Patterns of coding variation in the complete exomes of three Neandertals

Sergi Castellano[a,1], Genís Parra[a,2], Federico A. Sánchez-Quinto[b,2], Fernando Racimo[a,c,2], Martin Kuhlwilm[a,2], Martin Kircher[a,d], Susanna Sawyer[a], Qiaomei Fu[a,e], Anja Heinze[a], Birgit Nickel[a], Jesse Dabney[a], Michael Siebauer[a], Louise White[a], Hernán A. Burbano[a,f], Gabriel Renaud[a], Udo Stenzel[a], Carles Lalueza-Fox[b], Marco de la Rasilla[g], Antonio Rosas[h], Pavao Rudan[i], Dejana Brajković[j], Željko Kucan[i], Ivan Gušic[j], Michael V. Shunkov[k], Anatoli P. Derevianko[k], Bence Viola[a,l], Matthias Meyer[a], Janet Kelso[a], Aida M. Andrés[a], and Svante Pääbo[a,1]

[a]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany; [b]Institute of Evolutionary Biology, Consejo Superior de Investigaciones Científicas, Universitat Pompeu Fabra, 08003 Barcelona, Spain; [c]Department of Integrative Biology, University of California, Berkeley, CA 94720; [d]Department of Genome Sciences, University of Washington, Seattle, WA 98195; [e]Key Laboratory of Vertebrate Evolution and Human Origins of the Chinese Academy of Sciences, Institute of Vertebrate Palaeontology and Palaeoanthropology, Chinese Academy of Sciences, Beijing 100044, China; [f]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tuebingen, Germany; [g]Área de Prehistoria, Departamento de Historia, Universidad de Oviedo, 33011 Oviedo, Spain; [h]Departamento de Paleobiología, Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, 28006 Madrid, Spain; [i]Anthropology Center of the Croatian Academy of Sciences and Arts, HR-10000 Zagreb, Croatia; [j]Croatian Academy of Sciences and Arts, Institute for Quaternary Paleontology and Geology, HR-10000 Zagreb, Croatia; [k]Paleolithic Department, Institute of Archeology and Ethnography, Russian Academy of Sciences, Siberian Branch, Novosibirsk 630090, Russia; and [l]Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany

Contributed by Svante Pääbo, March 21, 2014 (sent for review January 17, 2014)

We present the DNA sequence of 17,367 protein-coding genes in two Neandertals from Spain and Croatia and analyze them together with the genome sequence recently determined from a Neandertal from southern Siberia. Comparisons with present-day humans from Africa, Europe, and Asia reveal that genetic diversity among Neandertals was remarkably low, and that they carried a higher proportion of amino acid-changing (nonsynonymous) alleles inferred to alter protein structure or function than present-day humans. Thus, Neandertals across Eurasia had a smaller long-term effective population than present-day humans. We also identify amino acid substitutions in Neandertals and present-day humans that may underlie phenotypic differences between the two groups. We find that genes involved in skeletal morphology have changed more in the lineage leading to Neandertals than in the ancestral lineage common to archaic and modern humans, whereas genes involved in behavior and pigmentation have changed more on the modern human lineage.

ancient DNA | exome capture | site frequency spectra | paleogenetics

Analyses of the coding regions of multiple present-day human individuals have uncovered many amino acid-changing SNPs segregating at low frequency in present-day human populations (1–7). It also has been shown that Europeans carry a larger proportion than Africans of amino acid-changing SNPs inferred to alter the structure or function of proteins and thus to be potentially deleterious (4), a fact attributed to the bottleneck and subsequent expansion of human populations during and after the migration out of Africa (4). In contrast, little is known about the coding variation in Neandertals, an extinct hominin group closely related to present-day humans. The main reasons for this are the rarity of Neandertal remains and the fact that >99% of the DNA extracted from typical Neandertal bones is derived from microbes (8, 9), making shotgun sequencing of the endogenous DNA impractical.

Here, we use a hybridization approach to enrich and sequence the protein-coding fractions of the genomes of two Neandertals from Spain and Croatia. We analyze them together with the genome sequence recently determined from a Neandertal from southern Siberia (10) and show that the genetic diversity, pattern of coding variation, and genes that may underlie phenotypic changes in Neandertals are remarkably different from those in present-day humans.

Results and Discussion

We used densely tiled oligonucleotide probes (9) and a recently described protocol (11) to capture the protein-coding exons from 17,367 genes in a ∼49,000-y-old (12) (uncalibrated radiocarbon date) Neandertal from Spain (SD1253, El Sidrón Cave) and a ∼44,000-y-old (uncalibrated radiocarbon date) Neandertal from Croatia (Vi33.15, Vindija Cave). The DNA libraries from these two Neandertals contain 0.2% and 0.5% endogenous DNA, respectively, and the capture approach enriched the endogenous DNA 325-fold and 153-fold, resulting in an average coverage of 12.5-fold and 42.0-fold for the El Sidrón and Vindija exomes, respectively (Table 1). Present-day human mitochondrial contamination in the enriched libraries is estimated to be between 0.24% [confidence interval (CI): 0.23–0.24%] and 0.40% (CI: 0.38–0.42%) for El Sidrón and between 0.28% (CI: 0.27–0.28%) and 1.08% (CI: 1.05–1.08%) for Vindija. Before genotypes were called, we lowered the base quality scores toward the ends of the Neandertal DNA sequences to compensate for substitutions potentially due to cytosine deaminations (*Methods* and Fig. S2).

Significance

We use a hybridization approach to enrich the DNA from the protein-coding fraction of the genomes of two Neandertal individuals from Spain and Croatia. By analyzing these two exomes together with the genome sequence of a Neandertal from Siberia we show that the genetic diversity of Neandertals was lower than that of present-day humans and that the pattern of coding variation suggests that Neandertal populations were small and isolated from one another. We also show that genes involved in skeletal morphology have changed more than expected on the Neandertal evolutionary lineage whereas genes involved in pigmentation and behavior have changed more on the modern human lineage.

Author contributions: S.C., F.R., M.M., J.K., A.M.A., and S.P. designed research; S.C., G.P., F.A.S.-Q., F.R., M. Kuhlwilm, S.S., Q.F., A.H., B.N., J.D., L.W., B.V., M.M., and A.M.A. performed research; S.S., Q.F., A.H., B.N., J.D., H.A.B., C.L.-F., M.d.l.R., A.R., P.R., D.B., Z.K., I.G., M.V.S., A.P.D., and M.M. contributed new reagents/analytic tools; S.C., G.P., F.A.S.-Q., F.R., M. Kuhlwilm, M. Kircher, M.S., G.R., and U.S. analyzed data; and S.C., F.A.S.-Q., F.R., B.V., J.K., A.M.A., and S.P. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The exome sequence capture data reported in this paper have been deposited in the European Nucleotide Archive (accession no. ERP002457). These data are also available at http://cdna.eva.mpg.de/neandertal/exomes/.

[1]To whom correspondence may be addressed. E-mail: sergi.castellano@eva.mpg.de or paabo@eva.mpg.de.

[2]G.P., F.A.S.-Q., F.R., and M. Kuhlwilm contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1405138111/-/DCSupplemental.

Table 1. Coverage comparison of the protein-coding regions in Neandertals

	Draft genome (8)	El Sidrón	Vindija	Altai genome (10)
Average coverage	1.3-fold	12.5-fold	42.0-fold	45.3-fold
Coding bases, *n* (%)*	14,289,004 (52.3)	25,651,281 (93.8)	26,876,083 (98.3)	27,313,554 (99.9)
Coding genes, *n* (%)†				
80%	177 (1.0)	15,502 (89.3)	16,944 (97.6)	17,322 (99.7)
90%	46 (0.3)	13,561 (78.1)	16,294 (93.8)	17,313 (99.7)
100%	19 (0.1)	5,476 (31.5)	11,696 (67.3)	17,264 (99.4)

*Number and percentage of bases in the exome covered by at least one read.
†Number and percentage of genes with at least 80%, 90%, and 100% of their length covered by one read.

We also retrieved the exomes from the genome sequences of a Neandertal (10) (Table 1) and a Denisovan (13) individual, both from Denisova Cave in the Altai Mountains, Siberia, as well as from nine present-day humans (Yoruba, Mandenka, and Dinka from Africa; French, Sardinian, and Italian American from Europe; Han and Dai from Asia; and Papuan from Oceania), which have been sequenced to a quality similar to that of the two archaic genomes (13, 14) (*SI Appendix* and Table S7). A neighbor-joining tree based on pairwise differences (Fig. 1*A*) shows that the three Neandertals form a clade, whereas the Denisovan individual is a sister group to the Neandertals, in agreement with previous results (8, 13, 15).

As expected for closely related groups, a substantial amount of the variation is shared between Neandertals and present-day humans. Thus, about 33% of the derived alleles seen in the Neandertals are shared with the present-day humans and about 19%, 24%, and 24% of the derived alleles seen in the Africans, Europeans, and Asians, respectively, are shared with the Neandertals (Table S14). The number of nucleotide differences within an individual (heterozygosity) is 0.143 per thousand coding bases in the El Sidrón Neandertal, 0.127 in the Vindija Neandertal, and 0.113 in the Altai Neandertal. The average heterozygosity among the three Neandertals (0.128) is 25%, 33%, and 36% of the average heterozygosity in the three African (0.507), European (0.387), and Asian (0.358) individuals, respectively (Table S12), and thus significantly smaller than that in humans today ($P = 4.5 \times 10^{-3}$; Mann–Whitney *U* test). The three Neandertals also have longer runs of homozygosity (Fig. S9), suggesting that mating among related individuals may have been more common in Neandertals than in present-day humans, compatible with the observations in the Altai Neandertal (10). Also compatible with these observations is that the genetic differentiation among individuals, as measured by the pairwise fixation index (F_{ST}) (Fig. 1*B*), is greater among Neandertals than among the Africans, Europeans, and Asians ($P = 1.3 \times 10^{-2}$; Mann–Whitney *U* test), suggesting that mating in small and isolated populations may have caused Neandertal populations to be more differentiated from one another than what is typical for present-day humans.

To investigate the extent to which the population history reflected in the low genetic diversity affected the evolution of coding regions, we investigated derived alleles present in Neandertals and compared them with those of the present-day Africans, Europeans, and Asians (Table 2). We found that the fraction of all derived SNPs inferred to change amino acids in Neandertals (51.1%) is larger than that in Africans (45.7%; $P = 3.1 \times 10^{-11}$; G-test), Europeans (46.4%; $P = 6.9 \times 10^{-10}$), and Asians (46.7%; $P = 4.5 \times 10^{-7}$) (Table 2). We used PolyPhen-2 (16) to assess whether the amino acid-changing SNPs in Neandertals are enriched in alleles inferred to affect protein function or structure and thus likely to be slightly deleterious. The proportion of such SNPs in the Neandertals (45.4%) is larger than that in Africans (36.4%; G-test; $P = 1.6 \times 10^{-14}$), Europeans (37.9%; $P = 4.5 \times 10^{-9}$), and Asians (38.1%; $P = 1.3 \times 10^{-9}$) (Table 2). Similarly, the proportion of such SNPs in conserved protein sites [PhastCons (17)] is larger in the Neandertals (50.1%) than in the Africans (38.1%; G-test; $P < 10^{-20}$), Europeans (40.2%; $P = 1.1 \times 10^{-15}$), and Asians (39.7%; $P < 10^{-20}$) (Table 2). Furthermore, a physicochemical classification of amino acid substitutions, the Grantham score (GS) (18), shows that the proportion of derived SNPs in the Neandertals that cause radical (GS >100) amino acid changes is larger than that in present-day humans (Table 2). Derived alleles in conserved protein sites that are homozygous in all individuals analyzed also are more frequent in Neandertals (39.6%) than in Africans (G-test; 29.4%; $P = 2.1 \times 10^{-3}$), Europeans (27.8%; $P = 1.6 \times 10^{-5}$), and Asians (26.9%; $P = 5.9 \times 10^{-6}$) (Table 2).

Although only six Neandertal chromosomes are available, we note that among derived amino acid-changing alleles likely to have been at low frequency in Neandertals because they are seen once among the six chromosomes (Fig. 2), a higher proportion is inferred to alter protein structure or function in Neandertals (49.1% by PolyPhen-2; similar results with PhastCons) than in Africans (G-test; 38.3%; $P = 2.2 \times 10^{-12}$), Europeans (40.6%; $P = 8.8 \times 10^{-7}$), and Asians (41.8%; $P = 5.1 \times 10^{-6}$). Furthermore, we note that these low-frequency, putatively deleterious alleles have a stronger inferred impact on proteins in Neandertals than in Africans ($P < 2.2 \times 10^{-16}$), Europeans ($P = 6.7 \times 10^{-9}$), or Asians ($P = 2.0 \times 10^{-7}$) (Fig. 2 and *SI Appendix*).

When analyzed individually, each of the three Neandertal individuals carries a larger fraction of derived amino acid-changing alleles ($P = 7.6 \times 10^{-3}$ for SNPs; $P = 7.0 \times 10^{-3}$ for homozygous; Mann–Whitney *U* test) as well as alleles inferred to change protein function (PhastCons; $P = 8.0 \times 10^{-3}$ for heterozygous, $P = 8.0 \times 10^{-3}$ for homozygous) than any of the present-day individuals analyzed (Tables S15 and S16). Thus, not only Neandertals as a group, but also each Neandertal individual analyzed, carried a larger fraction of putatively deleterious alleles than present-day humans. This also is the case for the Denisovan individual, who represents a little-known Asian population related to Neandertals.

We conclude that the patterns of coding variation in Neandertals differ strongly from those in present-day humans, a fact likely to be the result of differences in their demographic histories. Inferences from the complete Neandertal genome suggest that sometime after 0.5–1.0 Mya, the ancestral Neandertal population decreased in size, whereas the ancestors of present-day humans increased in number (10). A low population size over a long time would reduce the efficacy of purifying selection and contribute to a larger fraction of slightly deleterious alleles, particularly at low frequency, but not necessarily to the absolute number of slightly deleterious alleles per individual (19) (*SI Appendix*). The fact that the three Neandertals carry longer tracts of homozygosity and differ more from one another than present-day humans within continents further suggests that Neandertals may have been subdivided in small local populations. However, we note that the three Neandertal individuals analyzed are not contemporaneous but differ in time by perhaps as much as 20,000 y. It therefore is crucial for DNA sequences from more Neandertals to be generated when suitable specimens become available. It is also crucial for more complex demographic scenarios, for example involving recurrent population reductions and expansions, to be

explored to fully understand the differences in population history between Neandertals and modern humans.

We previously used the Neandertal and Denisova genomes (8, 10, 13) to identify amino acid changes that are shared by all humans today but do not occur in the archaic genomes and therefore are inferred to have risen to high frequency on the modern human lineage. The exomes of the three Neandertals and the Denisovan individual allow us, for the first time, to identify derived amino acid changes shared by the three Neandertals as well as the Denisovan individual that are not seen, or occur at low frequency, in present-day humans. Such changes are of interest because they may underlie phenotypes specific to the archaic populations. We calculated the fraction of all amino acid changes specific to either the archaic or present-day human lineages for each phenotype category of genes in the Human Phenotype Ontology (20), and compared it with the fraction of all amino acid changes in the same phenotype categories that occurred on the human lineage after its divergence from the chimpanzee lineage but before the divergence of modern and archaic humans (*SI Appendix*). Thus, unlike in previous approaches (8, 10, 13), we assessed the enrichment of amino acid changes in phenotypes across multiple archaic individuals while correcting for the size and base composition of genes across the genome (*Methods* and *SI Appendix*).

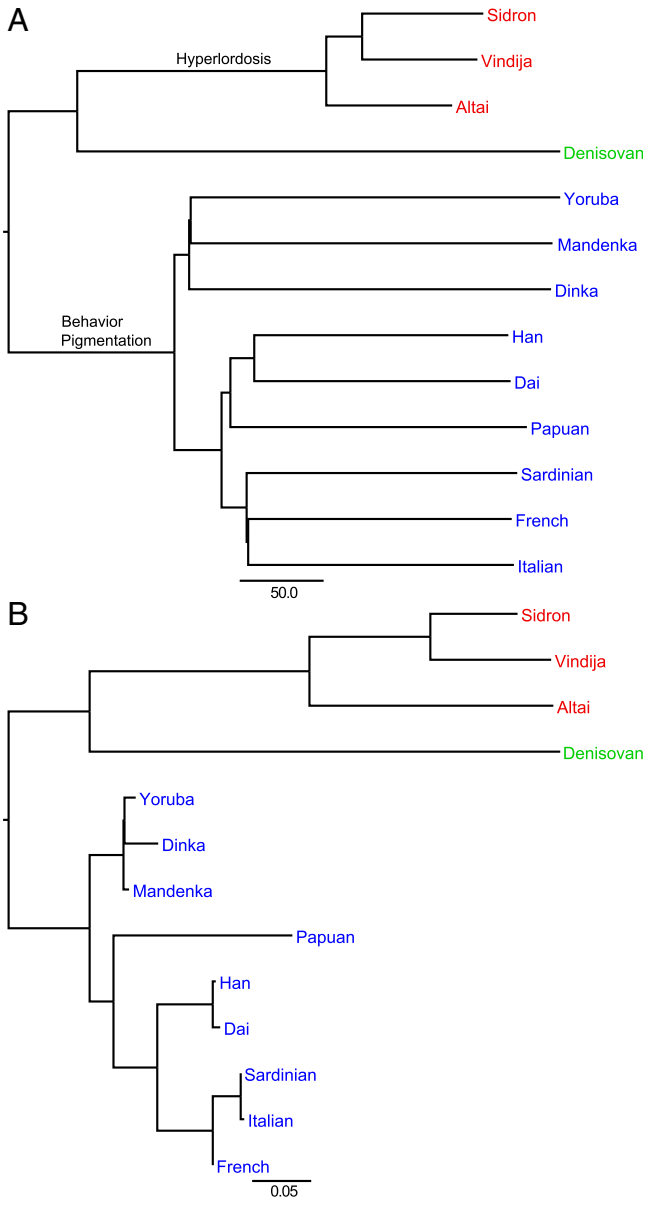


Fig. 1. (*A*) Neighbor-joining tree based on the number of pairwise differences between individuals. Phenotypic categories enriched in amino acid-changing substitutions on the lines to Neandertals and to modern humans, respectively, are indicated. The bar indicates 50 inferred substitutions per megabase. (*B*) Neighbor-joining tree of pairwise F_{ST} values for the same individuals. The bar indicates an F_{ST} of 0.05.

For the archaic lineage before the divergence of Neandertals and Denisovans, we find an overrepresentation of changes in genes related to metabolism, the cardiovascular system, hair distribution, and morphology (genitals, palate, face, extremities, joints, digits, thorax, orbital, and occipital skull regions and general mobility) [Table S28; binomial test; *P* values and false discovery rates (FDRs) in the table]. Thus, it may be that future functional investigations of the changes in the 22 genes identified in enriched categories (Table S29) will reveal the genetic basis for some features shared by Neandertals and Denisovans, although almost no morphologic information currently is available for the latter (15).

In the Neandertal lineage after the split from Denisovans, only the phenotype category "hyperlordosis" is overrepresented (Fig. 1*A*; binomial test; $P = 4.5 \times 10^{-5}$; FDR $= 2.1 \times 10^{-2}$). This is interesting because Neandertals had a reduced lordotic curvature relative to other archaic hominins as well as to modern humans (21). Changes in the six genes associated with this phenotype (*CUL7*, *GLB1*, *NEB*, *COL2A1*, *HSPG2*, and *VPS13B*) thus may have contributed to this feature of the Neandertal morphology.

Finally, the modern human lineage has an overrepresentation of amino acid-changing substitutions in genes related to pigmentation (Fig. 1*A*; $P = 3.2 \times 10^{-5}$; FDR $= 5.9 \times 10^{-2}$) and behavioral traits (Fig. 1*A*; $P < 5 \times 10^{-4}$; FDR $< 7.8 \times 10^{-2}$). The overrepresentation in the pigmentation category is the result of amino acid substitutions in two genes involved in melanosome function (*GPR143* and *LYST*), for which the derived alleles are at high frequency in Africa and fixed outside Africa. These mutations may contribute to differences in pigmentation among present-day humans (22, 23). In contrast, the behavioral category is the result of substitutions that seem to be fixed in all present-day humans and occur in the three genes adenylosuccinate lyase (*ADSL*), glycine dehydrogenase (*GLDC*), and transmembrane protein SLITRK1. These genes are associated with de novo purine biosynthesis, mitochondrial glycine degradation, and neurite outgrowth, respectively. In the case of *ADSL*, mutations known to impair the gene result in varying degrees of psychomotor retardation, autism, and muscle wasting, whereas in the case of *GLDC*, mutations result in glycine encephalopathy and in the case of *SLITRK1*, in Tourette syndrome, characterized by motoric and vocal ticks. In the Human Phenotype Ontology (20), all these genes are annotated as being involved in "hyperactivity" and "aggressive behavior." Thus, it is possible that the mutations in these genes may have affected some component of behavior unique to modern humans. However, we note that even if they did, the way in which they occurred is unknown. For example, if they affected activity or aggression levels, it is unclear whether they increased or decreased such traits. In any event, these changes clearly warrant further investigation.

In conclusion, we show that the genetic diversity of late Neandertals across their entire geographical range was lower than that of present-day humans and that there is evidence suggesting that Neandertal populations may have been small and isolated from one another. We also show that genes involved in morphology may have changed more on the Neandertal and Denisova lineages than on the preceding lineage from the common ancestor shared with chimpanzees, whereas genes involved in behavior may have changed more on the modern human lineage.

Methods

DNA Extraction and Library Preparation. We prepared DNA extracts from two Neandertal bones, SD1253 from El Sidrón Cave and Vi33.15 from Vindija Cave, as described in Rohland and Hofreiter (24) (Table S1), and prepared

Table 2. Distribution of derived alleles by functional class

	Derived alleles, *n* (%)							
	Polymorphic positions (SNPs)				Homozygous positions ("fixed")			
Category	Neandertals*	Africans†	Europeans†	Asians†	Neandertals*	Africans†	Europeans†	Asians†
Coding								
Synonymous	2,501 (48.9)	8,250 (54.3)	5,571 (53.6)	5,518 (53.3)	1,273 (56.1)	383 (59.3)	624 (57.2)	603 (57.8)
Nonsynonymous	2,611 (51.1)	6,947 (45.7)	4,820 (46.4)	4,839 (46.7)	996 (43.9)	263 (40.7)	467 (42.8)	441 (42.2)
PP-2								
Benign	1,354 (54.6)	3,596 (63.6)	2,266 (62.1)	2,292 (61.9)	600 (63.8)	16 (55.2)	53 (66.3)	59 (64.1)
Deleterious	1,128 (45.4)	2,057 (36.4)	1,385 (37.9)	1,409 (38.1)	340 (36.2)	13 (44.8)	31 (36.9)	33 (35.9)
PC								
Benign	1,296 (49.9)	4,280 (61.9)	2,872 (59.8)	2,903 (60.3)	601 (60.4)	185 (70.6)	335 (72.2)	321 (73.1)
Deleterious	1,303 (50.1)	2,635 (38.1)	1,931 (40.2)	1,913 (39.7)	394 (39.6)	77 (29.4)	129 (27.8)	118 (26.9)
GS								
Benign	2,112 (80.9)	5,673 (81.7)	3,936 (81.7)	3,954 (81.7)	833 (83.6)	231 (87.8)	413 (88.4)	374 (84.8)
Deleterious	499 (19.1)	1,274 (18.3)	884 (18.3)	885 (18.3)	163 (16.4)	32 (12.2)	54 (11.6)	67 (15.2)

For the coding category, the number and proportion of coding alleles that are either synonymous or nonsynonymous are shown. For the PolyPhen-2 (PP-2), PhastCons (PC) and Grantham score (GS) categories, the number and proportion of nonsynonymous alleles that are predicted to be either benign or deleterious are shown.

