBB3). Among these, RP1L1 is notable for having three fixed SNCs. There is evidence for selection in the region surrounding RP1L1 in modern humans (*114*) and a duplication of a region containing this gene (8p23) may result in developmental defects including delayed speech, difficulties in learning and autistic behavior (*128*). However, the same study pinpointed MCPH1 (located 4Mb away from RP1L1) as the most likely candidate for these defects.

In addition, we found one fixed non-synonymous SNC in the ATRX gene (Asp475His), associated with forebrain development and facial morphogenesis. Mutations in ATRX are associated with the alpha-thalassemia / mental retardation syndrome, which can lead to microcephaly and facial anomalies (*129*).

We also found two genes linked to autistic disorders that have non-synonymous fixed SNCs. One SNC (Ile345Val) is located in a position with a strong primate conservation score (0.95) inside the laminin G-like domain of the CNTNAP2 gene, associated with susceptibility to autism (*130, 131*) and language disorders (*27*). CNTNAP2 is particularly noteworthy for being one of the few known interactors of FOXP2 (*27*), a gene involved in language and speech development (*132*). CNTNAP2 codes for a neurexin that is specifically expressed in the human cortex and is involved in cortical development (*133*). The other SNC (Ala429Val) is located in a C-terminal helical domain of ADSL, in a position with a high primate conservation score (0.953) Mutations in the gene lead to adenylosuccinase deficiency, which can cause psychomotor retardation and autism (*134, 135*).

For each non-synonymous SNC in genes with OMIM entries, we also checked whether the resulting amino acid substitution affected a position in the protein where other amino acids are recorded as disease-associated in The Human Gene Mutation Database (*136*) (HGMD, retrieved on 03/10/2012). We found two fixed and ten high-frequency non-synonymous modern human derived SNCs that affect amino acid positions with HGMD entries (Table S50, right-most column). In all but one (SGCA) of the genes affected by these SNCs, the amino acid associated with the disease is also the ancestral amino acid that is present in Denisova and low-frequency or absent in modern humans.

One of the fixed substitutions is located in SGCA (Ile175Val), a gene associated with muscular dystrophy (OMIM). The ancestral allele is highly conserved among mammals (mammal conservation score = 0.933) as well as primates (primate conservation score = 0.974) and the derived allele has no human dbSNP entries. This site is heterozygous in Denisova: both the ancestral and the derived alleles are present. The mutation causes an amino acid substitution in position 175 of the adhalin protein, a component of the dystrophin-glycoprotein complex that is key to muscle fiber stability (*137*). The ancestral amino acid is an isoleucin, while the modern human amino acid is a valine. A non-synonymous substitution in a different nucleotide in the same codon (Val175Ala) has been linked to a recessive form of muscular dystrophy in modern humans (*138*). The amino acid is located in between Asn-174 and Thr-176, which together constitute one of two consensus sites for glycosylation of the protein (*138, 139*).

The other fixed SNC with a HGMD entry is located in ABCA4, a gene associated with cone-rod dystrophy (OMIM). The mutation causes an amino acid substitution (Gln223Lys) that was found to be a possible disease variant in a screen for Stargardt disease, a form of macular degeneration that leads to vision loss (*140*).

SNCs in splice sites

We found a total of 72 fixed SNCs and 116 high-frequency SNCs in splice sites within CCDS-verified genes. Four SNCs were located in "essential" splice sites (Table S51), defined by the VEP to be in the first two or last two base pairs of an intron. One of them is located in KCNJ16, which codes for Kir5.1, a potassium ion channel that is expressed in the human kidney and thyroid gland (*141*). We also found a fixed mutation in an essential splice site of IZUMO4, a sperm-egg fusion protein (*142*).

Insertions / deletions in genes

We found 4,169 fixed and 4,447 high-frequency derived InDels inside, upstream or downstream of CCDS genes. We classified these InDels by functional effect. Table S52 lists the number of InDels that fall in exons, splice sites and UTR regions. In Table S53, we list all fixed and high-frequency InDels that either cause a frameshift in a coding sequence, create an in-frame non-synonymous event (involving the substitution of more than 1 amino acid) or disrupt a splice site.

One of the genes in Table S53 is CLTCL1, which codes for a member of the clathrin family, known to coat cellular vesicles (*143*). A fixed modern human-specific InDel in this site is predicted to cause both a frame-shift and disruption in a splice site of this gene. DiGeorge syndrome - a disease with that may cause palate defects, schizophrenia and learning disabilities - has been associated with a chromosomal deletion that includes this gene (*144*).

Changes in regulatory sequences

Regulatory elements are modulators of gene expression and may influence phenotype in important ways (*145, 146*). Though our understanding of regulatory elements is still limited, we explored changes in well-characterized regulatory sequences with a view to determining those which have become fixed or have risen to high frequency recently during modern human evolution, thus suggesting that they may have been subject to selection in recent human history.

microRNAs

MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression post-transcriptionally through mRNA cleavage or repression of translation (*147, 148*). Single miRNAs may regulate a large number of target genes, and as such have the potential to broadly influence gene expression. miRNA sequences are generally highly conserved among species, reflecting their importance as regulators of gene expression (*149*). A number of studies have demonstrated the importance of miRNAs in development, particularly in the primate brain (*150*).

We identified one fixed derived SNC (chr12:79813049 A/G) in a miRNA gene (MIR1252). This SNC is in the mature sequence of the miRNA (Figure S37) and the change is therefore not expected to alter target specificity or change the folding of this miRNA. mIR1252 was identified in embryonic tissue (*151*) and expression has not been reported in adult tissues. TargetScan (http://www.targetscan.org Release 5.2) predicts 257 genes to be potential targets for MIR1252 in modern humans.

Based on the low coverage data of the Denisovan genome, two human-specific changes in miRNA genes were reported previously (*2*). One of them (chr14:77732622) was reported as a human-derived insertion in MIR1260, which is missing in Denisova and chimpanzees. The new human-chimpanzee alignments used in this study show that this is actually a Denisova-specific deletion, as both humans and chimpanzees appear to have the missing base. The other change (chr3:44903385) was reported as a human-specific G-to-A mutation in MIR564. The ancestral state is also observed in the 30X Denisova genome data, but with the addition of 1000G SNP

frequencies the derived state is now found not to be fixed in modern humans, but at 98% global frequency.

High-information sites in regulatory regions

We also looked for sites predicted by the VEP to be in regulatory regions identified via Chip-Seq and Dnase-Seq experiments and recorded in Ensembl's Regulatory Build. These are regions that contain experimentally identified transcription factor binding sites. Within these regions, we selected mutations found in high-information positions of predicted motif patterns from the JASPAR database (*152*). Motif patterns are short sequences where regulatory factors are thought to interact directly with DNA. High-information sites within these patterns are conserved across different motifs of the same factor and are considered important for recognition of target sequences. We therefore expect mutations in these sites to be strong candidates for differences in regulatory function.

We found 35 fixed SNCs, 3 fixed InDels, 52 high-frequency SNCs and 2 high-frequency InDels in high-information positions within motif features in regulatory regions. Table S54 contains the position of these sites, along with the transcription factor that is predicted to bind to the motif. We also report the genes lying within the two nearest CTCF insulator binding sites (obtained from CTCFBSDB (*153*)) on either side of the mutation. As insulators are known to restrict the effect of regulatory elements (*154*), these genes are the putative targets of factors binding to the motifs containing the mutations.

SNCs identified in GWAS studies

Genome-wide association studies (GWAS) are studies where two groups of individuals - one containing a phenotype of interest and a control - are genotyped to find candidates for associations between a SNP and the phenotype. These SNPs are thus located in regions of the genome that are likely to contribute to phenotype differences between individuals (*155, 156*). We used the Catalog of Published Genome-Wide Association Studies (www.genome.gov/gwastudies; accessed on 02/27/2012) to find phenotype associations for SNPs that have risen to high frequency in humans (not yet fixed) and where Denisova has an ancestral allele.

We found 28 high-frequency derived SNCs with GWAS entries (Table S55). In 11 of these sites, the ancestral (Denisova) allele is the risk allele for the disease or phenotype. The derived allele is the risk allele in 8 sites, while in 9 sites the risk allele is not specified or known. The associated phenotypes for these SNPs include height, coronary heart disease, bone mineral density and conduct disorder, among others.

Note 20: Catalog of features unique to the Denisovan genome

Fernando Racimo, Martin Kircher, Janet Kelso

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

Table of contents

Here we present the Denisovan state for sites that may have some phenotypic or disease-associated relevance.

Denisova-specific SNCs

Conversely to modern human-specific changes, we also looked for sites where the Denisovan genome is homozygous for a derived allele and modern humans possess the ancestral state (at high-frequency or fixed). These are candidates for changes having risen to high frequency or fixation on the Denisovan lineage. However, since only a single Denisovan genome is presently available, the majority of these changes are likely to have been polymorphic among Denisovans. We used the same genotype quality and coverage cutoffs as in Note 19.

We found 692,818 Denisova-specific SNCs where the ancestral state is fixed in modern humans and 280,553 Denisova-specific SNCs where modern humans have the ancestral state at high-frequency. These numbers are 6.2-fold and 1.47-fold larger, respectively, than the number of fixed and high-frequency derived SNCs in modern humans where Denisova is ancestral.

A likely explanation for these discrepancies is that private derived SNPs of the sequenced Denisovan individual dominate the observations. This is supported by Figure S38 which shows that for decreasing global frequency cutoffs for the modern human allele (used as the ancestral allele in Denisova-specific SNCs and as the derived allele in modern human-specific SNCs), the difference in number between modern human-specific and Denisova-specific SNCs in chromosome 21 becomes less pronounced. In other words, when modern human alleles segregating at intermediate frequencies are included, the number of Denisova-specific SNCs approaches the number of modern human-specific SNCs.

Non-synonymous sites found to be homozygous derived in the Denisovan individual and fixed in present-day global human populations can be found at:
http://bioinf.eva.mpg.de/download/HighCoverageDenisovaGenome/.

Denisova state of SNPs in GWAS and GAD databases

To make information available about the Denisovan state for all SNPs found in large phenotype association studies, we present here a catalog of Denisovan high-quality genotypes (derived, ancestral or polymorphic) for all positions where human SNPs are found in the NHGRI GWAS catalog (www.genome.gov/gwastudies; accessed on 02/27/2012) and the NIH Genetic Association Database (*157*) (accessed 03/17/2012). We used the same genotype quality, mapping quality and coverage filters as in Note 19. The list of SNPs with the Denisovan alleles and their corresponding phenotype-association entries can be found at:
http://bioinf.eva.mpg.de/download/HighCoverageDenisovaGenome/

Phenotypically interesting SNPs

We also manually looked at the "popular" human single-nucleotide polymorphisms (SNPs) in SNPedia (*158*) (retrieved on 03/18/2012) and checked their allelic state in Denisova. In Table 20.1, we list some of the most interesting sites.

SNPs associated with modern human pigmentation

Cerqueira et al. (*24*) analyzed 124 SNPs associated with skin, hair and eye pigmentation differences and checked their state in the genomes of both modern and archaic humans, including the Denisovan genome at 1.9X coverage (*2*), to predict pigmentation phenotypes. Here we re-check the same SNPs in the high-coverage Denisovan genome (Table S57). We are able to determine the alleles for 11 sites that were previously undetermined, and identify 3 sites where the predicted genotype is now known to be different from the state observed in the low-coverage genome. The predicted phenotype remains the same as reported before: the Denisovan individual is predicted to have had dark brown hair, dark skin, brown eyes and no freckles (Table S58). However, we note that predicted phenotypes in Cerqueira et al. did not show a strong concordance with observed phenotypes in modern humans (total percentage of agreement = 59%) (*24*) and the reliability of the inference for the Denisovan individual (belonging to an archaic group that might have had other pigmentation-associated SNPs not present in modern human variation) may be even lower.

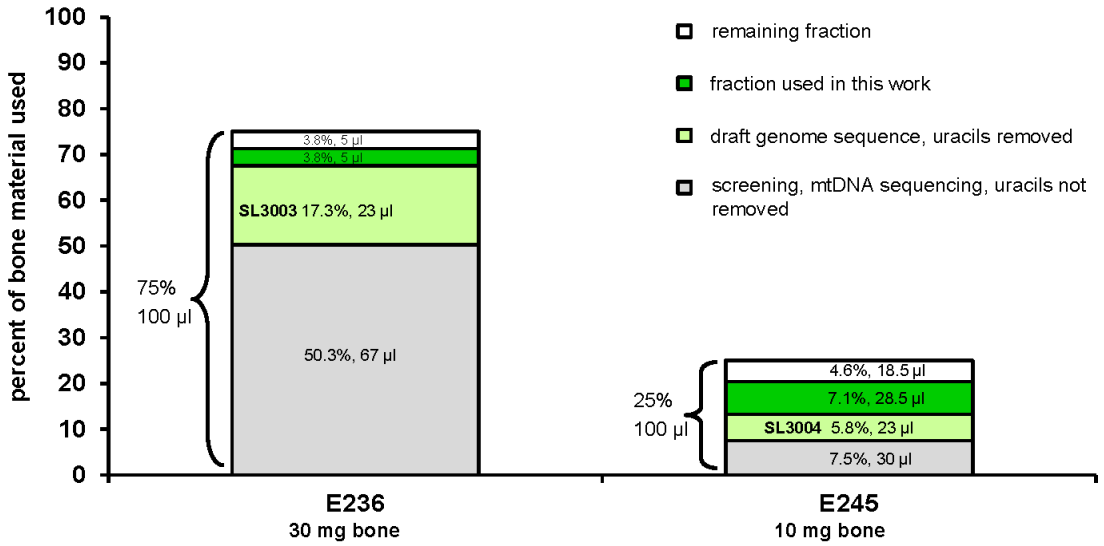


Figure S1: Usage of bone material from the Denisovan phalanx for previous and present work.

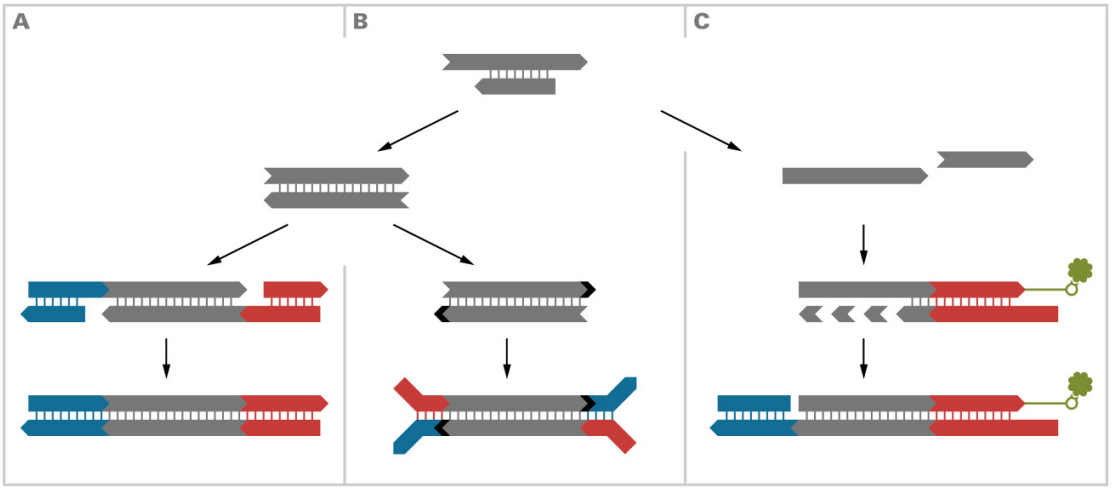


Figure S2: Library preparation methods used for ancient DNA **(A)** Following the method developed by 454 Life Sciences, template DNA is blunt-end repaired and two different adaptors (blue and red) are attached in a non-directional ligation reaction. Since 5'-phosphates are absent from the adaptors, only one strand is joined and the resulting nick is removed with a strand-displacing polymerase. Molecules with two identical adaptor sequences are lost, because they amplify poorly and cannot be sequenced. **(B)** With the method developed by Illumina, blunt end repair is followed by the addition of A-overhangs to the 3'-ends of the template molecules. Y-shaped adaptors with 3'-T-overhangs are added to both ends of the molecules in a sticky-end ligation reaction. **(C)** In the library preparation method described here, template DNA is first dephosphorylated and heat-denatured so it becomes single-stranded. A 5'-phosphorylated adaptor oligonucleotide with a biotinylated 3'-linker is attached to the template strands by single-stranded ligation. The ligated strands are immobilized on streptavidin-coated beads and copied by extending a primer, which is hybridized to the adaptor. One strand of a double-stranded adaptor is attached to the synthesized strand by blunt end ligation. Finally, the beads are destroyed by heat to release the library molecules. All methods are compatible with damage removal by deoxyuracil excision.

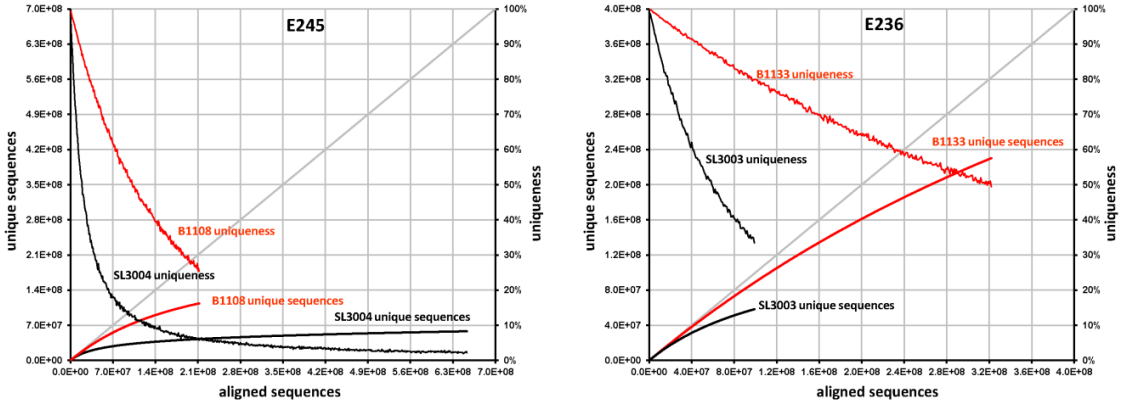


Figure S3: Sub-sampling plots showing the levels of saturation reached by sequencing libraries prepared from two extracts (E245 left, E236 right) with the double-stranded (black) and the single-stranded method (red). Only sequences with map quality >= 30 and length >= 35 bp were considered. In a library of infinite complexity, the number of unique sequences (Y-axis, left side of panels) would be identical to the total number of sequences obtained (X-axis) as indicated by the grey diagonal. Complete saturation is achieved if there is no more increase in the number of unique sequences. Uniqueness (Y-axis, right side of panels) is calculated as the percentage of new sequences found in the last 10,000 sequences sampled.

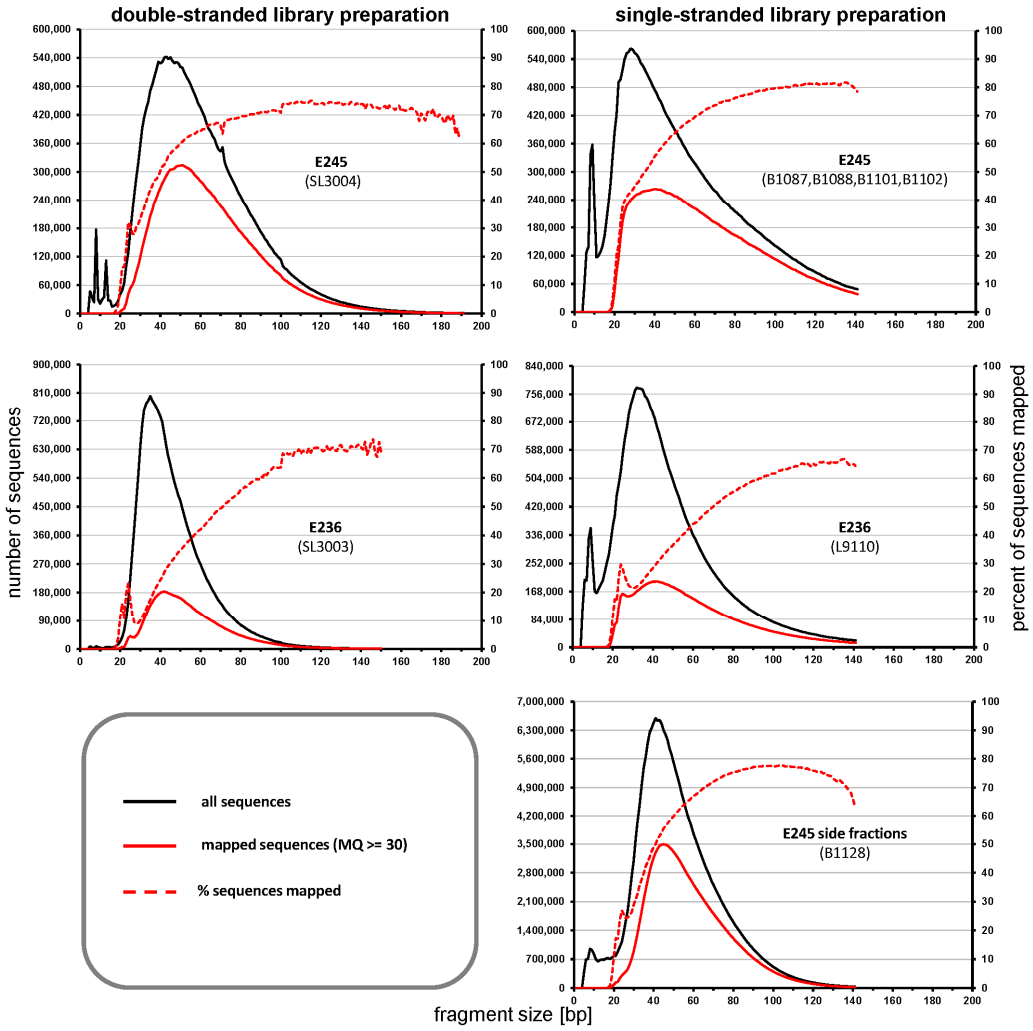


Figure S4: Fragment size distribution determined by sequencing in libraries prepared with the double-stranded and the single-stranded method. Only sequences merged from overlapping paired-end reads were considered. Thus, the upper fragment size is limited to double read length minus 11 bp (the minimum overlap required). The map quality filter (MQ >= 30) removes sequences that cannot be placed with certainty to a unique position in the human genome. The fraction of mapped sequences is therefore only roughly equivalent to the fraction of endogenous DNA.

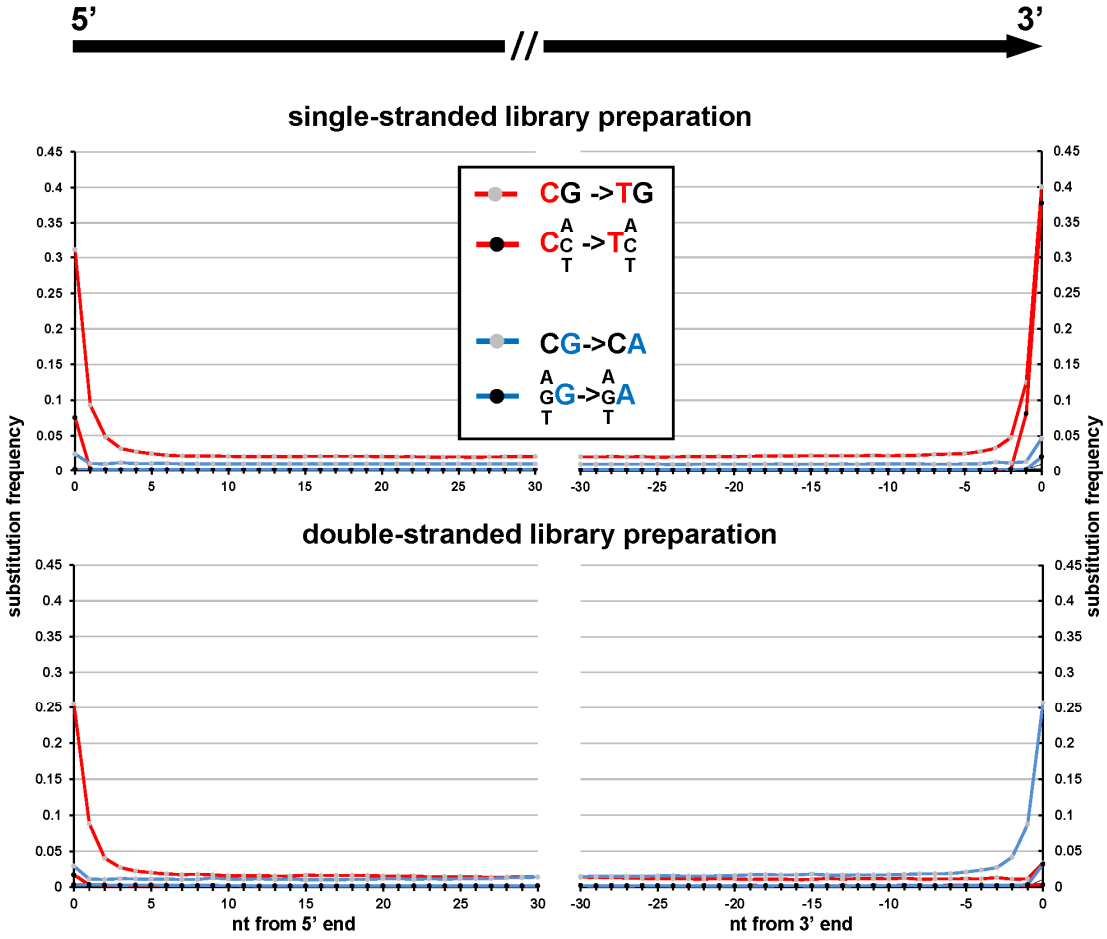


Figure S5: Position dependence of C->T and G->A substitution frequencies at 5'- and 3'-ends of Denisovan sequences. Only sequences of at least 35 bp that aligned to the human genome with a map quality of at least 30 were considered for this analysis. Substitution frequencies are shown both for CpG and non-CpG context. All other types of substitutions (shown in black) are indistinguishable from the base line.

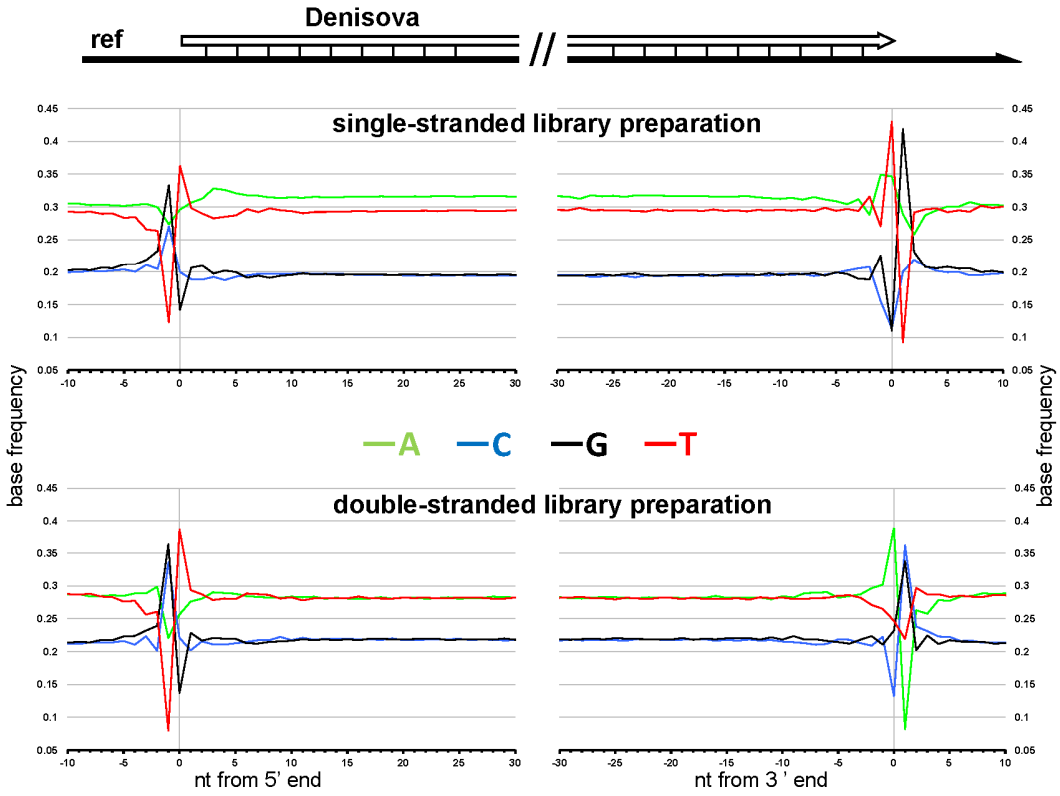


Figure S6: Base composition of the human reference genome around the 5'- and 3'-ends of Denisovan sequences.

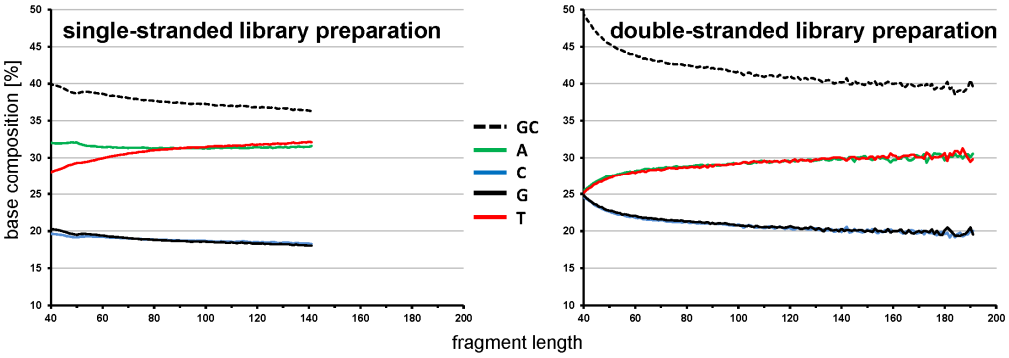


Figure S7: Base composition of Denisovan sequences of different length. For the double-stranded method, A and T (as well as C and G) are expected to occur at the same frequencies, because strand-information is not preserved.

```
loglik <- function(pad,pdd,con,err)
{
  E7.3 <- function(t,pad,pdd){ if (pad + pdd > 1) { pad <- 0; pdd <- 0; };
ifelse(t==0,1-pad-pdd,ifelse(t==1,pad,pdd)) };
  E7.4 <- function(t,total,derived,con,err) { choose(total,derived)*(1-
E7.5(t,con,err))^(derived)*E7.5(t,con,err)^(total-derived) };
  E7.5 <- function(t,con,err){ ifelse(t==0,1-con-
err+2*con*err,ifelse(t==1,(1-con)/2 + con*err,err)) };
  -sum(obs*log(rowSums(sapply(0:2,FUN=function(t){ E7.3(t,pad,pdd)*
E7.4(t,total,derived,con,err) })))));
}

m0 <- mle2(loglik,start=list(pad=0.3,pdd=0.3,con=0.5,err=0.01),method="L-
BFGS-
B",lower=list(pad=0.000001,pdd=0.000001,con=0,err=0),upper=list(pad=1,pdd=
1,con=0.5,err=0.5),control=list(parscale=rep(1e-6,4),maxit=10^6))
summary(m0)
p0 <- profile(m0)
confint(p0)
```

Figure S8: Code used for the MLE optimization procedure in the statistical software package R.