*Derived in Neandertals and homozygous ancestral in Africans. See Table S18 for additional details.

†Derived in Africans, Europeans, or Asians and homozygous ancestral in Neandertals.

DNA sequencing libraries containing a special 4-bp clean-room tag sequence to avoid contamination in later steps (25). During library preparation, we used a uracil-DNA glycosylase (UDG) and endonuclease VIII mix to remove uracils resulting from cytosine deaminations (26).

Gene Annotation and Array Design. We designed 22 arrays as described in Burbano et al. (9) to capture 173,086 coding exons annotated in the human reference genome (GRCh37/1000 Genome Project release). These exons encompass ~27 Mb and belong to 17,367 genes annotated in consensus coding sequences (CCDS) (27), RefSeq (28), and GENCODE (29). A total of 20,005,501 probes 60 bases long were generated, at 3-base tile density, to cover up to 100 bases upstream and downstream of each coding exon. Probes containing repetitive elements were discarded (Fig. S1).

Exome Capture Experiment. We used 40 libraries from the El Sidrón and 19 from the Vindija Neandertals for capture as described in Fu et al. (11). The pools of exome-enriched libraries from El Sidrón and Vindija were sequenced from both sides on 11 and 15 lanes of the Genome Analyzer IIx (Illumina), respectively. After sequences from the same DNA fragments were merged, sequences of at least 35 bases were aligned to the human reference genome with Burrows-Wheeler aligner (BWA) (30). Removal of PCR duplicates resulted in ~1.0 and ~3.4 Gb of sequence aligned to the human genome in El Sidrón and Vindija exomes, respectively (Table S4). The percentage of DNA sequences with a mapping quality of at least one that map on the tiled regions is 64.7% for the El Sidrón and 64.0% for the Vindija capture. Coverage is higher in longer exons with intermediate GC content (Figs. S5 and S6). Reference allele frequencies in El Sidrón and Vindija exomes are within the range of reference allele frequencies in present-day human DNA captures (*SI Appendix*). Note that a capture bias toward alleles present on the probes should not affect the nonsynonymous-to-synonymous and deleterious-to-benign ratios or the variance of the allele frequency distribution in Neandertals; that is, capture bias should not affect the test of proportions or the frequency spectra of Neandertals presented in this work.

Contamination Estimates. To measure the present-day human mtDNA contamination in the El Sidrón and Vindija exomes, we used the sequences aligned by BWA (30) to the human reference mitochondrial genome (31) and realigned them to the El Sidrón and Vindija mitochondrial genomes (32), using the mapping iterative assembler (33). We then computed a contamination upper estimate using 81 and 84 diagnostic positions at which El Sidrón or Vindija mitochondrial genomes, respectively, differed from all sequences in a worldwide panel of 311 present-day human mitochondrial genomes. Sequences overlapping at least one of the diagnostic positions were classified as contaminants if they showed the present-day human allele (Table S8). We computed a lower estimate using 483 and 631 diagnostic positions at which El Sidrón and Vindija mitochondrial genomes, respectively, differed from at least one sequence in the worldwide panel of 311 present-day human mitochondrial genomes. Sequences overlapping at least one of the diagnostic positions were classified as contaminants if they showed the differing present-day human allele (Table S8).

Computational Correction of Cytosine Deaminations. Sequences may carry residual cytosine deaminations in the first positions of the 5′ end and the last positions in the 3′ end despite the UDG treatment (26) (Fig. S2*A*). These bases are read as thymine and adenosine, respectively. We decreased the quality score to 2 of any "T" base occurring within the first five bases or "A" base within the last five positions in the El Sidrón and Vindija sequences. Because a 2 on the phred scale (34) represents a 63% chance of sequencing error, any downstream analysis will consider these bases as sequencing artifacts. This approach resulted in comparable C-to-T and G-to-A substitution patterns in ancient and present-day individuals after genotype calling (Fig. S2*B*), suggesting that cytosine deaminations do not significantly contribute SNPs at low frequency in Neandertals (Fig. 2) or to their proportion of putatively deleterious alleles (Table S17).

Variation Discovery. We called Neandertal genotypes with GATK (35) and obtained high-quality calls for 18,811,848 and 23,283,112 sites in El Sidrón and Vindija exomes, respectively (Table S10). Genotypes in the Altai Neandertal, Denisovan, and present-day human individuals in Fig. 1 were obtained similarly (Tables S10 and S11), and a combined file for all individuals was created and annotated as in Meyer et al. (13) (Table S9). High-quality genotype calls for the longest protein-coding transcript in each gene were used for analysis (*SI Appendix*).

Genetic Differentiation. We computed the pairwise F_{ST} among individuals using the pairwise number of differences between individuals ($\pi_{Between}$) and the number of heterozygous sites in each individual (π_{Within}). The F_{ST} value for individuals A and B is then

$$F_{ST} = \frac{\pi_{BetweenAB} - \frac{(\pi_{WithinA} + \pi_{WithinB})}{2}}{\pi_{BetweenAB}}.$$

Derived Alleles. Neandertals were compared with each of the three Africans, Europeans, and Asians separately. In each pairwise comparison, derived alleles present in either Neandertals or the three present-day humans were used (Table 2, Fig. 2, and Table S18). Derived states for alleles were determined from the inferred ancestral base in the Enredo Pecan Ortheus (EPO) six-primate genome alignments (36) and were classified as synonymous or nonsynonymous based on the gene annotation used in the array design (*SI Appendix*). When counting alleles in heterozygous and homozygous positions for each archaic or present-day individual (Tables S15 and S16), we used derived alleles that are ancestral in the three Africans or ancestral in the three Neandertals, respectively.

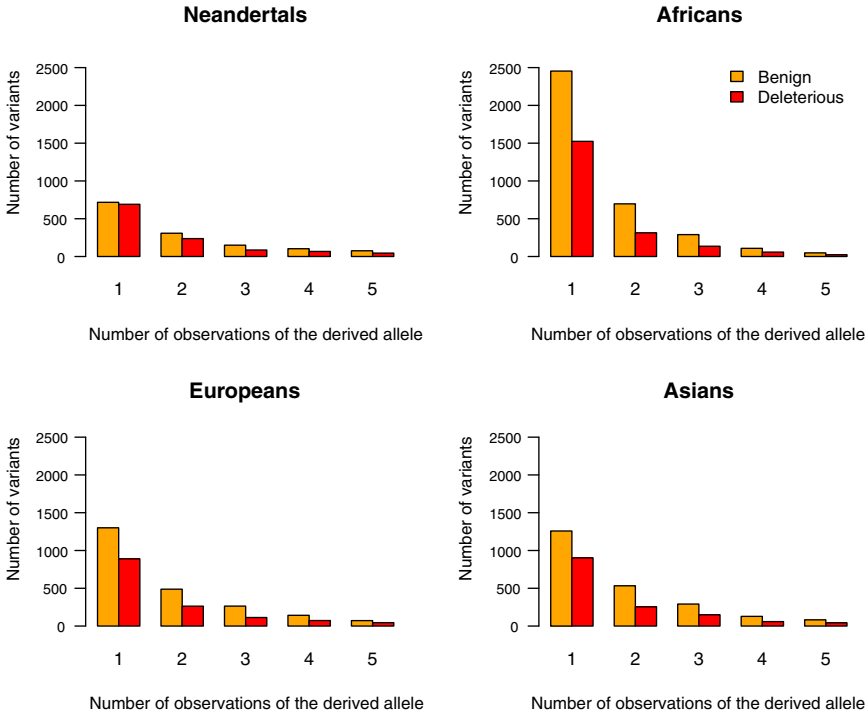


Fig. 2. Frequency spectra of nonsynonymous derived alleles classified by PolyPhen-2 as either benign or deleterious in Neandertals and present-day humans. The ratio of deleterious to benign derived alleles is higher in Neandertals than in the present-day humans. See Table S19 for the actual allele counts and Fig. S10 for the frequency spectra of nonsynonymous derived alleles assessed by PhastCons and GS.

Prediction of Functional Consequences. Nonsynonymous derived alleles classified in the "possibly" and "probably" damaging categories in PolyPhen-2 (16), in a position with a PhastCons (17) posterior probability larger than 0.9 or with a GS (18) of 101 or more, were considered to alter protein structure or function (Table 2 and Tables S15–S18). These alleles likely are deleterious. We used the HumDiv model of PolyPhen-2, which is based on the differences between human proteins and closely related mammalian homologs but does not incorporate present-day human polymorphism. Thus, this model is meant for the analysis of patterns of natural selection in which even slightly deleterious alleles must be considered. Furthermore, when counting putatively deleterious alleles per individual, we used only derived alleles that did not match the human reference sequence. This minimizes the bias in the counts of benign and deleterious alleles in polymorphic positions (Table 2) caused by the presence of present-day human, but not Neandertal, sequences in PolyPhen-2 alignments. It also results in the few counts of benign and deleterious alleles in homozygous positions in present-day humans using PolyPhen-2 (Table 2). The PhastCons conservation scores were computed on alignments of mammalian but not human proteins.

Phenotype Enrichment Analysis. We identified Human Phenotype Ontology (20) categories that are enriched for genes with amino acid-changing derived alleles in the archaic (Neandertal and Denisovan), the Neandertal, or the modern human lineages compared with amino acid-changing derived alleles shared among archaic and present-day humans in the same genes and phenotype categories using the program FUNC (37). Comparing the genes associated with a phenotype category with themselves at two different periods controls for differences in the number of genes in each category, as well as the length, sequence composition, and distribution across the genome of the genes associated with it. We restricted the analysis to phenotype ontology terms to which at least two genes are mapped and performed four separate comparisons for ontology enrichment (Tables S24, S26, and S28) that focus on different evolutionary time frames in the archaic, the Neandertal, and the modern human lineages. Our significance criteria are a P value <0.01 and an FDR ≤ 0.1 within each comparison in each lineage, but we also highlight categories that have a family-wise error rate ≤ 0.1 as a more conservative cutoff (Tables S24, S26, and S28). Derived alleles in the modern human lineage were required to be fixed or at high frequency (>90%) in the 1000 Genomes Project (14). Note that the high number of present-day humans makes the test more conservative in the modern than in the archaic human lineage.

ACKNOWLEDGMENTS. We thank David Reich and Montgomery Slatkin for comments on the manuscript, Oscar Lao for clarifying the ancestry of one present-day individual, Cesare de Filippo for help with the principal components analysis plots, Agilent Technologies for the capture arrays, and the Presidential Innovation Fund of the Max Planck Society for support. C.L.-F. and F.A.S.-Q. are supported by the Ministerio de Economía y Competitividad (Grant BFU2012-34157).

1. Cargill M, et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 22(3):231–238.
2. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. *Genetics* 158(3):1227–1234.
3. Bustamante CD, et al. (2005) Natural selection on protein-coding genes in the human genome. *Nature* 437(7062):1153–1157.
4. Lohmueller KE, et al. (2008) Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451(7181):994–997.
5. Ng SB, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461(7261):272–276.
6. Li Y, et al. (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* 42(11):969–972.
7. Tennessen JA, et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64–69.
8. Green RE, et al. (2010) A draft sequence of the Neandertal genome. *Science* 328(5979): 710–722.
9. Burbano HA, et al. (2010) Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* 328(5979):723–725.
10. Prüfer K, et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505(7481):43–49.
11. Fu Q, et al. (2013) DNA analysis of an early modern human from Tianyuan Cave, China. *Proc Natl Acad Sci USA* 110(6):2223–2227.
12. Wood RE, et al. (2013) A new date for the Neanderthals of El Sidrón Cave (Asturias, Northern Spain). *Archaeometry* 55(1):148–158.
13. Meyer M, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338(6104):222–226.
14. Abecasis GR, et al.; 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319): 1061–1073.

15. Reich D, et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468(7327):1053–1060.
16. Adzhubei IA, et al. (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7(4):248–249.
17. Siepel A, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15(8):1034–1050.
18. Grantham R (1974) Amino acid difference formula to help explain protein evolution. *Science* 185(4154):862–864.
19. Simons YB, Turchin MC, Pritchard JK, Sella G (2014) The deleterious mutation load is insensitive to recent population history. *Nat Genet* 46(3):220–224.
20. Robinson PN, Mundlos S (2010) The human phenotype ontology. *Clin Genet* 77(6):525–534.
21. Been E, Gómez-Olivencia A, Kramer PA (2012) Lumbar lordosis of extinct hominins. *Am J Phys Anthropol* 147(1):64–77.
22. Hider JL, et al. (2013) Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC Evol Biol* 13:150.
23. Liu F, et al. (2010) Digital quantification of human eye color highlights genetic association of three new loci. *PLoS Genet* 6(5):e1000934.
24. Rohland N, Hofreiter M (2007) Comparison and optimization of ancient DNA extraction. *Biotechniques* 42(3):343–352.
25. Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 40(1):e3.
26. Briggs AW, et al. (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res* 38(6):e87.
27. Pruitt KD, et al. (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* 19(7):1316–1323.
28. Pruitt KD, Tatusova T, Klimke W, Maglott DR (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* 37(Database issue): D32–D36.
29. Coffey AJ, et al. (2011) The GENCODE exome: Sequencing the complete human exome. *Eur J Hum Genet* 19(7):827–831.
30. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
31. Andrews RM, et al. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 23(2):147.
32. Briggs AW, et al. (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* 325(5938):318–321.
33. Green RE, et al. (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134(3):416–426.
34. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8(3):186–194.
35. McKenna A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303.
36. Paten B, et al. (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res* 18(11):1829–1843.
37. Prüfer K, et al. (2007) FUNC: A package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* 8:41.

Patterns of coding variation in the complete exomes of three Neandertals

SI Appendix

Sergi Castellano[a], Genís Parra[a*], Federico Sánchez-Quinto[b*], Fernando Racimo[a,c*], Martin Kuhlwilm[a*], Martin Kircher[a,d], Susanna Sawyer[a], Qiaomei Fu[a,e], Anja Heinze[a], Birgit Nickel[a], Jesse Dabney[a], Michael Siebauer[a], Louise White[a], Hernán A. Burbano[a,f], Gabriel Renaud[a], Udo Stenzel[a], Carles Lalueza-Fox[b], Marco de la Rasilla[g], Antonio Rosas[h], Pavao Rudan[i], Dejana Brajković[j], Željko Kucan[i], Ivan Gušic[j], Michael V. Shunkov[k], Anatoli P. Derevianko[k], Bence Viola[a,l], Matthias Meyer[a], Janet Kelso[a], Aida M. Andrés[a], Svante Pääbo[a]

[a] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany;

[b] Institute of Evolutionary Biology (UPF-CSIC), Dr. Aiguader 88, 08003 Barcelona, Spain;

[c] Department of Integrative Biology, University of California, Berkeley, California 94720, USA;

[d] Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA;

[e] Key Laboratory of Vertebrate Evolution and Human Origins of the Chinese Academy of Sciences, IVPP, CAS, Beijing,100049;

[f] Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tuebingen, Germany;

[g] Área de Prehistoria, Departamento de Historia, Universidad de Oviedo, 33011 Oviedo, Spain;

[h] Departamento de Paleobiología, Museo Nacional de Ciencias Naturales, CSIC, 28006 Madrid, Spain;

[i] Anthropology Center of the Croatian Academy of Sciences and Arts,. Ante Kovacica 5, HR-10000 Zagreb, Croatia;

[j] Croatian Academy of Sciences and Arts, Institute for Quaternary Paleontology and Geology, Ante Kovacica 5, HR-10000 Zagreb, Croatia;

[k] Paleolithic Department, Institute of Archeology and Ethnography, Russian Academy of Sciences, Siberian Branch, 630090 Novosibirsk, Russia;

[l] Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany;

[*]These authors contributed equally to this work.

Corresponding authors: Sergi Castellano (sergi.castellano@eva.mpg.de; +49 341 3550 806) and Svante Pääbo (paabo@eva.mpg.de; +49 341 3550 501).

SI Appendix

1. DNA Extraction and library preparation

1.1 El Sidrón and Vindija Neandertals

DNA extraction

We prepared DNA extracts from two Neandertal bones, SD1253 from El Sidrón Cave in Spain and Vi33.15 from Vindija Cave in Croatia, as described in Rohland *et al.* (1). For the SD1253 bone, we created seventeen 100 µl DNA extracts, where each extract was made from between 80 mg and 400 mg of bone powder (Table S1). For the Vi33.15 bone, we generated seventeen 100 µl DNA extracts (Table S1) from between 100 mg and 600 mg of bone powder. Three more Vindija extracts were re-extracted from the left over bone pellet of E930, E931 and E932 extracts (Table S1).

Library preparation

Most of El Sidrón and Vindija libraries were prepared from 20 µl and 74 µl of DNA extract, respectively, except L8249 from E428 (10 µl), L7855 from E565 (50 µl) and L7865 from E566 (50 µl). The method of library preparation is as described in Kircher *et al.* (2) using a uracil-DNA-glycosylase (UDG) and endonuclease (Endo VIII) enzyme mix to remove uracils (3). All libraries have a special four base pair clean room key that was added onto the Illumina Multiplex adapters (4) to avoid contamination in later steps: 5' - AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**GTCT** – insert - **AGAC**AGATCGGAAGAGCACACGTCTGAACTCCAGTCACIIIIIIIATCTCGTATGCCGTCTTCTGCTTG with the key sequence in bold. Between one and four libraries were produced using these extracts (see Table S1 for more details), meaning that we ultimately used 1.78 grams of El Sidrón bone and 4.62 grams of Vindija bone converted into libraries to capture these Neandertal exomes.

Each library was amplified using the following reagents: 1 µl of AmpliTaq Gold DNA Polymerase (5 U/µl Applied Biosystems), 10 µL 10X Gold Buffer (Applied Biosystems), 10 µl 25 mM $MgCl_2$ (Applied Biosystems), 1 µl of 25 mM each dNTPs (Fermentas), 4 µl of 10 µM Genomic R1 P5 primer (5' - ACACTCTTTCCCTACACGACGCTCTTCCGATCT), 4 µl of 10 µM Multiplex R2 P7 primer (5' - GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT), 68 µl of water and 2 µl of template. Cycling conditions were an activation step of 12 minutes at 95 °C, followed by 10 cycles of a 20 second denaturation step at 95 °C, a 30 second annealing step at 60 °C, and a 40 second elongation step at 72 °C. A final elongation step for 5 minutes at 72 °C was also included. PCR products were then purified using the MinElute purification kit by Qiagen and eluted in 60 µl EB.

Libraries L8249 to L8256 were not amplified as described above but were instead amplified with two indexing primers each. Each library was divided into six reactions to avoid adding too much template into the PCR reaction and was amplified as described in Kircher *et al.* (2) with 6 µl of template per library.

2. Gene annotation and array design

Exome annotation

The array was designed to capture the coding regions of genes annotated in the human genome. While some commercial and non-commercial exome capture designs are available (5), we designed an array with a custom set of human genes. Our custom array has two main advantages: 1) the probe design is optimized for the capture of ancient DNA(6); and 2) it contains only highly reliable human gene annotations, which results in a manageable number of physical arrays to be processed for the capture experiment.

Our exome annotation is based on three sets of well-characterized protein coding genes: 1) the consensus CDS database (CCDS, for build 37.1); 2) the human transcripts from the NCBI Reference sequence database (RefSeq, release 45); and 3) the Manual GENCODE V4 annotation (which contains manual gene annotations for some chromosomes). These datasets were obtained from the USCS genome browser human GRCh37/hg19 assembly. Table S2 summarizes the total number of genes, transcripts, exons and nucleotides for each source of annotation in our array.

Transcripts from the three datasets were clustered into genes using the ENSEMBL gene annotation. Incomplete transcripts were removed in the clustering process. We refer to the coding regions of this combined annotation as the primary target of our capture array design. Table S2 shows the total number of genes, transcripts, exons and nucleotides in the primary target.

Probe design

Probe design was similar to the one described in Burbano *et al*. (6). Using the human GRCh37/hg19 reference sequence as a template, we designed 60 bases length probes for: 1) the primary target regions; and 2) the 100 bases upstream and downstream of each primary target region. For increased capture efficiency in ancient DNA, the probes were generated with a tile density of 3 nucleotides (each probe was 3 bases apart). Probes containing repetitive elements – detected by 15-mer frequency counts – were discarded (7). See Figure S1 for a schematic representation of the probe design. The final tiled (primary) target in our array is the intersection of: 1) primary target positions covered by at least one probe; and 2) primary target positions with a Duke Uniqueness 20 bp score of 1 (uniquely mappable exome) from the UCSC genome browser (GRCh37/hg19). Table S2 shows the total number of genes, transcripts, exons and nucleotides in the tiled target.

Following this approach, a total of 20,005,501 probes were generated, and 22 Agilent custom one million feature capture arrays were used. The probes covered a total of 69,786,128 base pairs of genomic DNA.

3. Exome capture experiment

Mitochondrial and exonic DNA sequences were enriched from a total of 19 libraries from the Vindija sample and 40 libraries from the El Sidrón sample, as described in detail in Fu *et al*. (8). Briefly, all libraries were subjected to two rounds of amplification to obtain sufficient amounts of DNA (several micrograms). The probe library was stripped from 22 Agilent one million feature capture arrays and used to generate single-stranded biotinylated DNA probes. A mitochondrial probe library had been constructed previously(8). Two successive rounds of hybridization capture were performed for each library with both sets of probes. Capture eluates were amplified and barcoded with two indexes (2) and four library pools were generated, representing the exome and mtDNA captures from both Neandertal samples.

The pools of exome-enriched libraries from El Sidrón and Vindija were sequenced on 11 and 15 lanes of the Genome Analyzer IIx (Illumina), respectively. The mtDNA-enriched library pools were sequenced together with the other libraries for El Sidrón and on a single lane for Vindija. Paired end sequencing was carried out for 2x 76 cycles, and additional 7-cycle sequence reactions were performed to read the indexes in both adaptors (2).

4. Processing and mapping

Base calling and raw sequence processing

Analysis was based on the BCL and CIF intensity files from the Illumina Genome Analyzer RTA 1.9 software. Raw reads showing the control index ('TTGCCGC') in the first index read were aligned to the φX174 reference sequence to obtain a training data set for the Ibis base caller (9), which was then used to recall bases and quality scores of each run from the CIF files.

The reads were then filtered for the presence of the correct library indices, allowing for one substitution and/or the skipping the first base in each index (10). Minimum base quality score of 10 was required in both index reads (2). The remaining sequence reads were merged (and adapters were removed) by searching for an >= 11nt overlap between the forward and the reverse reads (2). 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(11) 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). Using BWA's samse command, alignments were converted to SAM format, and then via samtools(12) 0.1.18 to coordinate-sorted BAM files. The output 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 amplified library, reads which map to the same outer reference coordinates were replaced by a consensus sequence to collapse duplicate reads (2).

Local realignment for resolving insertions and deletion

After duplicate removal, the BAM files for all individual libraries were combined by bone. 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 sequences with an edit distance of more than 20% of their length were removed. This process resulted in a total of 1.03 gigabases of uniquely aligned El Sidrón sequences and 3.36 gigabases of uniquely aligned Vindija sequences against the human reference genome (Table S3).

Mitigation of cytosine deamination

Fragments may carry residual cytosine deamination in the first positions of the 5' end and the last positions in the 3' end in spite of the UDG treatment (Figure S2A). These bases are read as thymine and adenosine, respectively. Because the deamination process does not affect the qualities of these bases, deamination can potentially influence downstream analyses. To mitigate this problem, we lower the base quality score in the phred scale (14) to 2 for any 'T' nucleotide occurring within the first five bases or 'A' nucleotide within the last five positions in a sequence. This approach resulted in comparable C to T and G to A substitution patterns in the heterozygous positions in the archaic and present-day individuals after genotype calling (Figure S2B). The BAM files with cytosine deamination correction can be found at http://cdna.eva.mpg.de/neandertal/exomes/BAM.

5. Quality assessment

Capture efficiency

On target efficiency statistics are given in Table S4. Duplicate reads were removed before the calculation of these statistics. The percentage of sequences with mapping quality (MQ) of at least one is similar for El Sidrón (64.7%) and Vindija (64.0%)

exome captures. These results compare favorably with our previous capture of ancient DNA (6), which had an on-target percentage of 37% and to present-day DNA captures (15, 16).

For the coverage analyses below, only sequences with a mapping quality score of at least one (reads ≥ MQ1) were used. Note that a large proportion of these sequences were mapped with higher confidence (reads ≥ MQ30) in El Sidrón (95.1%) and Vindija exomes (95.9%). Thus, average coverage at MQ1 and MQ30 quality thresholds are similar in each exome (Table S4).

Coverage distribution

The average coverage for the captured El Sidrón and Vindija exomes is 12.5 and 42.0-fold, respectively (reads ≥ MQ1; Table S4). Figure S3 shows that the coverage distribution in the target regions resembles a truncated normal distribution (bounded at low coverage) in both exomes. Considering only positions covered at least once, coverage bins 2-33 and 4-95 encompass 95% of the exome-wide coverage in El Sidrón and Vindija captures, respectively (Table S5). Similarly, we computed the coverage distribution for the Altai exome (Figure S4). The average coverage for this exome is 45.3-fold, with bins 25-72 encompassing the central 95% of the data. We also computed the lower and higher coverage bins encompassing the central 95% of the data for the present-day human exomes used in this work: Yoruba (HGDP00927), Mandenka (HGDP01284) and Dinka (DNK2) from Africa, French (HGDP00521) and Sardinian (HGDP00665) from Europe, CEU (NA12891) of European (largely Italian) ancestry from Utah, Han (HGDP00778) and Dai (HGDP01307) from Asia and Papuan (HGDP00542) from Oceania. Data within these coverage ranges, with a minimum coverage of six, are used for analysis (Table S5). Because of the skewed coverage distribution in El Sidrón and Vindija exomes, the analyses in this work were repeated using only sites covered with at least 10, 15 or 20 sequences in each individual. No important differences in the results were observed (see for example Tables S13 and S17).

Coverage comparison with the Neandertal draft genome

The draft Neandertal genome is a mixture of three different samples (Vi33.16, Vi33.25 and Vi33.26). We obtained BAM files for the Neandertal sequences mapped to the human genome (GRCh37/hg19) from the UCSC browser: http://genome.ucsc.edu/Neandertal/. To filter these sequences we used the cutoffs described in the original paper(17). The coverage of the exome in the Neandertal draft genome is 1.3-fold, which is about 10, 32 and 35 times lower than El Sidrón, Vindija and Altai exome coverage presented here, respectively (sequences ≥ MQ1; Table S6).

Table S6 shows coverage per base, exon and gene for the Draft, El Sidrón, Vindija and Altai Neandertal exomes and the Denisovan exome. In brief, more than 90% of coding bases are covered in the captured exomes presented here, while just about 52% were covered in the draft Neandertal genome. Similarly, more than 90% of exons have at least 80% of their coding bases covered in El Sidrón, Vindija, Altai and Denisovan exomes, while less than 20% of exons in the draft Neandertal genome achieve this coverage. At the gene level, about 90%, 98%, 100% and 100% of genes have at least 80% of their bases covered in El Sidrón, Vindija, Altai and Denisovan exomes, respectively. This is in contrast to just 1% of genes covered at this level in the draft Neandertal genome.

GC content capture bias

Our results show a similar pattern of GC content dependence in both exome captures. Sequence reads are underrepresented to their probe density in regions of high and low GC content (Figure S5). A decrease in coverage at high and low GC content has been reported for exome capture in humans (18). Optimal coverage for GC content at the range of 40 to 60% is common for several capture technologies for modern DNA (18). Exome capture in Neandertals shows a similar GC content bias with coverage (Figure S5).

Length dependence capture bias

Interestingly, exons below 75 bases long having noticeably lower coverage (Figure S6). Two factors may explain this observation: 1) Short coding exons are mostly in the 5' of genes and, thus, tend to be in high GC content regions (19); and 2) short exons are tiled by less probes, resulting in less capture efficiency.

Reference allele capture bias

The human reference allele used to design the capture probes can influence the number of alternative alleles captured in heterozygous sites. Reference capture bias is non-negligible but small in the capture of present-day human exomes (18). The mean allele frequency on high-quality heterozygous sites typically shifts toward the reference allele from 0.5 to 0.54 (20). We find a mean reference allele frequency of 0.523 and 0.542 in El Sidrón and Vindija exomes, respectively. This frequency is 0.517 for the exome regions obtained from the Altai genome. This suggests that reference allele bias is not worse than in capture experiments with modern DNA. We also explored the distribution of reference to alternative ratios in the Neandertal exomes (Figure S7). As expected, the variance of reference allele frequencies is larger in the captured exomes. Note that the ability to call heterozygous positions is maintained at very high and low reference allele frequencies despite the extensive quality filtering we applied (see Variation discovery). This variation is within the range of reference allele frequencies in modern human DNA captures (18). Importantly, reference allele biases are systematic in nature and affect all types of sites. Thus, capture bias should not affect non-synonymous to synonymous ratios nor the variance of the allele frequency distribution (18) analyzed in this work.

Error rates

We computed the sequencing error rate in our data sets by using positions in the human reference genome where very little divergence is expected among Neandertals and present-day humans following the protocol in Meyer *et al*. (21). We selected positions from conserved regions taken from the 35-way GERP elements, and filtered them for a primate conservation score of at least 0.98. The primate conservation score was generated from the 6-primate EPO alignments (22, 23) using PhastCons (24). This resulted in about 1MB of sequence.

We found that, when comparing genotype calls to the reference human genome, 0.016 – 0.021% of genotypes in our exomes differ (Table S7). These values are a little bit lower than the genome-wide values previously reported (21). Exome regions are more conserved than other genomic regions and, thus, less discordance between genotypes is expected. We also observed that the largest genotype discordance with the reference human genome are found in the archaic exomes, suggesting that sequence divergence contributes to the estimates of per-base sequence error rates, which are obtained by counting differences between individual sequences 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.

6. Contamination estimation

To measure the percentage of library contamination by human sequences in El Sidrón and Vindija exomes, we used the BWA aligned mitochondrial sequences on the human revised Cambridge reference mitochondrial genome (25). We realigned these sequences using MIA (26) against the corresponding El Sidrón and Vindija mitochondrial genomes (27). Using these data we estimated upper and lower bounds of present-day human contamination. We computed an upper estimate using 81 and 84 diagnostic positions where El Sidrón or Vindija mitochondrial genomes differed from all sequences in a worldwide panel of 311 present-day human mitochondrial genomes. Sequences overlapping at least one of the diagnostic positions were classified as contaminants if they showed the present-day human allele (Table S8). We computed a lower estimate using 483 and 631 diagnostic positions where El Sidrón and Vindija mitochondrial genomes differed to at least one sequence in the worldwide panel of 311 present-day human mitochondrial genomes. Sequences overlapping at least one of the diagnostic positions were classified as contaminants if they showed the present-day human allele (Table S8). We conclude that present-day human contamination is no more than 1% in both El Sidrón and Vindija Neandertals (Table S8).

7. Variation discovery

Genotype calling

We called genotypes separately for El Sidrón, Vindija and Altai exomes using the UnifiedGenotyper from GATK(13) (v1.3-14-g348f2db). Both single nucleotide variants (SNVs) and insertions and deletions (InDels) were called. Calls for all sites in the primary target regions and 100 bases upstream and downstream of each primary target region were made.

We performed a second call on all SNV sites where at least one non-reference allele was found, replacing the reference for the alternative allele in the reference genome and re-running GATK UnifiedGenotyper for those sites with the modified reference (21). This re-call method allows calling heterozygous sites where both alleles differ from the reference allele, since this version of GATK recognizes only one alternative allele per site.

Variant annotation

Variant Call Format (VCF) files were annotated with supplementary information, using the extended VCF pipeline described in SOM 6 of Meyer *et al*. (21). Table S9 summarizes each of the additional terms included in the extended VCFs. The annotated VCFs can be found at http://cdna.eva.mpg.de/neandertal/exomes/VCF.

Systematic errors

Systematic errors are platform-specific sequencing errors that disproportionately affect specific genomic positions, regardless of the location of that genomic position within a read (28). This tendency for systematic error is thus an inherent property of particular genomic positions, resulting from their surrounding sequence background, and is not library dependent (as would be the case if the location within a read were important). Systematic error sites are expected to have good SNP quality scores because the (erroneous) alternative allele is seen in many independent reads. Species with high sequence similarity are therefore likely to exhibit a similar tendency towards systematic error caused by sequence motifs and/or composition at specific positions.

To generate an exome-wide list of candidate systematic error sites, we used a data set comprising 20 bonobo, 20 chimpanzee and 20 human exomes that had been sequenced to 20x coverage on Illumina GAIIx. These three species have high sequence similarity but little shared polymorphism. The reads from all species were mapped to the human genome (hg19) and only reads with a mapping quality (MQ) greater than 25 were used for SNP calling. SNPs were called separately for each species using GATK UnifiedGenotyper (v1.4) and using hg19 as the reference genome for all three species. Sites were excluded from further consideration if they lay in the 5bp on each side of InDels, were at the extreme ends of the coverage distribution (highest/lowest 2.5%) or had genotype calls in fewer than 5 individuals per species. Sites at which all chimpanzees or bonobos were homozygous non-human-reference were discarded as fixed differences. Positions were considered potential systematic errors if they were polymorphic in all three species, had a SNP quality (QUAL) >50 and a strand bias (SB) >-10. There were 6,517 sites that met these criteria in the three species exomes considered. For our analyses, we removed SNVs on these sites in both archaic and present-day exomes.

Variation discovery in Denisovans and humans

We previously produced variation calls as described above for a Denisovan and 8 modern human high-coverage genomes (21) from Africa (Mandenka, Yoruba and Dinka), Europe (French and Sardinian) and Asia (Han, Dai and Papuan from Oceania). In addition, we included an American individual from Utah of European (largely Italian) ancestry (29). We extracted genotype calls of the exomes (primary target regions plus 100 bases upstream and downstream of each region in our array) for the Denisovan individual and each of the present-day humans. A combined VCF file of the Vindija, El Sidrón and Altai Neandertal exomes, the Denisovan exome and the nine human exomes was created. For each individual, we obtained the coding positions for the longest transcript in each gene. We performed the majority of analyses on this set of coding positions. The coding sequence of the longest transcript of each of the genes analyzed in the archaic individuals can be found at

http://cdna.eva.mpg.de/neandertal/exomes/CDS. The sequence of the encoded proteins can be found at http://cdna.eva.mpg.de/neandertal/exomes/protein.

Variation assessment

We assessed genotype calls for each individual based on their annotation in the combined VCF file. A site in an exome was considered for analysis when the following set of filters is met: (1) a GATK call was made; (2) the genotype quality (GQ) is at least 20 and the SNP quality (QUAL) is at least 30; (3) a mapability score in the Duke 20mer uniqueness score (Map20) of 1; 4) the fraction of mapped reads with Mapping Quality (MQ) of zero is less than 10%; (5) coverage is within the 95% of the exome coverage with a minimum coverage of six; (6) the site is not flagged as a systematic error; (7) the site is not flagged as LowQuality; (8) the site has human-chimpanzee ancestry information; (9) the human-chimpanzee ancestral allele matches one of the two alleles at heterozygous sites; (10) human and chimpanzee appear no more than once in the EPO alignment block. Tables S10 and S11 show the number of coding genotype calls that met the filters above for each archaic and present-day human individual, respectively.

Comparable calls are obtained with genotype quality (GQ) of 30 for all individuals but El Sidrón Neandertal. The lower coverage in this archaic individual (average coverage of 12.5-fold) results in homozygous sites being undercalled at GQ20 and GQ30 (at least ten reads per position are needed to achieve this genotype quality). Therefore, El Sidrón Neandertal is called with GQ10. Importantly, the same number of heterozygous sites is called at GQ10, GQ20 and GQ30 in all individuals, including El Sidrón. Thus, the heterozygous genotypes used in our analysis are of high quality. In addition, we found 25 heterozygous sites with two alternative alleles (different from the hg19 human reference). These few sites were not included in our analyses.

8. Archaics and present-day humans relationships

Pairwise differences

We explored the relationships between the four archaic and the nine present-day human individuals building a tree from the number of differences between pairs of individuals. We first obtained autosomal positions from the combined VCF file that passed quality filters in the protein-coding regions of all individuals (no missing data). From this data, we removed positions where all individuals are either homozygous reference or alternative. The remaining 35,085 variable positions (both reference and alternative alleles are present) were used to compute the pairwise number of differences among individuals. A random allele was chosen in heterozygous sites.

We used the resulting distance matrix to infer a phylogenetic tree using the neighbor-joining method (30). The midpoint was chosen to root the tree. The tree, with branch lengths scaled at the pairwise number of differences per Mb, is presented in Figure

1A. Trees with similar topologies are obtained when only transitions or transversions are used.

Principal components analysis

We also explored the relationships between the four archaic and the nine present-day human individuals using Principal Components Analysis (PCA). We used the same 35,085 coding sites used for the neighbor-joining tree.

We encoded genotypes for each individual as follows: 1) '0' for homozygous reference; 2) '1' for heterozygous; 3) '2' for homozygous alternative. No missing data is present in this data set as genotype calls for all individuals were required above. We used the glPca function in the *adegenet* R package for the analysis (Figure S8).

The first principal component (PC1) separates the El Sidrón, Vindija, Altai and Denisovan archaic individuals from present-day humans (Figure S8A). "Fixed" derived alleles in the archaic and modern lineages are responsible for this separation. Because Europeans and Asians have "fixed" more alleles than Africans since the split with Neandertals (Table 2 and Table S16), they are more separated from the archaic individuals than Africans in PC1 (see also Figure 1B).

The second PC (PC2) splits the archaic group into a tight Neandertal cluster and the Denisovan individual (Figure S8A). Within present-day humans, the first two PCs separate Africans from Europeans and Asians. The third PC (PC3) separates Europeans from Asians.

F_{ST}

From the 35,085 sites above, we computed the pairwise fixation index (F_{ST}) among individuals using the pairwise number of differences between individuals A and B ($\pi_{BetweenAB}$) and the number of heterozygous sites in each individual A ($\pi_{WithinA}$) and B ($\pi_{WithinB}$), such that:

$$Fst = \frac{\pi_{BetweenAB} - \frac{\left(\pi_{WithinA} + \pi_{WithinB}\right)}{2}}{\pi_{BetweenAB}}$$

We used the pairwise F_{ST} values among individuals to build a neighbor-joining tree (Figure 1B). We find that Neandertals are more differentiated from each other than present-day humans within Africa, Europe and Asia, and that Neandertals and Eurasians are more differentiated more from each other than Neandertals and Africans or Eurasians and Africans (Table 2 and Table S16).

9. Heterozygosity estimates

We estimated heterozygosity in each archaic and present-day individual directly from the filtered GATK genotype calls in the coding regions of the autosomes. We divided the number of heterozygous genotypes by the total number of called genotypes per individual (Tables S10 and S11) to obtain the heterozygosity estimates in Table S12. The archaic exomes have remarkably lower genetic diversity than present-day humans. Within the archaic individuals, El Sidrón individual has the highest heterozygosity, while the Altai Neandertal has the lowest.

Because the differences in coverage distribution between the captured exomes (El Sidrón and Vindija) and the shotgun exomes (Altai, Denisovan and the nine present-day humans), we recalculated the heterozygosity estimates in coding positions with a depth of coverage of at least 10-, 15- and 20-fold (Table S13). This analysis supports the Neandertal and Denisovan individuals having lower heterozygosity than the present-day humans.

10. Heterozygous and Homozygous derived genotypes

Lineage-specific alleles

We computed the number of genotypes heterozygous and homozygous for the derived allele in each Neandertals and present-day human since the split of the two lineages. To do this, we compared El Sidrón, Vindija and Altai Neandertals with each group of three African, European and Asian individuals. In each comparison, we counted the number of heterozygous and homozygous derived sites in Neandertals that are homozygous ancestral in the present-day human group under comparison, and vice versa (Tables 2 and S17). We used the combined file of genotype calls to obtain the genotypes for each comparison. We used only those sites that passed all quality filters in the six individuals compared. The inferred ancestral base for humans and chimpanzees was used to identify derived sites.

In addition, Tables S15 and S16 show the number of heterozygous and homozygous derived genotypes in each archaic and present-day individual, respectively, since the split of the Neandertal and modern human lineages. Both private and shared alleles that differed from the inferred ancestral base in the EPO alignments are reported. Because the differences in coverage between individual exomes, we used only sites that passed all quality filters in all the archaic and present-day individuals. Hence, the absolute numbers are comparable among individuals. As before, the inferred ancestral base for humans and chimpanzees was used to identify derived sites.

Prediction of functional consequences

We inferred the functional consequences of non-synonymous derived alleles using three approaches: PolyPhen-2 (31), PhastCons (24) and Grantham scores (GS) (32). Non-synonymous derived alleles classified into the “possibly” and “probably” damaging categories in PolyPhen-2, with a PhastCons posterior probability larger

than 0.9 or with a Grantham scores (GS) of 101 or more were considered to be "deleterious" (Table 2 and Tables S15, S16 and S17). We used the HumDiv model of PolyPhen-2, which is based on the differences between present-day human proteins and their closely related mammalian homologs but does not incorporate present-day human polymorphism. Thus, this model is meant for the analysis of patterns of natural selection where even slightly deleterious alleles must be considered. Furthermore, when counting deleterious alleles per individual, we used only those sites in which the predicted "deleterious" allele was both derived and non-reference. This approach discards derived substitutions in the human reference sequence and minimizes any bias caused by the presence of present-day human but not Neandertal sequences in PolyPhen-2 alignments. The PhastCons conservation measured was based on alignments of mammalian but not human proteins and, thus, they could be obtained for all ancient or present-day derived positions found in the PhastCons alignments. The GS scores depend only on the physicochemical properties of amino acids and are available for all derived substitutions.

We tested whether the number of putatively deleterious alleles in Neandertals, as assessed by PhastCons, is significantly larger than in present-day humans. We summed the alleles in heterozygous (one allele) and homozygous (two alleles) positions (Tables S15 and S16) and found this not to be the case ($P = 0.15$; Mann-Whitney U-test). Indeed, four present-day humans have more putatively deleterious alleles than at least one Neandertal. This agrees with the suggestion that the overall deleterious load in humans is - in contrast to the proportion of deleterious alleles - mostly unaffected by population history (33). If we count only the putatively deleterious alleles in homozygosity (Table S16), we find evidence for an increase of the recessive load in Neandertals ($P = 0.016$; Mann-Whitney U-test) as predicted for human populations after a bottleneck (33).

We then tested whether a larger proportion of genes associated to autosomal recessive traits, as defined in the Human Phenotype Ontology database, have non-synonymous derived alleles inferred to be deleterious in Neandertal than in present-day human individuals. We find that 51.1% of such genes in Neandertals (Table S20, average on this group) have derived homozygous genotypes with likely deleterious effects, which is a higher proportion than in Africans (37.4%, $P = 0.07$; G-test), Europeans (38.2%; $P = 0.06$) and Asians (35.5%; $P = 0.02$). This result is suggestive of an enrichment of recessive disorders in Neandertals, but the health significance of this enrichment is unclear. Note that, while each Neandertal individual has a larger number of genes with putatively deleterious homozygous alleles than any of the present-day humans ($P = 4.5 \times 10^{-3}$; Mann-Whitney U-test) (Table S20), the difference is small and not larger than the two-fold range (from 22 to 46 genes) among humans today. This agrees with a non-negligible but slight effect of demographic history on the deleterious recessive load (33).

While one could argue that homozygous alleles are a better proxy for the deleterious genetic load, we also explore heterozygous alleles in genes associated to autosomal recessive traits as less severe but disease phenotypes can result from having a single functional allele with incomplete dominance. We find enrichment of the proportion of genes with heterozygous alleles classified as "deleterious" in Neandertals (56.0% versus <49% in present-day humans; Table S20), but the number of such genes is larger in present-day than in Neandertal individuals ($P = 8.0 \times 10^{-3}$; Mann-Whitney U-test) (Table S20). Thus, if we measure the deleterious load as the sum of genes with homozygous or heterozygous alleles in the "deleterious" category in Table S20), we find no difference between Neandertal and present-day human individuals ($P = 0.27$; Mann-Whitney U-test). So, we find no strong evidence for a role of recessive disorders in the demise of Neandertals.

Shared alleles

We computed the fraction of derived alleles in Neandertals that are shared with Africans, Europeans and Asians, as well as the fraction of derived alleles in these present-day human groups that are shared with Neandertals (Table S14). We also computed these fractions separately for SNPs and "fixed" alleles (Table S14). We see that the proportion of derived alleles in present-day humans not found in Neandertals is larger than the proportion of derived alleles in Neandertals not found in present-day humans. This is in agreement with the small genetic diversity of Neandertals.

We also find that 56.7% of derived alleles in the Neandertals are shared with present-day humans in the 1000G data. As expected, the proportion is higher for derived alleles "fixed" in the three Neandertals (83.4%) than for derived SNPs (33.9%).

11. Allele frequency distribution

We obtained the frequency distribution of the synonymous and non-synonymous derived polymorphism in Neandertals and present-day humans. Non-synonymous alleles were further classified into "deleterious" and "benign" according to PolyPhen-2, PhastCons and GS scores (Figure 2 and Figure S10).

We used those derived alleles segregating in Neandertals or the Africans, Europeans or Asians under comparison, and vice versa. Thus, three allele frequency distributions were obtained for Neandertals while only one for each group present-day humans (Figure S10). The three Neandertal SFS are comparable, irrespective of whether African, European or Asian individuals are used to identify derived alleles in Neandertals. In Figure 2, we show the distribution of derived alleles in Neandertals that are homozygous ancestral in the three African individuals.

We tested the correlation between the frequency of the non-synonymous derived SNPs in Figure 2 and their functional or structural impact on proteins as measured by their Polyphen-2 scores. We find a negative correlation between the allele frequencies

and the putatively deleterious effect of the alleles. However, the correlation is weak, with Neandertals (Spearman's rho = -0.129) having a somewhat stronger negative correlation than Africans (rho = -0.059), Europeans (rho = -0.073) and Asians (rho = -0.089). Although these results are suggestive of lower efficacy of purifying selection in Neandertals than in present-day humans, they do not allow any conclusions to be drawn, probably due to the small sample size.

The Polyphen-2 probabilities for the SNPs in Table 2, however, suggest a stronger deleterious effect on proteins in Neandertals than in Africans ($P < 4.6 \times 10^{-15}$; Mann-Whitney U-test), Europeans ($P = 7.2 \times 10^{-9}$) and Asians ($P = 8.0 \times 10^{-9}$). In particular, the SNPs seen once in the six chromosomes of Neandertals are more deleterious than at any other allele frequency ($P = 6.2 \times 10^{-10}$; Mann-Whitney U-test) and have a stronger impact on the function or structure of proteins than low frequency SNPs in Africans ($P < 2.2 \times 10^{-16}$), Europeans ($P = 6.7 \times 10^{-9}$) and Asians ($P = 2.0 \times 10^{-7}$). Thus, not only do Neandertals have a larger proportion of alleles inferred to be deleterious at low frequency, but these alleles are predicted to have a stronger deleterious effect than low frequency alleles in humans today.