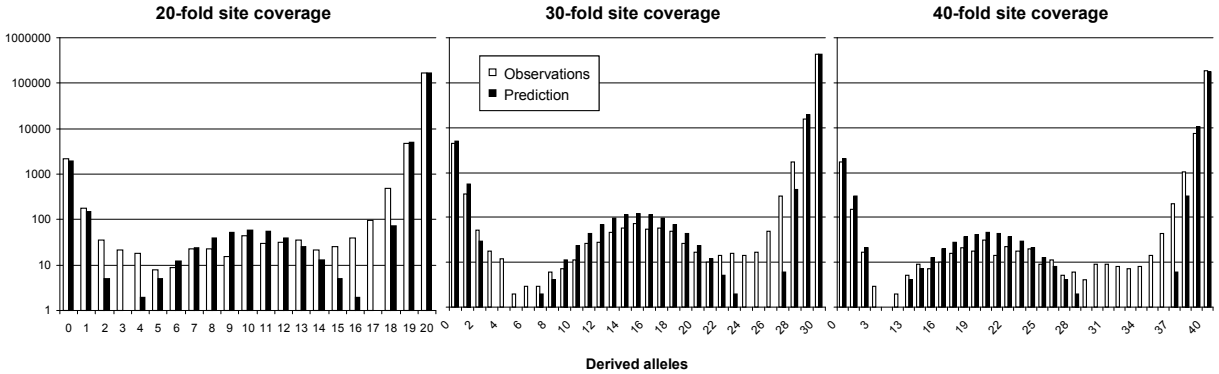


Figure S9: Observed number of sites in bins of derived allele counts for positions of 20-fold, 30-fold and 40-fold coverage compared to the number predicted sites when using the model instantiated with the values in Table S10.

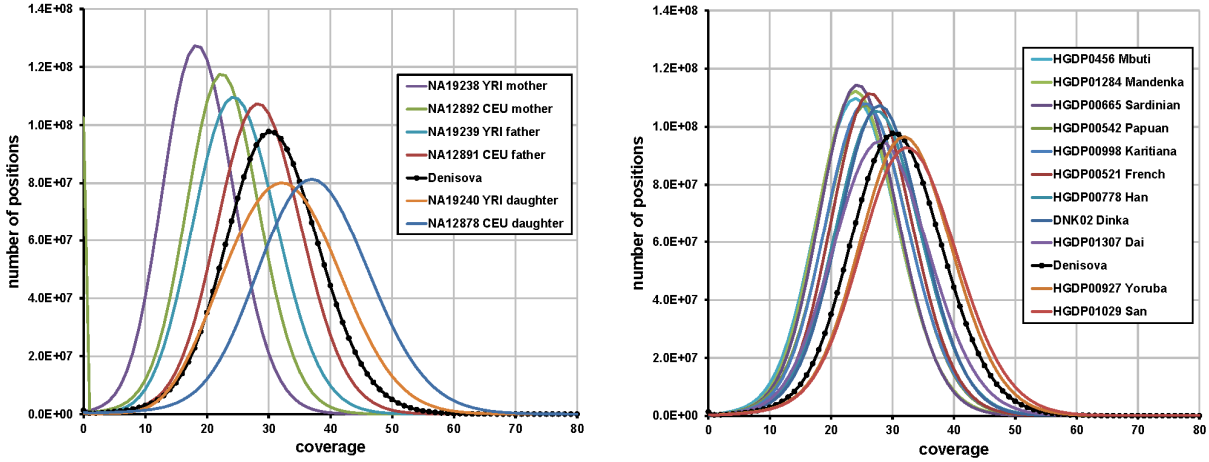


Figure S10: Comparison of coverage distributions in uniquely mappable parts of the autosomal genome for Denisova and the 1000 genomes trios (left) and the eleven present-day humans (right).

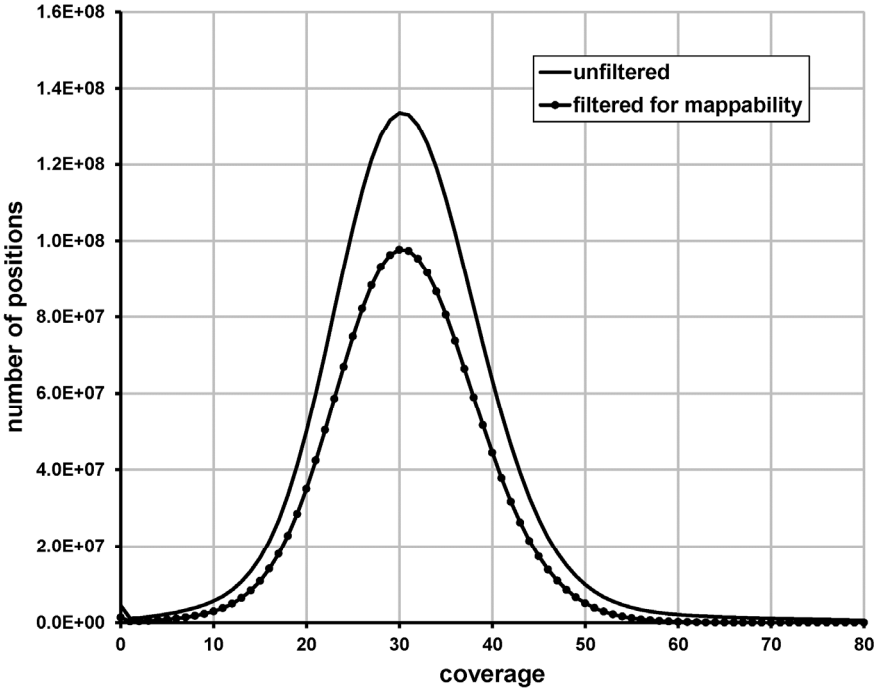


Figure S11: Comparison of the coverage distributions obtained for the autosomal genome (Denisova) with and without filtering for mappability (1.86 and 2.68 Gb, respectively).

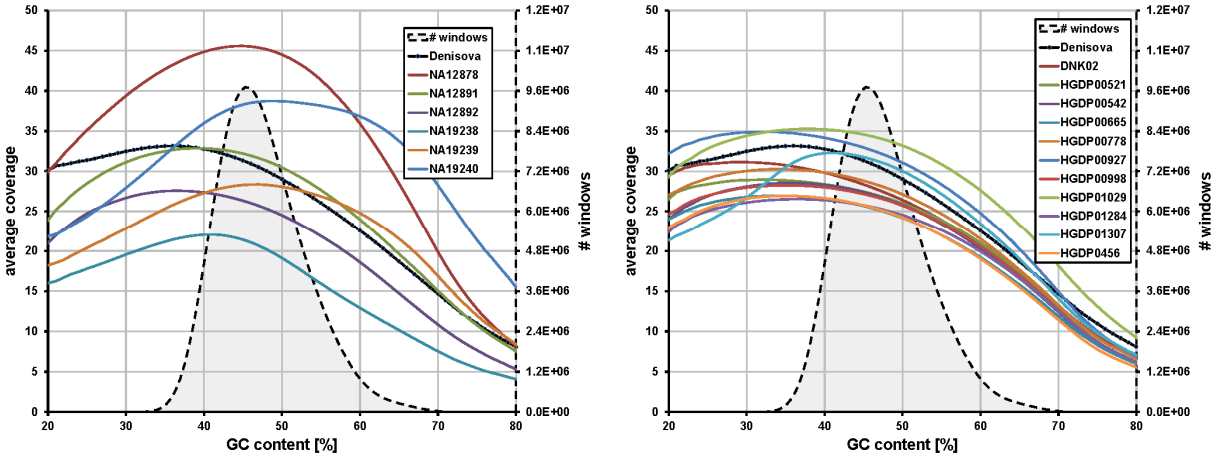


Figure S12: Average coverage of the autosomal genome as a function of GC-content for Denisova and the 1000 genomes trios (left) and the eleven present-day humans (right). The number of windows in each GC bin is indicated by a dashed line.

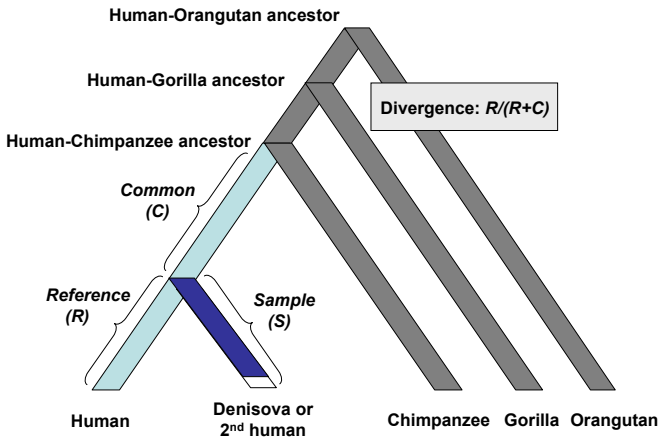


Figure S13: Schematic representation for determining average sequence divergence as the fraction of the branch leading from some human reference genome to the common ancestor of humans and chimpanzees. We count alleles common (C) to the reference (human) and a sample (Denisova or 2nd human), as well as reference-specific (R) and sample-specific alleles (S). Divergence is determined as R/(R+C). By comparing S and R, we determine branch length differences between sample and reference.

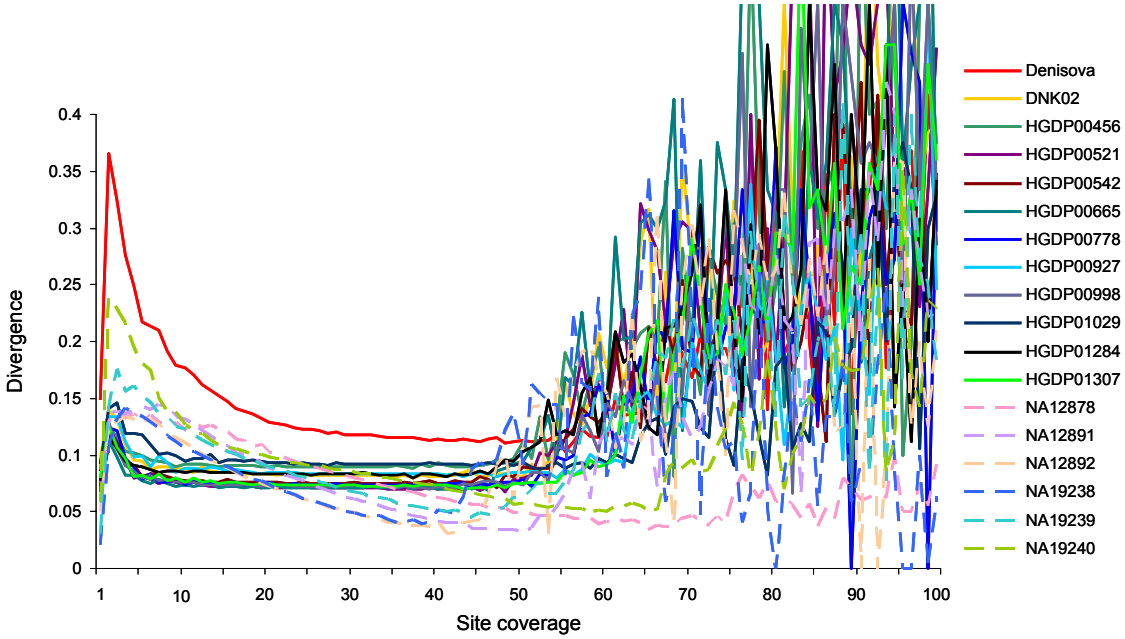


Figure S14: Divergence from the human reference sequence (GRCh37) for all autosomes by site coverage. Divergence estimates are unstable for low and high site coverage, probably reflecting an accumulation of false alignments in these site coverage bins.

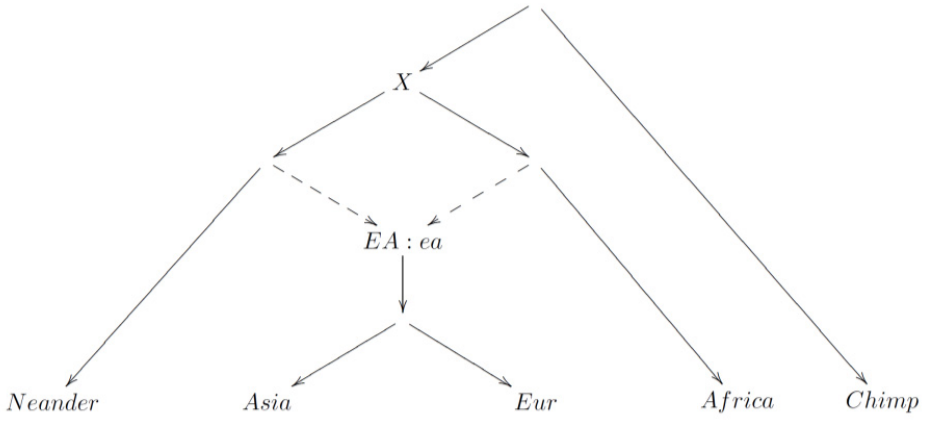


Figure S15: Enhanced *D*-statistics are valid tests for whether two populations share the same proportion of archaic ancestry. Conditioning on the African alleles being always ancestral does not change the expected rate of matching of Asians and Europeans to Neandertal, since for the null phylogeny shown, Africa has a symmetrical relationship to these two populations.

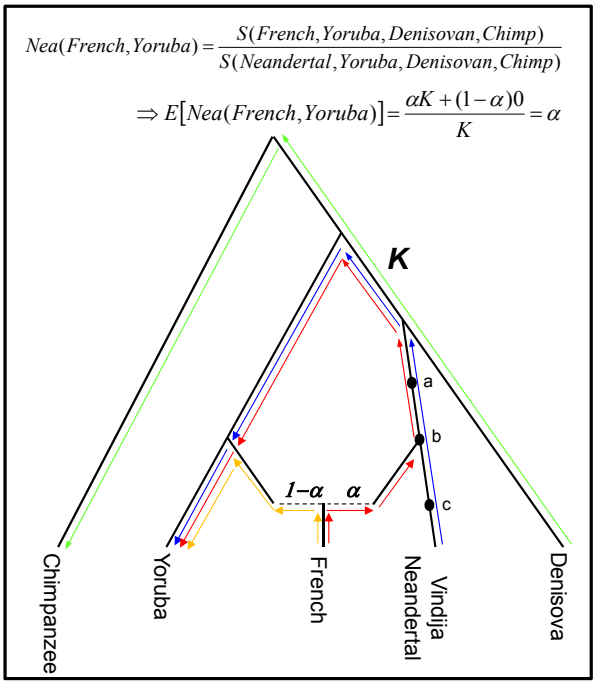


Figure S16: Estimating archaic ancestry. The *S*-statistic ratio *Nea(French,Yoruba)*, estimates the excess Neandertal ancestry in the French vs. Yoruba (assumed to have 0%). Denominator: The expection is *K,* the overlap between Denisova→Chimp (green) and Neandertal→Yoruba (blue). Numerator: The expectation is αK, the overlap *K* between Denisova→ Chimp (green) and the Neandertal ancestry in French (red), times the Neandertal proportion α. The modern human ancestry of French (orange) does not overlap Denisova→Chimpanzee, so does not contribute. This statistic is robust to how closely related the archaic population that contributed genes to French is to the Vindija Neandertal. As long as it is closer to Vindija than Denisova (i.e., it shares ancestry at *a, b,* or *c*), the red-green overlap *K* is the same. Analogous diagrams could be drawn to illustrate how *Nea(Han,French)* estimates the excess Neandertal ancestry in Han vs. French, and *Den(Papuan,Han)* the excess Denisovan ancestry in Papuans vs. Han.

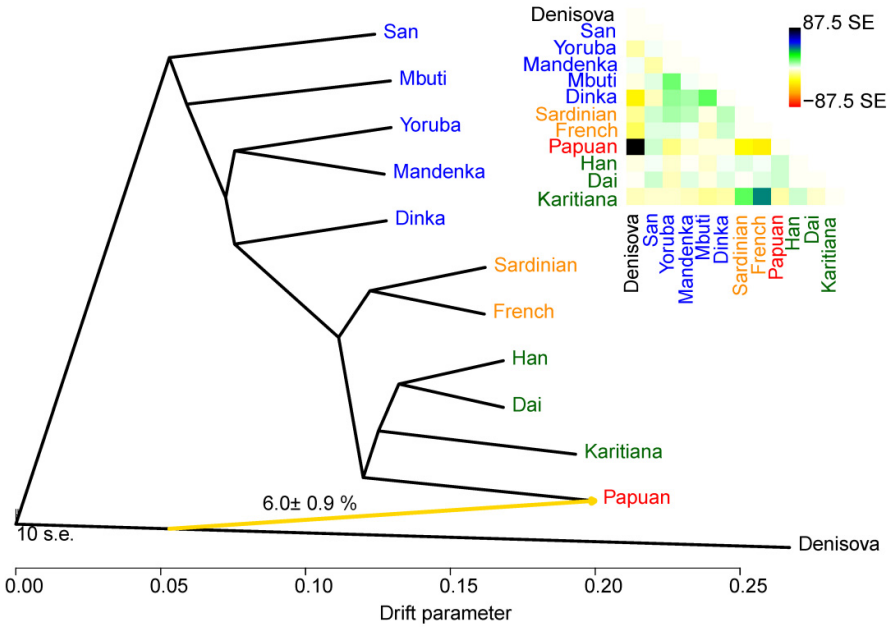


Figure S17: TreeMix maximum likelihood tree for Denisova and the eleven men. One migration event was allowed for the tree, and the residual matrix of the maximum likelihood tree without a migration event is shown in the upper right. Residuals above zero represent populations that are more closely related to each other in the data, with darker colors representing a stronger signal. The residual matrix shows that the Denisova-Papuan gene flow is the only strong migration signal in the data set. The arrow shows the Denisova-to-Papuan migration/admixture with a weight of 6.0% (±0.9% standard error). Populations are labeled in color (Denisova black, Africans blue, Europeans orange, mainland Asians and native South Americans green and Papuan red).

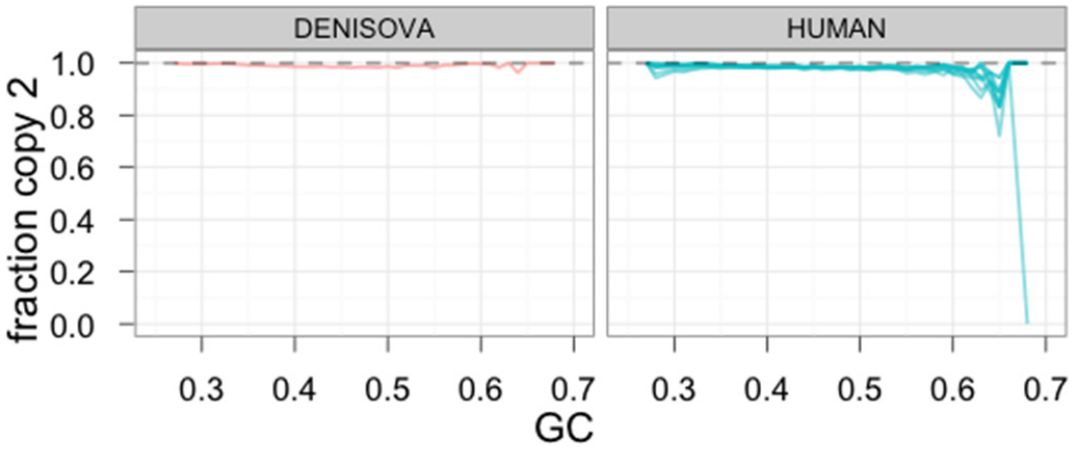


Figure S18: The number of unique 3 kbp regions correctly estimated to be copy number 2 plotted for the Denisovan and ten HGDP individuals plotted as a function of GC content. For both the Denisova and the ten humans, GC content does not affect copy number estimates until GC content exceeds 65%.

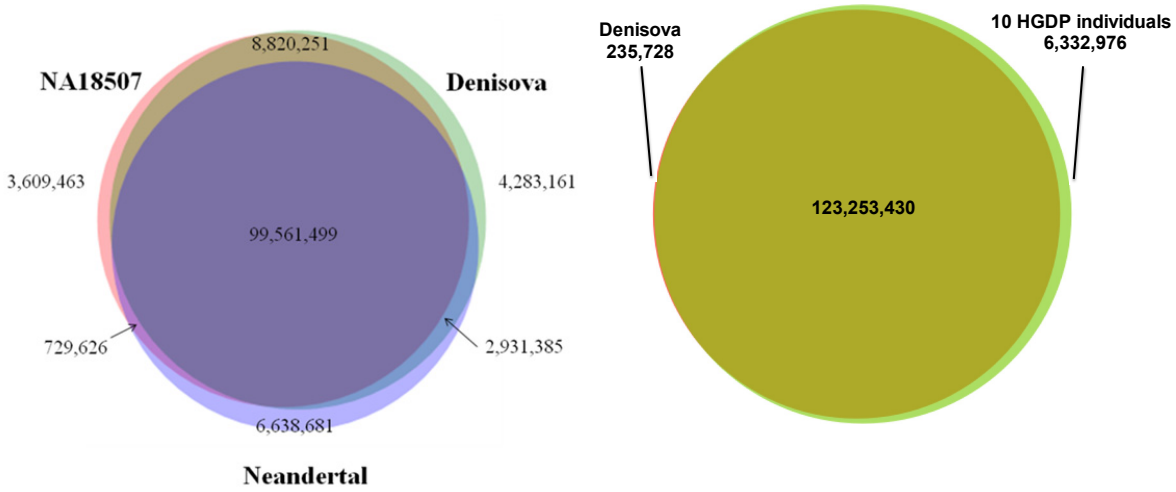
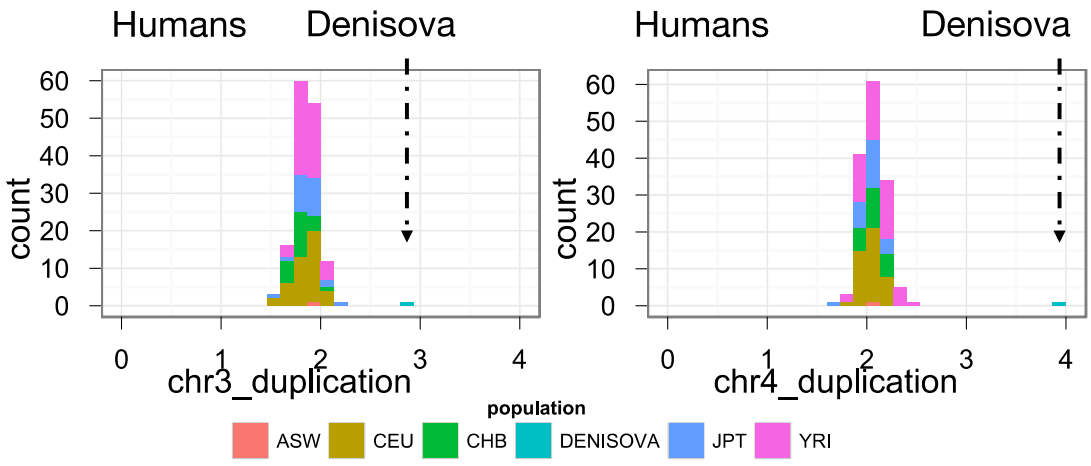


Figure S19: Three-way comparison of segmental duplications found in the human species and Denisova and a two-way comparison between duplications found in Denisova and ten humans (using relaxed thresholds) from the Human Genome Diversity Panel.



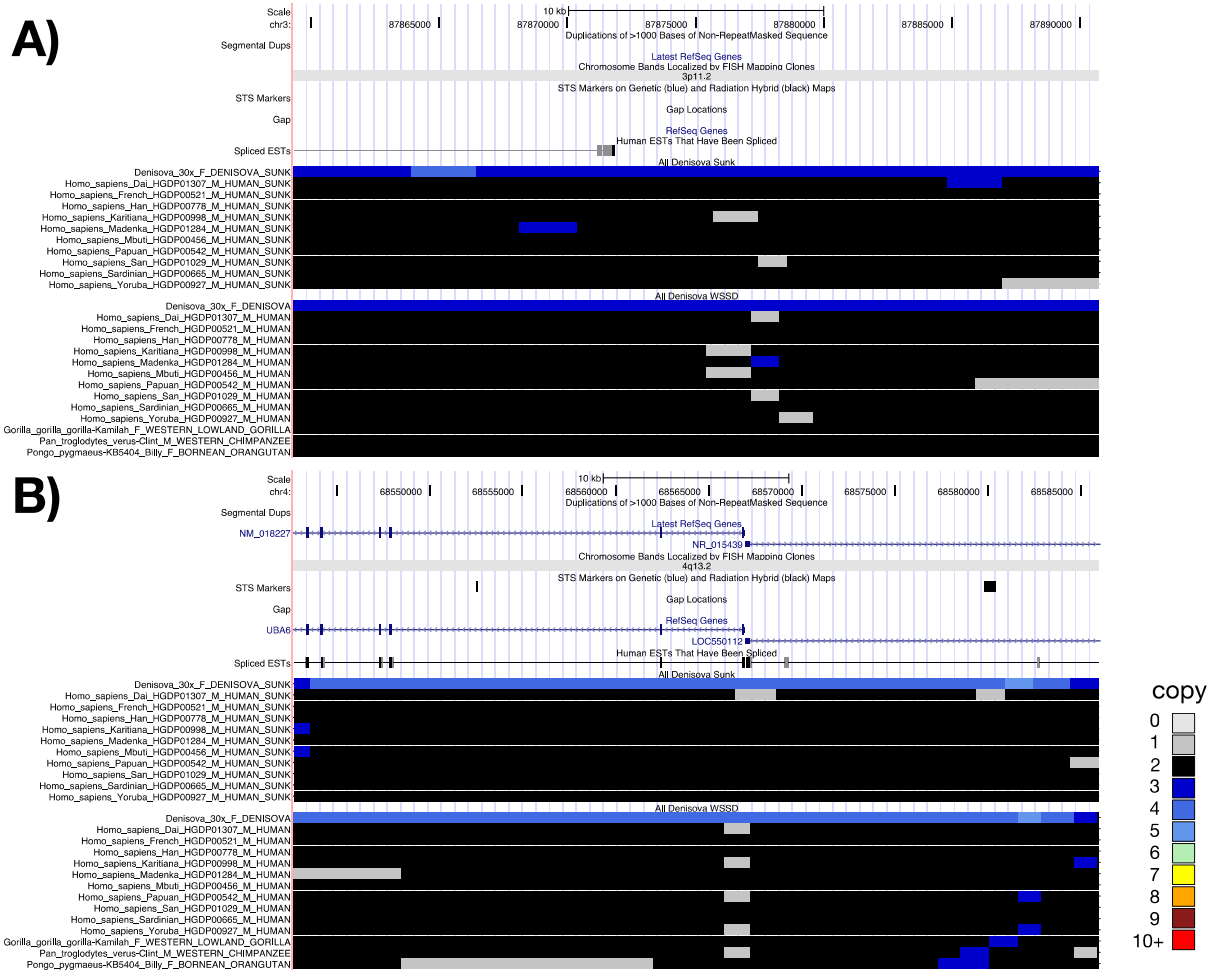


Figure S21: Heatmap representations of the Denisova-specific duplications located on chromosomes 3 and 4.

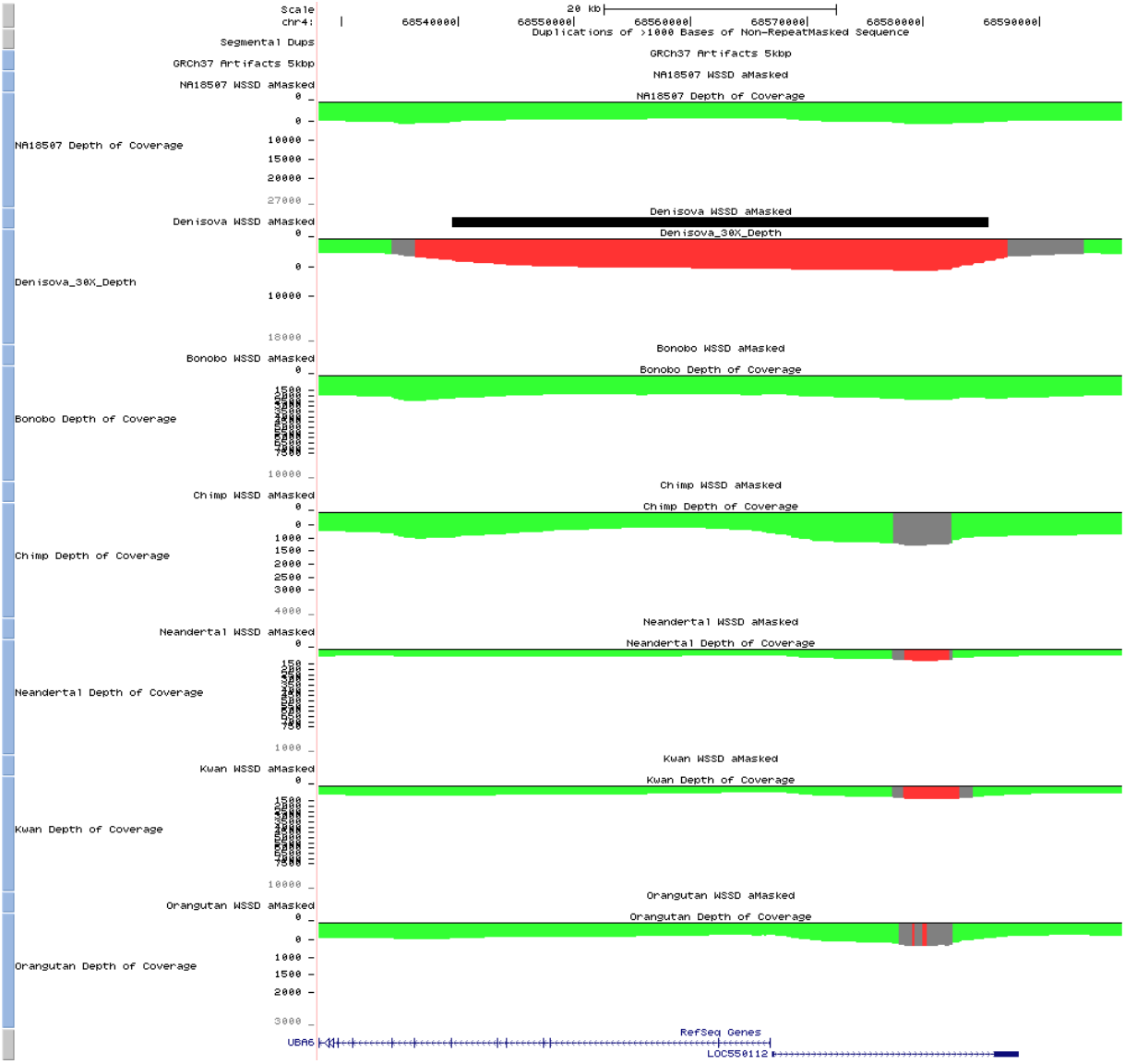


Figure S22: Read-depth in 4q13.2 region. Only the Denisovan sample shows an excess read-depth >20 kbp intersecting with two gene models *UBA6* and *LOC550112*. Grey coloring indicates read depth >3std from the mean, red coloring indicated read depth >4std from the mean.

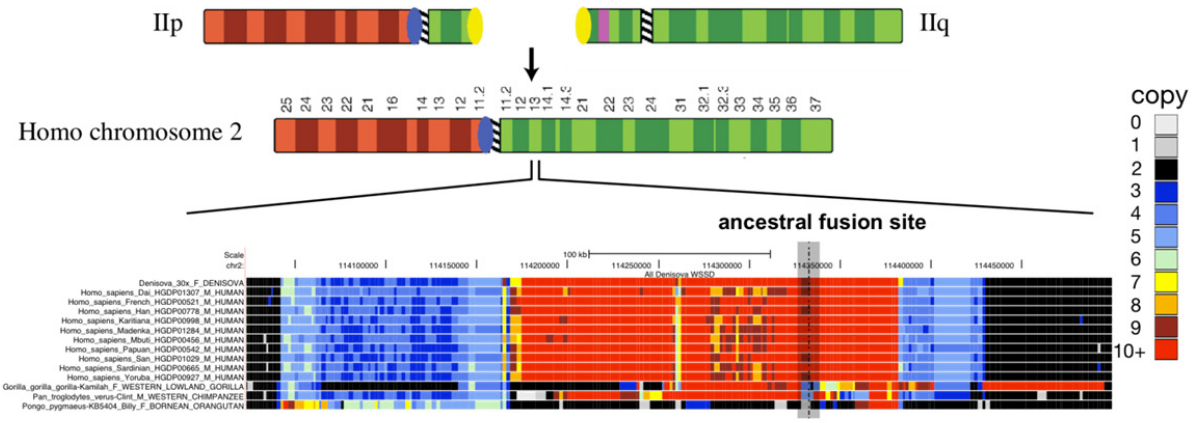


Figure S23: Chromosome 2 fusion in the human lineage involved juxtaposition of copy number polymorphic subtelomeric segmental duplications (yellow oval). A heatmap comparing the copy number for a 300 kbp region of the human chromosome 2p/2q fusion is shown. The boundaries and copy number of the segmental duplications are identical between Denisova and modern-day humans implying that there is no difference in the architecture of the region and consistent with this fusion being ancestral to both lineages.

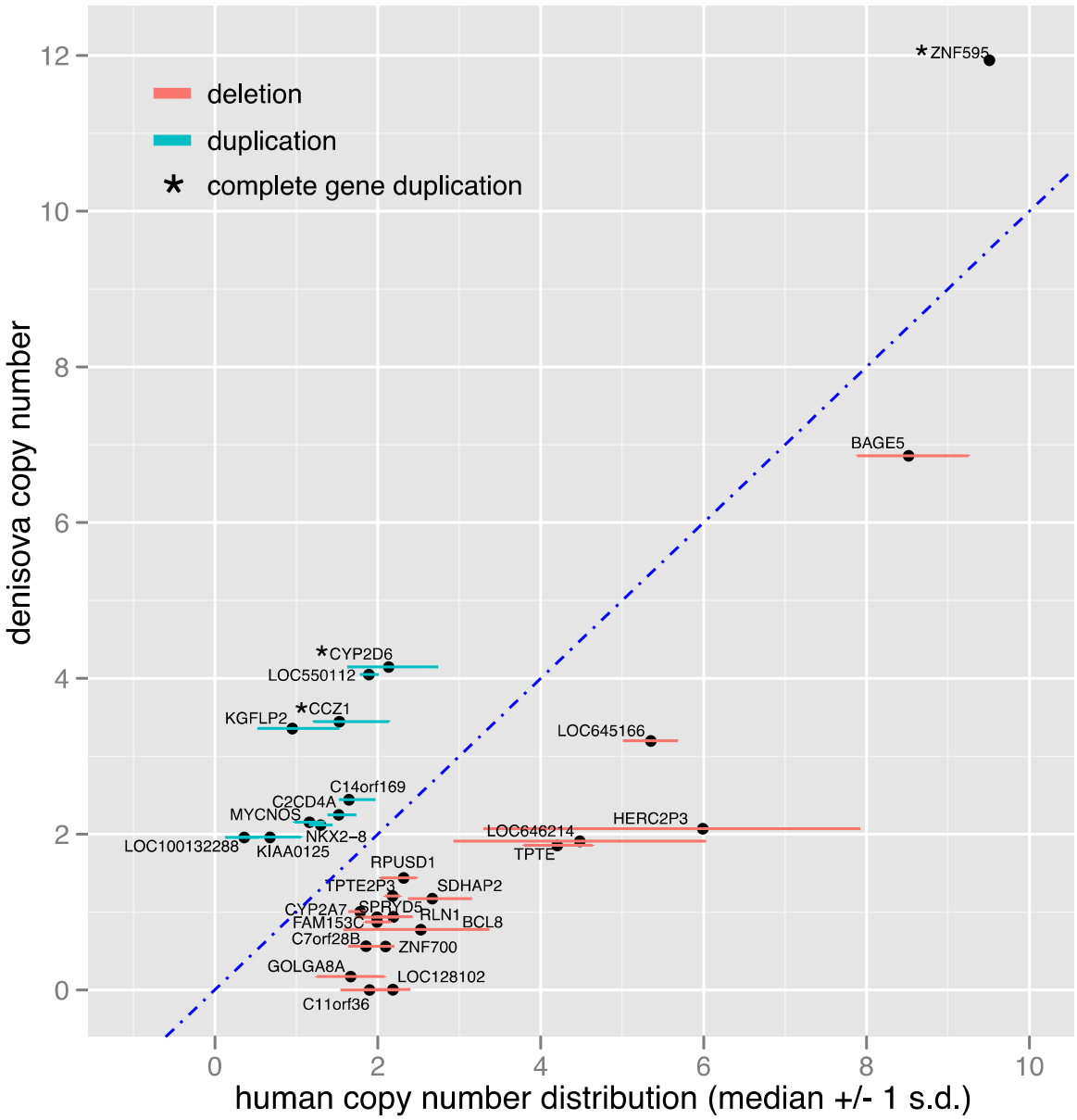


Figure S24: Genes duplicated and deleted in the Denisova that were not observed as copy number variant in the ten human genomes assayed. The Denisovan gene copy number is plotted versus the distribution of gene copy number estimates in the ten humans. Five genes were predicted to be duplicated in the Denisova but not in the 10 individuals analyzed and 19 genes deleted in the the Denisova but not in the 10 individuals analyzed. An examination of a larger sample of human genomes (1000 Genomes Pilot Project) revealed that with three exceptions, none of these were specific to Denisova when compared to contemporary human genomes.

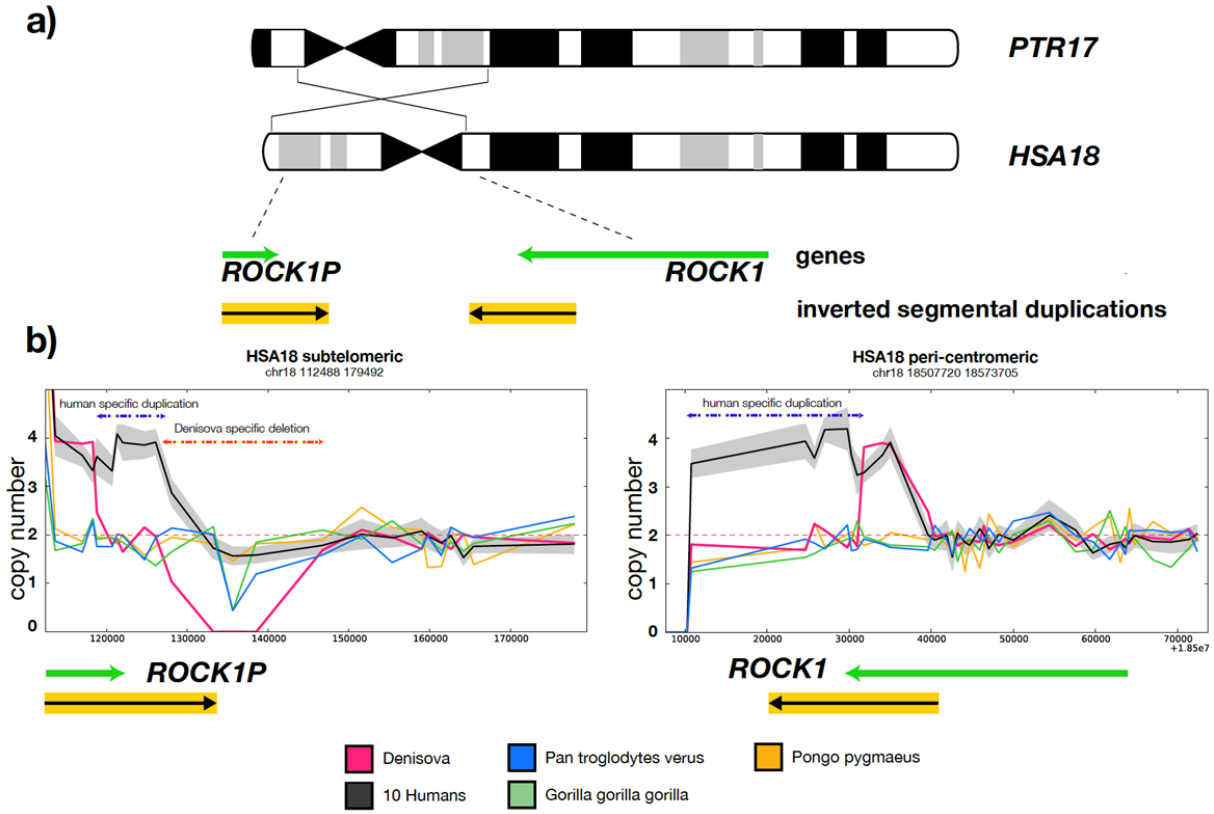


Figure S25: a) Chimpanzee chromosome 17 differs cytologically from its human homolog, chr18, by a pericentric inversion thought to be mediated in the human lineage by 19kb inverted segmental duplication which is not present in non-human primates. **b)** Copy number estimates at 500bp non-repeat masked resolution over the ROCK1 inversion breakpoint locus on the p and q arms of chromosome 18 demonstrate that the locus is duplicated to 4 copies in humans and diploid in non-human primates. The Denisovan specimen, however, shows only a partial duplication of the ROCK1 locus followed by a homozygous deletion, a pattern not observed in 146 human genomes sequenced in Pilot 1 of the 1000 genomes project

```
hg19      GGGTTGGGGTTGGGGTTGGGGTTAGGGTTAGCTAAACCTAACCCTAACCCCTAACCCCAACCC
Den(+)    GGGTTGGGGTTGGGGTTGGGGTTgGGtTTAGCTAAcCCTAACCCTAACCCC-AACCCC
Den(-)          GGGGTTGGGGTTaGGGTTAGGGTTAGCTAAACCcAACCCcAACCCC-AACCCCAACCC
```

Figure S26: Human reference sequence (hg19; chr2:114360474-114360537) and aligned Denisovan read (M_SOLEXA-GA03_00033_PEdi_MM_8:1:37:18364:19953) in forward and reverse orientation.

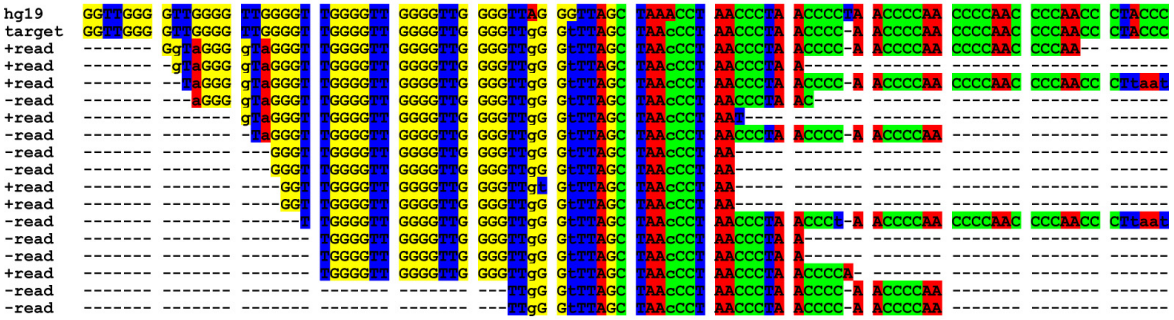


Figure S27: Denisovan reads aligned to the constructed target sequence (forward direction). Reads were reverse complemented if BWA reported a match to the reverse direction target. Read orientation is given as +/- on the left hand side.

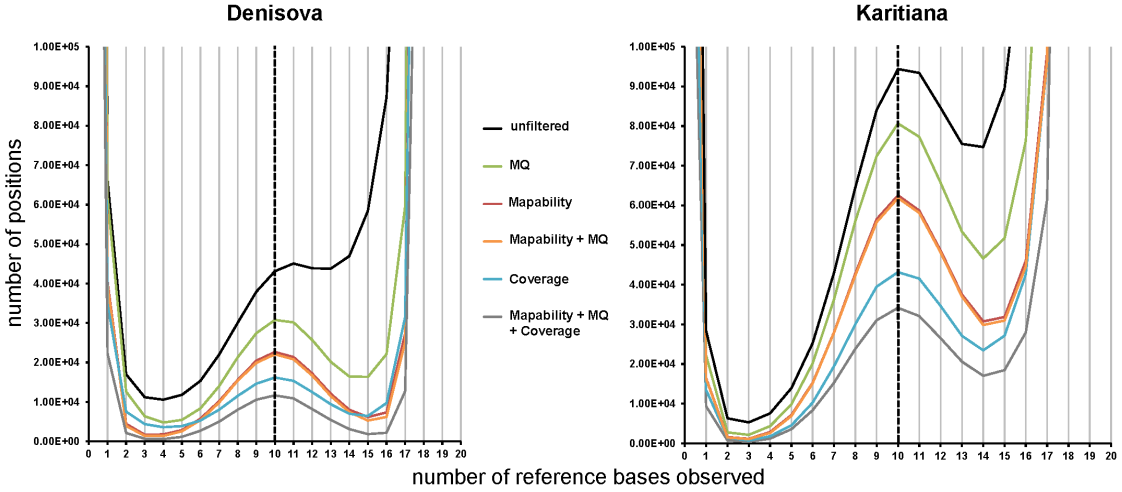


Figure S28: Effect of data filtering on base frequency spectra. Only Denisova and one modern human genome (HGDP00998, Karitiana) are shown. In this figure, changes in height of the central peak reflect differences in total number of sites remaining after filtering.

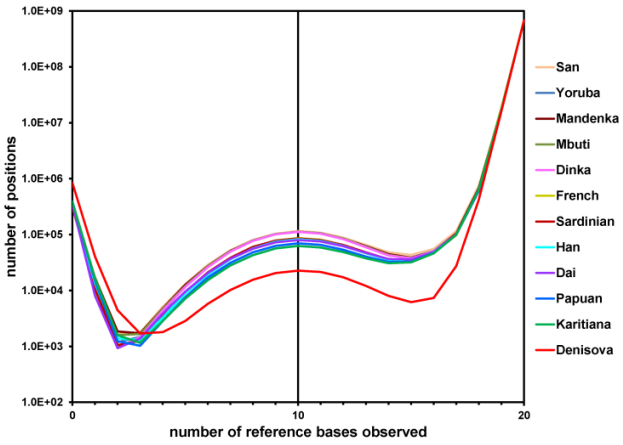


Figure S29: Comparison of base frequency spectra in logarithmic scale between Denisova and the 11 present-day human genomes (see Figure 5A of the main paper for the same plot in linear scale).

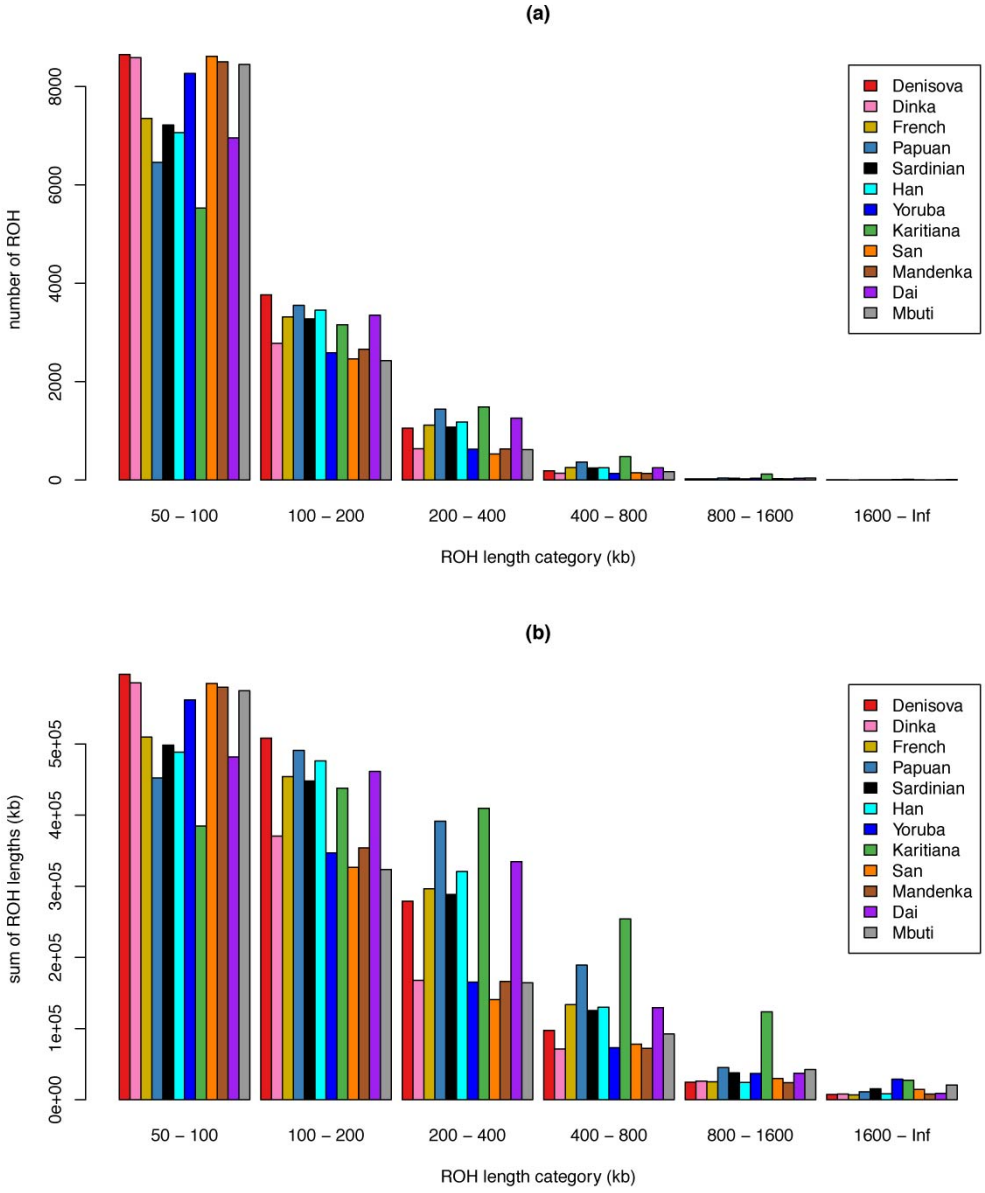


Figure S30: For each individual, **(a)** the number of ROH found for 6 length categories, and **(b)** the sum of ROH lengths found for each category are shown.

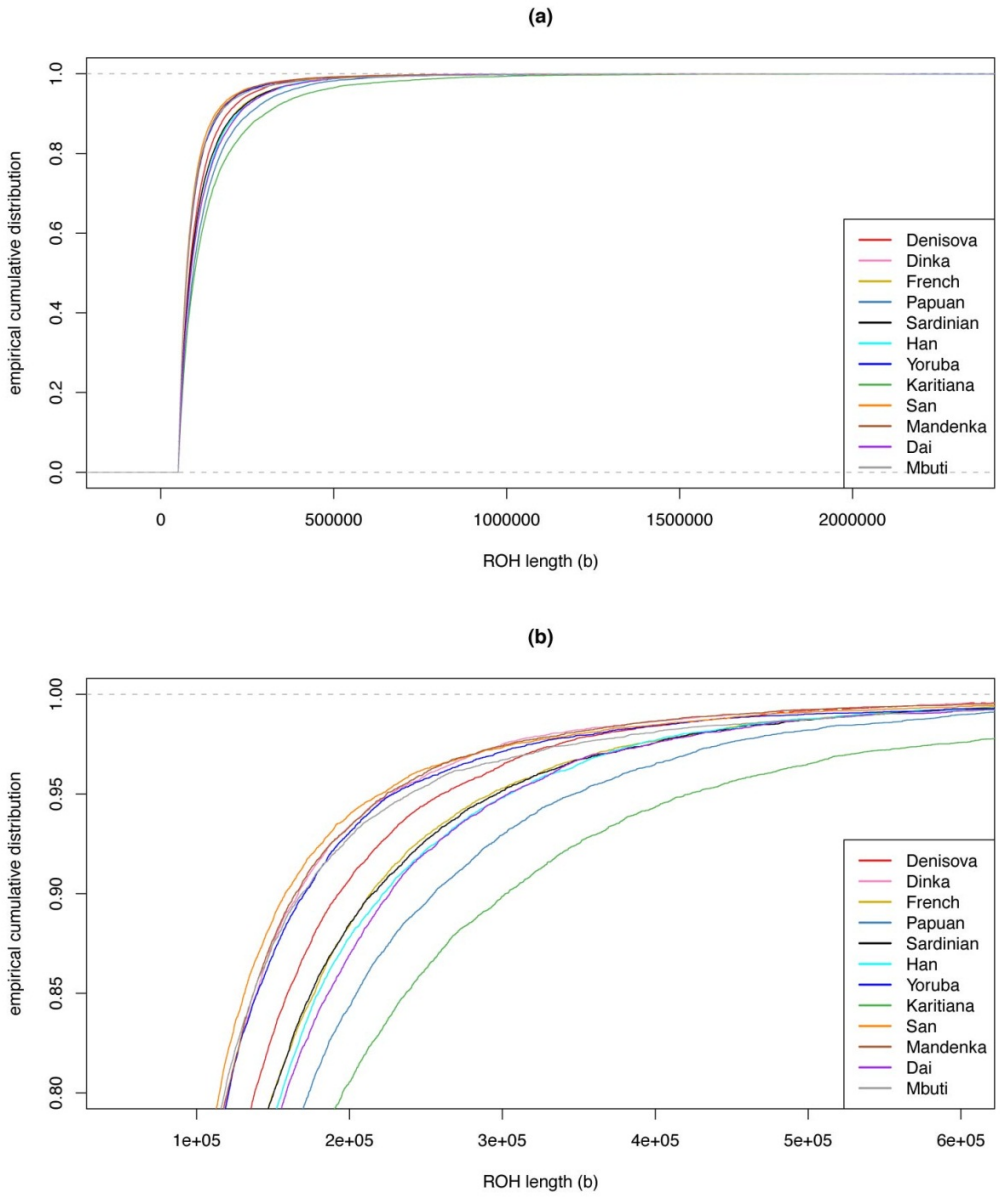


Figure S31: (a) Empirical cumulative distribution of the ROH lengths superior to 50 kb. **(b)** Zoom on this distribution.

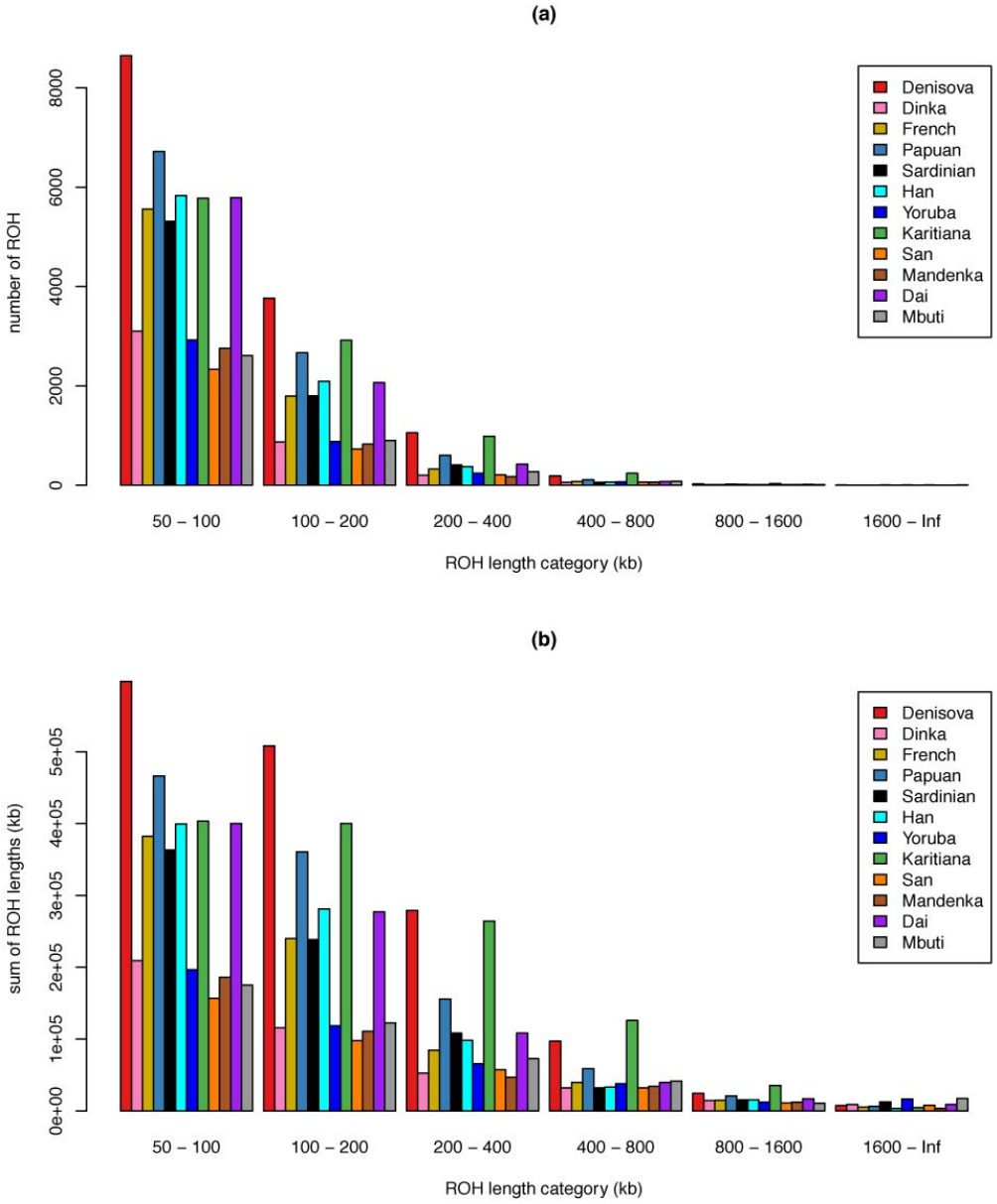


Figure S32: For each individual, **(a)** the number of ROH found for 6 length categories, and **(b)** the sum of ROH lengths found for each category are shown. For this plot, the data were NOT down-sampled to a common number of heterozygous sites.

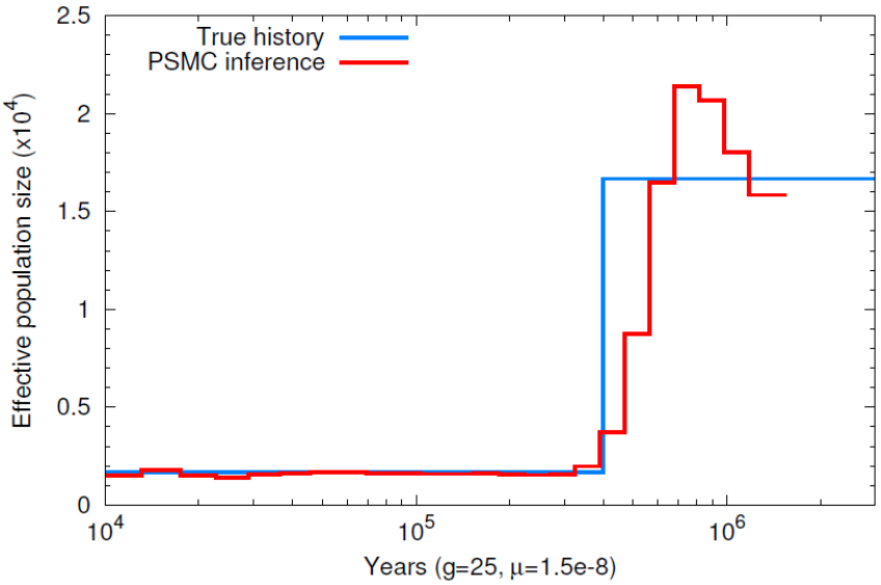


Figure S33: Simulations of a population crash in Denisovan history. The black line shows the true (simulated) history in which the ancestral population was constant at 16,667 diploid individuals until crashing to 1,667 at 400kya. Under this scenario, the PSMC smoothes out the population size change, thus estimating too-old a date for the crash and compensating for this by increasing the inferred population size prior to the date of the crash. A population crash just after the split from modern humans could explain the inference of a transiently larger population size in Denisovan ancestors than in modern humans in the period from 0.75×10^{-4} to 2×10^{-3} in divergence units (Figure 4).