12. Catalog and phenotype enrichment analysis

Using the exomes from the two Neandertals, Altai Neandertal genome and the 30-fold coverage Denisova genome (21) we identified coding sequence changes that specific to the present-day humans, to Neandertals and the Denisovan individual, and to Neandertals.

The availability of variation data for present-day humans allows us to identify sites where archaic human genomes carry the ancestral allele, while the derived allele is either fixed or at high frequency (>90%) in present-day human populations. We can also identify positions that are derived in the Denisovan individuals and three Neandertals, but ancestral in all present-day humans. The catalog is available online at http://cdna.eva.mpg.de/neandertal/exomes/catalog. A Human Phenotype Ontology enrichment analysis provides insights into the anatomical systems that may be affected by these changes.

Data Filtering for all catalogs

Using the variant call format annotation previously described (Variation discovery), we looked at positions where:

1. A GATK call was made
2. The inferred EPO human-chimpanzee ancestor allele is available and equal to at least one other human-ape ancestor (human-gorilla or human-orangutan).
3. Human and chimpanzee appear no more than once in the EPO alignment
4. The site is not flagged as a systematic error

We flagged positions in CpG and repeat masked (RM) regions, and nearby InDels (+/- 5bp), but did not exclude them from our analysis. Positions where at least one of the archaic humans is flagged as having low statistical evidence for having a SNP ("LowQual" in FILTER column of GATK VCF) were also flagged.

Catalog of present-day human-specific sites

To identify sites where all or nearly all present-day humans carry the derived allele, while the archaic genomes carry the ancestral allele (Figure S11), we used the Vindija Neandertal, Altai Neandertal and Denisovan exomes. We did not use the El Sidrón Neandertal for defining high-quality sites because the exome is of substantially lower coverage than the other three archaic individuals, and requiring a high-quality call for the El Sidrón individual leads to the exclusion of a large number of high-quality sites in the other individuals. For all sites that passed quality filters in the higher coverage individuals, we extracted the corresponding genotypes, including the genotype call for the El Sidrón individual. If the El Sidrón individual did not have a genotype call at those sites, we marked its genotype as "./.".

We kept only those sites:

1. That have a genotype quality (GQ) >= 30 in both Altai Neandertal and Denisova and GQ >= 10 in Vindija
2. Where the fraction of mapped reads with MQ = 0 is below 10% in Altai Neandertal, Vindija and Denisova
3. With a root mean square map quality (RMS MQ) >= 30 in Altai Neandertal, Vindija and Denisova
4. Where the coverage is within 95% of the genome-wide (exome-wide for Vindija) distribution
5. That, when heterozygous, the minor allele frequency is equal to or larger than 25% in Altai Neandertal, Vindija and Denisova

We then retained sites where present-day humans are either fixed, or carry the derived allele at high frequency (>90%) based on the allele frequencies in the 20110521 release of the 1000 Genomes project(34). Finally, we required that at least one of the eight possible alleles in the four archaic individuals be the same as the human-chimpanzee ancestor allele (Figure S11A). For downstream analyses, we also obtained the subset of these positions where all four archaic individuals are homozygous for the ancestral allele (Figure S11B).

Catalogs of archaic-specific and Neandertal-specific sites

To look for archaic-specific sites (shared by Neandertal+Denisova) and Neandertal-specific sites (Figures S12A and S12B), we required that:

1. Sites have a GQ >= 30 in both Altai Neandertal and Denisova and GQ >= 10 in both Vindija and Sidron
2. The fraction of mapped reads with MQ = 0 is below 10% in all individuals

3. the RMS MQ >= 30 in all archaic human individuals
4. Coverage is within 95% of the genome-wide (exome-wide for Vindija and Sidron) distribution
5. At heterozygous positions, the minor allele frequency is equal to or larger than 25% in all archaic human individuals

In the case of archaic-specific sites, we required that present-day humans be fixed or at high frequency (>90%) for the ancestral allele, and that all or at least seven of the eight alleles in the four archaic individuals be derived and the same as one another (Figure S12A). In the case of the Neandertal-specific sites, we required that present-day humans be fixed or at high frequency (>90%) for the ancestral allele, that both Denisovan alleles be equal to the ancestral allele, and that all or at least five of the six alleles in the three Neandertal individuals be derived and the same as one another (Figure S12B).

Annotation

We used Ensembl's Variant Effect Predictor v.2.5 (Ensembl 67 annotation) to annotate the changes identified in each of the catalogs. We filtered changes by their most severe predicted effect as in Supplementary Note 19 of Meyer et al. (2012), and restricted our ontology analyses to non-synonymous changes within the genic regions of the longest transcripts of CCDS-verified genes (Consensus Coding Sequence Project of EBI, NCBI, WTSI, and UCSC - Sep. 7th 2011 release).

Present-day human-specific changes

The number of present-day human-specific SNCs classified by different predicted consequences is detailed in Table S21.

Neandertal-specific changes

The number of Neandertal-specific SNCs classified by different predicted consequences in archaic and present-day humans is detailed in Table S22.

Archaic-specific changes

The number of archaic-specific SNCs classified by different predicted consequences in archaic and present-day humans is detailed in Table S23.

Human Phenotype Ontology analyses

To identify the potential impact of non-synonymous SNCs on phenotypes, we used a phenotype-ontology enrichment analysis test. We identified phenotype categories that are enriched for genes with non-synonymous changes along particular lineages (present-day human, Neandertal or archaic human). We used the program FUNC (35) and the Human Phenotype Ontology (HPO) (36), which contains standardized terms for phenotypes associated with human disorders and to which genes associated with

those disorders are mapped. We test for differences in the proportion of non-synonymous substitutions within a particular category before and after the human-Neandertal or human-archaic split using a binomial test. Even with low absolute numbers of changes after the split, the test may yield significant results in a particular category as long as the pre-split changes in that category are proportionally lower (relative to all pre-split changes) than the post-split changes in that category (relative to all post-split changes). For example, two derived non-synonymous changes in a category out of 100 changes in the modern human lineage after the split with Neandertals will be significantly larger than four derived changes in the same category out of 10,000 changes before the split of modern humans and Neandertals. We assume free recombination.

a) Present-day human-specific changes

We used a binomial test as in Supplementary Note 19 of Meyer *et al*. (2012) (21), and compared non-synonymous SNCs that are derived specifically in present-day humans (*i.e.* archaic humans have the ancestral state) with non-synonymous SNCs that occurred before the modern human - archaic human split (*i.e.* all archaic humans and all present-day humans have the derived state, while the human-chimpanzee ancestor has the ancestral state). This test is meant to look for particular ontology enrichments in a lineage while controlling for differences in gene length, nucleotide content and other factors that may differ among ontology categories. We restricted the analysis to phenotype ontology terms to which at least two genes are mapped, and performed four comparisons for ontology enrichment, shown in Table S24, which answer different biological questions. For example, looking at sites where all archaic humans are ancestral may serve to identify more recent putative selective events in modern humans than a test looking at changes where at least one archaic human is ancestral. We only list overrepresented terms that have a significance level of $P < 0.01$ and False Discovery Rate (FDR) <= 0.10 within each comparison. Terms listed in bold have also Family-Wise Error rate (FWER) <= 0.10 within each comparison. Genes with present-day human-specific non-synonymous SNCs that are mapped to enriched phenotype categories or to daughter terms of enriched categories are shown in Table S25.

b) Neandertal-specific changes

We also tested for enrichment in phenotypic categories of genes with non-synonymous changes specific to the Neandertal lineage. We used the binomial test as above, comparing derived sites specific to Neandertals (ancestral in present-day humans and Denisova) with derived sites shared by all archaic and present-day humans. As above, we restricted the analysis to terms that contained at least two genes, and performed four types of comparisons, as shown in Table S26. We only list overrepresented terms that have a significance level of $p < 0.01$ and FDR <= 0.1 within each comparison. Terms listed in bold have also Family-Wise Error rate (FWER) <= 0.10 within each comparison. Genes with Neandertal-specific non-

synonymous SNCs that are mapped to enriched categories or to daughter terms of enriched categories are shown in Table S27.

c) Archaic-specific changes

Finally, we tested for enrichment along the archaic (Neandertal+Denisova) lineage. In this case, we compared derived sites specific to Neandertals and Denisova (ancestral in present-day humans) with derived sites shared by all archaic and present-day humans. The four types of comparisons we performed are shown in Table S28. We only list overrepresented terms that have a significance level of P-value < 0.01 and FDR <= 0.1 within each comparison. Terms listed in bold have also Family-Wise Error rate (FWER) <= 0.10 within each comparison. Genes with non-synonymous SNCs that are mapped to enriched categories or to daughter terms of enriched categories are shown in Table S29.

Disruptive non-synonymous mutations

To find the set of disruptive missense mutations in the modern human, archaic humans (Neandertal+Denisova) and Neandertal lineages, we replaced the VEP's default PolyPhen-2 predictions (which use the HumVar model) for PolyPhen-2 predictions (31) obtained using the HumDiv model, which are better trained for evolutionary genetics analyses. We then looked for sites predicted to be deleterious by both PolyPhen-2 and SIFT(37), in CCDS-verified genes. We note, however, that these score predictions are based on multiple alignments that include present-day humans, and that all predictions are with respect to the human reference – not the ancestral allele –, which may result in potentially deleterious changes that are specific to present-day humans being predicted as benign. We found seven deleterious sites in the present-day human-specific catalog (0.01% of total non-synonymous SNCs), 76 in the archaic catalog (0.21%) and 174 in the Neandertal catalog (0.21%). The higher number of deleterious sites in the archaic and Neandertal catalogs may be due to three possible reasons: a) the human reference bias mentioned above, b) an excess of fixed deleterious alleles in Neandertals and Denisova, or c) the fact that we only have four archaic individuals to determine whether the alleles is fixed or at high-frequency in these populations. In Tables S30 and S31, we show the deleterious changes in the modern human and archaic human catalogs ranked by their Grantham scores.

Nonsense mutations

We also identified nonsense mutations shared among the available archaic and present-day humans in both CCDS and non-CCDS genes. Table S32 lists changes particular to the present-day human lineage where a STOP codon is introduced or removed at a site for which at least one of the archaic individuals has at least one ancestral allele and the derived state is fixed or at high frequency in present-day humans (1000 Genomes). In turn, Tables S33 and S34 list nonsense mutations particular to the Neandertal and archaic human lineages, respectively. We observe a

higher STOP loss / STOP gain ratio among the present-day human-specfic SNCs (6/6) than in SNCs specific to Neandertals (3/13) or both archaic human groups (0/8).

Lineage-specific changes

a) Modern humans

Among the fixed present-day human-specific non-synonymous SNCs, we see enrichment for changes in phenotype categories related to aggression and hyperactivity. The genes with present-day human-specific SNCs contained within these categories are *ADSL*, *GLDC* and *SLITRK1*, each having one human-specific non-synonymous change.

A particularly interesting SNC is a C-to-T mutation (chr22:40760978) in position 429 of the *ADSL* gene (C-terminal domain). The site is homozygous for the ancestral allele in Vindija, Sidron, Altai Neandertal and Denisova, and codes for alanine, while the derived allele in present-day humans codes for valine and is completely fixed. The position is highly conserved for the ancestral state in primates (conservation score = 0.953), has a highly positive GERP score (5.67), and the ancestral amino acid is conserved across multiple tetrapods, including rhesus, mouse, dog, elephant, opossum and chicken. The amino acid position is three residues away from the most common SNP known to cause adenylosuccinase deficiency, which produces psychomotor delay, autism, epilepsy and mental retardation (38, 39) suggesting that particular features of cognitive function could be associated with this change.

When also including high-frequency changes, we see enrichment for a term related to pigmentation, specifically in dermal melanosomes. Melanosomes are granules that synthesize, transport and store melanin pigments within the cell (40). The genes with present-day human-specific SNCs in these categories are *GPR143* and *LYST*. *GPR143* has one SNC that is 99% derived in present-day humans (the ancestral allele is present in Africans at 2% frequency in the 1000G data) and codes for a G-protein coupled receptor expressed in the eye and dermal melanocytes (RefSeq, Dec 2009). Mutations in this gene have been associated with ocular albinism and congenital nystagmus (41, 42). *LYST* has three SNCs that are 97% derived in present-day humans (each ancestral allele is present in Africans at 12% frequency in the 1000G data). This gene codes for a lysosomal transport regulator protein. It has been associated with Chédiak–Higashi syndrome: an autosomal recessive disorder that causes defects in melanosomes that lead to albinism, neuropathies and immune system problems (43). We also find that this gene has been associated to eye color in an association study (44) and that it has signatures of positive selection due to its unusual patterns of genetic variation in East Asia (45).

One of the high-frequency derived SNCs in present-day humans leads to the gain of a STOP codon in *CASP12* that disrupts the gene's function. *CASP12* is associated with responses to bacterial infections (46) and has been associated with cerebral infarction

in mice (47). The Denisovan individual had already been shown to carry the ancestral (functional) variant of this gene (21) and we can now confirm that all three Neandertals are also homozygous for the ancestral state. The ancestral variant is common in sub-Saharan Africa (48) and the STOP-gain is inferred to have occurred before the Neolithic, based on ancient DNA evidence (49). It is also hypothesized to have been subject to positive selection before the out-of-Africa expansion (50). We also find a second highly disruptive high-frequency SNC in the *CASP12* gene (rs115100183), deemed damaging/deleterious by PolyPhen-2 and SIFT, but its significance is unknown and its rise to high frequency could have been due to the pseudogenization of *CASP12* after the gain of the STOP codon.

b) Neandertals

Among the derived non-synonymous changes seen on the Neandertal lineage, but that are ancestral in Denisova and present-day humans, the only significantly enriched phenotypic term is "hyperlordosis", which is a pathogenic condition characterized by an accentuated inward curvature of the lumbar region (MedlinePlus). Neandertals had a reduced lordotic curvature, relative to both modern humans and other archaic humans (51), so these genetic changes may be candidates for causative mutations in genes affecting lordotic curvature that could have produced this particular skeletal morphology in the evolution of Neandertals.

The genes mapped to hyperlordosis that have changes on the Neandertal lineage are *CUL7, GLB1, COL2A1, HSPG2, VPS13B* and *NEB*, and several of them are associated with multiple muscoskeletal disorders. A non-synonymous mutation in *CUL7* is derived in all Neandertals, is predicted by SIFT to be deleterious, has a highly positive GERP score (4.9) and is found in its ancestral state in 99% of present-day humans (rs61732148). Other mutations in this gene are associated with 3-M syndrome, which leads to growth retardation and skeletal abnormalities (52). *HSPG2* contains three non-synonymous Neandertal-specific mutations that are high frequency ancestral in present-day humans (rs2229493, rs2228349, rs35669711) and this gene is associated with dyssegmental dysplasia(53) and Schwartz-Jampel syndrome (54), involving skeletal anomalies. *VPS13B* contains one Neandertal-specific change that is fixed ancestral in present-day humans. This gene has been associated with Cohen syndrome, leading to craniofacial dysmorphism, obesity and elongated extremities (55).

In addition, *GLB1* has one Neandertal-derived SNC that is fixed ancestral in present-day humans. This gene codes for a galactosidase that has been linked to GM1-gangliosidosis: a lysosomal storage disorder that affects the nervous system (56). Finally, the gene *NEB* codes for a very long protein (6,669 aa) and also has three Neandertal-specific SNCs that are 98%-99% ancestral in present-day humans (rs118191309, rs76767949, rs117271684) but derived in all Neandertals. All three of these changes have highly positive GERP scores (4.9, 5.97, 4.8). Both rs76767949

and rs117271684 are predicted by PolyPhen to be damaging. This protein encoded by this gene is thought to preserve sarcomere integrity (57).

c) Archaic humans

Among the non-synonymous changes that are derived in both Neandertals and the Denisovan individual, but are fixed or high-frequency ancestral in present-day humans, we see several enriched terms, many of which are related to abnormalities of extremities, joints, digits, thorax and general mobility. Fossils suggest that Neandertals had powerful forearms (58) and a broad barrel-shaped rib cage (59), suggesting some of these changes may be causative of these particular morphologies. The genes with non-synonymous archaic-specific SNCs that are associated with the enriched terms are *ACAN, ABCA12, COL10A1, ESCO2, FANCA, FBN2, FRAS1, FREM1, FREM2, GAA, GLI3, MGP, MLL2, NEB, NIPBLNSD1, MOGS, SPECC1L* and *WDPCP*. Four of these genes are also mapped to enriched terms related to hair pattern development (*FRAS1, FREM2, NIPBL, NSD1*), which could suggest changes in external phenotypes controlled by these genes may have allowed these groups of humans to survive cold Pleistocene environments. For example, *NIPBL* contains an archaic-specific non-synonymous SNC that is homozygous derived in all three Neandertals and heterozygous in the Denisovan individual (the ancestral state is fixed in present-day humans). Mutations in this gene are associated with Cornelia de Lange syndrome, which can lead to excessive body hair and thick eyebrows, among other symptoms (60, 61).

The proteins encoded by *FREM1, FREM2* and *FRAS1* are known to form a ternary complex that plays a role in epidermal membrane adhesion during embryonic development of the skin (62). Both *FREM2* and *FRAS1* contain changes predicted to be deleterious by PolyPhen and SIFT, and that are derived in all or almost all archaic humans and ancestral at high frequency in present-day humans (rs7699637 at 97% ancestral frequency and rs9548505 at 91% ancestral frequency). Mutations in both of these genes have been linked to Fraser Syndrome (63-65), which leads to several malformations including syndactyly, hypertelorism and malformed fallopian tubes.

We also see an enrichment of terms for genes involved in alanine/pyruvate metabolism which could reflect particular modifications in nutrition or metabolic demands. Gómez-Olivencia *et al*. (2009) suggests the body proportions of archaic human groups may have been linked to increased oxygen consumption due to more costly energy demands. The genes associated with pyruvate metabolism are *PDHX* and *PC*, each containing one SNC that is fixed ancestral in present-day humans and derived in all archaic humans (GERP score > 4 in both cases). *PDHX* codes for a component of the pyruvate dehydrogenase complex (66), while *PC* codes for pyruvate carboxylase (55,56). Both proteins are involved in key steps during glucose metabolism in the mitochondrial matrix.

SI Appendix references

1. Rohland N & Hofreiter M (2007) Comparison and optimization of ancient DNA extraction. *BioTechniques* 42(3):343-352.
2. Kircher M, Sawyer S, & Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 40(1):e3.
3. Briggs AW, *et al.* (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res* 38(6):e87.
4. Meyer M & Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. in *Cold Spring Harb Protoc 2010: pdb prot5448*.
5. Coffey AJ, *et al.* (2011) The GENCODE exome: sequencing the complete human exome. *European journal of human genetics : EJHG* 19(7):827-831.
6. Burbano HA, *et al.* (2010) Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* 328(5979):723-725.
7. Hodges E, *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nature genetics* 39(12):1522-1527.
8. Fu Q, *et al.* (2012) DNA analysis of an early modern human from Tianyuan Cave, China. *Proc Natl Acad Sci U S A*.
9. Kircher M, Stenzel U, & Kelso J (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome biology* 10(8):R83.
10. Meyer M & Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* 2010(6):pdb prot5448.
11. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
12. Li H, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
13. McKenna A, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297-1303.
14. Ewing B & Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8(3):186-194.
15. Li Y, *et al.* (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature genetics* 42(11):969-972.
16. Hvilsom C, *et al.* (2012) Extensive X-linked adaptive evolution in central chimpanzees. *Proc Natl Acad Sci U S A* 109(6):2054-2059.
17. Green RE, *et al.* (2010) A draft sequence of the Neandertal genome. *Science* 328(5979):710-722.
18. Asan, *et al.* (2011) Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome biology* 12(9):R95.
19. Kalari KR, *et al.* (2006) First exons and introns--a survey of GC content and gene structure in the human genome. *In silico biology* 6(3):237-242.
20. Heinrich V, *et al.* (2012) The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Res* 40(6):2426-2431.
21. Meyer M, *et al.* (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338(6104):222-226.
22. Paten B, Herrero J, Beal K, Fitzgerald S, & Birney E (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* 18(11):1814-1828.
23. Paten B, *et al.* (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res* 18(11):1829-1843.

24. Siepel A, *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15(8):1034-1050.
25. Andrews RM, *et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature genetics* 23(2):147.
26. Green RE, *et al.* (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134(3):416-426.
27. Briggs AW, *et al.* (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* 325(5938):318-321.
28. Meacham F, *et al.* (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC bioinformatics* 12:451.
29. Lao O, *et al.* (2008) Correlation between genetic and geographic structure in Europe. *Current biology : CB* 18(16):1241-1248.
30. Simonsen M, Mailund T, & C.N.S P (2008) Rapid Neighbour Joining. *8th Workshop in Algorithms in Bioinformatics (WABI)*, (Springer Verlag), pp 113-122.
31. Adzhubei IA, *et al.* (2010) A method and server for predicting damaging missense mutations. *Nature methods* 7(4):248-249.
32. Grantham R (1974) Amino acid difference formula to help explain protein evolution. *Science* 185(4154):862-864.
33. Simons YB, Turchin MC, Pritchard JK, & Sella G (2014) The deleterious mutation load is insensitive to recent population history. *Nature genetics* 46(3):220-224.
34. Durbin RM, *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061-1073.
35. Prüfer K, *et al.* (2007) FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC bioinformatics* 8:41.
36. Robinson PN & Mundlos S (2010) The human phenotype ontology. *Clinical genetics* 77(6):525-534.
37. Kumar P, Henikoff S, & Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* 4(7):1073-1081.
38. Sebesta I, *et al.* (1997) Adenylosuccinase deficiency: clinical and biochemical findings in 5 Czech patients. *Journal of inherited metabolic disease* 20(3):343-344.
39. Gitiaux C, *et al.* (2009) Misleading behavioural phenotype with adenylosuccinate lyase deficiency. *European journal of human genetics : EJHG* 17(1):133-136.
40. Wasmeier C, Hume AN, Bolasco G, & Seabra MC (2008) Melanosomes at a glance. *Journal of cell science* 121(Pt 24):3995-3999.
41. Hu J, Liang D, Xue J, Liu J, & Wu L (2011) A novel GPR143 splicing mutation in a Chinese family with X-linked congenital nystagmus. *Molecular vision* 17:715-722.
42. Yan N, *et al.* (2012) A novel nonsense mutation of the GPR143 gene identified in a Chinese pedigree with ocular albinism. *PLoS One* 7(8):e43177.
43. Nagle DL, *et al.* (1996) Identification and mutation analysis of the complete gene for Chediak-Higashi syndrome. *Nature genetics* 14(3):307-311.
44. Liu F, *et al.* (2010) Digital quantification of human eye color highlights genetic association of three new loci. *PLoS Genet* 6(5):e1000934.
45. Hider JL, *et al.* (2013) Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC evolutionary biology* 13:150.
46. Saleh M, *et al.* (2006) Enhanced bacterial clearance and sepsis resistance in caspase-12-deficient mice. *Nature* 440(7087):1064-1068.
47. Shimoke K, *et al.* (2011) Appearance of nuclear-sorted caspase-12 fragments in cerebral cortical and hippocampal neurons in rats damaged by autologous blood clot embolic brain infarctions. *Cellular and molecular neurobiology* 31(5):795-802.
48. Kachapati K, O'Brien TR, Bergeron J, Zhang M, & Dean M (2006) Population distribution of the functional caspase-12 allele. *Human mutation* 27(9):975.
49. Hervella M, *et al.* (2012) The loss of functional caspase-12 in Europe is a pre-neolithic event. *PLoS One* 7(5):e37022.