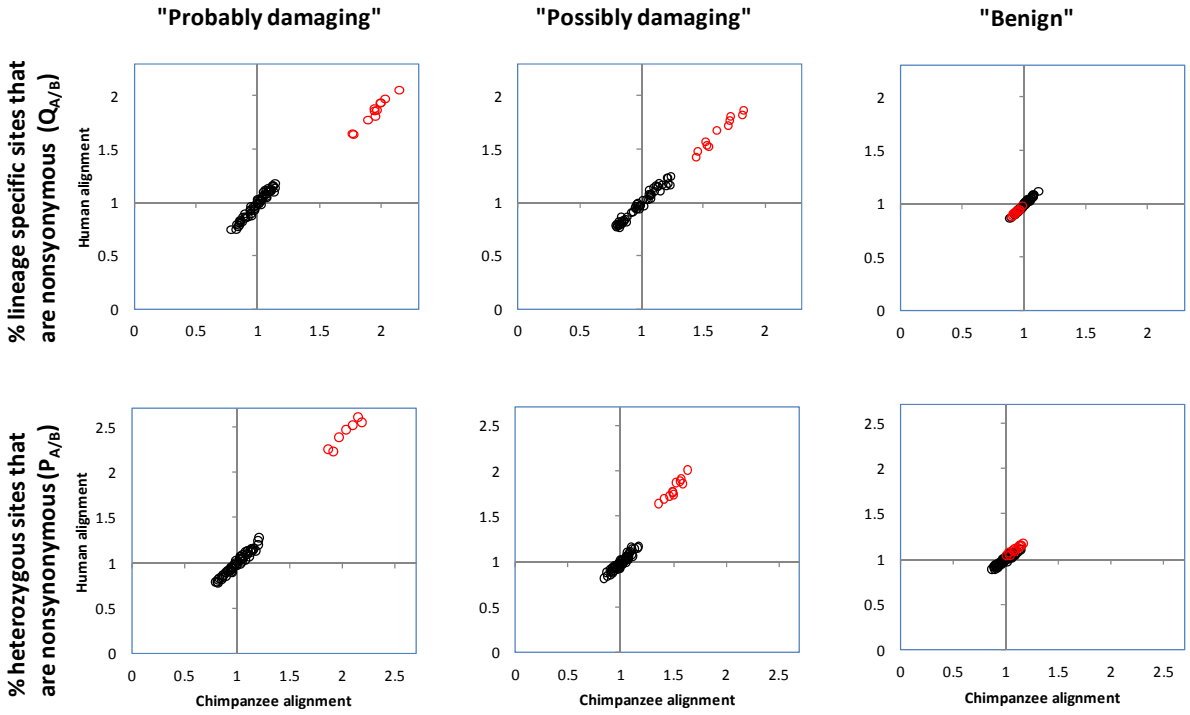


Figure S34: Non-synonymous-synonymous comparisons for all pairs of 11 modern humans and Denisova. The x-axis shows alignment to chimpanzee and the y alignment to human, showing evidence for less effective selection (higher non-synonymous-to-synonymous substitution rates) in Denisovans than in modern humans. In particular, the red points corresponding to Denisova/Modern are always >>1for (A) *qN/qS* ratios, and (B) *pN/pS* ratios for the more strongly constrained sites ("Probably damaging" and "Possibly damaging").

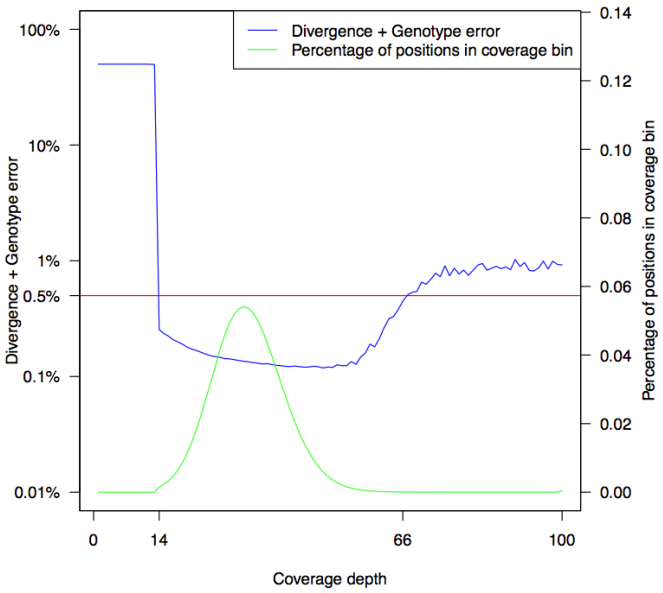


Figure S35: The sum of divergence and genotype error (blue curve) was calculated by counting the number of sites that differ from the human reference as a function of coverage. The percentage of total positions in each coverage bin (green curve) is also plotted for comparison. Genotype error was kept below 0.5% (red line) by filtering against sites with coverage below 14X or above 66X.

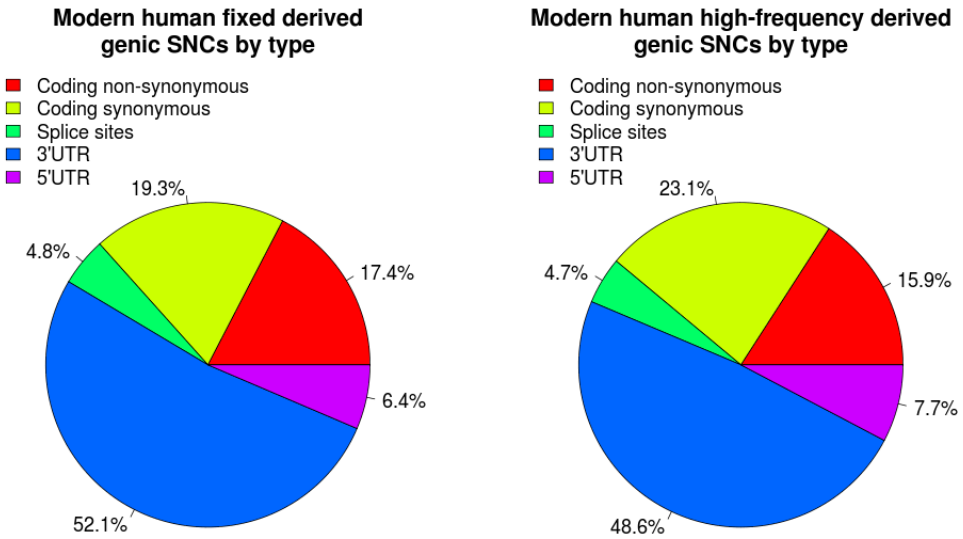


Figure S36: Modern-human-derived SNCs in CCDS-verified genes, classified by their predicted functional effect. The left pie-chart refers to SNCs fixed in modern humans, while the right pie-chart refers to SNCs that are not fixed but are at high-frequency (> 90%).

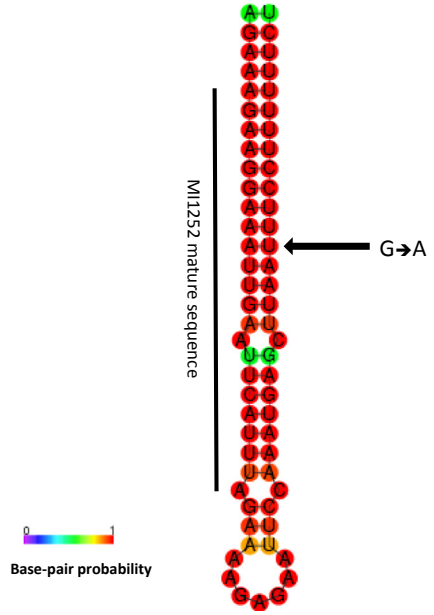


Figure S37: RNAfold prediction for the structure of MIR1252 (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi?). Colouring is according to base-pair probability. The fixed derived substitution (A/G) in the mature sequence in modern humans is indicated by an arrow.

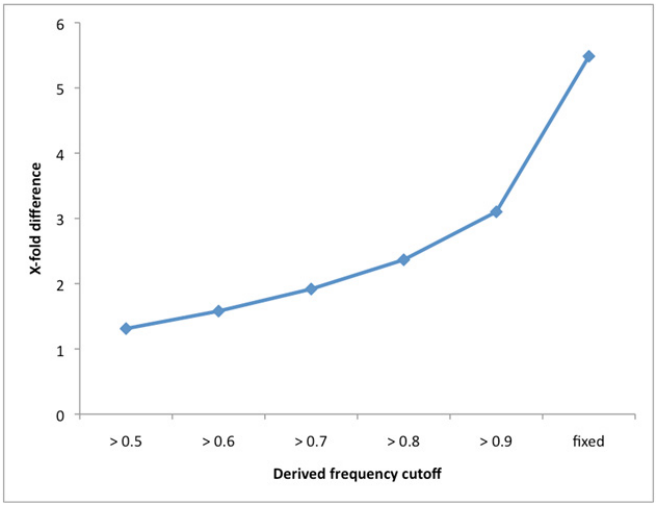


Figure S38: Number of Denisova-specific SNCs divided by the number of modern human-specific SNCs in chromosome 21, using different cutoffs for the modern human allele frequencies (derived in modern human-specific SNCs; ancestral in Denisova-specific SNCs).

Table S1: Libraries and sequences generated for this work.

extraction set	extract ID	volume of extract used for library preparation	molecules in library [qPCR copy count]	library ID after amplification (number of GAIIx lanes sequenced)	library ID after amplification and gel-excision (number of GAIIx lanes sequenced)
A	E236	5 µl	9.00E+09	L9110 (0.85)	B1133 (16)
	water	-	3.53E+08	L9114 (-)	- (-)
B	E245	28.5 µl	4.06E+09	B1087* (0.25) B1088* (0.25) B1101* (0.25) B1102* (0.25)	B1107 (8) B1108 (8) B1109 (6) B1110 (6)
	water	-	1.85E+08	B1082 (-)	- (-)
C	E245 side fractions	28 µl (all)	8.94E+09	B1128 (7)	B1130 (24)
	water	-	3.60E+08	B1126 (-)	- (-)
Further sequencing of libraries prepared in Reich et al. 2010	E236	23 µl		SL3003 (3)	
	E245	23 µl		SL3004 (40)	

* The library prepared from E245 was amplified in four separate reactions

Table S2: Oligonucleotides used in this work. Synthesis and purification (reverse-phase HPLC) were performed by Sigma-Aldrich (Steinheim, Germany)

oligo ID	sequence (5'-3')
CL78-2*	Phosphate-AGATCGGAAGXXXXXXXXXX-(TEG-biotin) (X = C3 spacer)
CL9	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
CL53	CGACGCTCTTC-ddC (ddC = dideoxy cytidine)
CL73	Phosphate-GGAAGAGCGTCGTGTAGGGAAAGAG*T*G*T*A (* = PTO bonds)
CL72	ACACTCTTTCCCTACACGACGCTCTTCC

* CL78-2 was double-purified by ion-exchange HPLC

Table S3: Comparison of sequence yield obtained from the double-stranded (gray) and the single-stranded (white) library preparation methods. Of the four libraries prepared from E245, B1108 was sequenced most deeply and is thus best suited to determine the minimum factor of improvement (bold). Sequences were filtered for length >= 35 bp and map quality >= 30.

Extract ID	Library ID	total number of mapped sequences	unique sequences	oversampling (total / unique sequences)	unique sequence [bp] obtained per microliter of extract	improvement [fold] over previous library preparation
E245	SL3004	6.53E+08	5.80E+07	11.26	2.12E+08	
	B1107	1.66E+08	1.04E+08	1.60	1.16E+09	5.6
	B1108	**2.13E+08**	**1.14E+08**	**1.87**	1.26E+09	**5.9**
	B1109	1.14E+08	7.61E+07	1.49	8.83E+08	4.2
	B1110	1.63E+08	9.50E+07	1.72	1.04E+09	4.9
E236	SL3003	9.96E+07	5.85E+07	1.70	1.54E+08	
	B1133	**3.22E+08**	**2.30E+08**	**1.40**	3.55E+09	**23.1**

Table S4: Index sequences expected in the first and second index read for each of the libraries sequenced. For the previously generated libraries (SL3003/SL3004), only a single index was used.

Extract ID	Library ID	First index sequence	Second index sequence
E245	SL3004	GTCGACT	-
	B1107	AATCTTC	TCGCAGG
	B1108	GGCGGAG	CTCTGCA
	B1109	ACCAACG	ATGGAGA
	B1110	AACCATG	CTCGATG
E236	SL3003	GTCGACT	-
	B1133	TGGACGT	TTGAAGT

Table S5: Amount of data generated from libraries prepared with the single-stranded (ssDNA) and double-stranded (dsDNA) methods. Numbers refer to trimmed reads aligning to the human genome.

Library	Number of mapped sequences	Number of unique sequences	Unique sequence total
ssDNA-1 (B1087, B1088, B1101, B1102, B1107, B1108, B1109, B1110)	734,975,883	433,274,222	30.69 Gb
ssDNA-2 (B1128, B1130)	785,182,632	573,373,035	32.92 Gb
ssDNA-3 (B1133)	363,597,807	257,529,083	18.58 Gb
dsDNA-1 (SL3003)	123,568,359	74,906,951	4.37 Gb
dsDNA-2 (SL3004)	783,579,414	79,874,407	5.96 Gb

Table S6: Number of read pairs generated for each of the 11 samples

Population	Sample ID	Passing read pairs	Barcode 1	Barcode 2	Barcode 3	Barcode 4
			TGTCTA	GGCGAC	CTGCTG	ACATGT
Dinka	**DNK02**	557,294,298	147,142,542	130,411,691	168,894,143	110,845,922
			CTATGA	TCCTCC	ATGTAG	GGTCGT
Mbuti	**HGDP00456**	470,074,459	118,067,290	179,353,003	599,81,681	112,672,485
			ACAATA	CACCAC	GATCGG	CGTTAT
French	**HGDP00521**	524,956,032	127,157,381	127,888,820	127,271,246	142,638,585
			GAGGCA	CTCAGC	GGTACG	ATGTGT
Papuan	**HGDP00542**	507,688,032	126,249,765	135,855,630	126,014,282	119,568,355
			TGAGTA	GCATTC	CCGACG	TAAGGT
Sardinian	**HGDP00665**	484,602,356	125,124,111	123,766,638	129,418,775	106,292,832
			CCTGTA	GTGAAC	AGGTCG	TAACAT
Han	**HGDP00778**	533,618,709	159,011,895	114,238,012	131,877,833	128,490,969
			TTCATA	GCTGCC	TTGGCG	CCGCGT
Yoruba	**HGDP00927**	623,180,427	195,042,398	134,030,523	159,865,376	134,242,130
			AGCGCA	TAGATC	CCATAG	GTGCCT
Karitiana	**HGDP00998**	531,376,838	134,613,560	128,496,477	136,904,364	131,362,437
			GCGCTA	AGGTTC	AGAAGG	TACGAT
San	**HGDP01029**	680,873,623	202,860,772	170,199,758	147,392,299	160,420,794
			CACTTA	ATACAC	TGTGTG	ACTGAT
Mandenka	**HGDP01284**	501,820,015	139,452,707	110,292,388	151,872,989	100,201,931
			AATACA	CGACTC	GCTAAG	TTCCAT
Dai	**HGDP01307**	528,381,063	117,995,360	140,502,391	128,743,238	141,140,074

Note: Each sample has 4 barcodes, each of which ends with a different nucleotide “A”, “C”, “G” and “T” which balances the nucleotide compositions of the libraries and improves the performance of the clustering.

Table S7: Mapping of reads to the human and chimpanzee genomes

Population	Sample ID	# mapped reads	# reads part of proper pairs	# reads after duplicate removal	% of total reads used	Gigabases used
Mapping to human (*hg19/GRCh37*)						
Dinka	DNK02	947,105,021	930,831,666	851,171,031	76.37%	82.56
Mbuti	HGDP00456	886,228,397	866,931,212	746,142,578	79.36%	72.38
French	HGDP00521	993,077,880	971,254,974	822,589,731	78.34%	79.79
Papuan	HGDP00542	957,209,275	935,417,074	802,115,913	79.00%	77.81
Sardinian	HGDP00665	916,389,106	892,615,658	767,127,809	79.15%	74.41
Han	HGDP00778	1,005,969,464	978,790,234	862,350,953	80.80%	83.65
Yoruba	HGDP00927	1,184,897,821	1,153,642,906	1,000,910,182	80.30%	97.09
Karitiana	HGDP00998	996,500,295	968,156,064	813,991,887	76.59%	78.96
San	HGDP01029	1,276,748,296	1,230,278,942	1,056,952,329	77.62%	102.52
Mandenka	HGDP01284	937,936,405	908,846,748	775,089,405	77.23%	75.18
Dai	HGDP01307	1,002,157,271	983,337,904	880,009,435	83.27%	85.36
	Total	**11,104,219,231**	**10,820,103,382**	**9,378,451,253**		**909.71**
Mapping to chimpanzee (*panTro2*)						
Dinka	DNK02	916,549,057	884,628,712	801,553,975	71.91%	77.75
Mbuti	HGDP00456	848,074,664	814,965,242	686,143,163	72.98%	66.56
French	HGDP00521	940,705,102	903,150,342	760,166,624	72.40%	73.74
Papuan	HGDP00542	915,884,235	878,398,804	734,275,412	72.32%	71.22
Sardinian	HGDP00665	873,543,296	835,455,770	700,043,844	72.23%	67.90
Han	HGDP00778	967,050,867	924,225,448	796,567,857	74.74%	77.27
Yoruba	HGDP00927	1,124,940,368	1,075,787,652	906,738,040	72.75%	87.95
Karitiana	HGDP00998	952,528,353	909,023,542	742,750,359	69.89%	72.05
San	HGDP01029	1,174,425,666	1,109,986,162	890,881,364	65.42%	86.42
Mandenka	HGDP01284	897,038,052	853,579,424	701,087,589	69.85%	68.01
Dai	HGDP01307	961,796,375	924,978,454	814,542,161	77.08%	79.01
	Total	**10,572,536,035**	**10,114,179,552**	**8,534,750,388**		**827.88**

Table S8: Comparison between the first and second iteration of GATK genotype calling for the Denisova genome with the same parameters but different reference genomes. In the first iteration, the human or chimpanzee reference genome was used, and for the second iteration the Denisovan alternative allele identified in the first iteration was used in a new reference sequence. Genotype Quality cutoffs (none, GQ ≥ 30 and GQ ≥ 40) were applied.

Genome	GQ cutoff	Genotype first call	Genotype second call (%)						#sites
			./.	0/0	0/1	1/1	1/2	2/2	
Chim-panzee	none	0/1	0.0012	0.1916	70.9358	28.6816	0.1896	0.0002	4,904,187
	30	0/1	0.0000	0.1523	99.7835	0.0037	0.0605	0.0000	3,454,674
	40	0/1	0.0000	0.0873	99.8647	0.0026	0.0454	0.0000	2,761,366
	none	1/1	0.0001	0.0000	0.0004	99.8643	0.1351	0.0001	30,198,961
	30	1/1	0.0000	0.0000	0.0001	99.9423	0.0576	0.0000	26,426,267
	40	1/1	0.0000	0.0000	0.0001	99.9602	0.0397	0.0000	24,647,342
Human	none	0/1	0.0028	0.2646	89.4394	10.2112	0.0817	0.0003	1,556,029
	30	0/1	0.0000	0.1740	99.7813	0.0052	0.0393	0.0000	1,382,953
	40	0/1	0.0000	0.1049	99.8579	0.0041	0.0329	0.0000	1,188,232
	none	1/1	0.0004	0.0002	0.0020	99.9028	0.0947	0.0000	3,907,833
	30	1/1	0.0000	0.0000	0.0008	99.9572	0.0419	0.0000	3,524,909
	40	1/1	0.0000	0.0000	0.0007	99.9701	0.0292	0.0000	3,335,158

Table S9: Mitochondrial contamination estimates.

extract	library ID	sequences identified as		contamination estimate		
		Denisovan	human	lower 95% C.I.	mean	upper 95% C.I.
E245	B1087+B1107	60,978	217	0.31%	0.35%	0.40%
	B1088+B1108	67,173	208	0.27%	0.31%	0.35%
	B1101+B1109	45,149	146	0.27%	0.32%	0.38%
	B1102+B1110	56,296	181	0.28%	0.32%	0.37%
	SL3004	38,148	135	0.30%	0.35%	0.42%
E236	B1133	148,409	402	0.25%	0.27%	0.30%
	SL3003	52,892	216	0.36%	0.41%	0.46%
E245 side fractions	B1128+B1130	230,649	928	0.38%	0.40%	0.43%
Combined		**699,694**	**2,433**	**0.33%**	**0.35%**	**0.36%**

Table S10: MLE fit of model parameters for the autosomal contamination estimate.

	Estimate	2.50%	97.50%	Std. Error	z value	Pr(z)
pad	0.001871	0.001843	0.001898	1.510E-05	124.0	< 2.2E116
pdd	0.986110	0.986040	0.986184	3.990E-05	24729.4	< 2.2E116
con	0.002243	0.002170	0.002317	3.780E-05	59.4	< 2.2E116
err	0.001472	0.001468	0.001477	2.400E-06	613.7	< 2.2E116

Table S11: Comparison of coverage statistics and error rates (autosomes only). The divergence-corrected per-base error rates are given in brackets.

data set	sample	population	*Uniquely mappable regions (1.86 Gb)* average coverage	% of positions covered >=1x	% of positions covered >=10x	% of positions covered >=20x	geno-type qual >= 40	*Conserved regions (5.6 Mb)* perBP error [%] (corrected)	*Conserved regions (5.6 Mb)* geno-type diff [%]
	Denisova		30.76	99.93	99.43	92.93	97.64	0.183 (0.133)	0.050
1000g trios	NA19238	YRI mother	18.87	99.94	94.92	44.51	70.14	1.717 (1.685)	0.032
	NA12892	CEU mother	22.68	94.48	92.99	66.14	85.18	1.493 (1.464)	0.029
	NA19239	YRI father	24.86	99.97	99.04	78.23	91.45	1.241 (1.208)	0.033
	NA12891	CEU father	28.59	99.96	99.44	90.55	95.68	1.446 (1.417)	0.029
	NA19240	YRI daughter	32.58	99.98	99.76	92.94	97.73	1.195 (1.161)	0.034
	NA12878	CEU daughter	37.47	99.97	99.61	97.31	98.70	1.203 (1.174)	0.029
11 male genomes, this study	HGDP0456	Mbuti	24.34	99.97	98.45	75.99	93.22	0.212 (0.175)	0.037
	HGDP01284	Mandenka	24.51	99.97	98.75	77.13	94.01	0.213 (0.179)	0.034
	HGDP00665	Sardinian	24.68	99.97	98.83	78.55	94.41	0.207 (0.177)	0.030
	HGDP00542	Papuan	25.93	99.96	98.85	82.27	95.03	0.206 (0.173)	0.033
	HGDP00998	Karitiana	26.02	99.96	98.98	82.74	95.41	0.209 (0.178)	0.031
	HGDP00521	French	26.73	99.97	99.19	86.10	96.40	0.216 (0.186)	0.030
	HGDP00778	Han	27.74	99.97	99.22	87.83	96.65	0.209 (0.178)	0.031
	DNK02	Dinka	27.98	99.99	99.31	88.79	96.78	0.213 (0.180)	0.033
	HGDP01307	Dai	28.31	99.97	99.14	87.03	96.31	0.212 (0.181)	0.031
	HGDP00927	Yoruba	32.12	99.98	99.50	94.58	98.17	0.219 (0.185)	0.034
	HGDP01029	San	32.74	99.98	99.57	94.89	98.30	0.233 (0.194)	0.039

Table S12: Coverage cutoffs for removing 2.5% of sites from either end of the coverage distribution for Denisova and each of the eleven present-day human datasets. Cutoffs are listed separately for autosomes and the X-chromosome.

	autosomes		X-chromosome	
individual	lowest coverage included	highest coverage included	lowest coverage included	highest coverage included
Denisova	16	46	16	46
DNK02	15	41	5	23
HGDP00521	14	39	5	22
HGDP00542	13	39	4	22
HGDP00665	13	37	4	21
HGDP00778	15	41	5	23
HGDP00927	17	47	6	26
HGDP00998	13	39	4	22
HGDP01029	18	48	6	27
HGDP01284	13	37	4	21
HGDP01307	14	43	4	23
HGDP0456	12	37	4	21

Table S13: Divergence and relative Denisova branch length estimated from the autosomes, based on the Human-Chimpanzee ancestor as outgroup. Branch length ratios are computed for the sample and reference-specific branches (R+S), and additionally for the branches leading to the Human-Chimpanzee ancestor ((C+S/(C+R)).

Reference	Population	Comment	C	R	S	Div	S/R	(C+S)/ (C+R)
DNK02	Dinka	this study	7215566	1018902	925645	12.4%	0.91	0.989
HGDP00456	Mbuti	this study	7216483	1017581	923776	12.4%	0.91	0.989
HGDP00521	French	this study	7231548	1001264	907295	12.2%	0.91	0.989
HGDP00542	Papuan	this study	7198335	1015310	919839	12.4%	0.91	0.988
HGDP00665	Sardinian	this study	7183739	1010147	916614	12.3%	0.91	0.989
HGDP00778	Han	this study	7256376	1029978	931924	12.4%	0.90	0.988
HGDP00927	Yoruba	this study	7227303	1017948	922743	12.3%	0.91	0.988
HGDP00998	Karitiana	this study	7176073	1025627	921628	12.5%	0.90	0.987
HGDP01029	San	this study	7152333	1011305	916857	12.4%	0.91	0.988
HGDP01284	Mandenka	this study	7207566	1012813	919676	12.3%	0.91	0.989
HGDP01307	Dai	this study	7191000	1018664	923898	12.4%	0.91	0.988
NA12878	CEU	1000G trio: daughter	7236669	1005347	921705	12.2%	0.92	0.990
NA12891	CEU	1000G trio: father	7228586	1004261	919670	12.2%	0.92	0.990
NA12892	CEU	1000G trio: mother	6827584	949233	870200	12.2%	0.92	0.990
NA19238	YRI	1000G trio: mother	7192467	998725	919344	12.2%	0.92	0.990
NA19239	YRI	1000G trio: father	7208609	1006342	921359	12.3%	0.92	0.990
NA19240	YRI	1000G trio: daughter	7189703	1014265	918701	12.4%	0.91	0.988
GRCh37	-	Human reference	7611529	1038443	974590	12.0%	0.94	0.993

Table S14: Matrix of all pairwise autosomal divergence estimates in percent. Divergence is reported twice for each pair of genomes (below and above the diagonal), because each genome can be used as reference (R) or sample (S) in divergence calculation (see also Figure S13). The daughters from the 1000 Genome trios are marked with an asterisk (*).

Reference \ Sample	Denisova	DNK02 (Dinka)	HGDP00456 (Mbuti)	HGDP00521 (French)	HGDP00542 (Papuan)	HGDP00665 (Sardinian)	HGDP00778 (Han)	HGDP00927 (Yoruba)	HGDP00998 (Karitiana)	HGDP01029 (San)	HGDP01284 (Mandenka)	HGDP01307 (Dai)	NA12878* (CEU)	NA12891 (CEU)	NA12892 (CEU)	NA19238 (YRI)	NA19239 (YRI)	NA19240* (YRI)	GRCh37
Denisova		11.4	11.4	11.3	11.1	11.3	11.3	11.4	11.3	11.4	11.4	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.4
DNK02	12.4		9.2	8.5	8.6	8.5	8.5	8.7	8.6	9.5	8.7	8.5	8.5	8.5	8.5	8.6	8.7	8.7	8.7
HGDP00456	12.4	9.2		9.3	9.4	9.3	9.3	9.3	9.3	9.6	9.4	9.3	9.3	9.3	9.3	9.3	9.3	9.3	9.4
HGDP00521	12.4	8.5	9.3		7.3	6.6	7.1	8.7	7.0	9.5	8.7	7.1	6.6	6.6	6.6	8.7	8.8	8.8	7.6
HGDP00542	12.2	8.6	9.4	7.3		7.3	7.0	8.9	7.0	9.6	8.8	7.0	7.3	7.3	7.3	8.8	8.9	8.9	8.0
HGDP00665	12.4	8.5	9.3	6.7	7.3		7.1	8.7	7.1	9.5	8.7	7.2	6.6	6.6	6.6	8.7	8.7	8.7	7.6
HGDP00778	12.3	8.5	9.3	7.1	7.0	7.1		8.8	6.5	9.5	8.7	6.3	7.1	7.1	7.1	8.7	8.8	8.8	7.8
HGDP00927	12.4	8.7	9.3	8.8	8.9	8.7	8.8		8.8	9.6	8.7	8.8	8.8	8.7	8.8	8.6	8.7	8.7	8.8
HGDP00998	12.3	8.6	9.3	7.0	7.0	7.1	6.5	8.8		9.6	8.8	6.6	7.0	6.9	7.0	8.7	8.8	8.8	7.7
HGDP01029	12.5	9.6	9.7	9.6	9.7	9.6	9.6	9.7	9.7		9.7	9.7	9.6	9.6	9.7	9.7	9.7	9.8	9.8
HGDP01284	12.4	8.7	9.4	8.7	8.9	8.7	8.8	8.7	8.8	9.6		8.8	8.8	8.7	8.8	8.7	8.7	8.7	8.8
HGDP01307	12.3	8.5	9.3	7.1	7.0	7.1	6.3	8.8	6.6	9.5	8.7		7.1	7.1	7.2	8.7	8.8	8.8	7.8
NA12878*	12.2	8.4	9.2	6.5	7.2	6.5	7.0	8.6	6.8	9.4	8.6	7.0		4.9	4.9	8.5	8.5	8.5	7.3
NA12891	12.2	8.3	9.2	6.5	7.2	6.5	7.0	8.6	6.8	9.4	8.6	7.0	4.9		6.4	8.5	8.5	8.5	7.3
NA12892	12.2	8.4	9.2	6.5	7.2	6.5	7.0	8.6	6.8	9.4	8.6	7.0	4.9	6.4		8.5	8.6	8.6	7.3
NA19238	12.2	8.4	9.1	8.5	8.6	8.5	8.5	8.4	8.5	9.4	8.5	8.5	8.4	8.4	8.4		8.3	6.4	8.4
NA19239	12.3	8.5	9.2	8.6	8.7	8.6	8.6	8.5	8.6	9.5	8.5	8.7	8.6	8.5	8.6	8.4		6.4	8.5
NA19240*	12.4	8.7	9.3	8.7	8.9	8.7	8.8	8.6	8.7	9.6	8.7	8.8	8.7	8.6	8.7	6.6	6.6		8.7
GRCh37	12.0	8.2	9.0	7.2	7.5	7.1	7.4	8.4	7.3	9.2	8.4	7.3	7.0	6.9	7.0	8.1	8.2	8.3	

Table S15: Matrix of all pairwise autosomal branch length ratios (S/R) based on the Human-Chimpanzee ancestor as outgroup. The daughters from the 1000 Genome trios are marked with an asterisk (*).

Reference \ Sample	Denisova	DNK02 (Dinka)	HGDP00456 (Mbuti)	HGDP00521 (French)	HGDP00542 (Papuan)	HGDP00665 (Sardinian)	HGDP00778 (Han)	HGDP00927 (Yoruba)	HGDP00998 (Karitiana)	HGDP01029 (San)	HGDP01284 (Mandenka)	HGDP01307 (Dai)	NA12878* (CEU)	NA12891 (CEU)	NA12892 (CEU)	NA19238 (YRI)	NA19239 (YRI)	NA19240* (YRI)	GRCh37
Denisova		1.10	1.10	1.10	1.10	1.10	1.10	1.11	1.10	1.11	1.10	1.10	1.09	1.09	1.09	1.09	1.09	1.10	1.07
DNK02	0.91		1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.01	1.00	1.00	0.98	0.99	0.98	0.98	0.98	1.00	0.95
HGDP00456	0.91	1.00		1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.00	0.98	0.99	0.98	0.98	0.98	1.00	0.95
HGDP00521	0.91	1.00	1.00		1.00	1.00	1.00	1.01	1.00	1.02	1.00	1.00	0.98	0.98	0.98	0.98	0.98	1.00	0.94
HGDP00542	0.91	1.00	1.00	1.00		1.00	1.00	1.01	1.00	1.01	1.00	1.00	0.98	0.98	0.98	0.98	0.98	1.00	0.94
HGDP00665	0.91	1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.01	1.00	1.00	0.97	0.98	0.98	0.98	0.98	1.00	0.94
HGDP00778	0.91	1.00	1.00	1.00	1.00	1.00		1.01	1.00	1.01	1.00	1.00	0.98	0.98	0.98	0.98	0.98	1.00	0.94
HGDP00927	0.90	0.99	1.00	0.99	0.99	1.00	0.99		0.99	1.01	1.00	0.99	0.98	0.98	0.98	0.97	0.98	0.99	0.94
HGDP00998	0.91	1.00	1.00	1.00	1.00	1.00	1.00	1.01		1.01	1.00	1.00	0.98	0.98	0.98	0.98	0.98	1.00	0.94
HGDP01029	0.90	0.99	0.99	0.98	0.99	0.99	0.99	0.99	0.99		0.99	0.98	0.97	0.98	0.97	0.97	0.97	0.98	0.94
HGDP01284	0.91	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01		1.00	0.98	0.99	0.98	0.98	0.98	1.00	0.95
HGDP01307	0.91	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.02	1.00		0.98	0.99	0.98	0.98	0.98	1.00	0.94
NA12878*	0.92	1.02	1.02	1.02	1.02	1.03	1.02	1.03	1.03	1.03	1.02	1.02		1.01	1.00	1.00	1.00	1.02	0.96
NA12891	0.92	1.01	1.01	1.02	1.02	1.02	1.02	1.02	1.02	1.03	1.01	1.02	0.99		1.00	0.99	1.00	1.01	0.95
NA12892	0.92	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.03	1.02	1.02	1.00	1.00		0.99	1.00	1.01	0.95
NA19238	0.92	1.02	1.02	1.02	1.02	1.03	1.02	1.03	1.02	1.04	1.03	1.02	1.00	1.01	1.01		1.01	1.03	0.96
NA19239	0.92	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.03	1.02	1.02	1.00	1.00	1.00	0.99		1.02	0.96
NA19240*	0.91	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.02	1.00	1.00	0.98	0.99	0.99	0.97	0.98		0.95
GRCh37	0.94	1.06	1.05	1.06	1.06	1.07	1.06	1.06	1.06	1.07	1.06	1.06	1.04	1.05	1.05	1.04	1.04	1.05	

Table S16: Matrix of all pairwise X chromosomal divergence estimates (%). The daughters from the 1000 Genome trios are marked with an asterisk (*).

Reference \ Sample	Denisova	DNK02 (Dinka)	HGDP00456 (Mbuti)	HGDP00521 (French)	HGDP00542 (Papuan)	HGDP00665 (Sardinian)	HGDP00778 (Han)	HGDP00927 (Yoruba)	HGDP00998 (Karitiana)	HGDP01029 (San)	HGDP01284 (Mandenka)	HGDP01307 (Dai)	NA12878* (CEU)	NA12891 (CEU)	NA12892 (CEU)	NA19238 (YRI)	NA19239 (YRI)	NA19240* (YRI)	GRCh37
Denisova		10.9	10.9	10.8	10.8	10.8	10.9	10.8	10.9	10.9	10.9	10.9	10.8	10.8	10.9	10.9	10.8	10.8	10.9
DNK02	12.1		8.0	7.4	7.3	7.4	7.4	7.6	7.5	8.8	7.6	7.3	7.4	7.3	7.4	7.1	7.5	7.2	7.4
HGDP00456	12.1	8.0		8.4	8.2	8.3	8.3	8.2	8.1	8.9	8.5	8.4	8.4	8.4	8.4	8.3	8.5	8.4	8.1
HGDP00521	12.3	7.7	8.7		5.5	5.0	5.5	7.9	5.5	8.9	7.9	5.6	5.0	5.1	4.9	7.7	7.9	7.8	5.6
HGDP00542	12.3	7.6	8.5	5.5		5.4	5.1	7.8	5.0	9.1	7.8	5.0	5.4	5.3	5.5	7.6	7.7	7.6	5.7
HGDP00665	12.5	7.9	8.8	5.2	5.6		5.7	8.0	5.4	9.2	7.9	5.8	5.0	5.1	5.0	7.8	7.7	7.7	5.7
HGDP00778	12.3	7.8	8.7	5.5	5.2	5.6		7.9	4.7	9.0	8.0	4.7	5.4	5.4	5.4	7.7	7.7	7.7	5.9
HGDP00927	12.1	7.7	8.4	7.7	7.6	7.7	7.7		7.4	8.9	7.8	7.7	7.6	7.5	7.7	7.5	7.5	7.5	7.5
HGDP00998	12.5	7.9	8.6	5.7	5.2	5.4	4.9	7.8		9.1	7.9	5.3	5.5	5.7	5.5	7.8	7.8	7.8	5.9
HGDP01029	12.4	9.1	9.3	9.0	9.1	9.1	9.1	9.2	9.0		9.1	9.2	9.1	9.0	9.2	9.2	9.3	9.3	9.1
HGDP01284	12.5	8.1	9.0	8.0	7.9	7.9	8.1	8.1	7.9	9.2		7.9	7.9	7.8	7.9	7.7	8.2	7.9	7.9
HGDP01307	12.2	7.5	8.6	5.5	4.9	5.5	4.6	7.8	5.0	9.0	7.7		5.4	5.3	5.3	7.5	7.6	7.5	5.7
NA12878*	12.0	7.4	8.4	4.6	5.0	4.5	5.0	7.5	4.9	8.7	7.5	5.1		2.6	3.5	7.2	7.3	7.2	5.1
NA12891	12.1	7.4	8.6	4.9	5.2	4.7	5.2	7.5	5.3	8.8	7.6	5.2	2.8		4.7	7.3	7.4	7.3	5.2
NA12892	12.1	7.4	8.4	4.6	5.2	4.6	5.0	7.6	5.0	8.8	7.5	5.1	3.6	4.5		7.3	7.3	7.3	5.1
NA19238	11.8	6.9	8.1	7.2	7.1	7.1	7.2	7.1	7.0	8.6	7.1	7.0	7.0	6.9	7.0		7.1	5.4	7.0
NA19239	12.0	7.6	8.6	7.6	7.4	7.3	7.5	7.5	7.4	9.0	7.8	7.4	7.4	7.3	7.4	7.4		4.1	7.5
NA19240*	12.0	7.2	8.4	7.4	7.3	7.2	7.3	7.3	7.3	8.8	7.4	7.2	7.2	7.1	7.3	5.6	4.0		7.3
GRCh37	11.9	7.2	8.0	5.1	5.2	5.1	5.4	7.2	5.2	8.6	7.3	5.3	4.9	4.8	4.8	7.0	7.2	7.2	

Table S17: Matrix of all pairwise X chromosomal branch length ratios (S/R). The daughters from the 1000 Genome trios are marked with an asterisk (*).

Reference \ Sample	Denisova	DNK02 (Dinka)	HGDP00456 (Mbuti)	HGDP00521 (French)	HGDP00542 (Papuan)	HGDP00665 (Sardinian)	HGDP00778 (Han)	HGDP00927 (Yoruba)	HGDP00998 (Karitiana)	HGDP01029 (San)	HGDP01284 (Mandenka)	HGDP01307 (Dai)	NA12878* (CEU)	NA12891 (CEU)	NA12892 (CEU)	NA19238 (YRI)	NA19239 (YRI)	NA19240* (YRI)	GRCh37
Denisova		1.13	1.12	1.15	1.16	1.17	1.15	1.13	1.17	1.16	1.17	1.14	1.12	1.14	1.13	1.10	1.13	1.12	1.11
DNK02	0.89		1.00	1.04	1.05	1.07	1.05	1.01	1.07	1.05	1.06	1.03	0.99	1.02	1.00	0.96	1.01	0.99	0.98
HGDP00456	0.89	1.00		1.04	1.05	1.07	1.05	1.02	1.07	1.05	1.06	1.03	1.00	1.02	1.01	0.97	1.01	1.00	0.99
HGDP00521	0.87	0.96	0.96		1.00	1.04	1.01	0.97	1.04	1.01	1.02	0.98	0.92	0.97	0.93	0.93	0.97	0.95	0.91
HGDP00542	0.87	0.96	0.96	1.00		1.03	1.00	0.97	1.04	1.01	1.02	0.98	0.93	0.97	0.94	0.92	0.97	0.95	0.91
HGDP00665	0.85	0.94	0.94	0.96	0.97		0.97	0.95	1.01	0.99	0.99	0.95	0.89	0.93	0.90	0.90	0.94	0.93	0.89
HGDP00778	0.87	0.95	0.95	0.99	1.00	1.03		0.97	1.04	1.01	1.02	0.97	0.92	0.97	0.93	0.92	0.96	0.95	0.92
HGDP00927	0.88	0.99	0.98	1.03	1.03	1.05	1.03		1.06	1.03	1.04	1.01	0.98	1.01	0.99	0.95	0.99	0.97	0.96
HGDP00998	0.85	0.94	0.94	0.96	0.96	0.99	0.96	0.95		0.99	0.99	0.94	0.89	0.93	0.91	0.90	0.94	0.93	0.88
HGDP01029	0.86	0.96	0.95	0.99	0.99	1.01	0.99	0.97	1.01		1.01	0.98	0.95	0.97	0.95	0.93	0.96	0.95	0.94
HGDP01284	0.86	0.94	0.94	0.98	0.98	1.01	0.98	0.96	1.01	0.99		0.97	0.94	0.96	0.94	0.91	0.95	0.93	0.92
HGDP01307	0.88	0.97	0.97	1.02	1.02	1.05	1.03	0.99	1.06	1.02	1.03		0.95	0.99	0.96	0.94	0.98	0.96	0.93
NA12878*	0.89	1.01	1.00	1.08	1.08	1.13	1.08	1.02	1.12	1.05	1.07	1.05		1.08	1.02	0.97	1.02	1.00	0.97
NA12891	0.88	0.98	0.98	1.03	1.04	1.08	1.04	0.99	1.07	1.03	1.04	1.01	0.93		0.97	0.95	0.99	0.97	0.93
NA12892	0.89	1.00	0.99	1.07	1.06	1.11	1.07	1.01	1.10	1.05	1.06	1.04	0.98	1.03		0.97	1.01	0.99	0.95
NA19238	0.91	1.04	1.03	1.08	1.08	1.11	1.08	1.05	1.11	1.08	1.10	1.06	1.03	1.06	1.04		1.05	1.04	1.00
NA19239	0.88	0.99	0.99	1.03	1.03	1.06	1.04	1.01	1.06	1.04	1.05	1.02	0.98	1.01	0.99	0.96		0.97	0.96
NA19240*	0.89	1.01	1.00	1.05	1.05	1.08	1.05	1.03	1.08	1.05	1.07	1.04	1.00	1.03	1.01	0.96	1.03		0.97
GRCh37	0.90	1.02	1.01	1.09	1.10	1.13	1.09	1.04	1.14	1.07	1.09	1.07	1.03	1.08	1.05	1.00	1.04	1.03	

Table S18: Number of pair-wise differences in one million bases as determined on the autosomes. Values above the diagonal refer to the number of transversion differences in one million base pairs, numbers below the diagonal refer to the number of transitions. The daughters from the 1000 Genome trios are marked with an asterisk (*).

transitions \ transversions	Denisova	DNK02	HGDP00456	HGDP00521	HGDP00542	HGDP00665	HGDP00778	HGDP00927	HGDP00998	HGDP01029	HGDP01284	HGDP01307	NA12878	NA12891	NA12892	NA19238	NA19239	NA19240	GRCh37
Denisova		507	510	507	498	506	505	511	506	514	508	505	507	508	508	509	511	516	492
DNK02	1018		409	379	384	378	380	388	381	426	386	381	383	383	385	388	391	396	371
HGDP00456	1021	778		413	417	412	414	416	415	431	416	415	419	419	421	417	421	426	405
HGDP00521	1019	720	790		328	299	320	391	315	426	388	321	304	303	305	390	395	401	324
HGDP00542	1000	729	796	616		327	314	397	316	430	394	315	333	332	335	395	401	405	341
HGDP00665	1018	718	789	559	616		319	389	318	425	388	320	302	303	304	388	394	399	324
HGDP00778	1014	719	789	601	587	599		392	296	426	389	285	324	324	326	391	397	401	335
HGDP00927	1021	732	788	739	750	738	741		392	432	389	393	397	398	398	389	394	397	379
HGDP00998	1017	723	792	591	591	596	549	742		428	391	298	319	319	320	390	397	401	331
HGDP01029	1022	807	814	809	816	809	809	815	812		431	427	433	433	434	435	438	443	419
HGDP01284	1019	733	791	739	749	738	739	732	742	816		390	394	394	395	389	392	398	377
HGDP01307	1014	721	790	601	588	602	526	742	552	810	739		326	328	328	391	398	403	333
NA12878	1015	718	791	559	617	557	598	742	589	810	742	602		240	240	392	400	405	326
NA12891	1016	717	791	556	615	557	599	741	587	810	740	603	421		305	392	399	405	324
NA12892	1019	722	795	562	621	561	602	744	591	814	744	606	424	556		391	400	406	323
NA19238	1013	723	782	729	740	727	730	720	731	811	726	730	729	728	728		390	313	368
NA19239	1018	731	790	740	750	737	741	728	743	816	731	742	742	740	743	719		314	379
NA19240	1021	733	791	742	752	740	742	730	743	819	735	744	745	743	747	552	554		388
GRCh37	1010	719	785	627	660	626	645	730	638	806	731	642	622	618	621	710	725	734	

Table S19: Number of pair-wise differences in one million bases as determined on the X chromosome. Values above the diagonal refer to the number of transversion differences in one million base pairs, numbers below the diagonal refer to the number of transitions. The daughters from the 1000 Genome trios are marked with an asterisk (*).

transversions / transitions	Denisova	DNK02 (Dinka)	HGDP00456 (Mbuti)	HGDP00521 (French)	HGDP00542 (Papuan)	HGDP00665 (Sardinian)	HGDP00778 (Han)	HGDP00927 (Yoruba)	HGDP00998 (Karitiana)	HGDP01029 (San)	HGDP01284 (Mandenka)	HGDP01307 (Dai)	NA12878* (CEU)	NA12891 (CEU)	NA12892 (CEU)	NA19238 (YRI)	NA19239 (YRI)	NA19240* (YRI)	GRCh37
Denisova		393	392	393	393	396	394	393	400	403	404	392	392	395	394	393	398	398	387
DNK02	757		287	269	264	271	272	280	277	322	285	263	265	268	270	256	280	267	260
HGDP00456	756	510		305	299	306	306	303	302	329	316	304	304	310	303	299	313	311	287
HGDP00521	766	491	552		202	186	203	281	207	323	288	200	182	189	179	274	283	282	192
HGDP00542	765	481	542	361		199	190	277	189	326	283	186	196	199	200	270	281	279	198
HGDP00665	770	497	553	330	355		206	282	198	330	285	203	178	184	179	272	276	276	193
HGDP00778	762	488	548	355	329	363		281	180	326	293	171	197	203	196	276	282	279	203
HGDP00927	757	498	541	510	499	511	502		276	328	293	281	276	275	282	273	279	278	264
HGDP00998	769	495	539	361	327	345	307	492		330	290	190	199	210	199	274	282	281	202
HGDP01029	763	580	585	583	590	596	582	586	588		334	327	324	328	327	327	340	336	317
HGDP01284	767	503	562	514	504	508	517	513	505	589		283	283	283	283	276	300	288	274
HGDP01307	761	477	551	357	320	362	292	503	326	586	500		193	197	192	266	278	275	199
NA12878*	754	475	539	310	335	305	327	488	334	575	495	332		112	142	267	275	272	186
NA12891	754	473	544	321	333	312	333	480	344	572	488	328	160		178	267	275	275	187
NA12892	757	479	543	302	343	306	329	499	336	579	496	332	221	288		266	276	274	180
NA19238	744	449	520	480	476	474	478	470	477	570	472	465	456	452	460		274	217	252
NA19239	747	484	541	498	482	482	486	483	482	585	510	477	467	465	474	460		167	267
NA19240*	749	462	537	491	482	482	481	475	482	579	485	475	464	463	471	347	240		270
GRCh37	760	480	528	352	361	355	364	481	361	579	494	360	329	330	324	453	475	474	

Table S20: Probability of a derived allele at a Yoruba heterozygous transversion

	Green	SNP array*			Deep genome sequences			
	2010	AT→GC	GC→AT	All	AT→GC	GC→AT	A↔T / C↔G	All
Yoruba	30.6%	31.4%	27.8%	29.3%				
Mandenka		31.2%	27.8%	29.2%	30.9%	28.4%	28.9%	29.3%
Dinka		31.1%	27.2%	28.8%	30.8%	28.4%	28.8%	29.2%
Sardinian		30.1%	26.5%	28.0%	30.6%	28.0%	28.5%	28.9%
French	29.7%	30.1%	26.4%	27.9%	30.5%	27.9%	28.4%	28.8%
Han	29.8%	29.8%	26.5%	27.9%	30.6%	27.9%	28.5%	28.9%
Dai		29.8%	26.5%	27.9%	30.6%	27.8%	28.5%	28.9%
Karitiana		29.8%	26.4%	27.8%	30.7%	27.8%	28.5%	28.9%
Papuan	29.3%	29.8%	26.0%	27.6%	30.2%	27.5%	28.1%	28.5%
Mbuti		28.5%	24.6%	26.2%	28.8%	25.9%	26.4%	26.9%
San	26.3%	27.3%	23.4%	25.0%	27.9%	25.0%	25.3%	25.9%
Neandertal	18.0%	17.5%	14.5%	15.8%	16.2%	17.3%	17.5%	17.1%
Denisova		17.4%	14.2%	15.5%	15.6%	15.9%	16.6%	16.1%

* The SNP array data is obtained by genotyping the samples in the CEPH-HGDP cell line panel using the Affymetrix Axiom® Human Origins SNP array. This array was developed for population genetic analyses. It includes SNPs that were discovered as heterozygous in sequencing reads from a single Yoruba individual, and where there is coverage from the draft genome sequences of Denisova and Neandertal, allowing us to report an allele for these archaic hominins.

Table S21: Difference between Yoruba derived probability and that in a test population

	Green	SNP array			Deep genome sequences			
	2010	AT→GC	GC→AT	All	AT→GC	GC→AT	A↔T / C↔G	All
Mandenka		-0.2%	0.0%	-0.1%	-0.1%	-0.1%	-0.1%	-0.1%
Dinka		-0.3%	-0.6%	-0.5%	-0.2%	-0.1%	-0.1%	-0.2%
Sardinian		-1.3%	-1.3%	-1.3%	-0.4%	-0.5%	-0.5%	-0.5%
French	-0.9%	-1.3%	-1.4%	-1.4%	-0.5%	-0.6%	-0.6%	-0.6%
Han	-0.8%	-1.6%	-1.3%	-1.4%	-0.4%	-0.5%	-0.5%	-0.5%
Dai		-1.6%	-1.3%	-1.4%	-0.4%	-0.7%	-0.5%	-0.5%
Karitiana		-1.6%	-1.4%	-1.5%	-0.3%	-0.6%	-0.5%	-0.5%
Papuan	-1.3%	-1.6%	-1.8%	-1.7%	-0.8%	-1.0%	-0.9%	-0.9%
Mbuti		-2.9%	-3.2%	-3.1%	-2.2%	-2.5%	-2.6%	-2.5%
San	-4.3%	-4.1%	-4.4%	-4.3%	-3.1%	-3.5%	-3.7%	-3.5%
Neandertal	-12.6%	-13.9%	-13.3%	-13.5%	-14.8%	-11.2%	-11.5%	-12.3%
Denisova		-14.0%	-13.6%	-13.8%	-15.4%	-12.6%	-12.4%	-13.2%

Notes: The table reports the difference between the value for Yoruba and each test population in Table S20. For the deep genome sequences, we do not have data from a second Yoruba, so we calibrate to the derived allele rate in Mandenka (assuming that its derived allele rate is -0.1% below Yoruba as in the SNP array data; the uncertainty in the true value of this number is far smaller than the 1.7% uncertainty (from -11.5% to -13.2%) in the probability that archaic samples carry the derived allele, so we do not directly account for it in our calculations).

Table S22: Inferred divergence of archaic humans from Yoruba as % of human-chimp

	Reduction vs. Yoruba	As % of Yoruba TMRCA			Range	Percent of human-chimpanzee divergence*
		Keinan	Wall	Li & Durbin		
Bound 1	-11.5%	51.4%	48.0%	35.7%	35.7-51.4%	3.13-4.51%
Bound 2	-13.2%	60.4%	56.2%	46.0%	46.0-60.4%	4.03-5.30%
Combined					35.7-60.4%	3.13-5.30%

Note: This table is based on Yoruba SNPs discovered by deep genome sequencing, from the two final columns of Table S21. The bounds we report correspond to the highest and lowest values inferred, for the union of Denisova and Neandertal.

Table S23: Estimated divergence time of humans and chimps for three clock calibrations

Calibration used	Mutation rate per year per bp	Implied human-chimp genetic divergence (Mya)	Implied Yoruba time to most recent ancestor (kya)	Denisova and Neandertal split date (kya)
Human-chimp genetic divergence = 5.6 Mya	1.22×10^{-9}	5.6	491	175-297
Human-chimp genetic divergence = 8.3 Mya	0.83×10^{-9}	8.3	728	260-440
$\mu=0.5\times10^{-9}$/ year and Yoruba heterozygosity = 0.00104	0.50×10^{-9}	13.1	1,149	410-694

Table S24: *D(H_1, H_2, Denisova, Chimp)* for all possible pairs of samples

Two humans analyzed		Computation using reads			Computation using genotypes		
H_1	H_2	D-stat	Std. err.	Z-score	D-stat	Std. err.	Z-score
African / African							
San	Dinka	-0.7%	0.4%	-1.7	-0.9%	0.4%	-2.4
San	Mandenka	-0.2%	0.4%	-0.5	-0.8%	0.4%	-2.2
San	Mbuti	0.0%	0.4%	0.0	-0.1%	0.4%	-0.2
San	Yoruba	-0.3%	0.4%	-0.7	-0.5%	0.4%	-1.2
Yoruba	Dinka	-0.4%	0.4%	-1.0	-0.5%	0.4%	-1.2
Yoruba	Mandenka	-0.1%	0.4%	-0.2	-0.4%	0.4%	-1.2
Yoruba	Mbuti	0.2%	0.4%	0.4	0.3%	0.4%	0.9
Mandenka	Dinka	-0.4%	0.4%	-1.1	0.0%	0.4%	0.1
Mandenka	Mbuti	0.2%	0.4%	0.6	0.8%	0.3%	2.3
Mbuti	Dinka	-0.7%	0.4%	-1.7	-0.8%	0.4%	-2.2
Non-African / Non-Afr.							
Papuan	Dai	6.1%	0.6%	9.8	6.0%	0.6%	10.1
Papuan	Han	6.8%	0.6%	10.9	6.4%	0.6%	10.6
Papuan	French	7.6%	0.6%	12.5	6.9%	0.6%	12.2
Papuan	Karitiana	6.9%	0.6%	11.2	6.4%	0.6%	11.0
Papuan	Sardinian	7.1%	0.6%	11.7	6.5%	0.6%	11.7
French	Dai	-1.7%	0.4%	-4.0	-1.2%	0.4%	-2.9
French	Han	-0.9%	0.5%	-1.9	-0.7%	0.5%	-1.6
French	Karitiana	-0.9%	0.5%	-1.7	-0.7%	0.5%	-1.6
French	Sardinian	-0.4%	0.5%	-1.0	-0.4%	0.4%	-1.0
Sardinian	Dai	-1.3%	0.5%	-2.7	-0.8%	0.4%	-1.9
Sardinian	Han	-0.6%	0.5%	-1.2	-0.4%	0.5%	-1.0
Sardinian	Karitiana	-0.4%	0.5%	-0.8	-0.4%	0.5%	-0.8
Dai	Han	0.8%	0.5%	1.7	0.5%	0.4%	1.2
Dai	Karitiana	1.0%	0.5%	2.1	0.5%	0.4%	1.2
Han	Karitiana	0.2%	0.5%	0.5	0.1%	0.5%	0.1
Non-African / African							
Papuan	Dinka	7.0%	0.6%	12.8	6.9%	0.5%	13.5
Papuan	Mandenka	7.1%	0.6%	12.8	6.7%	0.5%	13.2
Papuan	Mbuti	7.0%	0.5%	13.3	7.1%	0.5%	13.6
Papuan	San	6.8%	0.5%	12.9	7.0%	0.5%	14.3
Papuan	Yoruba	7.0%	0.5%	13.0	7.1%	0.5%	14.0
Dai	Dinka	2.0%	0.4%	5.0	2.1%	0.4%	5.5
Dai	Mandenka	2.4%	0.4%	5.6	2.0%	0.4%	5.0
Dai	Mbuti	2.3%	0.4%	5.5	2.6%	0.4%	6.8
Dai	San	2.5%	0.4%	6.0	2.7%	0.4%	6.8
Dai	Yoruba	2.4%	0.4%	6.0	2.4%	0.4%	6.3
Han	Dinka	1.4%	0.4%	3.3	1.7%	0.4%	4.0
Han	Mandenka	1.8%	0.4%	4.0	1.6%	0.4%	3.9
Han	Mbuti	1.9%	0.4%	4.4	2.3%	0.4%	5.7
Han	San	1.9%	0.4%	4.4	2.3%	0.4%	5.6
Han	Yoruba	1.8%	0.4%	4.1	2.0%	0.4%	4.7
Karitiana	Dinka	1.2%	0.4%	2.8	1.6%	0.4%	3.7
Karitiana	Mandenka	1.6%	0.4%	3.8	1.5%	0.4%	3.8
Karitiana	Mbuti	1.7%	0.4%	4.0	2.3%	0.4%	5.8
Karitiana	San	1.7%	0.4%	3.8	2.3%	0.4%	5.3
Karitiana	Yoruba	1.6%	0.4%	3.7	2.0%	0.4%	4.8
French	Dinka	0.6%	0.4%	1.4	1.1%	0.4%	2.7
French	Mandenka	1.0%	0.4%	2.5	1.0%	0.4%	2.5
French	Mbuti	1.1%	0.4%	2.6	1.7%	0.4%	4.3
French	San	1.2%	0.4%	2.7	1.8%	0.4%	4.6
French	Yoruba	1.0%	0.4%	2.6	1.5%	0.4%	3.9
Sardinian	Dinka	0.9%	0.4%	2.1	1.4%	0.4%	3.6
Sardinian	Mandenka	1.2%	0.4%	2.9	1.3%	0.4%	3.5
Sardinian	Mbuti	1.4%	0.4%	3.3	2.0%	0.4%	5.1
Sardinian	San	1.3%	0.4%	2.9	2.1%	0.4%	5.0
Sardinian	Yoruba	1.2%	0.4%	2.8	1.8%	0.4%	4.7

Notes: Computations are restricted to the autosomes. We highlight highly significant scores (|Z|>4 standard errors from 0) in red. Estimates are highly correlated whether we use reads or genotypes.