50. Wang X, Grus WE, & Zhang J (2006) Gene losses during human origins. *PLoS Biol* 4(3):e52.
51. Been E, Gomez-Olivencia A, & Kramer PA (2012) Lumbar lordosis of extinct hominins. *Am J Phys Anthropol* 147(1):64-77.
52. Huber C*, et al.* (2005) Identification of mutations in CUL7 in 3-M syndrome. *Nature genetics* 37(10):1119-1124.
53. Arikawa-Hirasawa E*, et al.* (2001) Dyssegmental dysplasia, Silverman-Handmaker type, is caused by functional null mutations of the perlecan gene. *Nature genetics* 27(4):431-434.
54. Nicole S*, et al.* (2000) Perlecan, the major proteoglycan of basement membranes, is altered in patients with Schwartz-Jampel syndrome (chondrodystrophic myotonia). *Nature genetics* 26(4):480-483.
55. Kolehmainen J*, et al.* (2003) Cohen syndrome is caused by mutations in a novel gene, COH1, encoding a transmembrane protein with a presumed role in vesicle-mediated sorting and intracellular protein transport. *Am J Hum Genet* 72(6):1359-1369.
56. Hinek A, Zhang S, Smith AC, & Callahan JW (2000) Impaired elastic-fiber assembly by fibroblasts from patients with either Morquio B disease or infantile GM1-gangliosidosis is linked to deficiency in the 67-kD spliced variant of beta-galactosidase. *Am J Hum Genet* 67(1):23-36.
57. Labeit S & Kolmerer B (1995) The complete primary structure of human nebulin and its correlation to muscle structure. *Journal of molecular biology* 248(2):308-315.
58. De Groote I (2011) The Neanderthal lower arm. *Journal of human evolution* 61(4):396-410.
59. Gomez-Olivencia A, Eaves-Johnson KL, Franciscus RG, Carretero JM, & Arsuaga JL (2009) Kebara 2: new insights regarding the most complete Neandertal thorax. *Journal of human evolution* 57(1):75-90.
60. Krantz ID*, et al.* (2004) Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B. *Nature genetics* 36(6):631-635.
61. Tonkin ET, Wang TJ, Lisgo S, Bamshad MJ, & Strachan T (2004) NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome. *Nature genetics* 36(6):636-641.
62. Kiyozumi D, Sugimoto N, & Sekiguchi K (2006) Breakdown of the reciprocal stabilization of QBRICK/Frem1, Fras1, and Frem2 at the basement membrane provokes Fraser syndrome-like defects. *Proc Natl Acad Sci U S A* 103(32):11981-11986.
63. McGregor L*, et al.* (2003) Fraser syndrome and mouse blebbed phenotype caused by mutations in FRAS1/Fras1 encoding a putative extracellular matrix protein. *Nature genetics* 34(2):203-208.
64. Shafeghati Y, Kniepert A, Vakili G, & Zenker M (2008) Fraser syndrome due to homozygosity for a splice site mutation of FREM2. *American journal of medical genetics. Part A* 146A(4):529-531.
65. van Haelst MM*, et al.* (2008) Molecular study of 33 families with Fraser syndrome new data and mutation review. *American journal of medical genetics. Part A* 146A(17):2252-2257.
66. Aral B*, et al.* (1997) Mutations in PDX1, the human lipoyl-containing component X of the pyruvate dehydrogenase-complex gene on chromosome 11p1, in congenital lactic acidosis. *Am J Hum Genet* 61(6):1318-1326.
67. McVicker G, Gordon D, Davis C, & Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* 5(5):e1000471.
68. McVicker G & Green P (2010) Genomic signatures of germline gene expression. *Genome Res* 20(11):1503-1511.
69. Rosenbloom KR*, et al.* (2012) ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res* 40(Database issue):D912-917.

70. Cooper GM, *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15(7):901-913.
71. Davydov EV, *et al.* (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 6(12):e1001025.
72. Li H, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.

SI Appendix Tables

Bone	Extract name	mg of bone used per extract	Libraries from extract
Sidron 1253	E608	100	L8201, L8204
Sidron 1253	E609	172	L8205-L8208
Sidron 1253	E600	240	L8211-L8213, L8245
Sidron 1253	E601	280	L8214-L8216, L8246
Sidron 1253	E612	133	L8219-L8221, L8241
Sidron 1253	E636	158	L8209, L8222-L8224
Sidron 1253	E238	160	L8229, L8254
Sidron 1253	E640	61	L8234, L8210
Sidron 1253	E641	95	L8237-L8239
Sidron 1253	E613	114	L8242-L8244, L8247
Sidron 1253	E239	70	L8255
Sidron 1253	E428	22	L8249
Sidron 1253	E429	16	L8250
Sidron 1253	E430	45	L8251
Sidron 1253	E431	16	L8252
Sidron 1253	E432	16	L8253
Sidron 1253	E414	78	L8256
Vindija 33.15	E565	53	L7855
Vindija 33.15	E566	66	L7865
Vindija 33.15	E929	311	L7849
Vindija 33.15	E930	383	L7850
Vindija 33.15	E931	414	L7851
Vindija 33.15	E932	459	L7852
Vindija 33.15	E933	259	L7856
Vindija 33.15	E934	260	L7857
Vindija 33.15	E935	260	L7858
Vindija 33.15	E936	260	L7859
Vindija 33.15	E937	279	L7860
Vindija 33.15	E938	260	L7861
Vindija 33.15	E939	261	L7862
Vindija 33.15	E940	270	L7863
Vindija 33.15	E941	268	L7864
Vindija 33.15	E942	266	L7866
Vindija 33.15	E943	289	L7853
Vindija 33.15	E966	re-extract from E931	L7867
Vindija 33.15	E967	re-extract from E932	L7868
Vindija 33.15	E968	re-extract from E933	L7869

Table S1. A list of the extracts and libraries made from Sidron 1253 and Vindija 33.15.

Annotation	Genes	Transcripts	Coding Exons	Non-ovelapping coding exons	Coding Nucleotides
CCDS		23,333	239,830	180,973	31,485,559
RefSeq	19,025	31,011	319,688	191,732	34,600,533
GENCODE	14,492	41,078	345,016	149,695	25,515,206
Primary Target	19,130	42,498	422,709	193,548	33,298,430
Tiled Target	17,367	39,099	396,952	173,086	27,332,312

Table S2. Summary of the gene annotation sets used to build the primary and tiled target regions of the capture array.

Sample	Reference	Aln. bases	Aln. reads	Q30 aln. bases	Q30 aln. reads
El Sidrón	*GRCh37*	1.026E+09	1.769E+07	6.443E+08	1.095E+07
El Sidrón	*pantro2*	8.666E+08	1.490E+07	6.166E+08	1.049E+07
Vindija	*GRCh37*	3.359E+09	5.580E+07	2.322E+09	3.774E+07
Vindija	*pantro2*	2.902E+09	4.804E+07	2.216E+09	3.601E+07

Table S3: Aligned unique sequence data obtained for the El Sidrón and Vindija captures against the human reference genome (*GRCh37*) and the chimpanzee reference genome (pantro2). Duplicate reads are removed before this calculation.

Neandertals	Mapping Quality	Number of sequences	% on tiled regions	% on tiled target	% bases covered	Average coverage
El Sidrón	All sequences	17,692,351	77.1	60.1	93.8	16.0
	Sequences ≥ MQ1	11,513,828	84.1	64.7	93.8	12.5
	Sequences ≥ MQ30	10,952,479	92.1	70.9	90.4	12.1
Vindija	All sequences	55,799,068	77.9	61.6	98.3	51.6
	Sequences ≥ MQ1	39,333,747	82.4	64.0	98.3	42.0
	Sequences ≥ MQ30	37,735,664	84.7	65.7	98.2	41.7

Table S4. Capture efficiency for El Sidrón and Vindija Neandertal exomes. The percentage of sequences that fall on the tiled target is used to assess the efficiency of the capture protocol. These sequences are also used to compute the average coverage per exome at different mapping quality (MQ) thresholds. MQ1 and MQ30 refer to sequences with a mapping quality score of at least one and 30, respectively. Duplicate reads are removed before these calculations.

Group	Population	Individual	Lower bin	Higher bin
Neandertals	El Sidrón	Sid1253	2 (6)	33
	Vindija	Vi33.15	4 (6)	95
	Altai		25	72
Denisovan	Denisova	Denisovan	10	46
Africans	Yoruba	HGDP00927	8	48
	Mandenka	HGDP01284	7	39
	Dinka	DNK02	7	42
Europeans	French	HGDP00521	7	41
	Sardinian	HGDP00665	6	38
	Italian (ancestry)	NA12891	9	45
Asians	Han	HGDP00778	7	42
	Dai	HGDP01307	8	45
	Papuan	HGDP00542	7	40

Table S5. Range of coverage bins encompassing the data used for analysis (central 95%) in each archaic and present-day human exome. A minimum coverage of six is used (in parenthesis). Duplicate reads are removed before this calculation.

		Draft Neandertal	El Sidrón Neandertal	Vindija Neandertal	Altai Neandertal	Denisovan
Average coverage		1.3-fold	12.5-fold	42.0-fold	45.3-fold	27.0-fold
Coding bases (27,332,312)		14,289,004 (52.3%)	25,651,281 (93.8%)	26,876,083 (98.3%)	27,313,554 (99.9%)	27,303,406 (99.9%)
Coding exons (173,086)	80%	32,277 (18.6%)	160,037 (92.5%)	169,810 (98.1%)	172,946 (99.9%)	172,858 (99.8%)
	90%	17,954 (10.4%)	154,814 (89.4%)	168,348 (97.2%)	172,932 (99.9%)	172,809 (99.8%)
	100%	8,612 (5.0%)	144,574 (83.5%)	165,235 (95.4%)	172,893 (99.9%)	172,479 (99.6%)
Coding genes (17,367)	80%	177 (1.0%)	15,502 (89.3%)	16,944 (97.6%)	17,322 (99.7%)	17,308 (99.7%)
	90%	46 (0.3%)	13,561 (78.1%)	16,294 (93.8%)	17,313 (99.7%)	17,289 (99.5%)
	100%	19 (0.1%)	5,476 (31.5%)	11,696 (67.3%)	17,264 (99.4%)	16,894 (97.3%)

Table S6. Coverage comparison per base, exon and gene for the Neandertal draft, El Sidrón, Vindija Altai and Denisovan exomes. Bases must be covered at least once to be considered. The number of exons and genes with at least 80%, 90% and 100% of their bases covered is shown. In parenthesis, the percentage of bases, exons or genes covered is given. Duplicate reads are removed before these calculations.

Group	Population	Per BP error [%] (corrected)	Genotype diff [%]	Per BP error [%] (corrected)	Genotype diff [%]
Neandertals	El Sidrón	-	-	0.188 (0.160)	0.028
	Vindija	-	-	0.171 (0.143)	0.028
	Altai	-	-	0.304 (0.275)	0.029
Denisovan	Denisovan	0.183 (0.133)	0.050	0.174 (0.145)	0.029
Africans	Yoruba	0.219 (0.185)	0.034	0.222 (0.203)	0.019
	Mandenka	0.213 (0.179)	0.034	0.209 (0.190)	0.019
	Dinka	0.213 (0.180)	0.033	0.213 (0.199)	0.019
Europeans	French	0.216 (0.186)	0.030	0.218 (0.200)	0.017
	Sardinian	0.207 (0.177)	0.030	0.206 (0.188)	0.018
	Italian (ancestry)	1.446 (1.417)	0.029	1.467 (1.450)	0.017
Asians	Han	0.209 (0.178)	0.031	0.212 (0.194)	0.018
	Dai	0.212 (0.181)	0.031	0.210 (0.193)	0.018
	Papuan	0.206 (0.173)	0.033	0.206 (0.187)	0.019

Table S7. Comparison of coverage statistics and error rates (autosomes only). The divergence-corrected per-base error rates are given in brackets.

Neandertals	Bounds	Neandertal fragments	Present-day human fragments	Lower 95% CI	Mean	Upper 95% CI
El Sidrón	Lower	1,997,710	4,726	0.23	0.24	0.24
	Upper	577,474	2,317	0.38	0.40	0.42
Vindija	Lower	2,298,077	6,378	0.27	0.28	0.28
	Upper	591,807	6,469	1.05	1.08	1.08

Table S8. Mitochondrial contamination estimates (%).

Additional information	Sub-field(s)	Field	Description	References / Data Sources
EPO primate ancestor bases	CAnc, GAnc, OAnc	INFO	Inferred ancestor bases from EPO primate blocks.	Ensembl Compara EPO 6 primate genome alignments (22, 23) (Ensembl release 64).
EPO alignment codes	TS	INFO	One-letter code of each of the species aligned in the EPO block for the site: 'H' (human), 'P' (chimpanzee), 'G' (gorilla), 'O' (orangutan), 'M' (macaque), 'C' (callitrix).	
EPO reference primate bases	TSseq	INFO	All primate reference bases aligned in the EPO alignment corresponding to the site, in the same order as the species codes in the TS subfield.	
CpG flag	CpG	INFO	Flag added if the site is within a CpG dinucleotide context in the inferred human-chimpanzee ancestor or the human reference genome.	
Repeat masking flag	RM	INFO	Flag added if the site appears as soft-masked in the EPO block.	
Primate conservation score	pSC	INFO	Primate phastCons conservation scores calculated using 6-primates EPO alignments, excluding humans.	Meyer et al. 2012 (21)
Mammal conservation score	mSC	INFO	Mammalian phastCons conservation scores calculated using EPO alignments of 35 eutherian mammals, excluding humans.	Meyer et al. 2012 (21)
1000 Genomes dbSNP ID	-	ID	dbSNP ID obtained from the 1000 Genomes Project.	1000 Genomes Project 20110521 release (34)
1000 Genomes alternative allele	1000gALT	INFO	Alternative alleles obtained from the 1000 Genomes Project.	
1000 Genomes allele frequencies	AF1000g, AMR_AF, ASN_AF, EUR_AF, AFR_AF	INFO	Global and population-wide average allele frequencies corresponding to the alternative allele in 1000gALT.	
Systematic error flag #1	SysErr	INFO	Flag added if site is in a region where the sequence context causes high error rates in Illumina sequencing.	Meyer et al. 2012 (21)
Systematic error flag #2	SysErrHCB	INFO	Flag added if site is in a region inferred to be prone to systematic errors due to shared SNPs across humans, chimpanzees and bonobos.	See this SOM.
Background selection score	bSC	INFO	Score that reflects information about local density of conserved elements and recombination rate.	McVicker et al. 2009 (67), McVicker and Green 2010 (68)
Mapability score	Map20	INFO	Score that reflects the mapability of the region in which the site is located, based on uniqueness of 20-mer sequences across the genome.	UCSC Duke Uniqueness track from ENCODE project (69)
Structural variation control	UR	INFO	Flag added if the site is in a region not considered a candidate for structural variation.	Meyer et al. 2012 (21)
Genomic Evolutionary Rate Profiling (GERP) scores	GRP	INFO	Score that reflects a per-base estimate of evolutionary constraint based on a 35-mammal alignment.	UCSC GERP track (Cooper et al. 2005 (70), Davydov et al. 2010 (71))
Number of A,C,G,T bases and reads starting with InDels	A, C, G, T, IR	FORMAT	Number of each type of base observed for each strand and the number of reads starting an insertion or a deletion at the position.	Obtained using samtools mpileup (72)

Table S9. Additional annotation included in extended Variant Call Format files, using the pipeline first described in Meyer *et al*.(21).

Group	Population	Individual	0/0	0/1	1/1	Homozygous derived	Total
Neandertals	El Sidrón	SD1253	18,799,261	2,690	9,897	53,566	18,811,848
	Vindija	Vi33.15	23,264,913	2,954	15,245	68,757	23,283,112
	Altai	Toe bone	25,630,957	2,908	17,400	77,326	25,651,265
Denisovan	Denisova	Finger bone	24,691,231	3,578	16,874	73,059	24,711,683

Table S10. The number of high-quality homozygous reference (0/0), heterozygous (0/1) and homozygous alternative (1/1) coding genotypes in the autosomes of each archaic exome. The homozygous derived column contains 0/0 and 1/1 derived genotypes with respect to the human-chimpanzee ancestral allele.

Groups	Population	Individual	0/0	0/1	1/1	Homozygous derived	Total
Africans	Mandenka	HGDP01284	24,638,601	12,591	6,428	68,062	24,657,620
	Yoruba	HGDP00927	24,766,451	12,591	6,662	68,819	24,785,704
	Dinka	DNK02	24,793,220	12,482	6,509	68,927	24,812,211
Europeans	French	HGDP00521	24,670,111	9,878	5,994	69,645	24,685,983
	Sardinian	HGDP00665	24,569,299	9,600	5,996	69,275	24,584,895
	Italian (ancestry)	NA12891	24,638,486	9,158	5,874	69,314	24,653,518
Asians	Han	HGDP00778	24,718,396	9,347	6,635	70,191	24,734,378
	Dai	HGDP01307	24,746,356	9,263	6,595	70,121	24,762,214
	Papuan	HGDP00542	24,612,370	7,900	7,524	70,357	24,627,794

Table S11. The number of high-quality homozygous reference (0/0), heterozygous (0/1) and homozygous alternative (1/1) coding genotypes in the autosomes of each of the nine present-day human exomes. The homozygous derived column contains 0/0 and 1/1 derived genotypes with respect the human-chimpanzee ancestral allele.

Group	Population	Individual	Heterozygosity
Neandertals	El Sidrón	SD1253	0.143
	Vindija	Vi33.15	0.127
	Altai	Toe bone	0.113
Denisovan	Denisova	Finger bone	0.145
Africans	Mandenka	HGDP01284	0.511
	Yoruba	HGDP00927	0.508
	Dinka	DNK02	0.503
Europeans	French	HGDP00521	0.400
	Sardinian	HGDP00665	0.390
	Italian (ancestry)	NA12891	0.371
Asians	Han	HGDP00778	0.378
	Dai	HGDP01307	0.374
	Papuan	HGDP00542	0.321

Table S12. Heterozygosity is given as the number of heterozygotes calls per thousand coding sites in the autosomes (from counts in Tables S10 and S11). These heterozygosities are discussed in the main text.

			Heterozygosity		
Group	**Population**	**Individual**	**10-fold**	**15-fold**	**20-fold**
Neandertals	El Sidrón	SD1253	0.148	0.160	0.163
	Vindija	Vi33.15	0.126	0.125	0.122
	Altai	Toe bone	0.113	0.113	0.113
Denisovan	Denisova	Finger bone	0.145	0.142	0.131
Africans	Mandenka	HGDP01284	0.510	0.503	0.478
	Yoruba	HGDP00927	0.508	0.500	0.487
	Dinka	DNK02	0.502	0.490	0.467
Europeans	French	HGDP00521	0.399	0.393	0.379
	Sardinian	HGDP00665	0.389	0.380	0.365
	Italian (ancestry)	NA12891	0.371	0.364	0.347
Asians	Han	HGDP00778	0.378	0.372	0.358
	Dai	HGDP01307	0.373	0.371	0.360
	Papuan	HGDP00542	0.321	0.316	0.305

Table S13. Heterozygosity is given as the number of heterozygotes calls per thousand coding sites in the autosomes. Coverage per position is at least 10-, 15- or 20-fold.

Group comparisons	Derived alleles	Derived SNPs	Derived "fixed"
Derived alleles in Neandertals shared with Africans	33.0%	10.9%	12.7%
Derived alleles in Neandertals shared with Europeans	31.8%	9.3%	19.4%
Derived alleles in Neandertals shared with Asians	32.7%	9.9%	19.7%
Derived alleles in Africans shared with Neandertals	18.8%	3.6%	42.7%
Derived alleles in Europeans shared with Neandertals	23.6%	4.5%	41.6%
Derived alleles in Asians shared with Neandertals	24.3%	4.7%	43.1%

Table S14. Proportion of derived alleles in one group that are shared with another group.

Group	Population	Total	Non-synonymous	PP-2 "deleterious"	PC "deleterious"	GS "deleterious"
Neandertals	El Sidrón	1,689	877 (51.9%)	418 (50.0%)	482 (55.3%)	173 (19.7%)
	Vindija	1,227	638 (52.0%)	281 (46.4%)	311 (48.8%)	124 (19.4%)
	Altai	1,032	518 (50.2%)	222 (44.8%)	260 (50.4%)	100 (19.3%)
	Group average	1,333	678 (50.8%)	307 (47.1%)	351 (51.5%)	132 (19.5%)
Denisovan	Denisova	1,629	899 (55.2%)	403 (47.4%)	413 (46.2%)	195 (21.7%)
Africans	Yoruba	5,139	2,261 (44.0%)	631 (35.5%)	833 (37.0%)	411 (18.2%)
	Mandenka	5,198	2,338 (45.0%)	669 (35.7%)	865 (37.1%)	431 (18.4%)
	Dinka	5,084	2,308 (45.4%)	643 (35.0%)	871 (37.9%)	403 (17.5%)
	Group average	5,140	2,302 (44.8%)	648 (35.4%)	856 (37.3%)	415 (18.0%)
Europeans	French	3,846	1,765 (45.9%)	463 (35.5%)	658 (37.5%)	321 (18.2%)
	Sardinian	3,793	1,721 (45.4%)	477 (38.0%)	686 (40.0%)	299 (17.4%)
	Italian (ancestry)	3,721	1,668 (44.8%)	451 (26.9%)	632 (38.0%)	290 (17.4%)
	Group average	3,787	1,718 (45.4%)	464 (33.5%)	659 (38,5%)	303 (17.7%)
Asians	Han	3,601	1,614 (44.8%)	443 (36.9%)	611 (38.0%)	302 (18.7%)
	Dai	3,541	1,588 (44.8%)	462 (37.9%)	614 (38.8%)	271 (17.1%)
	Papuan	3,098	1,443 (46.6%)	409 (37.1%)	561 (39.1%)	280 (19.4%)
	Group average	3,413	1,538 (45.1%)	438 (37.3%)	595 (38.6%)	284 (18.4%)

Table S15. Derived heterozygous genotypes in each individual since the split of the archaic and modern human lineages by functional class. For the PolyPhen-2 (PP-2), PhastCons (PC) and Grantham scores (GS) categories, the number and proportion of non-synonymous alleles that are predicted to be "deleterious" are shown. Because the different individual exomes have different sequence coverage, we only use sites that pass all quality filters in all individuals.