Table S25: Enhanced D-statistics document less archaic ancestry in Europeans than in Eastern non-Africans

		$D_{basic}(H_1, H_2, Denisova, Chimpanzee)$					$D_{enhanced}(H_1, H_2, Denisova, Chimpanzee)$				
H_1	H_2	n_{BABA}	n_{ABBA}	D	Err.	Z	n_{BABA}	n_{ABBA}	D-stat	Err.	Z
Papuan – other											
Papuan	Dai	152,871	135,656	6.0%	0.6%	10.1	11,907	3,668	52.9%	2.2%	23.6
Papuan	Han	153,784	135,374	6.4%	0.6%	10.6	11,873	3,782	51.7%	2.3%	22.2
Papuan	Karitiana	153,140	134,610	6.4%	0.6%	11.0	11,832	3,316	56.2%	2.1%	26.3
Papuan	French	159,665	139,186	6.9%	0.6%	12.2	12,032	3,080	59.2%	2.1%	27.8
Papuan	Sardinian	159,001	139,490	6.5%	0.6%	11.7	12,021	2,928	60.8%	2.1%	29.5
Papuan	3 East	159,755	141,431	6.1%	0.5%	11.6	12,051	3,668	53.3%	2.1%	25.8
Papuan	2 Europe	164,984	144,656	6.6%	0.5%	12.5	12,230	3,071	59.9%	1.9%	31.2
Papuan	5 non-African	163,193	144,023	6.2%	0.5%	12.4	12,143	3,439	55.9%	1.9%	29.6
East – West											
Dai	French	144,447	140,950	1.2%	0.4%	2.9	3,877	3,115	10.9%	2.7%	4.0
Dai	Sardinian	143,814	141,414	0.8%	0.4%	1.9	3,932	3,018	13.2%	2.7%	4.9
Dai	2 Europe	149,704	146,675	1.0%	0.4%	2.7	3,984	3,137	11.9%	2.5%	4.8
Han	French	144,004	141,884	0.7%	0.5%	1.6	4,075	3,148	12.8%	2.7%	4.7
Han	Sardinian	143,124	141,895	0.4%	0.5%	1.0	4,049	2,956	15.6%	2.8%	5.7
Han	2 Europe	149,031	147,398	0.6%	0.4%	1.4	4,143	3,124	14.0%	2.6%	5.5
Karitiana	French	140,197	138,172	0.7%	0.5%	1.6	3,547	3,095	6.8%	2.8%	2.5
Karitiana	Sardinian	140,823	139,828	0.4%	0.5%	0.8	3,506	2,895	9.6%	2.7%	3.5
Karitiana	2 Europe	145,908	144,383	0.5%	0.4%	1.3	3,601	3,061	8.11%	2.5%	3.3
3 East	French	149,417	146,778	0.9%	0.4%	2.4	3,921	3,190	10.3%	2.5%	4.1
3 East	Sardinian	148,978	147,400	0.5%	0.4%	1.5	3,915	3,024	12.8%	2.5%	5.2
3 East	2 Europe	156,594	154,449	0.7%	0.3%	2.2	4,011	3,191	11.4%	2.2%	5.3
East – East											
Dai	Han	127,862	126,561	0.5%	0.4%	1.2	3,448	3,638	-2.7%	2.6%	-1.0
Dai	Karitiana	132,541	131,203	0.5%	0.4%	1.2	3,646	3,355	4.2%	2.6%	1.6
Han	Karitiana	131,450	131,300	0.1%	0.5%	0.1	3,782	3,306	6.7%	2.7%	2.5
West - West											
French	Sardinian	131,485	132,569	-0.4%	0.4%	-1.0	2,982	2,812	2.9%	2.6%	1.1

Note: To increase resolution to study differences in archaic ancestry across populations, we not only analyze single samples, but also pools of "2 Europe" (Sardinian+French), "3 East" (Dai+Han+Karitiana) and "5 non-African" (2 Europe + 3 East). We highlight very significant signals in red (|Z|>4 standard errors from 0).

Table S26: Estimates of archaic ancestry

Quantity we are estimating	Statistic used	As fraction of the reference population		As an absolute proportion of ancestry*	
		Basic estimate (95% C.I.)	**Enhanced estimate (95% C.I.)**	**Basic estimate (95% C.I.)**	**Enhanced estimate (95% C.I.)**
Excess Denisova ancestry in Dai as % of Papuans (Han baseline)	$\frac{S(Dai, Han; Denisova, Chimp)}{S(Papuan, Han; Denisova, Chimp)}$	-0.6% (-10.7% to 9.5%)	-3.5% (-9.5% to 2.5%)	-0.03% (-0.53% to 0.48%)	-0.18% (-0.48% to 0.12%)
Excess Neandertal ancestry in French as % of Han (Yoruba baseline)	$\frac{S(French, Yoruba; Neandertal, Chimp)}{S(Han, Yoruba; Neandertal, Chimp)}$	67% (51% to 83%)	n/a	1.67% (1.27% to 2.07%)	n/a
Excess Neandertal ancestry in Europe as % of East (Africa baseline)	$\frac{S(Europe, Africa; Neandertal, Chimp)}{S(East, Africa; Neandertal, Chimp)}$	76% (64% to 88%)	n/a	1.90% (1.59 to 2.21%)	n/a

Note: To represent "Africa" we use a pool of Mandenka+Yoruba+Dinka+Mbuti; to represent "Europe" we use a pool of Sardinian+French; and to represent "East" we use a pool of Dai+Han+Karitiana. We only compute an enhanced statistic (which requires that all African samples carry the ancestral allele) for the estimate of the excess Denisovan ancestry, since this is the only computation that does not directly involve analysis of African samples.

* The *S*-statistic ratio gives us an estimate of archaic ancestry proportion as a fraction of that in reference population known to have archaic introgression. To convert these numbers to an absolute estimate of archaic ancestry proportion (final two columns), we assume that Papuans have 5% Denisovan ancestry (for the top row of the table), and that East Asians have 2.5% Neandertal ancestry (for the last two rows of the table).

Table S27: Tests for whether gene flows are from populations more closely related to Neandertals or to Denisovans

		H_1 = Europe **H_2 = Africa**	**H_1 = East** **H_2 = Europe**	**H_1 = Papuan** **H_2 = East**
Basic	$S_{Denisova}$	491 (±157)	512 (±146)	2921 (±248)
	$S_{Neandertal}$	2677 (±228)	837 (±257)	1434 (±326)
	Z-score for $S_{Denisova}$-$S_{Neandertal}$)	-13.1	-1.7	5.8
Enhanced	$S_{Denisova}$	n/a	204 (±40)	1664 (±98)
	$S_{Neandertal}$	n/a	554 (±119)	799 (±159)
	Z-score for ($S_{Denisova}$-$S_{Neandertal}$)	n/a	-3.6	5.8
	Comments	$S_{Neandertal} > S_{Denisova}$ (as expected from known Neandertal introgression)	$S_{Neandertal} > S_{Denisova}$ (excess archaic ancestry in eastern non-Africans shares more derived alleles with Neandertal than Denisova)	$S_{Denisova} > S_{Neandertal}$ (as expected from known Denisova introgression)

Note: To represent "Africa" we use a pool of Mandenka+Yoruba+Dinka+Mbuti; to represent "Europe" we use a pool of Sardinian+French; and to represent "East" we use a pool of Dai+Han+Karitiana. The Z-score is the number of standard errors by which $\Delta S = S_{Denisova}$-$S_{Neandertal}$ differs from zero (a positive score indicates greater relatedness to Denisova and negative greater relatedness to Neandertal). For the "Enhanced" *S*-statistic, we restrict to sites where >99% of reads from 35 sub-Saharan Africans carry the ancestral allele, thus increasing the chance that these are variants that introgressed from archaic populations (amplifying our signal). This invalidates the analysis when H_1 = Europe and H_2 = Africa, and so this computation is not performed.

Table S28: Comparison of archaic ancestry estimates on the autosomes and chromosome X

		Autosomes		Chromosome X		Autosomes - X		
H_1	H_2	Anc. Est.	Std. Err.	Anc. Est.	Std. Err.	Anc. Diff.	Std. Err.	Z-score
Nea(Europe, Africa): Excess Neandertal ancestry *(no enhancement is used in the ancestry estimation because Africans are involved in the computation)*								
Europe	Africa	1.0%	0.3%	0.3%	1.5%	0.7%	1.5%	0.5
Sardinian	Yoruba	1.2%	0.4%	1.7%	2.6%	-0.5%	2.7%	-0.2
Sardinian	Dinka	1.6%	0.4%	1.0%	2.2%	0.6%	2.2%	0.3
Sardinian	Mandenka	0.4%	0.4%	5.0%	2.1%	-4.6%	2.1%	-2.1
Sardinian	Mbuti	1.3%	0.4%	-0.6%	2.3%	1.9%	2.3%	0.8
French	Yoruba	1.0%	0.4%	-1.6%	2.7%	2.5%	2.7%	0.9
French	Dinka	1.3%	0.4%	-1.7%	2.2%	3.0%	2.3%	1.3
French	Mandenka	0.2%	0.4%	1.9%	2.4%	-1.7%	2.5%	-0.7
French	Mbuti	1.1%	0.4%	-3.2%	2.1%	4.3%	2.1%	2.0
Nea(East, Europe): **Excess Neandertal ancestry** *(we enhance the power of the ancestry estimation by requiring that all reads from 35 Africans are ancestral)*								
East	Europe	0.7%	0.2%	0.4%	0.4%	0.4%	0.5%	0.8
Han	Sardinian	1.0%	0.3%	0.7%	0.6%	0.3%	0.6%	0.5
Han	French	0.7%	0.2%	0.8%	0.5%	0.0%	0.5%	0.0
Dai	Sardinian	0.8%	0.2%	0.5%	0.6%	0.3%	0.6%	0.4
Dai	French	0.6%	0.2%	0.5%	0.5%	0.1%	0.5%	0.2
Karitiana	French	0.7%	0.2%	-0.2%	0.6%	0.9%	0.7%	1.3
Karitiana	Sardinian	0.4%	0.3%	-0.2%	0.7%	0.6%	0.7%	0.8
Den(Papuan, East): **Excess Denisova ancestry** *(we enhance the power of the ancestry estimation by requiring that all reads from 35 Africans are ancestral)*								
Papuan	East	3.0%	0.8%	0.0%	0.9%	3.1%	1.2%	2.6
Papuan	Han	2.0%	0.9%	-1.2%	1.5%	3.2%	1.7%	1.8
Papuan	Dai	3.3%	0.8%	0.3%	0.9%	3.0%	1.2%	2.4
Papuan	Karitiana	3.8%	0.8%	0.8%	0.7%	3.1%	1.0%	3.0

Note: The first line in each section analyzes pools of samples. To represent "Africa" we use Mandenka+Yoruba+Dinka+Mbuti; to represent "Europe" we Sardinian+French; and to represent "East" Dai+Han+Karitiana.

Table S29: Ancestry estimates show no clear effect of proximity to genes as measured by McVicker's *B*-statistic

B-statistic (larger means less selection)	Europe -Africa (Neandertal excess)				East-Europe (Neandertal excess)				Papuan-East (Denisova excess)			
	Autosomes		Chromosome X		Autosomes		Chromosome X		Autosomes		Chromosome X	
	Est.	Std. Err.	Est.	Std. Err.	Est.	Std. Err.	Est.	Std. Err.	Est.	Std. Err.	Est.	Std. Err.
All data	1.0%	0.3%	0.3%	1.5%	0.7%	0.2%	0.4%	0.4%	3.0%	0.8%	0.0%	0.9%
B<0.4	1.0%	0.9%	-1.2%	2.5%	0.9%	0.3%	0.7%	0.6%	6.0%	1.6%	0.4%	1.4%
0.4<B<0.5	0.7%	1.1%	-2.6%	4.0%	0.5%	0.4%	-0.1%	0.1%	4.7%	1.3%	-0.1%	0.2%
0.5<B<0.6	3.4%	1.0%	5.2%	5.4%	1.5%	0.4%	-3.2%	2.3%	1.9%	1.2%	0.7%	0.5%
0.6<B<0.7	-0.8%	0.8%	-2.6%	2.9%	0.8%	0.3%	0.1%	0.7%	2.6%	1.2%	-0.3%	1.5%
0.7<B<0.8	1.6%	0.7%	1.8%	3.6%	0.0%	0.3%	-1.2%	1.1%	5.8%	1.1%	5.1%	3.6%
0.8<B<0.9	1.5%	0.5%	1.7%	4.6%	0.6%	0.3%	1.9%	1.1%	3.0%	1.0%	-2.6%	1.5%
0.9<B<1	0.6%	0.4%	0.6%	3.1%	0.9%	0.2%	1.1%	1.0%	1.8%	0.9%	-1.4%	1.7%

Note: "Africa" is a pool of Yoruba+Mandenka+Dinka+Mbuti, "Europe" is a pool of Sardinian+French, and "East" is a pool of Dai+Han+Karitiana. For the Europe-African comparison, we do not restrict to sites that are ancestral in Africans because Africans are directly involved in the comparison. For the other comparisons, we enhance sensitivity by restricting to sites where >99% of reads from 35 African samples carry the ancestral allele.

Table S30: Ancestry estimates and *D*-statistics by chromosome

Chr.	Europe-Africa (Neandertal excess)			East-Europe (Neandertal excess)				Papuan-East (Denisova excess)			
	Anc. Est.	Std. Err.	D_{basic} (Z)	Anc. Est.	Std. Err.	D_{basic} (Z)	$D_{enhanced}$ (Z)	Anc. Est.	Std. Err.	D_{basic} (Z)	$D_{enhanced}$ (Z)
1	1.7%	0.9%	1.1	0.8%	0.5%	-0.3	1.4	5.2%	3.1%	2.5	7.1
2a	0.1%	1.3%	0.1	0.4%	0.4%	0.5	1.9	-0.9%	1.6%	1.3	2.1
2b	0.3%	1.6%	0.9	1.6%	0.9%	1.0	2.3	1.8%	3.0%	0.5	3.2
3	1.0%	1.0%	0.3	-0.6%	0.5%	0.5	-1.4	4.7%	2.5%	4.8	18.6
4	-0.2%	1.2%	0.8	0.9%	0.8%	1.7	1.8	0.9%	2.9%	2.0	1.8
5	3.3%	1.2%	2.9	0.6%	0.8%	-0.2	1.9	6.3%	2.8%	4.2	9.6
6	1.1%	0.9%	1.2	1.5%	0.6%	1.4	1.8	3.0%	3.4%	4.1	10.6
7	-0.2%	1.1%	1.2	0.8%	0.7%	-0.9	1.8	1.8%	2.1%	2.7	4.9
8	0.9%	1.2%	0.5	1.0%	0.5%	1.3	0.7	-1.0%	1.5%	2.8	5.6
9	1.1%	1.3%	1.5	1.3%	0.8%	1.5	2.2	3.0%	4.0%	1.6	4.2
10	1.3%	1.6%	2.5	0.6%	0.7%	-1.4	0.7	-0.9%	2.7%	2.3	1.8
11	0.5%	1.5%	1.4	1.1%	0.6%	1.4	2.8	-4.9%	1.9%	-1.1	0.4
12	0.4%	1.3%	1.1	1.5%	1.0%	2.2	1.3	2.5%	4.0%	1.4	7.2
13	0.3%	1.5%	0.7	0.9%	0.7%	0.9	1.2	-1.1%	2.6%	2.4	3.5
14	1.5%	2.1%	1.2	-0.5%	0.9%	-0.4	-0.8	16.5%	5.4%	3.3	16.5
15	1.0%	1.8%	2.9	0.8%	0.5%	-0.8	3.0	11.7%	4.8%	4.2	15.3
16	0.2%	1.7%	3.3	0.1%	0.6%	-1.0	1.5	1.0%	3.4%	4.0	8.4
17	2.7%	1.4%	2.5	-0.1%	0.8%	0.8	0.8	0.3%	1.7%	4.9	12.3
18	-0.6%	1.6%	0.8	1.2%	0.6%	0.2	2.5	6.4%	3.9%	2.8	8.5
19	1.4%	2.1%	1.5	0.8%	0.9%	0.6	2.6	9.4%	4.9%	0.7	5.3
20	2.8%	1.5%	1.9	-1.6%	1.1%	-1.1	-1.3	4.4%	2.4%	4.6	9.6
21	0.9%	2.2%	-0.2	1.1%	0.9%	0.4	3.7	9.4%	2.8%	3.4	9.1
22	3.6%	2.3%	1.5	1.8%	1.3%	1.3	2.2	8.8%	4.4%	2.3	7.7
1-22	1.0%	0.3%	5.8	0.7%	0.2%	2.2	5.3	3.0%	0.8%	11.6	25.8
X	0.3%	1.5%	0.5	0.4%	0.4%	1.2	1.1	0.0%	0.9%	-0.1	2.3

Note: Standard errors are computed from a Block Jackknife, with 500 equally sized blocks for all autosomes and 50 blocks for each individual chromosome. "Africa" is a pool of Yoruba+Mandenka+Dinka+Mbuti, "Europe" is a pool of Sardinian+French, and "East" is a pool of Dai+Han+Karitiana. For the Europe-African comparison, we do not restrict to sites that carry the ancestral alleles in Africans because Africans are directly involved in the comparison (that is, we do not compute enhanced *D*-statistics). For the other two population comparisons, we enhance sensitivity in the ancestry estimates (and also report the *Z*-score from an enhanced *D*-statistic) by restricting to sites where >99% of reads from 35 African samples carry the ancestral allele. We highlight very significant signals in red (|Z|>4 standard errors from 0).

Table S31: Summary of segmental duplication analysis.

Sample	Coverage	Average RD (5 kbp)	STD (5 kbp)	Average RD (1 kbp)	STD (1 kbp)	Duplicated (>10 kbp)	Duplicated (>20 kbp)	Duplicated* (>10 kbp)	Duplicated* (>20 kbp)
Denisova	23X	2,730.70	122.85	542.94	58.23	127,048,991	122,415,587	119,810,470	115,596,296
Neandertal	1X	79.62	8.84	15.95	4.58	149,085,259	116,222,458	141,121,521	109,861,191
Human (NA18507)†	42X	5,183.58	279.63	1,031.22	127.35	139,408,563	135,409,699	116,246,500	112,720,839
Chimpanzee†	7X	780.03	148.47	154.15	52.65	169,055,549	125,689,882	155,323,987	113,352,208
Bonobo	27X	2,015.55	221.77	401.90	101.71	132,942,704	110,544,491	124,768,431	103,790,117
Gorilla†	10X	945.39	69.15	192.43	37.52	133,893,160	126,971,263	116,632,244	110,335,854
Orangutan†	19X	500.71	47.30	100.59	22.40	97,415,408	90,226,192	83,549,453	76,928,588

GC corrected. 5 kbp windows overlap (1 kbp slide); 1 kbp windows are discrete. RD = read-depth. STD = standard deviation.*Autosomal only. † Male samples.

Table S32: Coverage and QC statistics for genomes analyzed.

Genome	mrsFAST 36 bp mapping coverage (X)	Known copy correlation*
Denisova	23.4	0.906
Dai_HGDP01307	9.6	0.935
French_HGDP00521	12.8	0.941
Han_HGDP00778	13.2	0.932
Karitiana_HGDP00998	9.5	0.935
Mandenka_HGDP01284	12.2	0.939
Mbuti_HGDP00456	8.7	0.924
Papuan_HGDP00542	9.0	0.935
San_HGDP01029	17.7	0.942
Sardinian_HGDP00665	11.9	0.937
Yoruba_HGDP00927	15.5	0.939

*Correlations were calculated between read-depth and 32 known diploid and ancestrally duplicated regions.

Table S33: Pairwise genome comparisons

Denisova vs.	Shared*			Specific			
	SD Regions	# basepairs (Denisova)	# basepairs (comparison)	SD Regions (Denisova)	# basepairs (Denisova)	SD Regions (comparison)	# basepairs (comparison)
Neandertal	714	94,261,598	82,068,418	218	21,334,698	600	27,792,773
NA18507	783	107,430,621	106,713,749	149	8,165,675	55	6,007,090
Chimpanzee	536	65,781,227	56,462,551	396	49,815,069	834	56,889,657
Bonobo	665	78,853,581	73,017,406	267	36,742,715	431	30,772,711
Gorilla	647	84,433,930	77,527,513	285	31,162,366	373	32,808,341
Orangutan	428	42,356,878	38,223,555	504	73,239,418	478	38,705,033

* >50% reciprocal overlap. We note a significant undercalling within Neandertal. SD = segmental duplication. The number of basepairs shared is distinct between Denisova and the 'comparison' individual as SD regions are considered shared if they overlap by 50% reciprocal overlap, however, still contain unique basepairs. All venn diagrams are based on shared basepairs only.

Table S34: Paralog-specific gene copy number polymorphisms in the Denisova compared to ten HGDP individuals

chr	start	end	gene	number of SUNK identifiers	Denisova copy number	Denisova specific event compared to 146 human genomes
chr11	3239561	3244361	C11orf36	2486	0	No
chr1	120106502	120115199	LOC128102	3118	0	Yes
chr15	34671269	34729667	GOLGA8A	5272	0.2	No
chr7	6838565	6865926	C7orf28B	1822	0.6	No
chr19	12035899	12061578	ZNF700	1541	0.6	No
chr15	20874796	20961480	BCL8	2525	0.8	No
chr19	41381343	41388657	CYP2A7	1917	0.9	No
chr5	177435688	177474656	FAM153C	2029	0.9	No
chr9	5334968	5339873	RLN1	1940	0.9	No
chr11	55650772	55659284	SPRYD5	2342	1	No
chr3	195384909	195415735	SDHAP2	2577	1.2	No
chr13	53063127	53161225	TPTE2P3	3937	1.2	No
chr16	834973	838383	RPUSD1	1802	1.4	No
chr15	21932513	21940739	LOC646214	3041	1.9	No
chr21	10906742	10990920	TPTE	11211	1.9	No
chr15	20613649	20711433	HERC2P3	3187	2.1	No
chr1	148930404	148953054	LOC645166	1573	3.2	No
chr9	41958801	42019584	KGFLP2	2443	3.4	No
chr7	5938340	5965603	CCZ1	1708	3.5	No
chr4	68566995	68588222	LOC550112	7479	4.1	Yes
chr22	42522500	42526883	CYP2D6	1761	4.2	No
chr21	11020841	11098925	BAGE5	12638	6.7	No
chr4	53226	88099	ZNF595	5165	11.9	Yes

Table S35: Human specific expansions

chrom	start	end	Genes
chr4	69584986	69663793	
chr10	46377245	46389583	
chr10	46397795	46409339	
chr10	51274200	51293392	*LOC728407*
chr10	51301586	51313166	*LOC728407*
chr15	23617303	23631796	
chr15	23639920	23670668	
chr16	14488808	14537594	*PARN*
chr16	30198636	30210741	*CORO1A; LOC606724; BOLA2B; SLX1B; SLX1B-SULT1A4; SULT1A4*
chr18	18524675	18535016	*ROCK1*

Table S36: Heterozygosity estimates obtained using the three different methods, showing consistently lower heterozygosity in Denisova. Absolute heterozygosity (bold) is reported as the number of heterozygotes per thousand sites. Relative heterozygosity (italic) reports the ratio of heterozygosity in Denisova relative to the heterozygosity of the present-day individual.

Individual		[1] base frequency spectra		[2] mlRho		[3] genotype calls	
		abs. [‰]	*rel.*[%]	abs. [‰]	*rel.*[%]	abs. [‰]	*rel.*[%]
	Denisova	-	-	**0.22**	-	**0.22**	-
HGDP01029	San	-	*19.7*	**1.15**	*19.4*	**1.05**	*20.5*
HGDP00927	Yoruba	-	*20.1*	**1.10**	*20.3*	**1.00**	*21.6*
HGDP01284	Mandenka	-	*20.2*	**1.11**	*20.1*	**1.01**	*21.4*
HGDP0456	Mbuti	-	*20.3*	**1.10**	*20.3*	**1.00**	*21.5*
DNK02	Dinka	-	*20.4*	**1.07**	*20.8*	**0.98**	*22.0*
HGDP00521	French	-	*26.2*	**0.84**	*26.4*	**0.76**	*28.2*
HGDP00665	Sardinian	-	*26.6*	**0.84**	*26.4*	**0.76**	*28.2*
HGDP00778	Han	-	*27.4*	**0.81**	*27.6*	**0.73**	*29.5*
HGDP01307	Dai	-	*28.5*	**0.81**	*27.7*	**0.73**	*29.5*
HGDP00542	Papuan	-	*32.6*	**0.70**	*32.0*	**0.63**	*34.4*
HGDP00998	Karitiana	-	*36.2*	**0.63**	*35.3*	**0.57**	*38.1*

Table S37: Heterozygosity of the Denisova genome expressed as percentage of the heterozygosity seen in eleven present-day human genomes. The values are influenced little by different filtering strategies as long as at least one filter is applied.

Mapability		+	-	-	+	+	-
Map quality		-	+	-	+	+	-
Coverage		-	-	+	-	+	-
#sites		*696 Mb*	*864 Mb*	*487 Mb*	*694 Mb*	*385 Mb*	*948 Mb*
HGDP01029	San	19.7	21.1	19.9	19.4	18.3	26.3
HGDP00927	Yoruba	20.1	21.5	20.3	19.7	18.7	26.9
HGDP01284	Mandenka	20.2	21.6	20.4	19.8	18.7	26.8
HGDP0456	Mbuti	20.3	21.6	20.4	19.7	18.7	26.8
DNK02	Dinka	20.4	21.9	20.6	20.0	18.9	27.2
HGDP00521	French	26.2	27.9	26.6	25.8	24.3	34.2
HGDP00665	Sardinian	26.6	28.3	27.1	26.2	24.7	34.8
HGDP00778	Han	27.4	29.3	28.2	27.0	25.9	35.7
HGDP01307	Dai	28.5	30.4	29.1	28.1	26.5	37.0
HGDP00542	Papuan	32.6	34.6	33.0	32.0	30.6	41.9
HGDP00998	Karitiana	36.2	38.2	37.5	35.7	34.1	45.7

Table S38: Absolute and relative heterozygosity (expressed as θ in per-mille) as well as estimates of per-base sequencing error rate inferred from mlRho. Note that error rate estimates are not comparable to the ones estimated elsewhere (Note 7 and Note 8), because a base quality filter was applied.

Individual		**MQ 30 + BQ 30**			**MQ 30 + BQ 30 + mappability**		
		Θ [‰]	**$\theta_{Denisova}$/ θ_{Human}**	**error rate**	**θ**	**$\theta_{Denisova}$ /θ_{Human}**	**error rate**
	Denisova	0.32	1.00	7.86e-04	0.22	1.00	7.50e-04
HGDP01029	San	1.33	0.24	3.81e-04	1.15	0.19	3.14e-04
HGDP00927	Yoruba	1.26	0.26	3.31e-04	1.10	0.20	2.88e-04
HGDP01284	Mandenka	1.26	0.26	3.41e-04	1.11	0.20	2.87e-04
HGDP0456	Mbuti	1.24	0.26	2.78e-04	1.10	0.20	2.48e-04
DNK02	Dinka	1.22	0.26	3.14e-04	1.07	0.21	2.90e-04
HGDP00521	French	0.97	0.33	3.27e-04	0.84	0.26	2.95e-04
HGDP00665	Sardinian	0.97	0.33	3.04e-04	0.84	0.26	2.62e-04
HGDP00778	Han	0.94	0.34	3.10e-04	0.81	0.28	2.66e-04
HGDP01307	Dai	0.93	0.35	2.88e-04	0.81	0.28	2.64e-04
HGDP00542	Papuan	0.82	0.39	2.95e-04	0.70	0.32	2.60e-04
HGDP00998	Karitiana	0.75	0.43	3.30e-04	0.63	0.35	2.83e-04

Table S39: Number of genotype calls retained in the Denisova genome when applying various filters. The coverage filter was used in all combinations of filters (it eliminated 5% of the initial genotype calls). The final filter set used for determining absolute heterozygosity is marked in bold.

FILTERS[§]							
Map	RM	SurrInd	SysErr	GQ	**Sites**	**Sites retained [%]**	**# heterozygotes**
					2,431,275,256	-	882,828
+					1,709,263,335	72.47	421,948
	+				1,243,885,562	51.15	332,524
		+			2,424,802,282	82.28	857,219
			+		2,431,172,155	97.07	879,274
				+	2,335,345,230	96.05	741,647
		+	+	+	2,329,078,665	89.51	717,159
		+	+		2,424,701,647	80.29	853,846
+		+	+		1,706,986,716	71.53	415,228
+	+				1,103,584,523	45.37	249,815
+	+	+	+		**1,100,259,824**	**44.84**	**246,029**
+	+	+	+	+	1,093,109,146	44.55	229,038

[§]*Map* = Mappability of 20mer; *RM* = Repeat Masking; *SurrInd* = positions surrounding indels +/- 5 bp; *SysErr* = systematic errors; *GQ* = Genotype Quality >= 40

Table S40: Effect of filtering on estimates of heterozygosity. Absolute heterozygosity is reported in per-mille and relative heterozygosity in percent (in brackets).