Group	Population	Total	Non-synonymous	PP-2 "deleterious"	PC "deleterious"	GS "deleterious"
Neandertals	El Sidrón	2,624	1,168 (44.5%)	403 (36.7%)	471 (40.4%)	188 (16.1%)
	Vindija	2,750	1,241 (45.1%)	430 (36.8%)	498 (40.3%)	202 (16.3%)
	Altai	2,666	1,211 (45.4%)	414 (36.1%)	502 (41.5%)	194 (16.0%)
	Group average	2,680	1,207 (45.0%)	416 (36.5%)	490 (40.7%)	195 (16.1%)
Denisovan	Denisova	3,131	1,459 (46.6%)	497 (36.1%)	559 (38.4%)	246 (16.9%)
Africans	Yoruba	1,717	695 (40.5%)	100 (35.2%)	218 (31.5%)	101 (14.5%)
	Mandenka	1,645	686 (41.7%)	78 (28.8%)	222 (32.6%)	98 (14.3%)
	Dinka	1,731	726 (41.9%)	91 (31.3%)	205 (28.3%)	95 (13.1%)
	Group average	1,698	702 (41.3%)	90 (31.8%)	215 (30.8%)	98 (14.0%)
Europeans	French	2,062	857 (41.6%)	96 (32.4%)	269 (31.5%)	130 (15.2%)
	Sardinian	2,058	861 (41.8%)	133 (40.4%)	284 (33.1%)	123 (14.3%)
	Italian (ancestry)	2,102	890 (42.4%)	117 (37.3%)	284 (32.1%)	143 (16.1%)
	Group average	2,074	869 (41.9%)	115 (36.7%)	279 (32.2%)	132 (15.2%)
Asians	Han	2,133	891 (41.8%)	112 (31.7%)	277 (31.2%)	137 (15.4%)
	Dai	2,162	918 (42.5%)	115 (34.0%)	271 (29.8%)	154 (16.8%)
	Papuan	2,460	1,053 (42.8%)	154 (32.4%)	356 (34.1%)	166 (15.8%)
	Group average	2,252	954 (42.4%)	127 (32.7%)	301 (31.7%)	152 (16.0%)

Table S16. Derived homozygous genotypes in each individual since the split of the archaic and modern human lineage by functional class. For the PolyPhen-2 (PP-2), PhastCons (PC) and Grantham scores (GS) categories, the number and proportion of non-synonymous alleles that are predicted to be "deleterious" are shown. Because the different individual exomes have different sequence coverage, we only use sites that pass all quality filters in all individuals.

		Heterozygous				Homozygous			
Group	**Population**	**10-fold**	**15-fold**	**20-fold**	**Without potential deaminations**	**10-fold**	**15-fold**	**20-fold**	**Without potential deaminations**
Neandertals	El Sidrón	56.6%	59.4%	61.8%	59.2%	40.9%	40.6%	43.3%	44.2%
	Vindija	48.9%	48.6%	48.4%	50.2%	40.2%	40.4%	40.6%	45.0%
	Altai	50.4%	50.4%	50.4%	55.0%	41.5%	41.5%	41.5%	45.4%
	Group average	52.0%	52.8%	53.5%	54.8%	40.9%	40.8%	41.8%	44.9%
Denisovan	Denisova	46.2%	46.8%	45.6%	47.9%	38.4%	38.4%	39.3%	41.9%
Africans	Yoruba	37.0%	36.9%	37.3%	41.5%	31.5%	31.5%	31.8%	34.0%
	Mandenka	37.2%	37.6%	39.1%	38.6%	32.6%	32.3%	32.6%	34.0%
	Dinka	38.0%	38.1%	38.9%	41.0%	28.5%	29.1%	29.5%	32.2%
	Group average	37.4%	37.5%	38.4%	40.4%	30.9%	31.0%	31.3%	33.4%
Europeans	French	37.5%	38.1%	38.6%	41.4%	31.5%	32.3%	33.6%	34.3%
	Sardinian	40.1%	40.5%	42.9%	42.9%	33.2%	33.5%	34.3%	36.7%
	Italian (ancestry)	38.0%	38.4%	39.1%	41.0%	32.1%	32.1%	32.2%	34.1%
	Group average	38.5%	39.0%	40.2%	41.8%	32.3%	32.6%	33.4%	35.0%
Asians	Han	38.0%	38.0%	38.8%	38.5%	31.3%	31.8%	31.7%	34.1%
	Dai	38.8%	38.9%	39.2%	40.8%	29.8%	30.0%	30.0%	31.4%
	Papuan	39.1%	39.4%	40.6%	42.0%	34.0%	34.5%	34.6%	36.4%
	Group average	38.6%	38.8%	39.5%	40.4%	31.7%	32.1%	32.1%	34.0%

Table S17. Proportion of non-synonymous alleles that are predicted to be "deleterious" by PhastCons with (at 10-, 15- or 20-fold minimum sequence coverage) or without potentially deaminated sites (ancestral C to derived T and ancestral G to derived A) in the archaic individuals. These sites are also removed from present-day humans for this comparison.

Category		Derived alleles							
		Polymorphic positions (SNPs)				Homozygous positions ("fixed")			
		Neandertals*	Africans†	Europeans†	Asians†	Neandertals*	Africans†	Europeans†	Asians†
Co-ding	Synonymous	2,501 (48.9%) 2,460 (48.3%) 2,483 (40.0%)	8,250 (54.3%)	5,571 (53.6%)	5,518 (53.3%)	1,273 (56.1%) 1,319 (56.1%) 1,270 (56.5%)	383 (59.3%)	624 (57.2%)	603 (57.8%)
	Non-synonymous	2,611 (51.1%) 2,629 (51.6%) 2,589 (51.0%)	6,947 (45.7%)	4,820 (46.4%)	4,839 (46.7%)	996 (43.9%) 1,032 (43.9%) 979 (43.5%)	263 (40.7%)	467 (42.8%)	441 (42.2%)
PP-2	"Benign"	1,354 (54.6%) 1,358 (54.6%) 1,332 (54.1%)	3,596 (63.6%)	2,266 (62.1%)	2,292 (61.9%)	600 (63.8%) 606 (63.7%) 579 (64.8%)	16 (55.2%)	53 (66.3%)	59 (64.1%)
	"Deleterious"	1,128 (45.4%) 1,131 (45.4%) 1,128 (45.9%)	2,057 (36.4%)	1,385 (37.9%)	1,409 (38.1%)	340 (36.2%) 345 (36.3%) 315 (35.2%)	13 (44.8%)	31 (36.9%)	33 (35.9%)
PC	"Benign"	1,296 (49.9%) 1,309 (50.1%) 1,277 (49.6%)	4,280 (61.9%)	2,872 (59.8)	2,903 (60.3%)	601 (60.4%) 626 (60.8%) 595 (60.8%)	185 (70.6%)	335 (72.2%)	321 (73.1%)
	"Deleterious"	1,303 (50.1%) 1,303 (50.1%) 1,299 (50.4%)	2,635 (38.1%)	1,931 (40.2%)	1,913 (39.7%)	394 (39.6%) 404 (39.2%) 383 (39.2%)	77 (29.4%)	129 (27.8%)	118 (26.9%)
GS	"Benign"	2,112 (80.9%) 2,134 (81.2%) 2,097 (81.0%)	5,673 (81.7%)	3,936 (81.7%)	3,954 (81.7%)	833 (83.6%) 862 (83.5%) 822 (84.0%)	231 (87.8%)	413 (88.4%)	374 (84.8%)
	"Deleterious"	499 (19.1%) 495 (18.8%) 492 (19.0%)	1,274 (18.3%)	884 (18.3%)	885 (18.3%)	163 (16.4%) 170 (16.5%) 157 (16.0%)	32 (12.2%)	54 (11.6%)	67 (15.2%)

Table S18. Distribution of derived alleles by functional class. For the coding category, the number and proportion of coding alleles that are either synonymous or non-synonymous are shown. For the PolyPhen-2 (PP-2), PhastCons (PC) and Grantham scores (GS) categories, the number and proportion of non-synonymous alleles that are predicted to be either "benign" or "deleterious" are shown.

* Derived in Neandertals and homozygous ancestral in Africans (1st row), Europeans (2nd row) and Asians (3rd row).
† Derived in Africans, Europeans or Asians and homozygous ancestral in Neandertals.

Group	Category	1/6		2/6		3/6		4/6		5/6		Total	
		PP-2	PC	PP-2	PC	PP-2	PC	PP-2	PC	PP-2	PC	PP-2	PC
Neandertals	**"Benign"**	717	652	308	301	150	151	103	115	76	77	1,354	1,296
	"Deleterious"	691	837	237	256	86	93	67	68	47	49	1,128	1,303
Africans	**"Benign"**	2,454	2,636	697	807	290	429	108	252	47	156	3,596	4,280
	"Deleterious"	1,524	1,794	314	475	136	202	58	102	24	63	2,057	2,635
Europeans	**"Benign"**	1,300	1,375	488	635	264	415	142	252	72	195	2,266	2,872
	"Deleterious"	889	1,136	263	362	113	189	73	158	47	86	1,385	1,931
Asians	**"Benign"**	1,258	1,379	533	666	291	416	128	236	82	206	2,292	2,903
	"Deleterious"	902	1,068	254	372	149	253	59	118	45	102	1,409	1,913

Table S19. Frequency distribution of the non-synonymous derived alleles inferred by PolyPhen-2 (PP-2) or PhastCons (PC) to be either "benign" or "deleterious" in Neandertals and present-day humans from Table 2. These counts are used in the frequency spectra in Figures 2 and S10.

		Derived alleles in genes associated to recessive traits			
		Heterozygous positions		Homozygous positions	
Group	**Population**	**"Benign"**	**"Deleterious"**	**"Benign"**	**"Deleterious"**
Neandertals	El Sidrón	38	45	54	58
	Vindija	22	34	60	61
	Altai	22	25	56	59
	Group average	(44.0%) 27.3	(56.0%) 34.7	(49.9%) 56.7	(51.1%) 59.3
Denisovan	Denisova	35	36	88	66
Africans	Yoruba	117	83	34	24
	Mandenka	102	74	47	27
	Dinka	107	95	48	26
	Group average	(56.4%) 108.7	(43.6%) 84.0	(65.3%) 43	(34.7%) 25.7
Europeans	French	79	61	53	32
	Sardinian	78	62	56	39
	Italian (ancestry)	82	57	62	35
	Group average	(57.1%) 79.7	(42.9%) 60.0	(61.8%) 57	(38.2%) 35.3
Asians	Han	85	72	64	37
	Dai	68	74	61	22
	Papuan	63	55	66	46
	Group average	(51.8%) 72.0	(48.2%) 67.0	(64.5%) 63.7	(35.5%) 35.0

Table S20. The proportion (per group) and number of genes associated to autosomal recessive traits (36) with "benign" or "deleterious" non-synonymous derived alleles, as predicted by PhastCons, since the split of the archaic and modern human lineages. Genes with only "benign" or one or more "deleterious" alleles are included in the "benign" or "deleterious" categories, respectively. Genes with heterozygous and homozygous "deleterious" or heterozygous and homozygous "benign" alleles are included in the corresponding homozygous category. See SI text for details.

	Present-day human-specific high-frequency (not fixed) SNCs						Present-day human-specific fixed SNCs					
	Total	*Non-synonymous*	*Synonymous*	*Splice sites*	*3' UTR*	*5' UTR*	*Total*	*Non-synonymous*	*Synonymous*	*Splice sites*	*3' UTR*	*5' UTR*
At least one ancestral allele found in at least 1 archaic human	3,167	432	608	114	85	45	1,760	241	308	65	63	27
All archaic humans are homozygous ancestral	1,208	169	256	51	30	17	348	58	71	16	9	4

Table S21. Classification of present-day human-specific changes by level of fixation in present-day humans, ancestral state in archaic humans and predicted consequence. High-frequency SNCs are sites with >90% global derived allele frequency in present-day humans but that are not yet fixed. Fixed SNCs are sites that either do not have a 1000G SNP or have a SNP with a derived allele frequency = 100% (within the resolution of 1000G). Numbers for non-synonymous, synonymous, 3' UTR and 5' UTR changes are restricted to those in genes with a CCDS ID.

	SNCs with ancestral state at high frequency (not fixed) in present-day humans and homozygous in Denisova						SNCs with ancestral state fixed in present-day humans and homozygous in Denisova					
	Total	*Non-synonymous*	*Synonymous*	*Splice sites*	*3' UTR*	*5' UTR*	*Total*	*Non-synonymous*	*Synonymous*	*Splice sites*	*3' UTR*	*5' UTR*
At least five out of six Neandertal alleles are derived	1,903	335	424	67	52	21	2,480	477	576	89	66	38
All Neandertals are homozygous derived	1,649	294	370	61	42	20	2,068	409	479	76	57	30

Table S22. Derived changes in Neandertals in sites where the ancestral allele is homozygous ancestral in Denisova and fixed or at high frequency (>90% but not fixed) in present-day humans. Numbers for non-synonymous, synonymous, 3' UTR and 5' UTR changes are restricted only to genes with a CCDS ID.

	SNCs with ancestral state at high frequency (not fixed) in present-day humans						SNCs with ancestral state fixed in present-day humans					
	Total	*Non-synonymous*	*Synonymous*	*Splice sites*	*3' UTR*	*5' UTR*	*Total*	*Non-synonymous*	*Synonymous*	*Splice sites*	*3' UTR*	*5' UTR*
At least seven out of eight archaic human alleles are derived	1223	202	272	49	29	12	944	165	206	37	27	16
All archaic humans are homozygous derived	1061	177	243	44	24	12	809	145	174	35	24	14

Table S23. Derived changes in archaic humans in sites where the ancestral allele is fixed or at high frequency (>90% but not fixed) in present-day humans. Numbers for non-synonymous, synonymous, 3' UTR and 5' UTR changes are restricted only to genes with a CCDS ID.

List 1: present-day human derived after present-day-archaic split	List 2: present-day human derived before present-day-archaic split	Human phenotype overrepresentation after the present-day-archaic split
Present-day human fixed derived SNCs in sites where at least one archaic human has at least one ancestral allele	Present-day human fixed derived SNCs in sites that are homozygous derived in all archaic humans	None
Present-day human fixed derived SNCs in sites that are homozygous ancestral in all archaic humans	Present-day human fixed derived SNCs in sites that are homozygous derived in all archaic humans	- Hyperactivity (P=0.00031; FWER=0.12; FDR=0.048) - Self-mutilation (P=0.00047; FWER=0.17; FDR=0.051) - Aggressive behavior (P=0.00019; FWER=0.078; FDR=0.078) - Abnormal aggressive, impulsive or violent behavior (P=0.00023; FWER=0.093; FDR=0.048) - Autoaggression (P=0.00047; FWER=0.17; FDR=0.051)
Present-day human fixed and high-frequency derived SNCs in sites where at least one archaic human has at least one ancestral allele	Present-day human fixed and high-frequency derived SNCs in sites that are homozygous derived in all archaic humans	None
Present-day human fixed and high-frequency derived SNCs in sites that are homozygous ancestral in all archaic humans	Present-day human fixed and high-frequency derived SNCs in sites that are homozygous derived in all archaic humans	- **Giant melanosomes in melanocytes** (P=3.2e-5; FWER=0.059; FDR=0.059)

Table S24. Four types of FUNC binomial tests between present-day human-specific non-synonymous changes and non-synonymous changes having occurred before the modern-archaic human split, to find overrepresented Human Phenotype Ontology terms among the present-day human-specific changes. Listed in parentheses are the raw P-values, the family-wise error rate (FWER) and the false discovery rate (FDR). Colored in blue are terms that have a refined P-value < 0.01 after exclusion of categories that are significant only because their subtree contains categories that are significant. Terms listed in bold blue have also a FWER <= 0.10.

Gene	Ensembl IDs	Protein changes and HGNC description	Enriched HPO terms	HPO IDs
ADSL	ENSG00000239900 ENST00000216194 ENSP00000216194	A429V adenylosuccinate lyase	Hyperactivity; Self-mutilation; Aggressive behavior; Abnormal aggressive, impulsive or violent behavior	HP:0000752; HP:0000742; HP:0000718; HP:0006919
GLDC	ENSG00000178445 ENST00000321612 ENSP00000370737	F220L glycine dehydrogenase (decarboxylating)	Hyperactivity; Aggressive behavior; Abnormal aggressive, impulsive or violent behavior	HP:0000752; HP:0000718; HP:0006919
SLITRK1	ENSG00000178235 ENST00000377084 ENST00000377084	A330S SLIT and NTRK-like family, member 1	Hyperactivity; Self-mutilation; Aggressive behavior; Abnormal aggressive, impulsive or violent behavior	HP:0000752; HP:0000742; HP:0000718; HP:0006919
GPR143	ENSG00000101850 ENST00000467482 ENSP00000417161	V346M G protein-coupled receptor 143	Giant melanosomes in melanocytes	HP:0005592
LYST	ENSG00000143669 ENST00000389794 ENSP00000374444	V192L N1017S D2804G Lysosomal trafficking regulator	Giant melanosomes in melanocytes	HP:0005592

Table S25. IDs and description of amino acid changes (ancestral to derived) and genes that are mapped either directly to enriched HPO terms or to a daughter term of an enriched HPO term in the modern human lineage, and that have modern-human-specific non-synonymous SNCs (see Table S24).

List 1: Neandertal derived after present-day-archaic split	List 2: Neandertal derived before present-day-archaic split	Human phenotype overrepresentation after the present-day-archaic split
Present-day human fixed ancestral SNCs in sites where at least five out of six Neandertal alleles are derived	Present-day human fixed derived SNCs in sites that are homozygous derived in all archaic humans	None
Present-day human fixed ancestral SNCs in sites where all six Neandertal alleles are derived	Present-day human fixed derived SNCs in sites that are homozygous derived in all archaic humans	None
Present-day human fixed and high-frequency ancestral SNCs in sites where at least five out of six Neandertal alleles are derived	Present-day human fixed and high-frequency derived SNCs in sites that are homozygous derived in all archaic humans	None
Present-day human fixed and high-frequency ancestral SNCs in sites where all six Neandertal alleles are derived	Present-day human fixed and high-frequency derived SNCs in sites that are homozygous derived in all archaic humans	- **Hyperlordosis** (P=4.5e-5; FWER=0.036; FDR=0.021)

Table S26. Four types of FUNC binomial tests between Neandertal-specific non-synonymous changes and non-synonymous changes having occurred before the modern-archaic human split, to find overrepresented Human Phenotype Ontology terms among the Neandertal-specific changes. Listed in parentheses are the raw P-values, the family-wise error rate (FWER) and the false discovery rate (FDR). Colored in blue are terms that have a refined P-value < 0.01 after exclusion of categories that are significant only because their subtree contains categories that are significant. Terms listed in bold blue have also a FWER <= 0.10.

Gene	Ensembl IDs	Protein changes and HGNC description	Enriched HPO terms	HPO IDs
CUL7	ENSG00000044090 ENST00000535468 ENSP00000438788	A955V cullin 7	Hyperlordosis	HP:0003307
GLB1	ENSG00000170266 ENST00000307363 ENSP00000306920	A79T galactosidase, beta 1	Hyperlordosis	HP:0003307
NEB	ENSG00000183091 ENST00000397345 ENSP00000380505	V1701A D2730G K7867Q nebulin	Hyperlordosis	HP:0003307
COL2A1	ENSG00000139219 ENST00000380518 ENSP00000369889	V1331I collagen, type II, alpha 1	Hyperlordosis	HP:0003307
HSPG2	ENSG00000142798 ENST00000374695 ENSP00000363827	G2225S A3168T R3632Q heparan sulfate proteoglycan 2	Hyperlordosis	HP:0003307
VPS13B	ENSG00000132549 ENST00000358544 ENSP00000351346	L2065V vacuolar sorting protein 13 homolog B (yeast)	Hyperlordosis	HP:0003307

Table S27. IDs and description of amino acid changes (ancestral to derived) and genes that are mapped either directly to an enriched HPO term or to a daughter term of an enriched term in the Neandertal lineage, and that have Neandertal-specific non-synonymous SNCs (see Table S26).

List 1: archaic human derived after modern-archaic split	List 2: archaic human derived before modern-archaic split	Human phenotype overrepresentation after the modern-archaic split
Present-day human fixed ancestral SNCs in sites where at least seven out of eight archaic alleles are derived	Present-day human fixed derived SNCs in sites that are homozygous derived in all archaic humans	- Hyperalaninemia (P=0.00021; FWER=0.15; FDR=0.062) - Increased serum pyruvate (P=0.00043; FWER=0.23; FDR=0.082) - Abnormality of alanine metabolism (P=0.00021; FWER=0.15; FDR=0.062) - Abnormality of pyruvate family amino acid metabolism (P=0.00021; FWER=0.15; FDR=0.062) - Abnormality of glycolysis (P=0.00043; FWER=0.23; FDR=0.082)
Present-day human fixed ancestral SNCs in sites where all eight archaic alleles are derived	Present-day human fixed derived SNCs in sites that are homozygous derived in all archaic humans	- Hyperalaninemia (P=0.00021; FWER=0.15; FDR=0.062) - Increased serum pyruvate (P=0.00043; FWER=0.23; FDR=0.082) - Abnormality of alanine metabolism (P=0.00021; FWER=0.145; FDR=0.062) - Abnormality of pyruvate family amino acid metabolism (P=0.00021; FWER=0.15; FDR=0.062) - Abnormality of glycolysis (P=0.00043; FWER=0.225; FDR=0.082)
Present-day human fixed and high-frequency ancestral SNCs in sites where at least seven out of eight archaic alleles are derived	Present-day human fixed and high-frequency derived SNCs in sites that are homozygous derived in all archaic humans	- **Ventricular septal defect** (P=5.8e-5; FWER=0.02; FDR=0.0096) - Patent ductus arteriosus (P=0.00022; FWER=0.11; FDR=0.025) - Abnormality of the thorax (P=0.00041; FWER=0.22; FDR=0.036) - Metatarsus adductus (P=0.00083; FWER=0.37; FDR=0.039) - Hyperalaninemia (P=0.001; FWER=0.47; FDR=0.04) - Aplasia/Hypoplasia of the thumb (P=0.0012; FWER=0.51; FDR=0.049) - Abnormality of the frontal hairline (P=0.0013; FWER=0.51; FDR=0.049) - Elbow flexion contracture (P=0.0013; FWER=0.54; FDR=0.053) - High-arched palate (P=0.0016; FWER=0.57; FDR=0.055) - Renal hypoplasia (P=0.0017; FWER=0.61; FDR=0.06) - Cleft eyelid (P=0.00187825; FWER=0.63; FDR=0.058) - Increased serum pyruvate (P=0.0021; FWER=0.68; FDR=0.065)