		FILTERS[§]								
Map	RM	SurrInd	SysErr	GQ	**Denisova**	**San**	**Mbuti**	**Yoruba**	**Mandenka**	**Dinka**
					0.41 (-)	1.47 (27.75)	1.33 (30.67)	1.38 (29.55)	1.36 (30.06)	1.32 (30.87)
+					0.26 (-)	1.25 (20.75)	1.18 (21.99)	1.19 (21.88)	1.19 (21.78)	1.16 (22.43)
	+				0.31 (-)	1.27 (24.66)	1.18 (26.66)	1.21 (26.04)	1.19 (26.32)	1.17 (26.87)
		+			0.26 (-)	1.15 (22.98)	1.09 (24.26)	1.09 (24.34)	1.11 (23.88)	1.06 (24.87)
			+		0.40 (-)	1.45 (27.76)	1.32 (30.68)	1.36 (29.60)	1.34 (30.08)	1.31 (30.90)
				+	0.26 (-)	1.14 (22.96)	1.10 (23.78)	1.07 (24.45)	1.10 (23.73)	1.05 (24.94)
		+	+	+	0.22 (-)	0.98 (22.06)	0.97 (22.32)	0.93 (23.38)	0.97 (22.30)	0.92 (23.59)
		+	+		0.26 (-)	1.14 (23.07)	1.08 (24.36)	1.08 (24.46)	1.10 (23.98)	1.05 (24.98)
+		+	+		0.23 (-)	1.11 (20.55)	1.05 (21.58)	1.05 (21.63)	1.06 (21.36)	1.03 (22.05)
+	+				0.24 (-)	1.17 (20.56)	1.11 (21.71)	1.11 (21.68)	1.12 (21.56)	1.09 (22.15)
+	**+**	**+**	**+**		**0.22 (-)**	**1.05 (20.50)**	**1.00 (21.51)**	**1.00 (21.59)**	**1.01 (21.35)**	**0.98 (21.99)**
+	+	+	+	+	0.19 (-)	0.94 (20.74)	0.92 (21.03)	0.89 (21.92)	0.92 (21.03)	0.88 (22.11)
Map	RM	SurrInd	SysErr	GQ	**Sardinian**	**French**	**Dai**	**Han**	**Papuan**	**Karitiana**
					1.04 (39.35)	1.06 (38.36)	1.01 (40.51)	1.02 (40.12)	0.88 (46.56)	0.82 (49.99)
+					0.91 (28.63)	0.91 (28.49)	0.87 (29.90)	0.87 (29.79)	0.75 (34.67)	0.68 (38.04)
	+				0.91 (34.62)	0.93 (33.79)	0.88 (35.61)	0.89 (35.43)	0.76 (41.44)	0.70 (45.16)
		+			0.83 (31.64)	0.83 (31.63)	0.79 (33.31)	0.80 (32.98)	0.69 (38.47)	0.63 (42.09)
			+		1.02 (39.44)	1.05 (38.46)	0.99 (40.61)	1.00 (40.23)	0.86 (46.71)	0.80 (50.19)
				+	0.81 (32.16)	0.82 (31.92)	0.77 (33.97)	0.78 (33.69)	0.67 (39.21)	0.60 (43.32)
		+	+	+	0.72 (30.23)	0.71 (30.37)	0.67 (32.21)	0.68 (31.98)	0.58 (37.22)	0.52 (41.37)
		+	+		0.83 (31.80)	0.83 (31.80)	0.79 (33.48)	0.79 (33.15)	0.68 (38.69)	0.62 (42.35)
+		+	+		0.81 (28.11)	0.81 (28.13)	0.77 (29.47)	0.77 (29.43)	0.66 (34.32)	0.60 (37.79)
+	+				0.85 (28.42)	0.85 (28.29)	0.81 (29.63)	0.81 (29.58)	0.70 (34.46)	0.63 (38.07)
+	**+**	**+**	**+**		**0.76 (28.16)**	**0.76 (28.17)**	**0.73 (29.45)**	**0.73 (29.48)**	**0.63 (34.37)**	**0.57 (38.07)**
+	+	+	+	+	0.69 (28.35)	0.68 (28.52)	0.65 (30.09)	0.65 (30.08)	0.56 (34.89)	0.50 (38.99)

[§]*Map* = Mappability of 20mer; *RM* = Repeat Masking; *SurrInd* = positions surrounding indels +/- 5 bp; *SysErr* = systematic errors; *GQ* = Genotype Quality >= 40

Table S41: Summary information for 12 genomes as processed for PSMC

Population	Sample	DNA Source	Depth*	Heterozygosity¶
Denisova	Denisova	Ancient bone	27.71	0.216
Karitiana	HGDP00998	Cell line	26.22	0.588
Papuan	HGDP00542	Cell line	26.09	0.651
Dai	HGDP01307	Cell line	29.00	0.770
Han	HGDP00778	Cell line	27.89	0.772
Sardinian	HGDP00665	Cell line	24.79	0.795
French	HGDP00521	Cell line	26.68	0.810
Dinka	DNK02	Mouthwash	27.71	1.040
Mbuti	HGDP0456	Cell line	24.54	1.044
Mandenka	HGDP01284	Cell line	24.81	1.057
Yoruba	HGDP00927	Cell line	32.14	1.065
San	HGDP01029	Cell line	33.26	1.111

* Sequencing depth is computed based on coverage of HapMap3 SNP sites.
¶ Heterozygosity (expressed per-mille) inferred from SAMtools as described in Note S17. This agrees qualitatively with the inferences of Note 15.

Table S42: Non-synonymous/synonymous comparisons

	$Q_{Denisova/Modern}$	$P_{Denisova/Modern}$
Probably damaging	1.86	2.56
Possibly damaging	1.66	1.81
Benign	0.93	1.11
Probably + Possibly	1.78	2.28
All sites	1.15	1.52

Note: This table is based on the alignments to the human genome

Table S43: Classification of single-nucleotide changes (SNCs) in protein-coding sequences by their predicted functional effects using Ensembl's Variant Effect Predictor (VEP). SNCs were classified according to their most severe predicted effect. STOP gains and losses were also included in the "non-synonymous" category. Splice site SNCs that occur in 3' and 5' UTRs were only included in the "splice site" category. Ensembl genes includes all genes in the Ensembl 65 annotation, including CCDS-verified genes.

Single nucleotide changes (SNCs)		**Fixed derived in modern humans**			**High-frequency derived in modern humans**		
		Only one ancestral allele in Denisova	Denisova is homozygous ancestral	Total	Only one ancestral allele in Denisova	Denisova is homozygous ancestral	Total
CCDS-verified coding genes (longest transcript per gene)	Non-synonymous	46	214	260	25	368	393
	Synonymous	34	254	288	23	549	572
	STOP gained	0	0	0	0	2	2
	STOP lost	0	1	1	0	1	1
	Splice site	6	66	72	8	108	116
	Essential splice site	0	3	3	0	1	1
	3′ UTR	150	630	780	70	1135	1205
	5′ UTR	16	80	96	19	172	191
Ensembl genes (all coding transcripts per gene)	Non-synonymous	83	269	352	29	467	496
	Synonymous	55	275	330	24	586	610
	STOP gained	0	1	1	0	5	5
	STOP lost	0	1	1	0	6	6
	Splice site	10	90	100	10	141	151
	Essential splice site	2	5	7	1	8	9
	3′ UTR	218	900	1118	89	1527	1616
	5′ UTR	36	200	236	30	396	426

Table S45: Fixed SNCs in positions with a high primate conservation score (above or equal to 0.95) ranked by this score. The function or description of the protein coded by the gene affected by each SNC is listed in the right-most column. SNCs that are inferred as fixed in modern humans using the 1000G data but have dbSNP entries are marked with an asterisk in the 1000G Frequency column. GS = Grantham score. STAS = Sulphate Transporter and AntiSigma factor antagonist domain.

Position	Modern human derived	Denisova	Chimpanzee -human ancestor	Gorilla- human ancestor	Orangutan -human ancestor	1000G Frequency	Gene	Amino acid change	Protein domain (UniProtKB)	GS	Primate conservation score	Protein function / description
chr14:26918100	C	T/T	T	T	T	Fixed	NOVA1	I197V	KH (RNA binding)	29	0.985	Neuron-specific RNA-binding protein (*159*).
chr18:51889262	C	A/A	A	A	A	Fixed	C18orf54	E237D	-	0	0.984	May be involved in cell proliferation (UniProtKB by similarity)
chr11:18339402	C	T/T	T	T	T	Fixed	HPS5	T2A	-	58	0.983	Skin pigmentation (*107*).
chr5:149431516	G	A/A	A	A	A	Fixed*	HMGXB3	S1214G	-	56	0.98	HMG box domain containing protein.
chr11:129772293	T	G/G	G	G	G	Fixed	PRDM10	T1133N	-	65	0.977	Craniofacial development (*160*).
chr12:1937340	G	A/A	A	A	A	Fixed*	LRTM2	E9G	signal peptide	98	0.974	Leucine-rich repeat and transmembrane domain-containing protein.
chr17:48245872	G	G/A	A	A	A	Fixed	SGCA	I175V	-	29	0.974	Muscle fiber stability (*138*).
chr1:179380284	A	G/G	G	G	G	Fixed	AXDND1	M371I	coiled coil	10	0.974	Axonemal dynein.
chr8:92378897	G	A/A	A	A	A	Fixed	SLC26A7	I526M	STAS	10	0.971	Anion exchange in kidney (*161*).
chr17:45232079	G	A/A	A	A	A	Fixed	CDC27	S306P	-	74	0.967	Mitotic regulation (*162*).
chr15:65983405	A	C/C	C	C	C	Fixed*	DENND4A	G1175V	-	109	0.966	DENN-domain containing protein.
chr11:28119295	T	C/C	C	C	C	Fixed	KIF18A	R67K	kinesin motor	26	0.966	Mitotic regulation (*163*).
chr13:84454655	A	C/C	C	C	C	Fixed*	SLITRK1	A330S	-	99	0.965	Regulates neuronal dendrite growth; associated with Tourette syndrome and trichotillomania (*164*).
chr6:149918766	T	C/C	C	C	C	Fixed*	KATNA1	A343T	-	58	0.963	ATPase activity in microtubule transport required for axonal growth (UniProtKB by similarity).
chr7:134642991	A	G/G	G	G	G	Fixed*	CALD1	V671I	-	29	0.961	Regulation of muscle contraction (UniProtKB by similarity).
chr1:23418576	T	C/C	C	C	C	Fixed*	LUZP1	A727T	-	58	0.956	Involved in neural tube closure during development of the brain (*165*).
chr5:54585213	C	T/T	T	T	T	Fixed*	DHX29	I317M	-	10	0.954	RNA helicase involved in translation initiation (*166*).
chr11:128840599	A	T/T	T	T	T	Fixed	ARHGAP32	E1489D	-	0	0.954	Regulates dendritic spine morphology (*167*).
chr22:40760978	T	C/C	C	C	C	Fixed	ADSL	A429V	-	64	0.953	Associated with adenylosuccinase deficiency, leading to psychomotor retardation and autism (*134, 135*).
chr2:231974031	C	T/T	T	T	T	Fixed	HTR2B	N216D	extracellular	23	0.953	Serotonin receptor associated with severe impulsivity (*168*). Plays a role in presynaptic inhibition (*169*).
chr1:158612618	C	G/G	G	G	G	Fixed*	SPTA1	P1531A	spectrin repeat	27	0.952	Actin filament organization in cytoskeleton (*170*).
chr19:11491606	C	G/G	G	G	G	Fixed	EPOR	L261V	helical	32	0.95	Erythropoietin receptor (*171*) associated with erythrocytosis and anemia (*172, 173*).
chr7:146825878	G	A/A	A	A	A	Fixed	CNTNAP2	I345V	laminin G-like	29	0.95	Neurexin expressed during cortical development (*133*); associated with susceptibility to autism (*130, 131*) and language disorders (*27*).

Table S46: CCDS genes with more than one non-synonymous SNC where modern humans are fixed derived and Denisova is homozygote ancestral.

Number of non-synonymous SNCs	Genes
2	ADAM18, ANKRD30A, C18orf54, C5orf20, CASC5, HERC5, HPS5, IFI44L, MAGEA4, OR5K4, SETD2, SPAG17, SPTA1, SSH2, TP53TG5, ZNF185, ZNF333
3	ITGB4, RP1L1, SPAG5, TTF1

Table S47: STOP gains and losses in both CCDS and non-CCDS genes. Marked in blue are SNCs in CCDS-verified genes.

Position	Modern human derived	Denisova	Chimpanzee ancestral	Gorilla ancestral	Orangutan ancestral	1000G Freq.	Transcripts	STOP	Gene	Flag
chr1:161967680	C	T/T	T	T	T	Fixed	ENST00000294794 ENST00000367940	lost	OLFML2B	InDel nearby
chr1:171178090	T	C/C	C	C	C	96%	ENST00000209929 ENST00000441535	gained	FMO2	-
chr1:183592594	A	G/G	G	G	G	92%	ENST00000367534	gained	ARPC5	CpG
chr2:27551325	G	A/A	A	A	A	93%	ENST00000415683	lost	GTF3C2	CpG
chr2:198593260	A	C/C	C	C	C	99%	ENST00000430004	lost	BOLL	-
chr6:154360569	C	T/T	T	T	T	97%	ENST00000434900 ENST00000520282	lost	OPRM1	-
chr11:64893151	T	C/C	C	C	C	Fixed	ENST00000526171	gained	MRPL49	-
chr11:104763117	A	G/G	G	G	G	96%	ENST00000375726 ENST00000422698 ENST00000433738 ENST00000441710 ENST00000446862 ENST00000447913 ENST00000448103 ENST00000494737 ENST00000508062	gained	CASP12	CpG
chr12:57003964	T	A/A	A	A	A	96%	ENST00000551996	lost	BAZ2A	-
chr14:31952754	A	G/G	G	G	G	98%	ENST00000399285	gained	GPR33	CpG
chr14:50798969	C	G/G	G	G	G	98%	ENST00000534267	gained	CDKL1C	-
chr15:62932556	G	C/C	C	C	C	96%	ENST00000558940	lost	RP11-625H11.1	-
chr19:7705502	T	A/A	A	A	A	99%	ENST00000320400	lost	STXBP2	-

Table S51: Fixed and high-frequency essential splice site SNCs.

Position	Modern human derived	Denisova	Chimpanzee	Gorilla	Orangutan	1000G Freq.	Gene Name	Gene Function	Flag
chr3:9594532	C	A/A	A	-	A	Fixed	LHFPL4	Unknown	InDel nearby
chr17:68127117	G	A/A	A	A	A	Fixed	KCNJ16	Potassium ion channel expressed in kidney and thyroid gland	-
chr19:2098974	G	A/A	A	A	A	Fixed	IZUMO4	Sperm-egg fusion protein	-
chr19:50879835	T	C/C	C	C	C	98%	NR1H2	Reglates uptake of cholesterol (UniProtKB by similarity)	-

Table S52: Classification of modern human-derived InDels in exons, splice sites and UTR regions of CCDS genes according to the VEP's predicted functional effect. Changes in "essential splice sites" are also included in the "splice site" category.

InDels		Fixed derived in modern humans			High-frequency derived in modern humans		
		Only one ancestral allele in Denisova	Denisova is homozygous ancestral	Total	Only one ancestral allele in Denisova	Denisova is homozygote ancestral	Total
CCDS-verified coding genes (longest transcript per gene)	Frameshift	0	3	3	0	2	2
	In-frame non-synonymous	0	2	2	0	0	0
	Splice site	1	7	8	0	7	7
	Essential splice site	0	0	0	0	2	2
	3′ UTR	1	89	90	3	76	79
	5′ UTR	0	21	21	0	13	13

Table S53: Fixed and high-frequency modern human-derived, Denisova-ancestral InDels that cause a frameshift in a coding sequence, an in-frame non-synonymous event or a disruption of a splice site. The base upstream of each InDel is also included for reference. Changes that are inferred as fixed using the 1000G data but have dbSNP entries are marked with an asterisk in the 1000G Frequency column. The RM flag means that the change is found in a repeat-masked region.

Position	Modern human derived	Denisova	Chimpanzee	Gorilla	Orangutan	1000G Freq.	Gene	CCDS longest transcript	Consequence	Flag
chr1:156565049	A	AAC/AAC	AAC	AAC	AAC	Fixed	GPATCH4	ENST00000438976	Frameshift	-
chr1:197576307	AT	A/A	A	A	A	98%	DENND1B	ENST00000367396	Splice site disrupted	-
chr2:10808776	C	CAA/CAA	CAA	CAA	CAA	Fixed	NOL10	ENST00000381685	Splice site disrupted	-
chr2:219692790	AG	A/A	A	A	A	Fixed*	PRKAG3	ENST00000529249	Splice site disrupted	-
chr4:77018837	A	A/AC	AC	AC	AC	Fixed	ART3	ENST00000355810	Splice site disrupted	-
chr5:71528390	C	CA/CA	CA	CA	CA	91%	MRPS27	ENST00000261413	Splice site disrupted	-
chr5:111500816	CTAAA	C/C	C	C	C	97%	EPB41L4A	ENST00000261486	Essential splice site disrupted	-
chr8:94147031	A	ATTG/ ATTG	ATTG	ATTG	ATTG	Fixed	RP11-88J22.1.1	ENST00000521906	In-frame non-synonymous (L16PM)	RM
chr8:101206459	A	AGAC/ AGAC	AGAC	AGAC	AGAC	Fixed	SPAG1	ENST00000388798	In-frame non-synonymous (K353KD)	-
chr12:10217326	C	CTT/CTT	CTT	CTT	CTT	Fixed	CLEC9A	ENST00000355819	Splice site disrupted	-
chr12:50829263	T	TTATTC/ TTATTC	TTATTC	TTATTC	TTATTC	Fixed	LARP4	ENST00000398473	Splice site disrupted	-
chr12:117273983	A	AAT/AAT	AAT	AAT	AAT	93%	RNFT2	ENST00000257575	Splice site disrupted	-
chr13:60385060	A	ATTAC/ ATTAC	ATTAC	ATTAC	ATTAC	Fixed	DIAPH3	ENST00000400324	Splice site disrupted	-
chr13:79916792	TA	T/T	T	T	T	98%	RBM26	ENST00000267229	Splice site disrupted	-
chr17:15343524	CCTT	C/C	C	C	C	98%	FAM18B2-CDRT4	ENST00000522212	Essential splice site disrupted	-
chr17:26692224	AG	A/A	A	A	A	Fixed*	SEBOX	ENST00000431468	Frameshift	RM
chr17:43318777	GC	G/G	G	G	-	96%	FMNL1	ENST00000331495	Frameshift	-
chr18:61326628	AT	A/A	A	A	A	Fixed*	SERPINB3	ENST00000283752	Splice site disrupted	-
chr20:590541	AG	A/A	A	A	A	91%	TCF15	ENST00000246080	Frameshift	-
chr22:19189003	A	AC/AC	AC	AC	AC	Fixed	CLTCL1	ENST00000263200	Frameshift, splice site disrupted	-
chr22:36006923	GC	G/G	G	G	-	Fixed*	MB	ENST00000406324	Splice site disrupted	RM

Table S56: Denisovan state for highly-cited SNPs.

SNP ID(s) - Reference (hg19) / Alternative	Risk allele(s)	Description	Denisovan genotype	Global human frequency of the Denisovan allele (dbSNP)
rs429358 - T/C rs7412 - C/T	C, C (ApoE4 haplotype)	Associated with risk of Alzheimer's disease (*174*)	C/C, C/C (ancestral)	15.4%, 91.6%
rs1800497 - C/T	T	Associated with a reduced number of dopamine binding sites in the brain (*175*), increased risk for alcoholism (*176*) and nicotine dependence (*177*)	T/T (ancestral)	29.7%
rs9939609 - T/A	A	Associated with risk of diabetes (*178*) and obesity (*179*)	A/A (ancestral)	35.6%
rs7903146 - C/T	T	Associated with risk of diabetes (*180*)	T/T (ancestral)	21.9%
rs4680 - G/A	"warrior": G "worrier": A	Associated with increased attention and memory ("worrier strategy") vs. exploratory behavior and fast response to aversive stimuli ("warrior strategy") (*181-183*). Also, associated with risk of schizophrenia (*184*).	G/G (ancestral)	60.9%
rs7495174 - A/G rs4778241 (rs6497268) - A/C rs4778138 (rs11855019) - A/G	different haplotypes	Associated with eye pigmentation (*185*)	G/G, A/A, G/G (ancestral haplotype - predicted brown eye color)	25.6%, 47.3%, 44.6%
rs12913832 - A/G	blue: G brown: A	Associated with brown vs. blue eye pigmentation (*88*)	A/A (ancestral)	70.8%
rs1805007 - C/T	T	13 - 20X higher likelihood of red hair color (MC1R gene) (*186*)	C/C (ancestral)	97.1%
rs17822931 - C/T	wet: C dry: T	Associated with wet vs. dry earwax (*187*)	C/C (ancestral)	69%
rs4988235 - C/T rs182549 - C/T	tolerance: T, T intolerance: C, C	Associated with lactose tolerance in European populations (*188, 189*)	C/C, C/C (ancestral)	76.6%, 76.6%
rs4988234 - C/T	tolerance: T intolerance: C	Associated with lactose intolerance in sub-Saharan Africa (*190, 191*)	C/C (ancestral)	100%
rs3827760 - C/T	C	Associated with hair morphology (*192*) and incisor shape (*193*)	T/T (ancestral)	70.9%
rs53576 - A/G	A	Associated with decreased empathy and other personality traits (*194*)	G/G (ancestral)	59.1%
rs1815739 - T/C	T	Associated with impaired muscle performance (*195*)	C/C (ancestral)	62.5%
rs6152 - G/A	G	Associated with male pattern baldness (*196*)	G/G (ancestral)	78.8%

Table S57: Allele states for pigmentation-predictive SNPs in the Denisovan individual at 1.9X coverage as reported in Cerqueira et al. and the Denisova genome at high coverage. We are able to identify 11 SNPs whose alleles were previously undetermined, as well as 3 SNPs where the predicted genotype is now known to be different from the state observed in the low-coverage genome.

Genes	SNPs from Cerqueira et al. (2012)	Position (GRCh37)	Denisova 1.9X	Denisova 30X
SLC45A2 (MATP)	rs26722	chr5:33963870	nd	G/G
SLC45A2 (MATP)	rs6867641	chr5:33985857	C	C/C
IRF4	rs12203592	chr6:396321	C	C/C
TPCN2	rs3829241	chr11:68855363	G	G/G
TPCN2	rs3750965	chr11:68840160	A	A/A
TPCN2	rs35264875	chr11:68846399	A	A/A
TPCN2	rs896978	chr11:68828929	C	C/C
TYR	rs1042602	chr11:88911696	C	C/C
TYR	rs1126809	chr11:89017961	G	G/G
TYR	rs1393350	chr11: 89011046	G	G/G
SLC24A4	rs2402130	chr14:92801203	G	G/G
OCA2	rs1498519	chr15:28011651	C	C/C
OCA2	rs1800414	chr15: 28197037	A	A/A
OCA2	rs1800407	chr15:28230318	G	G/G
OCA2	rs1800401	chr15:28260053	C	C/C
OCA2	rs7495174	chr15:28344238	G	G/G
OCA2	rs4778241	chr15:28338713	A	A/A
OCA2	rs4778138	chr15:28335820	nd	G/G
OCA2	rs1584407	chr15:28157259	C	C/C
OCA2	rs2703952	chr15:28181981	A	A/A
OCA2	rs2594935	chr15:28185038	A	A/A
OCA2	rs728405	chr15:28199853	G/T	G/G
OCA2	rs1448488	chr15:28216857	A	A/A
OCA2	rs4778220	chr15:28221138	nd	A/A
OCA2	rs7170869	chr15:28288748	A	A/A
OCA2	rs1545397	chr15:28187772	A	A/A
HERC2	rs1129038	chr15:28356859	G	G/G
HERC2	rs12913832	chr15:28365618	A	A/A
HERC2	rs1667394	chr15:28530182	nd	G/G
HERC2	rs8039195	chr15:28516084	C	C/C
HERC2	rs7183877	chr15:28365733	C	C/C
HERC2	rs1635168	chr15:28535266	T	T/T
HERC2	rs8028689	chr15:28488888	C	C/C
HERC2	rs16950987	chr15:28526228	A	A/A
HERC2	rs916977	chr15:28513364	A	A/A
HERC2	rs7494942	chr15:28364059	A	A/A
HERC2	rs3935591	chr15:28374012	A	A/A
HERC2	rs7170852	chr15:28427986	T	T/T
HERC2	rs2238289	chr15:28453215	C	C/C
HERC2	rs2240203	chr15:28494202	G	G/G
HERC2	rs2240204	chr15:28494032	C	C/C
HERC2	rs16950979	chr15:28520506	A	A/A
SLC24A5 (NCKX5)	rs1426654	chr15:48426484	G	G/G
MC1R	rs1805007	chr16:89986117	C	C/C
MC1R	rs1805008	chr16:89986144	nd	C/C
MC1R	rs3212346	chr16:89982358	A	A/A
MC1R	rs885479	chr16:89986154	G	G/G
DPEP1	rs164741	chr16:89692298	C	C/C
C16orf55	rs7188458	chr16:89726484	G	G/G
C16orf55	rs459920	chr16:89730827	T	T/T
ZNF276	rs7204478	chr16:89795485	C	C/C
ZNF276	rs6500437	chr16:89789898	T	T/T
ZNF276	rs1800359	chr16:89805261	C	C/C
ZNF778	rs9921361	chr16:89294439	G	G/G
PRDM7	rs2078478	chr16:90130136	T	T/T
PRDM7	rs7196459	chr16:90141477	G	G/G
ACSF3	rs12599126	chr16:89206483	C	C/C
ANKRD11	rs2353033	chr16:89385561	T	T/T
ANKRD11	rs1466540	chr16:89344477	C	C/C
ANKRD11	rs2353028	chr16:89352678	G	G/G
ANKRD11	rs2306633	chr16:89355278	A	A/A
ANKRD11	rs3096304	chr16:89373707	A	A/A
ANKRD11	rs889574	chr16:89386808	C	C/C
ANKRD11	rs2965946	chr16:89516612	T	C/C
SPG7	rs382745	chr16:89603586	C/T	C/C
CPNE7	rs455527	chr16:89644001	T	T/T
CPNE7	rs352935	chr16:89648580	G	C/C
CPNE7	rs464349	chr16:89656251	T	T/T
CHMP1A	rs460879	chr16:89712889	T	T/T
CDK10	rs258322	chr16:89755903	T	T/T
CDK10	rs258324	chr16:89754255	C	C/C
CDK10	rs3751700	chr16:89752194	G	G/G
CDK10	rs1946482	chr16:89762410	T	T/T
SPATA2L	rs3751695	chr16:89764549	C	C/C
FANCA	rs7195066	chr16:89836323	A	A/A
FANCA	rs8058895	chr16:89814807	T	T/T
FANCA	rs2011877	chr16:89814818	A	A/A
FANCA	rs2239359	chr16:89849480	A	A/A
FANCA	rs16966142	chr16:89851033	C	C/C
FANCA	rs1800286	chr16:89869761	G	G/G
FANCA	rs11861084	chr16:89875710	C	C/C
SPIRE2	rs8060934	chr16:89920025	C	C/C
SPIRE2	rs3803688	chr16:89934886	nd	T/T
TCF25	rs2270460	chr16:89972416	T	T/T
CENPBD1	rs4785755	chr16:90037828	C	C/C
DBNDD1	rs8059973	chr16:90079534	G	G/G
DBNDD1	rs11648785	chr16:90084561	C	C/C
GAS8	rs2241039	chr16:90088437	C	C/C
GAS8	rs3785181	chr16:90105333	G	G/G
GAS8	rs1048149	chr16:90110950	C	C/C
AFG3L1	rs4785763	chr16:90066936	A	A/A
AFG3L1	rs4408545	chr16:90044028	nd	C/C
DYNLRB1	rs2281695	chr20:33129164	nd	C/C
PIGU	rs2378199	chr20:33186480	C	C/C
PIGU	rs2378249	chr20:33218090	A	A/A
NCOA6	rs6060034	chr20:33351864	nd	C/C
NCOA6	rs6060043	chr20:33364584	T	T/T
EIF6	rs619865	chr20:33867697	G/A	G/A
ASIP	rs6058017	chr20:32856998	G	G/G
Intergenic regions	rs9328192	chr6:434364	A	A/A
	rs9405681	chr6:449358	T	T/T
	rs4959270	chr6:457748	A	A/A
	rs9378805	chr6:417727	A	A/A
	rs1540771	chr6:466033	A	A/A
	rs1011176	chr11:68933897	G	G/G
	rs2305498	chr11:68866914	C	C/C
	rs12821256	chr12:89328335	T	T/T
	rs8016079	chr14:92758445	G	G/G
	rs4904864	chr14:92764519	G	G/G
	rs4904868	chr14:92781001	T	T/T
	rs12896399	chr14:92773663	G	G/G
	rs438702	chr15:28999168	nd	A/A
	rs9932354	chr16:89052565	nd	C/C
	rs11076747	chr16:89057025	G	G/G
	rs4785648	chr16:89328477	A	A/A
	rs4347628	chr16:89570635	C	C/C
	rs12443954	chr16:89741496	G	G/G
	rs9936896	chr16:90069059	C	C/C
	rs4785612	chr16:90113107	C	C/C
	rs7201721	chr16:90058746	A	A/A
	rs4238833	chr16:90050689	G	G/G
	rs6119471	chr20:32785212	G	G/G
	rs1015362	chr20:32738612	A	T/T
	rs4911414	chr20:32729444	G	G/G

Table S58: Pigmentation-predictive SNP counts for the Denisovan individual at 1.9X coverage as reported in Cerqueira et al. and the Denisova genome at 30X coverage. All the predicted phenotypes remain the same as in Cerqueira et al.

Phenotype category	Denisova 1.9X SNP counts	Denisova 30X SNP counts	Predicted phenotype
Fairer skin	1	2	Darker skin
Darker skin	13	13	
Darker brown hair	3	4	Darker brown hair
Lighter brown hair	0	1	
Brown hair	10	11	Brown hair
Not-brown hair	1	1	
Blond hair	4	4	Not-blond hair
Not-blond hair	10	11	
Red hair	24	27	Not-red hair
Not-red hair	31	34	
Brown eyes	19	21	Brown eyes
Not-brown eyes	6	8	
Green eyes	4	6	Not-green eyes
Not-green eyes	8	9	
Blue eyes	5	5	Not-blue eyes
Not-blue eyes	10	11	
Freckles	18	18	Not-freckles
Not-freckles	23	28	

References and Notes

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

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

3. J. J. Hublin, Out of Africa: Modern human origins special feature: The origin of Neandertals. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 16022 (2009). doi:10.1073/pnas.0904119106 Medline

4. J. Krause *et al*., The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464**, 894 (2010). doi:10.1038/nature08976 Medline

5. A. Gibbons, Who were the Denisovans? *Science* **333**, 1084 (2011). doi:10.1126/science.333.6046.1084 Medline

6. D. Reich *et al*., Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* **89**, 516 (2011). doi:10.1016/j.ajhg.2011.09.005 Medline

7. H. A. Burbano *et al*., Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* **328**, 723 (2010). doi:10.1126/science.1188046 Medline

8. Materials and methods are available as supplementary materials on *Science* Online.

9. 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). doi:10.1073/pnas.0704665104 Medline

10. L. Orlando *et al*., True single-molecule DNA sequencing of a pleistocene horse bone. *Genome Res.* **21**, 1705 (2011). doi:10.1101/gr.122747.111 Medline

11. M. Kircher, S. Sawyer, M. Meyer, Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3 (2012). doi:10.1093/nar/gkr771 Medline

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

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

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

15. M. Goodman, The genomic record of humankind's evolutionary roots. *Am. J. Hum. Genet.* **64**, 31 (1999). doi:10.1086/302218 Medline

16. J. Pickrell, J. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *Nature Precedings* (2012); http://precedings.nature.com/documents/6956/version/1.