		- Partial agenesis of the corpus callosum (P=0.0021; FWER=0.7; FDR=0.067) - Abnormality of the tongue (P=0.0023; FWER=0.71; FDR=0.068) - Abnormality of the umbilicus (P=0.0025; FWER=0.72; FDR=0.068) - Abnormality of the nipple (P=0.003; FWER=0.76; FDR=0.077) - Cryptorchidism (P=0.003; FWER=0.77; FDR=0.075) - Facial cleft (P=0.003; FWER=0.77; FDR=0.076) - Abnormality of the anus (P=0.0032; FWER=0.8; FDR=0.075) - Defect in the atrial septum (P=0.0032; FWER=0.8; FDR=0.075) - Limited elbow extension (P=0.0031; FWER=0.8; FDR=0.077) - Prominent occiput (P=0.0031; FWER=0.8; FDR=0.077) - Short finger (P=0.0033; FWER=0.8; FDR=0.074) - Calf muscle hypoplasia (P=0.0034; FWER=0.89; FDR=0.095) - Congenital kyphoscoliosis (P=0.0034; FWER=0.89; FDR=0.095) - Crumpled ear (P=0.0034; FWER=0.89; FDR=0.095) - Facial capillary hemangioma (P=0.0034; FWER=0.89; FDR=0.095) - Patellar subluxation (P=0.0034; FWER=0.89; FDR=0.095) - Hypertelorism (P=0.004; FWER=0.9; FDR=0.098) - Aplasia/Hypoplasia of the sternum (P=0.0041; FWER=0.9; FDR=0.097) - Abnormality of the ventricular septum (P=5.8e-5; FWER=0.02; FDR=0.0096) - Aplasia/Hypoplasia of fingers (P=5.9e-5; FWER=0.022; FDR=0.0081) - Abnormality of the thumb (P=0.00011; FWER=0.057; FDR=0.017) - Abnormality of finger (P=0.00012; FWER=0.058; FDR=0.015) - Limited elbow movement (P=0.00017; FWER=0.093; FDR=0.022) - Abnormality of the palate (P=0.00026; FWER=0.14; FDR=0.029) - Abnormality of cardiac ventricle (P=0.00031; FWER=0.15; FDR=0.029) - Congenital malformation of the great arteries (P=0.00031; FWER=0.16; FDR=0.027) - Abnormality of the cardiac septa (P=0.00035; FWER=0.17; FDR=0.028) - Coloboma (P=0.00042; FWER=0.23; FDR=0.035) - Abnormality of the hairline (P=0.00043; FWER=0.23; FDR=0.033) - Aplasia/Hypoplasia involving bones of the hand (P=0.00046; FWER=0.24; FDR=0.032) - Congenital abnormal hair pattern (P=0.00048; FWER=0.25; FDR=0.032) - Abnormality of calvarial morphology (P=0.00059; FWER=0.29; FDR=0.035) - Abnormal hair pattern (P=0.00059; FWER=0.29; FDR=0.037) - High palate (P=0.0006; FWER=0.29; FDR=0.034) - Brachydactyly (hand) (P=0.00063; FWER=0.3; FDR=0.033) - Abnormality of the digits (P=0.00077; FWER=0.35; FDR=0.039) - Aplasia/Hypoplasia involving bones of the upper limbs (P=0.00081; FWER=0.36; FDR=0.039) - Aplasia/Hypoplasia of the extremities (P=0.00091; FWER=0.38; FDR=0.039) - Contractures of the joints of the upper limbs (P=0.0001; FWER=0.42; FDR=0.044) - Abnormality of alanine metabolism (P=0.001; FWER=0.47; FDR=0.045) - Abnormality of pyruvate family amino acid metabolism (P=0.001; FWER=0.47; FDR=0.045) - Abnormality of the joints of the lower limbs (P=0.0014; FWER=0.54; FDR=0.052) - Brachydactyly syndrome (P=0.0015; FWER=0.55; FDR=0.052) - Abnormality of the elbow (P=0.0017; FWER=0.605; FDR=0.0587394) - Deviation of finger (P=0.0017; FWER=0.61; FDR=0.058) - Abnormality of globe location (P=0.0019; FWER=0.62; FDR=0.059) - Abnormality of glycolysis (P=0.0021; FWER=0.68; FDR=0.065) - Abnormal localization of kidneys (P=0.0023; FWER=0.71; FDR=0.07) - Limb joint contracture (P=0.0023; FWER=0.71; FDR=0.069) - Aplasia/Hypoplasia involving the skeleton (P=0.0024; FWER=0.716; FDR=0.069) - Abnormality of the orbital region (P=0.0027; FWER=0.76; FDR=0.076) - Limitation of joint mobility (P=0.003; FWER=0.76; FDR=0.077) - Narrow palate (P=0.0032; FWER=0.8; FDR=0.075) - Deviation of the hand or of fingers of the hand (P=0.0034; FWER=0.81; FDR=0.074) - Aortic root dilatation (P=0.0034; FWER=0.89; FDR=0.095) - Dilatation of the ascending aorta (P=0.0034; FWER=0.89; FDR=0.095) - Syndactyly (P=0.0036; FWER=0.9; FDR=0.096) - Abnormality of cardiac atrium (P=0.0036; FWER=0.89; FDR=0.097) - Abnormality of the aorta (P=0.0037; FWER=0.9; FDR=0.096) - Abnormality of the scalp hair (P=0.0041; FWER=0.9; FDR=0.097) - Abnormality of the scalp (P=0.0041; FWER=0.9; FDR=0.097) - Abnormality of the testis (P=0.0038; FWER=0.9; FDR=0.099) - Abnormality of male internal genitalia (P=0.0039; FWER=0.9; FDR=0.099)
Present-day human fixed and high-frequency ancestral SNCs in sites where all eight archaic alleles are derived	Present-day human fixed and high-frequency derived SNCs in sites that are homozygous derived in all archaic humans	- Hyperalaninemia (P=0.00081; FWER=0.408; FDR=0.063) - **Patent ductus arteriosus** (P=0.00011; FWER=0.074; FDR=0.074) - **Ventricular septal defect** (P=0.00013; FWER=0.082; FDR=0.0279466) - Metatarsus adductus (P=0.00051; FWER=0.3; FDR=0.084) - Abnormality of the ventricular septum (P=0.00013; FWER=0.082; FDR=0.028) - Congenital malformation of the great arteries (P=0.00015; FWER=0.11; FDR=0.029) - Abnormality of cardiac ventricle (P=0.00049; FWER=0.26; FDR=0.084) - Abnormality of finger (P=0.00057; FWER=0.31; FDR=0.079) - Abnormality of the cardiac septa (P=0.00064; FWER=0.33; FDR=0.079) - Limited elbow movement (P=0.0007; FWER=0.36; FDR=0.08) - Abnormality of calvarial morphology (P=0.00076; FWER=0.38; FDR=0.079) - Abnormality of the palate (P=0.0008; FWER=0.39; FDR=0.075) - Abnormality of alanine metabolism (P=0.0008; FWER=0.41; FDR=0.063) - Abnormality of pyruvate family amino acid metabolism (P=0.00081; FWER=0.41; FDR=0.063) - Abnormality of the thumb (P=0.00090; FWER=0.455; FDR=0.066) - High palate (P=0.0009; FWER=0.46; FDR=0.07)

Table S28. Four types of FUNC binomial tests between archaic human-specific non-synonymous changes and non-synonymous changes having occurred before the modern-archaic human split, to find overrepresented Human Phenotype Ontology terms among the archaic human-specific changes. Listed in parentheses are the raw p-values, the family-wise error rate (FWER) and the false discovery rate (FDR). Colored in blue are terms that have a refined p-value < 0.01 after exclusion of categories that are significant only because their subtree contains categories that are significant. Terms listed in bold blue have also a FWER <= 0.10.

Gene	Ensembl IDs	Protein changes and HGNC description	HPO terms	HPO IDs
ABCA12	ENSG00000144452 ENST00000272895 ENSP00000272895	W199C ATP-binding cassette, sub-family A (ABC1), member 12	Aplasia/Hypoplasia of fingers; Abnormality of finger; Aplasia/Hypoplasia involving bones of the hand; Brachydactyly (hand); Abnormality of the digits; Aplasia/Hypoplasia involving bones of the upper limbs; Aplasia/Hypoplasia of the extremities; Brachydactyly syndrome; Abnormality of globe location; Aplasia/Hypoplasia involving the skeleton; Abnormality of the orbital region; Short finger	HP:0006265; HP:0001167; HP:0005927; HP:0100667; HP:0011297; HP:0006496; HP:0009815; HP:0001156; HP:0100886; HP:0009115; HP:0000315; HP:0009381
ACAN	ENSG00000157766 ENST00000439576 ENSP00000387356	I404V aggrecan	Abnormality of the joints of the lower limbs	HP:0100491
COL10A1	ENSG00000123500 ENST00000327673 ENSP00000327368	D128N collagen, type X, alpha 1	Aplasia/Hypoplasia of fingers; Abnormality of finger; Abnormality of the thorax; Aplasia/Hypoplasia involving bones of the hand; Brachydactyly (hand); Abnormality of the digits; Aplasia/Hypoplasia involving bones of the upper limbs; Aplasia/Hypoplasia of the extremities; Abnormality of the joints of the lower limbs; Brachydactyly syndrome; Aplasia/Hypoplasia involving the skeleton; Short finger	HP:0006265; HP:0001167; HP:0000765; HP:0005927; HP:0100667; HP:0011297; HP:0006496; HP:0009815; HP:0100491; HP:0001156; HP:0009115; HP:0009381
ESCO2	ENSG00000171320 ENST00000305188 ENSP00000306999	I508V establishment of cohesion 1 homolog 2 (*S. cerevisiae*)	Abnormality of the ventricular septum; Ventricular septal defect; Aplasia/Hypoplasia of fingers; Abnormality of the thumb; Abnormality of finger; Limited elbow movement; Patent ductus arteriosus; Abnormality of the palate; Abnormality of cardiac ventricle; Congenital malformation of the great arteries; Abnormality of the cardiac septa; Coloboma; Aplasia/Hypoplasia involving bones of the hand; Abnormality of calvarial morphology; Brachydactyly (hand); Abnormality of the digits; Aplasia/Hypoplasia involving bones of the upper limbs; Aplasia/Hypoplasia of the extremities; Contractures of the joints of the upper limbs; Aplasia/Hypoplasia of the thumb; Abnormality of the joints of the lower limbs; Elbow flexion contracture; Brachydactyly syndrome; Abnormality of the elbow; Deviation of finger; Abnormality of globe location; Cleft eyelid; Abnormal localization of kidneys; Limb joint contracture; Aplasia/Hypoplasia involving the skeleton; Abnormality of the orbital region; Limitation of joint mobility; Cryptorchidism; Defect in the atrial septum; Short finger; Deviation of the hand or of fingers of the hand; Facial capillary hemangioma; Syndactyly; Abnormality of cardiac atrium; Hypertelorism; Abnormality of the testis; Abnormality of male internal genitalia	HP:0010438; HP:0001629; HP:0006265; HP:0001172; HP:0001167; HP:0002996; HP:0001643; HP:0000174; HP:0001713; HP:0011603; HP:0001671; HP:0000589; HP:0005927; HP:0002648; HP:0100667; HP:0011297; HP:0006496; HP:0009815; HP:0100360; HP:0009601; HP:0100491; HP:0002987; HP:0001156; HP:0009811; HP:0004097; HP:0100886; HP:0000625; HP:0100542; HP:0003121; HP:0009115; HP:0000315; HP:0001376; HP:0000028; HP:0001631; HP:0009381; HP:0009484; HP:0000996; HP:0001159; HP:0005120; HP:0000316; HP:0000035; HP:0000022
FANCA	ENSG00000187741 ENST00000389301 ENSP00000373952	P643A Fanconi anemia, complementation group A	Aplasia/Hypoplasia of fingers; Abnormality of the thumb; Abnormality of finger; Aplasia/Hypoplasia involving bones of the hand; Brachydactyly (hand); Abnormality of the digits; Aplasia/Hypoplasia involving bones of the upper limbs; Aplasia/Hypoplasia of the extremities; Aplasia/Hypoplasia of the thumb; Brachydactyly syndrome; Abnormal localization of kidneys; Aplasia/Hypoplasia involving the skeleton; Abnormality of the orbital region; Cryptorchidism; Short finger; Abnormality of the testis; Abnormality of male internal genitalia	HP:0006265; HP:0001172; HP:0001167; HP:0005927; HP:0100667; HP:0011297; HP:0006496; HP:0009815; HP:0009601; HP:0001156; HP:0100542; HP:0009115; HP:0000315; HP:0000028; HP:0009381; HP:0000035; HP:0000022
FBN2	ENSG00000138829 ENST00000508053 ENSP00000424571	T1322A T1428N Fibrillin 2	Abnormality of the ventricular septum; Ventricular septal defect; Abnormality of the thumb; Abnormality of finger; Limited elbow movement; Patent ductus arteriosus; Abnormality of the palate; Abnormality of cardiac ventricle; Congenital malformation of the great arteries; Abnormality of the cardiac septa; Abnormality of the thorax; Coloboma; Abnormality of calvarial morphology; High palate;	HP:0010438; HP:0001629; HP:0001172; HP:0001167; HP:0002996; HP:0001643; HP:0000174; HP:0001713; HP:0011603; HP:0001671; HP:0000765; HP:0000589; HP:0002648; HP:0000218;

			Abnormality of the digits; Metatarsus adductus; Contractures of the joints of the upper limbs; Abnormality of the joints of the lower limbs; Elbow flexion contracture; High-arched palate; Abnormality of the elbow; Deviation of finger; Limb joint contracture; Aplasia/Hypoplasia involving the skeleton; Limitation of joint mobility; Defect in the atrial septum; Narrow palate; Deviation of the hand or of fingers of the hand; Aortic root dilatation; Calf muscle hypoplasia; Congenital kyphoscoliosis; Crumpled ear; Dilatation of the ascending aorta; Patellar subluxation; Abnormality of cardiac atrium; Abnormality of the aorta	HP:0011297; HP:0001840; HP:0100360; HP:0100491; HP:0002987; HP:0000156; HP:0009811; HP:0004097; HP:0003121; HP:0009115; HP:0001376; HP:0001631; HP:0000189; HP:0009484; HP:0002616; HP:0008962; HP:0008453; HP:0009901; HP:0005111; HP:0010499; HP:0005120; HP:0001679
FRAS1	ENSG00000138759 ENST00000264895 ENSP00000264895	S147G P209S Fraser syndrome 1	Aplasia/Hypoplasia of fingers; Abnormality of the thumb; Abnormality of finger; Abnormality of the palate; Abnormality of the thorax; Coloboma; Abnormality of the hairline; Aplasia/Hypoplasia involving bones of the hand; Congenital abnormal hair pattern; Abnormality of calvarial morphology; Abnormal hair pattern; Brachydactyly (hand); Abnormality of the digits; Aplasia/Hypoplasia involving bones of the upper limbs; Aplasia/Hypoplasia of the extremities; Aplasia/Hypoplasia of the thumb; Abnormality of the frontal hairline; Abnormality of the joints of the lower limbs; Brachydactyly syndrome; Renal hypoplasia; Abnormality of globe location; Cleft eyelid; Abnormality of the tongue; Abnormality of the umbilicus; Aplasia/Hypoplasia involving the skeleton; Abnormality of the orbital region; Abnormality of the nipple; Cryptorchidism; Facial cleft; Abnormality of the anus; Syndactyly; Abnormality of the scalp hair; Abnormality of the scalp; Hypertelorism; Abnormality of the testis; Abnormality of male internal genitalia; Aplasia/Hypoplasia of the sternum	HP:0006265; HP:0001172; HP:0001167; HP:0000174; HP:0000765; HP:0000589; HP:0009553; HP:0005927; HP:0011361; HP:0002648; HP:0010720; HP:0100667; HP:0011297; HP:0006496; HP:0009815; HP:0009601; HP:0000599; HP:0100491; HP:0001156; HP:0000089; HP:0100886; HP:0000625; HP:0000157; HP:0001551; HP:0009115; HP:0000315; HP:0004404; HP:0000028; HP:0002006; HP:0004378; HP:0001159; HP:0100037; HP:0001965; HP:0000316; HP:0000035; HP:0000022; HP:0006714
FREM1	ENSG00000164946 ENST00000380880 ENSP00000370262	A1358S FRAS1 related extracellular matrix 1	Abnormality of the thorax; Coloboma; Abnormality of globe location; Cleft eyelid; Abnormality of the orbital region; Abnormality of the anus; Hypertelorism	HP:0000765; HP:0000589; HP:0100886; HP:0000625; HP:0000315; HP:0004378; HP:0000316
FREM2	ENSG00000150893 ENST00000280481 ENSP00000280481	R2066C FRAS1 related extracellular matrix 1	Aplasia/Hypoplasia of fingers; Abnormality of the thumb; Abnormality of finger; Abnormality of the palate; Abnormality of the thorax; Coloboma; Abnormality of the hairline; Aplasia/Hypoplasia involving bones of the hand; Congenital abnormal hair pattern; Abnormality of calvarial morphology; Abnormal hair pattern; Brachydactyly (hand); Abnormality of the digits; Aplasia/Hypoplasia involving bones of the upper limbs; Aplasia/Hypoplasia of the extremities; Aplasia/Hypoplasia of the thumb; Abnormality of the frontal hairline; Abnormality of the joints of the lower limbs; Brachydactyly syndrome; Renal hypoplasia; Abnormality of globe location; Cleft eyelid; Abnormality of the tongue; Abnormality of the umbilicus; Aplasia/Hypoplasia involving the skeleton; Abnormality of the orbital region; Abnormality of the nipple; Cryptorchidism; Facial cleft; Abnormality of the anus; Syndactyly; Abnormality of the scalp hair; Abnormality of the scalp; Hypertelorism; Abnormality of the testis; Abnormality of male internal genitalia; Aplasia/Hypoplasia of the sternum	HP:0006265; HP:0001172; HP:0001167; HP:0000174; HP:0000765; HP:0000589; HP:0009553; HP:0005927; HP:0011361; HP:0002648; HP:0010720; HP:0100667; HP:0011297; HP:0006496; HP:0009815; HP:0009601; HP:0000599; HP:0100491; HP:0001156; HP:0000089; HP:0100886; HP:0000625; HP:0000157; HP:0001551; HP:0009115; HP:0000315; HP:0004404; HP:0000028; HP:0002006; HP:0004378; HP:0001159; HP:0100037; HP:0001965; HP:0000316; HP:0000035; HP:0000022; HP:0006714
GAA	ENSG00000171298 ENST00000390015 ENSP00000374665	S448T glucosidase, alpha; acid	Abnormality of the thorax; Abnormality of the tongue	HP:0000765; HP:0000157
GLI3	ENSG00000106571 ENST00000395925 ENSP00000379258	R1537C GLI family zinc finger 3	Abnormality of the ventricular septum; Ventricular septal defect; Aplasia/Hypoplasia of fingers; Abnormality of the thumb; Abnormality of finger; Patent ductus arteriosus; Abnormality of cardiac ventricle; Congenital malformation of the great arteries; Abnormality of the cardiac septa; Abnormality of the thorax; Aplasia/Hypoplasia involving bones of the hand; Abnormality of calvarial morphology; Brachydactyly (hand); Abnormality of the digits; Aplasia/Hypoplasia involving bones of the upper limbs; Aplasia/Hypoplasia of the extremities; Contractures of the joints of the upper limbs; Aplasia/Hypoplasia of the thumb; Abnormality of the joints of the lower limbs; Brachydactyly syndrome; Abnormality of the elbow; Renal hypoplasia; Abnormality of globe location; Abnormal localization of kidneys; Limb joint contracture; Abnormality of the tongue; Abnormality of the umbilicus; Aplasia/Hypoplasia involving the skeleton; Abnormality of the orbital region; Cryptorchidism; Abnormality of the anus; Short finger;	HP:0010438; HP:0001629; HP:0006265; HP:0001172; HP:0001167; HP:0001643; HP:0001713; HP:0011603; HP:0001671; HP:0000765; HP:0005927; HP:0002648; HP:0100667; HP:0011297; HP:0006496; HP:0009815; HP:0100360; HP:0009601; HP:0100491; HP:0001156; HP:0009811; HP:0000089; HP:0100886; HP:0100542; HP:0003121; HP:0000157; HP:0001551; HP:0009115; HP:0000315; HP:0000028; HP:0004378; HP:0009381; HP:0000996; HP:0001159; HP:0001679; HP:0000316;

			Facial capillary hemangioma; Syndactyly; Abnormality of the aorta; Hypertelorism; Abnormality of the testis; Abnormality of male internal genitalia	HP:0000035; HP:0000022
LAMB3	ENSG00000196878 ENST00000367030 ENSP00000355997	N690S A926D laminin, beta 3	Abnormality of finger; Abnormality of the digits; Contractures of the joints of the upper limbs; Abnormality of the joints of the lower limbs; Limb joint contracture	HP:0001167; HP:0011297; HP:0100360; HP:0100491; HP:0003121
MGP	ENSG00000111341 ENST00000228938 ENSP00000228938	A8V matrix Gla protein	Abnormality of the ventricular septum; Ventricular septal defect; Aplasia/Hypoplasia of fingers; Abnormality of the thumb; Abnormality of finger; Abnormality of cardiac ventricle; Abnormality of the cardiac septa; Abnormality of the thorax; Aplasia/Hypoplasia involving bones of the hand; Brachydactyly (hand); Abnormality of the digits; Aplasia/Hypoplasia involving bones of the upper limbs; Aplasia/Hypoplasia of the extremities; Aplasia/Hypoplasia of the thumb; Brachydactyly syndrome; Aplasia/Hypoplasia involving the skeleton; Short finger	HP:0010438; HP:0001629; HP:0006265; HP:0001172; HP:0001167; HP:0001713; HP:0001671; HP:0000765; HP:0005927; HP:0100667; HP:0011297; HP:0006496; HP:0009815; HP:0009601; HP:0001156; HP:0009115; HP:0009381
MLL2	ENSG00000167548 ENST00000301067 ENSP00000301067	A476T myeloid / lymphoid or mixed-lineage leukemia 2	Abnormality of the ventricular septum; Ventricular septal defect; Aplasia/Hypoplasia of fingers; Abnormality of finger; Abnormality of the palate; Abnormality of cardiac ventricle; Abnormality of the cardiac septa; Aplasia/Hypoplasia involving bones of the hand; High palate; Brachydactyly (hand); Abnormality of the digits; Aplasia/Hypoplasia involving bones of the upper limbs; Aplasia/Hypoplasia of the extremities; Abnormality of the joints of the lower limbs; Brachydactyly syndrome; High-arched palate; Abnormal localization of kidneys; Aplasia/Hypoplasia involving the skeleton; Cryptorchidism; Abnormality of the anus; Defect in the atrial septum; Short finger; Narrow palate; Abnormality of cardiac atrium; Abnormality of the aorta; Abnormality of the testis; Abnormality of male internal genitalia	HP:0010438; HP:0001629; HP:0006265; HP:0001167; HP:0000174; HP:0001713; HP:0001671; HP:0005927; HP:0000218; HP:0100667; HP:0011297; HP:0006496; HP:0009815; HP:0100491; HP:0001156; HP:0000156; HP:0100542; HP:0009115; HP:0000028; HP:0004378; HP:0001631; HP:0009381; HP:0000189; HP:0005120; HP:0001679; HP:0000035; HP:0000022
MOGS	ENSG00000115275 ENST00000233616 ENSP00000233616	R495Q mannosyl-oligosaccharide glucosidase	Abnormality of finger; Abnormality of the palate; Abnormality of the thorax; Abnormality of calvarial morphology; High palate; Abnormality of the digits; High-arched palate; Deviation of finger; Prominent occiput; Narrow palate; Deviation of the hand or of fingers of the hand	HP:0001167; HP:0000174; HP:0000765; HP:0002648; HP:0000218; HP:0011297; HP:0000156; HP:0004097; HP:0000269; HP:0000189; HP:0009484
NEB	ENSG00000183091 ENST00000397345 ENSP00000380505	G7635E nebulin	Abnormality of the palate; Abnormality of the thorax; High palate; High-arched palate; Narrow palate	HP:0000174; HP:0000765; HP:0000218; HP:0000156; HP:0000189
NIPBL	ENSG00000164190 ENST00000282516 ENSP00000282516	D2512E Nipped-B homolog (*Drosophila*)	Abnormality of the ventricular septum; Ventricular septal defect; Aplasia/Hypoplasia of fingers; Abnormality of the thumb; Abnormality of finger; Limited elbow movement; Abnormality of the palate; Abnormality of cardiac ventricle; Abnormality of the cardiac septa; Abnormality of the thorax; Coloboma; Abnormality of the hairline; Aplasia/Hypoplasia involving bones of the hand; Congenital abnormal hair pattern; Abnormal hair pattern; High palate; Abnormality of the digits; Aplasia/Hypoplasia involving bones of the upper limbs; Aplasia/Hypoplasia of the extremities; Contractures of the joints of the upper limbs; Elbow flexion contracture; High-arched palate; Abnormality of the elbow; Renal hypoplasia; Deviation of finger; Abnormality of globe location; Abnormal localization of kidneys; Limb joint contracture; Abnormality of the umbilicus; Aplasia/Hypoplasia involving the skeleton; Abnormality of the orbital region; Abnormality of the nipple; Limitation of joint mobility; Cryptorchidism; Limited elbow extension; Narrow palate; Deviation of the hand or of fingers of the hand; Syndactyly; Abnormality of the scalp hair; Abnormality of the scalp; Abnormality of the testis; Abnormality of male internal genitalia; Aplasia/Hypoplasia of the sternum	HP:0010438; HP:0001629; HP:0006265; HP:0001172; HP:0001167; HP:0002996; HP:0000174; HP:0001713; HP:0001671; HP:0000765; HP:0000589; HP:0009553; HP:0005927; HP:0011361; HP:0010720; HP:0000218; HP:0011297; HP:0006496; HP:0009815; HP:0100360; HP:0002987; HP:0000156; HP:0009811; HP:0000089; HP:0004097; HP:0100886; HP:0100542; HP:0003121; HP:0001551; HP:0009115; HP:0000315; HP:0004404; HP:0001376; HP:0000028; HP:0001377; HP:0000189; HP:0009484; HP:0001159; HP:0100037; HP:0001965; HP:0000035; HP:0000022; HP:0006714
NSD1	ENSG00000165671 ENST00000439151 ENSP00000395929	A1036P M2250I nuclear receptor binding SET domain protein 1	Abnormality of the ventricular septum; Ventricular septal defect; Abnormality of finger; Limited elbow movement; Patent ductus arteriosus; Abnormality of the palate; Abnormality of cardiac ventricle; Congenital malformation of the great arteries; Abnormality of the cardiac septa; Abnormality of the thorax; Abnormality of the hairline; Congenital abnormal hair pattern; Abnormality of calvarial	HP:0010438; HP:0001629; HP:0001167; HP:0002996; HP:0001643; HP:0000174; HP:0001713; HP:0011603; HP:0001671; HP:0000765; HP:0009553; HP:0011361; HP:0002648; HP:0010720;