17. P. Skoglund, M. Jakobsson, Archaic human ancestry in East Asia. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18301 (2011). doi:10.1073/pnas.1108181108 Medline

18. M. Currat, L. Excoffier, Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15129 (2011). doi:10.1073/pnas.1107450108 Medline

19. R. J. Petit, L. Excoffier, Gene flow and species delimitation. *Trends Ecol. Evol.* **24**, 386 (2009). doi:10.1016/j.tree.2009.02.011 Medline

20. J. A. Coyne, H. A. Orr, in *Speciation and its Consequences*, D. Otte, and J. A. Endler, Eds. (Wiley, New York, 1989), pp. 180–207.

21. J. R. Kidd, F. L. Black, K. M. Weiss, I. Balazs, K. K. Kidd, Studies of three Amerindian populations using nuclear DNA polymorphisms. *Hum. Biol.* **63**, 775 (1991). Medline

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

23. D. F. Conrad *et al*.; 1000 Genomes Project, Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712 (2011). doi:10.1038/ng.862 Medline

24. C. C. Cerqueira *et al*., Predicting *homo* pigmentation phenotype through genomic data: From neanderthal to James Watson. *Am. J. Hum. Biol.* **24**, 705 (2012). doi:10.1002/ajhb.22263 Medline

25. J. W. IJdo, A. Baldini, D. C. Ward, S. T. Reeders, R. A. Wells, Origin of human chromosome 2: An ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 9051 (1991). doi:10.1073/pnas.88.20.9051 Medline

26. R. M. Durbin *et al*.; 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (2010). doi:10.1038/nature09534 Medline

27. S. C. Vernes *et al*., A functional genetic link between distinct developmental language disorders. *N. Engl. J. Med.* **359**, 2337 (2008). doi:10.1056/NEJMoa0802828 Medline

28. W. Enard *et al*., A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. *Cell* **137**, 961 (2009). doi:10.1016/j.cell.2009.03.041 Medline

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

30. N. Rohland, M. Hofreiter, Ancient DNA extraction from bones and teeth. *Nat. Protoc.* **2**, 1756 (2007). doi:10.1038/nprot.2007.247 Medline

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

32. D. R. Bentley *et al*., Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53 (2008). doi:10.1038/nature07517 Medline

33. T. Blondal *et al*., Isolation and characterization of a thermostable RNA ligase 1 from a *Thermus scotoductus* bacteriophage TS2126 with good single-stranded DNA ligation properties. *Nucleic Acids Res.* **33**, 135 (2005). doi:10.1093/nar/gki149 Medline

34. T. W. Li, K. M. Weeks, Structure-independent and quantitative ligation of single-stranded DNA. *Anal. Biochem.* **349**, 242 (2006). doi:10.1016/j.ab.2005.11.002 Medline

35. M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protoc.*, 10.1101/pdb.prot5448 (2010).

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

37. J. Dabney, M. Meyer, Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* **52**, 87 (2012). Medline

38. U. Varshney, J. H. van de Sande, Specificities and kinetics of uracil excision from uracil-containing DNA oligomers by *Escherichia coli* uracil DNA glycosylase. *Biochemistry* **30**, 4055 (1991). doi:10.1021/bi00230a033 Medline

39. J. Krause *et al*., A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr. Biol.* **20**, 231 (2010). doi:10.1016/j.cub.2009.11.068 Medline

40. S. Sawyer, J. Krause, K. Guschanski, V. Savolainen, S. Pääbo, Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE* **7**, e34131 (2012). doi:10.1371/journal.pone.0034131 Medline

41. D. Jiang, Z. Hatahet, R. J. Melamede, Y. W. Kow, S. S. Wallace, Characterization of *Escherichia coli* endonuclease VIII. *J. Biol. Chem.* **272**, 32230 (1997). doi:10.1074/jbc.272.51.32230 Medline

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

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

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

45. H. Li *et al*.; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (2009). doi:10.1093/bioinformatics/btp352 Medline

46. H. M. Cann *et al*., A human genome diversity cell line panel. *Science* **296**, 261 (2002). doi:10.1126/science.296.5566.261b Medline

47. N. Rohland, D. Reich, Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939 (2012). doi:10.1101/gr.128124.111 Medline

48. 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). doi:10.1101/gr.076554.108 Medline

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

50. G. McVicker, D. Gordon, C. Davis, P. Green, Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009). doi:10.1371/journal.pgen.1000471 Medline

51. G. McVicker, P. Green, Genomic signatures of germline gene expression. *Genome Res.* **20**, 1503 (2010). doi:10.1101/gr.106666.110 Medline

52. P. A. Fujita *et al*., The UCSC Genome Browser database: Update 2011. *Nucleic Acids Res.* **39**(Database issue), D876 (2011). doi:10.1093/nar/gkq963 Medline

53. A. Siepel *et al*., Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034 (2005). doi:10.1101/gr.3715005 Medline

54. K. R. Rosenbloom *et al*., ENCODE whole-genome data in the UCSC Genome Browser: Update 2012. *Nucleic Acids Res.* **40**(Database issue), D912 (2012). doi:10.1093/nar/gkr1012 Medline

55. R Development Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2010).

56. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009). doi:10.1186/gb-2009-10-3-r25 Medline

57. T. Lindahl, Instability and decay of the primary structure of DNA. *Nature* **362**, 709 (1993). doi:10.1038/362709a0 Medline

58. A. Hodgkinson, A. Eyre-Walker, Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12**, 756 (2011). doi:10.1038/nrg3098 Medline

59. D. C. Presgraves, S. V. Yi, Doubts about complex speciation between humans and chimpanzees. *Trends Ecol. Evol.* **24**, 533 (2009). doi:10.1016/j.tree.2009.04.007 Medline

60. A. Scally *et al*., Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169 (2012). doi:10.1038/nature10842 Medline

61. K. D. Makova, W.-H. Li, Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624 (2002). doi:10.1038/416624a Medline

62. J. E. Pool, R. Nielsen, Population size changes reshape genomic patterns of diversity. *Evolution* **61**, 3001 (2007). doi:10.1111/j.1558-5646.2007.00238.x Medline

63. L. Duret, P. F. Arndt, The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* **4**, e1000071 (2008). doi:10.1371/journal.pgen.1000071 Medline

64. 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). doi:10.1038/ng2116 Medline

65. 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). doi:10.1093/molbev/msp096 Medline

66. S. F. Schaffner *et al*., Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576 (2005). doi:10.1101/gr.3709305 Medline

67. J. L. Bischoff *et al*., High-resolution U-series dates from the Sima de los Huesos hominids yields 600(-66)(+infinity) kyrs: Implications for the 66 evolution of the early Neanderthal lineage. *J. Archaeol. Sci.* **34**, 763 (2007). doi:10.1016/j.jas.2006.08.003

68. C. Stringer, The status of *Homo heidelbergensis* (Schoetensack 1908). *Evol. Anthropol.* **21**, 101 (2012). doi:10.1002/evan.21311 Medline

69. F. M. T. A. Busing, E. Meijer, R. Van Der Leeden, Delete-m jackknife for unequal m. *Stat. Comput.* **9**, 3 (1999). doi:10.1023/A:1008800423698

70. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239 (2011). doi:10.1093/molbev/msr048 Medline

71. P. Moorjani *et al*., The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet.* **7**, e1001373 (2011). doi:10.1371/journal.pgen.1001373 Medline

72. M. T. Seielstad, E. Minch, L. L. Cavalli-Sforza, Genetic evidence for a higher female migration rate in humans. *Nat. Genet.* **20**, 278 (1998). doi:10.1038/3088 Medline

73. M. Kayser *et al*., Independent histories of human Y chromosomes from Melanesia and Australia. *Am. J. Hum. Genet.* **68**, 173 (2001). doi:10.1086/316949 Medline

74. J. F. Wilkins, F. W. Marlowe, Sex-biased migration in humans: What should we expect from genetic data? *Bioessays* **28**, 290 (2006). doi:10.1002/bies.20378 Medline

75. P. K. Tucker, R. D. Sage, J. Warner, A. C. Wilson, E. M. Eicher, Abrupt cline for sex-chromosomes in a hybrid zone between 2 species of mice. *Evolution* **46**, 1146 (1992). doi:10.2307/2409762

76. 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). doi:10.1038/nature04789 Medline

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

78. S. Gottipati, L. Arbiza, A. Siepel, A. G. Clark, A. Keinan, Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nat. Genet.* **43**, 741 (2011). doi:10.1038/ng.877 Medline

79. J. A. Bailey *et al*., Recent segmental duplications in the human genome. *Science* **297**, 1003 (2002). doi:10.1126/science.1072047 Medline

80. C. Alkan *et al*., Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061 (2009). doi:10.1038/ng.437 Medline

81. P. H. Sudmant *et al*.; 1000 Genomes Project, Diversity of human copy number variation and multicopy genes. *Science* **330**, 641 (2010). doi:10.1126/science.1197005 Medline

82. D. F. Conrad *et al*.; Wellcome Trust Case Control Consortium, Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704 (2010). doi:10.1038/nature08516 Medline

83. T. C. S. and Analysis Consortium; Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69 (2005). doi:10.1038/nature04072 Medline

84. K. Prüfer *et al*., The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**, 527 (2012). Medline

85. M. Ventura *et al*., Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res.* **21**, 1640 (2011). doi:10.1101/gr.124461.111 Medline

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

87. R. A. Sturm *et al*., A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am. J. Hum. Genet.* **82**, 424 (2008). doi:10.1016/j.ajhg.2007.11.005 Medline

88. H. Eiberg *et al*., Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum. Genet.* **123**, 177 (2008). doi:10.1007/s00439-007-0460-x Medline

89. B. K. Dennehey, D. G. Gutches, E. H. McConkey, K. S. Krauter, Inversion, duplication, and changes in gene context are associated with human chromosome 18 evolution. *Genomics* **83**, 493 (2004). doi:10.1016/j.ygeno.2003.08.017 Medline

90. V. Goidts, J. M. Szamalek, H. Hameister, H. Kehrer-Sawatzki, Segmental duplication associated with the human-specific inversion of chromosome 18: A further example of the impact of segmental duplications on karyotype and genome evolution in primates. *Hum. Genet.* **115**, 116 (2004). doi:10.1007/s00439-004-1120-z Medline

91. J. J. Yunis, O. Prakash, The origin of man: A chromosomal pictorial legacy. *Science* **215**, 1525 (1982). doi:10.1126/science.7063861 Medline

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

93. D. Karolchik *et al*., The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.* **36**(Database issue), D773 (2008). doi:10.1093/nar/gkm966 Medline

94. D. P. Howrigan, M. A. Simonson, M. C. Keller, Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics* **12**, 460 (2011). doi:10.1186/1471-2164-12-460 Medline

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

96. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987 (2011). doi:10.1093/bioinformatics/btr509 Medline

97. R. D. Hernandez *et al*.; 1000 Genomes Project, Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920 (2011). doi:10.1126/science.1198878 Medline

98. K. E. Lohmueller *et al*., Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet.* **7**, e1002326 (2011). doi:10.1371/journal.pgen.1002326 Medline

99. I. A. Adzhubei *et al*., A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248 (2010). doi:10.1038/nmeth0410-248 Medline

100. R. A. Gibbs *et al*.; Rhesus Macaque Genome Sequencing and Analysis Consortium, Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222 (2007). doi:10.1126/science.1139247 Medline

101. K. Wang, M. Li, H. Hakonarson, ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010). doi:10.1093/nar/gkq603 Medline

102. M. A. DePristo *et al*., A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491 (2011). doi:10.1038/ng.806 Medline

103. W. McLaren *et al*., Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069 (2010). doi:10.1093/bioinformatics/btq330 Medline

104. P. C. Ng, S. Henikoff, Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863 (2001). doi:10.1101/gr.176601 Medline

105. J. L. Desseyn, J. P. Aubert, N. Porchet, A. Laine, Evolution of the large secreted gel-forming mucins. *Mol. Biol. Evol.* **17**, 1175 (2000). doi:10.1093/oxfordjournals.molbev.a026400 Medline

106. Y. Niimura, M. Nei, Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. *Gene* **346**, 23 (2005). doi:10.1016/j.gene.2004.09.027 Medline

107. Q. Zhang *et al*., Ru2 and Ru encode mouse orthologs of the genes mutated in human Hermansky-Pudlak syndrome types 5 and 6. *Nat. Genet.* **33**, 145 (2003). doi:10.1038/ng1087 Medline

108. P. W. Faber *et al*., Huntingtin interacts with a family of WW domain proteins. *Hum. Mol. Genet.* **7**, 1463 (1998). doi:10.1093/hmg/7.9.1463 Medline

109. M. E. MacDonald *et al*.The Huntington's Disease Collaborative Research Group, A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971 (1993). doi:10.1016/0092-8674(93)90585-E Medline

110. L. A. Durfee, N. Lyon, K. Seo, J. M. Huibregtse, The ISG15 conjugation system broadly targets newly synthesized proteins: Implications for the antiviral function of ISG15. *Mol. Cell* **38**, 722 (2010). doi:10.1016/j.molcel.2010.05.002 Medline

111. Y. Tang *et al*., Herc5 attenuates influenza A virus by catalyzing ISGylation of viral NS1 protein. *J. Immunol.* **184**, 5777 (2010). doi:10.4049/jimmunol.0903588 Medline

112. J. B. Wang *et al*., Human μ opiate receptor. cDNA and genomic clones, pharmacologic characterization and chromosomal assignment. *FEBS Lett.* **338**, 217 (1994). doi:10.1016/0014-5793(94)80368-4 Medline

113. E. E. Quillen *et al*., OPRM1 and EGFR contribute to skin pigmentation differences between Indigenous Americans and Europeans. *Hum. Genet.* **131**, 1073 (2012). doi:10.1007/s00439-011-1135-1 Medline

114. L. B. Barreiro, G. Laval, H. Quach, E. Patin, L. Quintana-Murci, Natural selection has driven population differentiation in modern humans. *Nat. Genet.* **40**, 340 (2008). doi:10.1038/ng.78 Medline

115. C. T. Dolphin, E. A. Shephard, S. Povey, R. L. Smith, I. R. Phillips, Cloning, primary sequence and chromosomal localization of human FMO2, a new member of the flavin-containing mono-oxygenase family. *Biochem. J.* **287**, 261 (1992). Medline

116. C. T. Dolphin *et al*., The flavin-containing monooxygenase 2 gene (FMO2) of humans, but not of other primates, encodes a truncated, nonfunctional protein. *J. Biol. Chem.* **273**, 30599 (1998). doi:10.1074/jbc.273.46.30599 Medline

117. M. Saleh *et al*., Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature* **429**, 75 (2004). doi:10.1038/nature02451 Medline

118. Y. Xue *et al*., Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am. J. Hum. Genet.* **78**, 659 (2006). doi:10.1086/503116 Medline

119. K. Prüfer *et al*., FUNC: A package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* **8**, 41 (2007). doi:10.1186/1471-2105-8-41 Medline

120. P. Kumar, S. Henikoff, P. C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073 (2009). doi:10.1038/nprot.2009.86 Medline

121. M. Galdzicka *et al*., A new gene, EVC2, is mutated in Ellis-van Creveld syndrome. *Mol. Genet. Metab.* **77**, 291 (2002). doi:10.1016/S1096-7192(02)00178-6 Medline

122. V. L. Ruiz-Perez *et al*., Mutations in two nonhomologous genes in a head-to-head configuration cause Ellis-van Creveld syndrome. *Am. J. Hum. Genet.* **72**, 728 (2003). doi:10.1086/368063 Medline

123. W. Shen, D. Han, J. Zhang, H. Zhao, H. Feng, Two novel heterozygous mutations of EVC2 cause a mild phenotype of Ellis-van Creveld syndrome in a Chinese family. *Am. J. Med. Genet. A.* **155A**, 2131 (2011). doi:10.1002/ajmg.a.34125 Medline

124. H. Brkić, I. Filipović, [The meaning of taurodontism in oral surgery—case report]. *Acta Stomatol. Croat.* **25**, 123 (1991). Medline

125. B. W. Neville, D. D. Damm, C. M. Allen, J. E. Bouquot, *Oral & Maxillofacial Pathology* (Saunders, Philadelphia, ed. 5, 2002).

126. H. Jafarzadeh, A. Azarpazhooh, J. T. Mayhall, Taurodontism: A review of the condition and endodontic treatment challenges. *Int. Endod. J.* **41**, 375 (2008). doi:10.1111/j.1365-2591.2008.01388.x Medline

127. B. C. Barker, Taurodontism: The incidence and possible significance of the trait. *Aust. Dent. J.* **21**, 272 (1976). doi:10.1111/j.1834-7819.1976.tb05763.x Medline

128. M. Glancy *et al*., Transmitted duplication of 8p23.1-8p23.2 associated with speech delay, autism and learning difficulties. *Eur. J. Hum. Genet.* **17**, 37 (2009). doi:10.1038/ejhg.2008.133 Medline

129. R. J. Gibbons, G. K. Suthers, A. O. Wilkie, V. J. Buckle, D. R. Higgs, X-linked alpha-thalassemia/mental retardation (ATR-X) syndrome: Localization to Xq12-q21.31 by X inactivation and linkage analysis. *Am. J. Hum. Genet.* **51**, 1136 (1992). Medline

130. M. Alarcón *et al*., Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am. J. Hum. Genet.* **82**, 150 (2008). doi:10.1016/j.ajhg.2007.09.005 Medline

131. D. E. Arking *et al*., A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *Am. J. Hum. Genet.* **82**, 160 (2008). doi:10.1016/j.ajhg.2007.09.015 Medline

132. C. S. Lai, S. E. Fisher, J. A. Hurst, F. Vargha-Khadem, A. P. Monaco, A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**, 519 (2001). doi:10.1038/35097076 Medline

133. K. A. Strauss *et al*., Recessive symptomatic focal epilepsy and mutant contactin-associated protein-like 2. *N. Engl. J. Med.* **354**, 1370 (2006). doi:10.1056/NEJMoa052773 Medline

134. J. Jaeken, G. Van den Berghe, An infantile autistic syndrome characterised by the presence of succinylpurines in body fluids. *Lancet* **2**, 1058 (1984). Medline

135. R. L. Stone *et al*., A mutation in adenylosuccinate lyase associated with mental retardation and autistic features. *Nat. Genet.* **1**, 59 (1992). doi:10.1038/ng0492-59 Medline

136. P. D. Stenson *et al*., The Human Gene Mutation Database: Providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum. Genomics* **4**, 69 (2009). doi:10.1186/1479-7364-4-2-69 Medline

137. K. Matsumura, J. M. Ervasti, K. Ohlendieck, S. D. Kahl, K. P. Campbell, Association of dystrophin-related protein with dystrophin-associated proteins in *mdx* mouse muscle. *Nature* **360**, 588 (1992). doi:10.1038/360588a0 Medline

138. S. L. Roberds *et al*., Missense mutations in the adhalin gene linked to autosomal recessive muscular dystrophy. *Cell* **78**, 625 (1994). doi:10.1016/0092-8674(94)90527-4 Medline

139. J. M. Ervasti, K. P. Campbell, Membrane organization of the dystrophin-glycoprotein complex. *Cell* **66**, 1121 (1991). doi:10.1016/0092-8674(91)90035-W Medline

140. S. Stenirri *et al*., Denaturing HPLC profiling of the ABCA4 gene for reliable detection of allelic variations. *Clin. Chem.* **50**, 1336 (2004). doi:10.1373/clinchem.2004.033241 Medline

141. Y. Liu *et al*., The human inward rectifier K^+ channel subunit kir5.1 (KCNJ16) maps to chromosome 17q25 and is expressed in kidney and pancreas. *Cytogenet. Cell Genet.* **90**, 60 (2000). doi:10.1159/000015662 Medline

142. D. A. Ellerman *et al*., Izumo is part of a multiprotein family whose members form large complexes on mammalian sperm. *Mol. Reprod. Dev.* **76**, 1188 (2009). doi:10.1002/mrd.21092 Medline

143. K. R. Long, J. A. Trofatter, V. Ramesh, M. K. McCormick, A. J. Buckler, Cloning and characterization of a novel human clathrin heavy chain gene (CLTCL). *Genomics* **35**, 466 (1996). doi:10.1006/geno.1996.0386 Medline

144. C. Desmaze *et al*., Physical mapping by FISH of the DiGeorge critical region (DGCR): Involvement of the region in familial cases. *Am. J. Hum. Genet.* **53**, 1239 (1993). Medline

145. G. A. Wray, The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8**, 206 (2007). doi:10.1038/nrg2063 Medline

146. J. D. Gruber, K. Vogel, G. Kalay, P. J. Wittkopp, Contrasting properties of gene-specific regulatory, coding, and copy number mutations in *Saccharomyces cerevisiae*: Frequency, effects, and dominance. *PLoS Genet.* **8**, e1002497 (2012). doi:10.1371/journal.pgen.1002497 Medline

147. L. P. Lim *et al*., Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769 (2005). doi:10.1038/nature03315 Medline

148. M. Selbach *et al*., Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58 (2008). doi:10.1038/nature07228 Medline

149. A. E. Pasquinelli *et al*., Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86 (2000). doi:10.1038/35040556 Medline

150. M. Somel *et al*., MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. *PLoS Biol.* **9**, e1001214 (2011). doi:10.1371/journal.pbio.1001214 Medline

151. R. D. Morin *et al*., Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.* **18**, 610 (2008). doi:10.1101/gr.7179508 Medline

152. J. C. Bryne *et al*., JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update. *Nucleic Acids Res.* **36**(Database issue), D102 (2008). doi:10.1093/nar/gkm955 Medline

153. L. Bao, M. Zhou, Y. Cui, CTCFBSDB: A CTCF-binding site database for characterization of vertebrate genomic insulators. *Nucleic Acids Res.* **36**(Database issue), D83 (2008). doi:10.1093/nar/gkm875 Medline

154. J. R. Raab, R. T. Kamakaka, Insulators and promoters: Closer than we think. *Nat. Rev. Genet.* **11**, 439 (2010). doi:10.1038/nrg2765 Medline

155. J. Hardy, A. Singleton, Genomewide association studies and human disease. *N. Engl. J. Med.* **360**, 1759 (2009). doi:10.1056/NEJMra0808700 Medline

156. T. A. Manolio, Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363**, 166 (2010). doi:10.1056/NEJMra0905980 Medline

157. K. G. Becker, K. C. Barnes, T. J. Bright, S. A. Wang, The genetic association database. *Nat. Genet.* **36**, 431 (2004). doi:10.1038/ng0504-431 Medline

158. M. Cariaso, G. Lennon, SNPedia: A wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res.* **40**(Database issue), D1308 (2012). doi:10.1093/nar/gkr798 Medline

159. R. J. Buckanovich, Y. Y. Yang, R. B. Darnell, The onconeural antigen Nova-1 is a neuron-specific RNA-binding protein, the activity of which is inhibited by paraneoplastic antibodies. *J. Neurosci.* **16**, 1114 (1996). Medline

160. J. A. Park, K. C. Kim, Expression patterns of PRDM10 during mouse embryonic development. *BMB Rep.* **43**, 29 (2010). doi:10.5483/BMBRep.2010.43.1.029 Medline

161. H. Lohi *et al*., Functional characterization of three novel tissue-specific anion exchangers SLC26A7, -A8, and -A9. *J. Biol. Chem.* **277**, 14246 (2002). doi:10.1074/jbc.M111802200 Medline

162. L. Jin, A. Williamson, S. Banerjee, I. Philipp, M. Rape, Mechanism of ubiquitin-chain formation by the human anaphase-promoting complex. *Cell* **133**, 653 (2008). doi:10.1016/j.cell.2008.04.012 Medline

163. M. I. Mayr *et al*., The human kinesin Kif18A is a motile microtubule depolymerase essential for chromosome congression. *Curr. Biol.* **17**, 488 (2007). doi:10.1016/j.cub.2007.02.036 Medline

164. J. F. Abelson *et al*., Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* **310**, 317 (2005). doi:10.1126/science.1116502 Medline

165. C. Y. Hsu *et al*., LUZP deficiency affects neural tube closure during brain development. *Biochem. Biophys. Res. Commun.* **376**, 466 (2008). doi:10.1016/j.bbrc.2008.08.170 Medline

166. V. P. Pisareva, A. V. Pisarev, A. A. Komar, C. U. Hellen, T. V. Pestova, Translation initiation on mammalian mRNAs with structured 5'UTRs requires DExH-box protein DHX29. *Cell* **135**, 1237 (2008). doi:10.1016/j.cell.2008.10.037 Medline

167. T. Okabe *et al*., RICS, a novel GTPase-activating protein for Cdc42 and Rac1, is involved in the beta-catenin-N-cadherin and N-methyl-D-aspartate receptor signaling. *J. Biol. Chem.* **278**, 9920 (2003). doi:10.1074/jbc.M208872200 Medline

168. L. Bevilacqua *et al*., A population-specific HTR2B stop codon predisposes to severe impulsivity (Erratum re: vol 468, pg 1061, 2010). *Nature* **470**, 424 (2011). doi:10.1038/nature09808

169. S. Doly *et al*., Serotonin 5-HT2B receptors are required for 3,4-methylenedioxymethamphetamine-induced hyperlocomotion and 5-HT release in vivo and in vitro. *J. Neurosci.* **28**, 2933 (2008). doi:10.1523/JNEUROSCI.5723-07.2008 Medline

170. D. W. Speicher, L. Weglarz, T. M. DeSilva, Properties of human red cell spectrin heterodimer (side-to-side) assembly and identification of an essential nucleation site. *J. Biol. Chem.* **267**, 14775 (1992). Medline

171. S. S. Jones, A. D. D'Andrea, L. L. Haines, G. G. Wong, Human erythropoietin receptor: Cloning, expression, and biologic characterization. *Blood* **76**, 31 (1990). Medline

172. G. D. Longmore, H. F. Lodish, An activating mutation in the murine erythropoietin receptor induces erythroleukemia in mice: a cytokine receptor superfamily oncogene. *Cell* **67**, 1089 (1991). doi:10.1016/0092-8674(91)90286-8 Medline

173. X. Yu, C. S. Lin, F. Costantini, C. T. Noguchi, The human erythropoietin receptor gene rescues erythropoiesis and developmental defects in the erythropoietin receptor null mouse. *Blood* **98**, 475 (2001). doi:10.1182/blood.V98.2.475 Medline

174. D. C. Rubinsztein, D. F. Easton, Apolipoprotein E genetic variation and Alzheimer's disease. a meta-analysis. *Dement. Geriatr. Cogn. Disord.* **10**, 199 (1999). doi:10.1159/000017120 Medline

175. T. Pohjalainen *et al*., The A1 allele of the human D2 dopamine receptor gene predicts low D2 receptor availability in healthy volunteers. *Mol. Psychiatry* **3**, 256 (1998). doi:10.1038/sj.mp.4000350 Medline

176. M. Lucht *et al*., Influence of DRD2 and ANKK1 genotypes on apomorphine-induced growth hormone (GH) response in alcohol-dependent patients. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **34**, 45 (2010). doi:10.1016/j.pnpbp.2009.08.024 Medline

177. S. P. David *et al*., Bupropion efficacy for smoking cessation is influenced by the DRD2 Taq1A polymorphism: Analysis of pooled data from two clinical trials. *Nicotine Tob. Res.* **9**, 1251 (2007). doi:10.1080/14622200701705027 Medline

178. P. Burton *et al*.; Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661 (2007). doi:10.1038/nature05911 Medline

179. T. M. Frayling *et al*., A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889 (2007). doi:10.1126/science.1141634 Medline

180. V. Lyssenko *et al*., Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *J. Clin. Invest.* **117**, 2155 (2007). doi:10.1172/JCI30706 Medline

181. A. Heinz, M. N. Smolka, The effects of catechol *O*-methyltransferase genotype on brain activation elicited by affective stimuli and cognitive tasks. *Rev. Neurosci.* **17**, 359 (2006). doi:10.1515/REVNEURO.2006.17.3.359 Medline

182. M. J. Frank, B. B. Doll, J. Oas-Terpstra, F. Moreno, Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nat. Neurosci.* **12**, 1062 (2009). doi:10.1038/nn.2342 Medline

183. H. Pálmason *et al*., Attention-deficit/hyperactivity disorder phenotype is influenced by a functional catechol-*O*-methyltransferase variant. *J. Neural Transm.* **117**, 259 (2010). doi:10.1007/s00702-009-0338-2 Medline

184. M. Gupta *et al*., Genetic susceptibility to schizophrenia: role of dopaminergic pathway gene polymorphisms. *Pharmacogenomics* **10**, 277 (2009). doi:10.2217/14622416.10.2.277 Medline

185. D. L. Duffy *et al*., A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am. J. Hum. Genet.* **80**, 241 (2007). doi:10.1086/510885 Medline

186. P. A. Frändberg, M. Doufexis, S. Kapas, V. Chhájlani, Human pigmentation phenotype: A point mutation generates nonfunctional MSH receptor. *Biochem. Biophys. Res. Commun.* **245**, 490 (1998). doi:10.1006/bbrc.1998.8459 Medline

187. K.-i. Yoshiura *et al*., A SNP in the ABCC11 gene is the determinant of human earwax type. *Nat. Genet.* **38**, 324 (2006). doi:10.1038/ng1733 Medline

188. N. S. Enattah *et al*., Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233 (2002). doi:10.1038/ng826 Medline

189. T. Bersaglieri *et al*., Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111 (2004). doi:10.1086/421051 Medline

190. C. A. Mulcare *et al*., The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. *Am. J. Hum. Genet.* **74**, 1102 (2004). doi:10.1086/421050 Medline

191. S. A. Tishkoff *et al*., Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**, 31 (2007). doi:10.1038/ng1946 Medline

192. C. Mou *et al*., Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form. *Hum. Mutat.* **29**, 1405 (2008). doi:10.1002/humu.20795 Medline

193. R. Kimura *et al*., A common variation in EDAR is a genetic determinant of shovel-shaped incisors. *Am. J. Hum. Genet.* **85**, 528 (2009). doi:10.1016/j.ajhg.2009.09.006 Medline

194. S. M. Rodrigues, L. R. Saslow, N. Garcia, O. P. John, D. Keltner, Oxytocin receptor genetic variation relates to empathy and stress reactivity in humans. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 21437 (2009). doi:10.1073/pnas.0909579106 Medline

195. S. M. Roth *et al*., The ACTN3 R577X nonsense allele is under-represented in elite-level strength athletes. *Eur. J. Hum. Genet.* **16**, 391 (2008). doi:10.1038/sj.ejhg.5201964 Medline

196. J. A. Ellis, M. Stebbing, S. B. Harrap, Polymorphism of the androgen receptor gene is associated with male pattern baldness. *J. Invest. Dermatol.* **116**, 452 (2001). doi:10.1046/j.1523-1747.2001.01261.x Medline