			morphology; Abnormal hair pattern; High palate; Abnormality of the digits; Metatarsus adductus; Aplasia/Hypoplasia of the extremities; Contractures of the joints of the upper limbs; Abnormality of the frontal hairline; Abnormality of the joints of the lower limbs; High-arched palate; Abnormality of the elbow; Deviation of finger; Abnormality of globe location; Partial agenesis of the corpus callosum; Limb joint contracture; Abnormality of the tongue; Abnormality of the umbilicus; Aplasia/Hypoplasia involving the skeleton; Abnormality of the orbital region; Abnormality of the nipple; Limitation of joint mobility; Cryptorchidism; Defect in the atrial septum; Limited elbow extension; Prominent occiput; Narrow palate; Deviation of the hand or of fingers of the hand; Abnormality of cardiac atrium; Abnormality of the scalp hair; Abnormality of the scalp; Hypertelorism; Abnormality of the testis; Abnormality of male internal genitalia	HP:0000218; HP:0011297; HP:0001840; HP:0009815; HP:0100360; HP:0000599; HP:0100491; HP:0000156; HP:0009811; HP:0004097; HP:0100886; HP:0001338; HP:0003121; HP:0000157; HP:0001551; HP:0009115; HP:0000315; HP:0004404; HP:0001376; HP:0000028; HP:0001631; HP:0001377; HP:0000269; HP:0000189; HP:0009484; HP:0005120; HP:0100037; HP:0001965; HP:0000316; HP:0000035; HP:0000022
PC	ENSG00000173599 ENST00000393955 ENSP00000377527	R830H pyruvate carboxylase	Abnormality of alanine metabolism; Abnormality of pyruvate family amino acid metabolism; Hyperalaninemia; Abnormality of glycolysis; Increased serum pyruvate	HP:0010916; HP:0010915; HP:0003348; HP:0004366; HP:0003542
PDHX	ENSG00000110435 ENST00000227868 ENSP00000227868	N270S pyruvate dehydrogenase complex, component X	Abnormality of the palate; Abnormality of the thorax; Abnormality of calvarial morphology; High palate; Abnormality of alanine metabolism; Abnormality of pyruvate family amino acid metabolism; Hyperalaninemia; Abnormality of globe location; Abnormality of glycolysis; Increased serum pyruvate; Partial agenesis of the corpus callosum; Abnormality of the orbital region; Hypertelorism	HP:0000174; HP:0000765; HP:0002648; HP:0000218; HP:0010916; HP:0010915; HP:0003348; HP:0100886; HP:0004366; HP:0003542; HP:0001338; HP:0000315; HP:0000316
SPECC1L	ENSG00000100014 ENST00000437398 ENSP00000393363	L200F sperm antigen with calponin homology and coiled-coil domains 1-like	Facial cleft	HP:0002006
WDPCP	ENSG00000143951 ENST00000272321 ENSP00000272321	Y502C WD repeat containing planar cell polarity effector	Abnormality of finger; Abnormality of the palate; Abnormality of cardiac ventricle; High palate; Abnormality of the digits; Brachydactyly syndrome; High-arched palate; Deviation of finger; Narrow palate; Deviation of the hand or of fingers of the hand; Syndactyly; Abnormality of the testis; Abnormality of male internal genitalia	HP:0001167; HP:0000174; HP:0001713; HP:0000218; HP:0011297; HP:0001156; HP:0000156; HP:0004097; HP:0000189; HP:0009484; HP:0001159; HP:0000035; HP:0000022

Table S29. IDs and description of amino acid changes (ancestral to derived) and genes that are mapped either directly to an enriched HPO term or to the daughter term of an enriched HPO term in the archaic human (Neandertal+Denisova) lineage, and that have archaic-human-specific non-synonymous SNCs (see Table S28).

Position	Ancestral / derived alleles	Archaic genotypes (Vindija, Sidron, Altai, Denisova)	1000G derived frequency	Gene	Amino acid change	PolyPhen-2 prediction (score)	SIFT prediction (score)	Grantham score	Gene function / description (UniProtKB, EntrezGene)
chr9:125563200	T/C	T/T,T/C,T/T,T/T	fixed	OR1K1	C267R	possibly damaging (0.762)	deleterious (0.04)	180	Olfactory receptor
chr14:50298962	T/A	T/T,T/T,T/T,T/T	93%	NEMF	S257C	probably damaging (0.999)	deleterious (0.03)	112	Involved in nuclear export
chr10:11789382	G/A	A/A,A/A,A/A,G/G	93%	ECHDC3	A69T	probably damaging (0.989)	deleterious (0)	58	-
chr2:128381861	A/G	A/A,A/A,A/A,G/G	99%	MYO7B	Q1312R	possibly damaging (0.776)	deleterious (0.01)	43	Actin-based motor protein
chr11:104761921	C/T	T/T,T/T,T/T,C/C	99%	CASP12	V215I	possibly damaging	deleterious (0.05)	29	Reduces cytokine release during

						(0.866)			bacterial infection. Pseudogenized in most present-day humans (see Table 8 and Discussion).
chr9:6328947	T/C	T/T,T/T,T/T,T/T	94%	TPD52L3	F118L	probably damaging (0.999)	deleterious (0.03)	22	May have a function in spermatogenesis
chr19:44470189	T/A	A/A,A/A,A/A,T/T	95%	ZNF221	F179I	possibly damaging (0.664)	deleterious (0.01)	21	May have a function in transcriptional regulation

Table S30. List of most disruptive non-synonymous derived SNCs in the present-day human-specific catalog. Changes marked as "fixed*" are fixed in 1000G but have a dbSNP ID.

Position	Ancestral / derived alleles	Archaic genotypes (Vindija, Sidron, Altai, Denisova)	1000G derived frequency	Gene	Amino acid change	PolyPhen-2 prediction (score)	SIFT prediction (score)	Grantham score	Gene function / description (UniProtKB, EntrezGene)
Chr22:40161439	C/G	G/G,G/G,G/G,G/G	fixed	ENTHD1	W336C	probably damaging (0.998)	deleterious (0)	215	-
Chr11:119058712	C/T	T/T,T/T,T/T,T/T	fixed	PDZD3	R241C	probably damaging (0.998)	deleterious (0.01)	180	Acts as a regulatory protein that associates with GUCY2C and negatively modulates its heat-stable enterotoxin-mediated activation
Chr8:88365925	T/A	A/A,A/A,A/A,A/A	fixed	CNBD1	I405N	probably damaging (1)	deleterious (0)	149	-
Chr7:23313149	C/T	T/T,T/T,T/T,T/T	fixed	GPNMB	S492L	possibly damaging (0.951)	deleterious (0.01)	145	Could be a melanogenic enzyme
Chr9: 125424224	T/G	G/G,G/G,G/G,G/G	fixed	OR1L1	I77S	possibly damaging (0.528)	deleterious (0.03)	142	Odorant receptor
Chr11:119044199	C/T	T/T,T/T,T/T,T/T	fixed	NLRX1	R81W	probably damaging (0.978)	deleterious (0.02)	101	Participates in antiviral signaling
Chr12:120306870	G/A	A/A,A/A,A/A,A/A	fixed	CIT	R78W	possibly damaging (0.916)	deleterious (0.01)	101	-
Chr8:134024152	T/C	C/C,C/C,C/C,C/C	fixed	TG	L2090P	probably damaging (0.993)	deleterious (0)	98	Precursor of the iodinated thyroid hormones thyroxine (T4) and triiodothyronine (T3)
Chr11: 134183303	T/C	C/C,C/C,C/C,C/C	fixed	GLB1L3	L505P	probably damaging (0.999)	deleterious (0)	98	Belongs to the glycosyl hydrolase 35 family
Chr5: 159821755	A/T	T/T,T/T,T/T,T/T	fixed	C5orf54	M248K	probably damaging	deleterious (0)	95	May be derived from an ancient transposon that

						(0.999)			has lost its ability to translocate
chr1:247150674	T/G	G/G,G/G,G/G,G/G	fixed	ZNF695	K381N	possibly damaging (0.712)	deleterious (0)	94	May be involved in transcriptional regulation
Chr2:228758555	C/T	T/T,T/T,T/T,T/T	fixed	WDR69	T121I	possibly damaging (0.865)	deleterious (0.04)	89	May play a role in axonemal outer row dynein assembly
chrX:154020521	A/G	G/G,G/G,G/G,G/G	fixed	MPP1	Y48H	possibly damaging (0.955)	deleterious (0.01)	83	Essential regulator of neutrophil polarity
Chr6:13612048	C/T	T/T,T/T,T/T,T/T	fixed	SIRT5	T295M	possibly damaging (0.67)	deleterious (0.02)	81	NAD-dependent lysine demalonylase and desuccinylase that specifically removes malonyl and succinyl groups on target proteins
Chr10:115922567	T/G	G/G,G/G,G/G,G/G	fixed	C10orf118	K154T	possibly damaging (0.634)	deleterious (0.01)	78	-
Chr20:57769097	C/A	A/A,A/A,A/A,A/A	fixed	ZNF831	P10008H	possibly damaging (0.938)	deleterious (0.01)	77	Contains 2 C2H2-type zinc fingers
Chr12:16347327	G/A	A/A,A/A,A/A,A/A	fixed	SLC15A5	A515V	possibly damaging (0.86)	deleterious (0.01)	64	Proton oligopeptide cotransporter
Chr7:11676519	G/C	C/C,C/C,C/C,C/C	fixed	THSD7A	A87G	probably damaging (0.998)	deleterious (0.02)	60	The soluble form promotes endothelial cell migration and filopodia formationduring angiogenesis via a FAK-dependent mechanism
Chr12:11463280	C/G	G/G,G/G,G/G,G/G	fixed	PRB4	S18T	possibly damaging (0.736)	deleterious (0)	58	Salivary proline-rich protein
Chr17:39538600	T/C	C/C,C/C,C/C,C/C	fixed	KRT34	T9A	possibly damaging (0.688)	deleterious (0)	58	Hair keratin
chr1:156909625	C/T	T/T,T/T,T/T,T/T	fixed	ARHGEF11	E1271K	probably damaging (0.964	deleterious (0.01)	56	May play a role in the regulation of RhoA GTPase by guanine nucleotide-binding alpha-12 (GNA12) and alpha-13 (GNA13)
chr2:74689432	C/T	T/T,T/T,T/T,T/T	fixed	MOGS	R495Q	possibly damaging (0.728)	deleterious (0.04)	43	mannosyl-oligosaccharide glucosidase activity
Chr4:77231149	A/G	G/G,G/G,G/G,G/G	fixed	STBD1	H358R	possibly damaging (0.766)	deleterious (0)	29	May have the capability to bind to carbohydrates

Chr13:46935652	C/T	T/T,T/T,T/T,T/T	fixed	KIAA0226L	R348H	probably damaging (0.999)	deleterious (0)	29	-
Chr13:52951949	C/T	T/T,T/T,T/T,T/T	fixed	THSD1	R719H	possibly damaging (0.865)	deleterious (0.04)	29	Membrane protein
Chr19:18377835	C/T	T/T,T/T,T/T,T/T	fixed	KIAA1683	R172H	probably damaging (1)	deleterious (0.01)	29	-
Chr11:123994464	C/T	T/T,T/T,T/T,T/T	fixed	VWA5A	L373F	possibly damaging (0.803)	deleterious (0.02)	22	May play a role in tumorigenesis as a tumor suppressor
Chr11:126432775	C/T	T/T,T/T,T/T,T/T	fixed	KIRREL3	V30M	probably damaging (0.99)	deleterious (0.01)	21	-

Table S31. List of most disruptive non-synonymous derived SNCs in the archaic-specific catalog.

Position	Ancestral / derived alleles	Archaic genotypes (Vindija, Sidron, Altai, Denisova)	1000G derived frequency	Consequence	Gene	Transcripts affected	Gene function / description (UniProtKB, EntrezGene)
chr1:19597328	G/A	G/G,G/G,G/G,A/A	97%	STOP gained	AKR7L	ENST00000211454	Pseudogene in present-day humans
chr1:161967680	T/C	C/C,C/C,C/C,T/T	fixed	STOP lost	OLFML2B	ENST00000294794 ENST00000367940	-
chr1:183592594	G/A	A/A,A/A,A/A,G/G	92%	STOP gained	ARPC5	ENST00000367534	Regulation of actin polymerization
chr2:27551325	A/G	A/A,A/A,A/A,A/A	93%	STOP lost	GTF3C2	ENST00000415683	General transcription factor
chr2:198593260	C/A	A/A,A/A,A/A,C/C	99%	STOP lost	BOLL	ENST00000430004	RNA-binding protein, may be required during spermatogenesis
chr6:154360569	T/C	T/T,T/T,T/T,T/T	97%	STOP lost	OPRM1	ENST00000434900 ENST00000520282	Opioid receptor
chr11:64893151	C/T	C/C,C/C,C/C,C/C	fixed	STOP gained	MRPL49	ENST00000526171	Mitochondrial ribosomal protein
chr11:104763117	G/A	G/G,G/G,G/G,G/G	96%	STOP gained	CASP12	ENST00000375726 ENST00000422698 ENST00000433738 ENST00000441710 ENST00000446862 ENST00000447913 ENST00000448103 ENST00000494737 ENST00000508062	Reduces cytokine release during bacterial infection. The non-STOP variant is common in sub-Saharan African populations (Kachapati et al. 2006) and occurred before the Neolithic (Hervella et al. 2012)
chr12:57003964	A/T	A/A,A/A,A/A,A/A	96%	STOP lost	BAZ2A	ENST00000551996	Essential component of nuclear remodeling complex; involved in transcriptional

							silencing
chr14:50798969	G/C	G/G,G/C,G/G,G/G	98%	STOP gained	CDKL1	ENST00000534267	-
chr14:74763086	C/G	G/G,G/G,G/G,C/C	94%	STOP lost	ABCD4	ENST00000554453 (NMD)	ATP binding
chr16:48134784	G/A	G/G,G/G,G/G,G/G	93%	STOP gained	ABCC12	ENST00000497206 (NMD)	Multidrug-resistance associated protein

Table S32. STOP gains and loses particular to the modern human lineage (where at least 1 archaic has at least 1 ancestral allele). Colored in blue are changes in non-NMD, CCDS-verified transcripts. NMD = transcript subject to nonsense mediated decay.

Position	Ancestral / derived alleles	Archaic genotypes (Vindija, Sidron, Altai,Denisova)	1000G ancestral frequency	Consequence	Gene	Transcripts affected	Gene function / description (UniProtKB, EntrezGene)
chr1:23240250	A/T	T/T,T/T,T/T,A/A	99%	STOP gained	EPHB2	ENST00000400191	Receptor tyrosine kinase which binds ephrin-B family ligands
chr1:31414769	A/G	G/G,G/G,G/G,A/A	99%	STOP lost	PUM1	ENST00000530669 (NMD)	Regulation of translation
chr1:152127741	G/A	A/A,A/A,A/A,G/G	fixed	STOP gained	RPTN	ENST00000316073 ENST00000541545	Epidermal matrix protein
chr1:159785413	C/T	T/T,T/T,T/T,C/C	92%	STOP gained	FCRL6	ENST00000368106	-
chr2:27353219	T/C	C/C,C/C,C/C,T/T	fixed	STOP lost	ABHD1	ENST00000448950 (NMD)	-
chr3:9867484	C/T	T/T,T/T,T/T,C/C	99%	STOP gained	TTL3	ENST00000430390 (NMD)	-
chr3:69053587	T/A	A/A,A/A,A/A,T/T	fixed*	STOP gained	C3orf64	ENST00000295571 ENST00000383701 ENST00000403140 ENST00000540764 ENST00000479765	-
chr3:167284676	G/A	A/A,A/A,A/A,G/G	96%	STOP gained	WDR49	ENST00000479765	-
chr5:156768132	C/T	T/T,T/T,T/T,C/C	fixed*	STOP gained	CYFIP2	ENST00000442283	T-cell adhesion; induction of apoptosis
chr9:35819909	C/A	A/A,A/A,A/A,C/C	fixed	STOP lost	C9orf128	ENST00000388950 (NMD)	-
chr11:63991494	G/A	A/A,A/A,A/A,G/G	fixed	STOP gained	TRPT1	ENST00000544286	Catalyzes tRNA splicing
chr11:65651874	G/A	A/A,A/A,A/A,G/G	fixed*	STOP gained	FIBP	ENST00000533045	May be involved in fibroblast mitogenic function
chr12:10978253	G/A	A/A,A/A,A/A,G/G	fixed	STOP gained	TAS2R10	ENST00000240619	Taste receptor for bitter compounds
chr12:56236602	G/A	A/A,A/A,A/A,G/G	fixed	STOP gained	MMP19	ENST00000322569 ENST00000409200	Degradation of the extracellular matrix during development

						ENST00000548629 ENST00000548882 (NMD) ENST00000552763 (NMD) ENST00000552872 (NMD)	and haemostasis
chr12:123213874	C/T	T/T,T/T,T/T,C/C	fixed	STOP gained	HCAR1	ENST00000356987 ENST00000432564 ENST00000436083	Mediates L-lactate's anti-lipolytic effect.
chr15:99696328	G/A	A/A,A/A,A/A,G/G	fixed	STOP gained	TTC23	ENST00000394129	-
chr19:7743869	C/T	T/T,T/T,T/T,C/C	fixed*	STOP gained	C19orf59	ENST00000333598	-

Table S33. STOP gains and loses particular to the Neandertal lineage (where all or at least 5 out of the 6 Neandertal alleles are derived and equal to each other) while present-day humans are fixed or at high frequency for the ancestral allele in 1000G and Denisova is homozygous ancestral. Colored in blue are changes in non-NMD, CCDS-verified transcripts. Changes marked as "fixed*" are fixed in 1000G but have a dbSNP ID. NMD = transcript subject to nonsense mediated decay.

Position	Ancestral / derived alleles	Archaic genotypes (Vindija, Sidron, Altai, Denisova)	1000G ancestral frequency	Consequence	Gene	Transcripts affected	Gene function / description (UniProtKB, EntrezGene)
chr4:77296802	C/T	T/T,T/T,T/T,T/T	96%	STOP gained	CCDC158	ENST00000434846	-
chr6:25969631	C/T	T/T,T/T,T/T,T/T	97%	STOP gained	TRIM38	ENST00000349458 ENST00000357085 ENST00000540262	-
chr9:4626449	C/T	T/T,T/T,T/T,T/T	99%	STOP gained	SPATA6L	ENST00000485615	-
chr9:21481483	G/A	A/A,A/G,A/A,A/A	94%	STOP gained	IFNE	ENST00000448696	-
chr17:37034365	C/T	T/T,T/T,T/T,T/T	fixed*	STOP gained	LASP1	ENST00000433206	Regulation of cortical cytoskeleton
chr19:55898080	G/A	A/A,A/A,A/A,A/A	fixed	STOP gained	RPL28	ENST00000431533	Ribosomal protein
chr20:44511257	G/A	A/A,A/A,A/A,A/A	99%	STOP gained	ZSWIM1	ENST00000372520 ENST00000372523	-
chr22:32643460	C/A	A/A,A/A,A/A,A/A	99%	STOP gained	SLC5A4	ENST00000266086	Sodium-dependent glucose transporter

Table S35. STOP gains and loses particular to the archaic human lineage (where all or at least 7 out of the 8 archaic human alleles are derived and equal to each other) while present-day humans are fixed or at high frequency for the ancestral allele in 1000G. Colored in blue are changes in non-NMD, CCDS-verified transcripts. Changes marked as "fixed*" are fixed in 1000G but have a dbSNP ID. NMD = transcript subject to nonsense mediated decay.

SI Appendix Figures

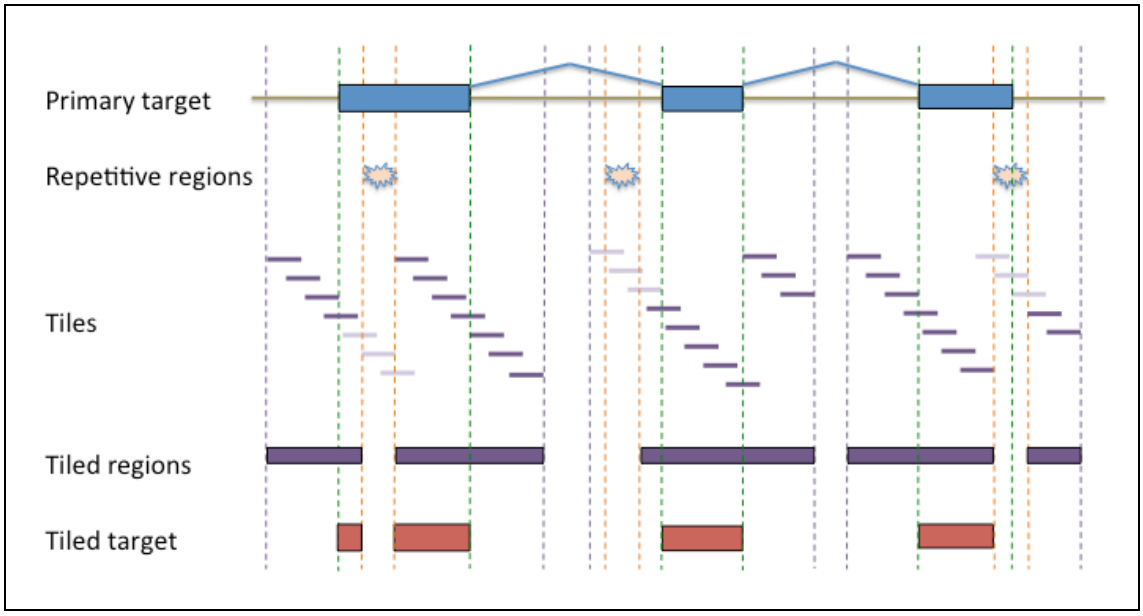


Figure S1. Schema of the probe design process. On top there is the primary target containing the coding exons to be tiled (blue boxes). Overlapping probes of 60 bases were generated in the non-repetitive regions starting 100 bases upstream and ending 100 bases downstream (purple tiles). At the bottom there is the final tiled target regions in the array design (red boxes).

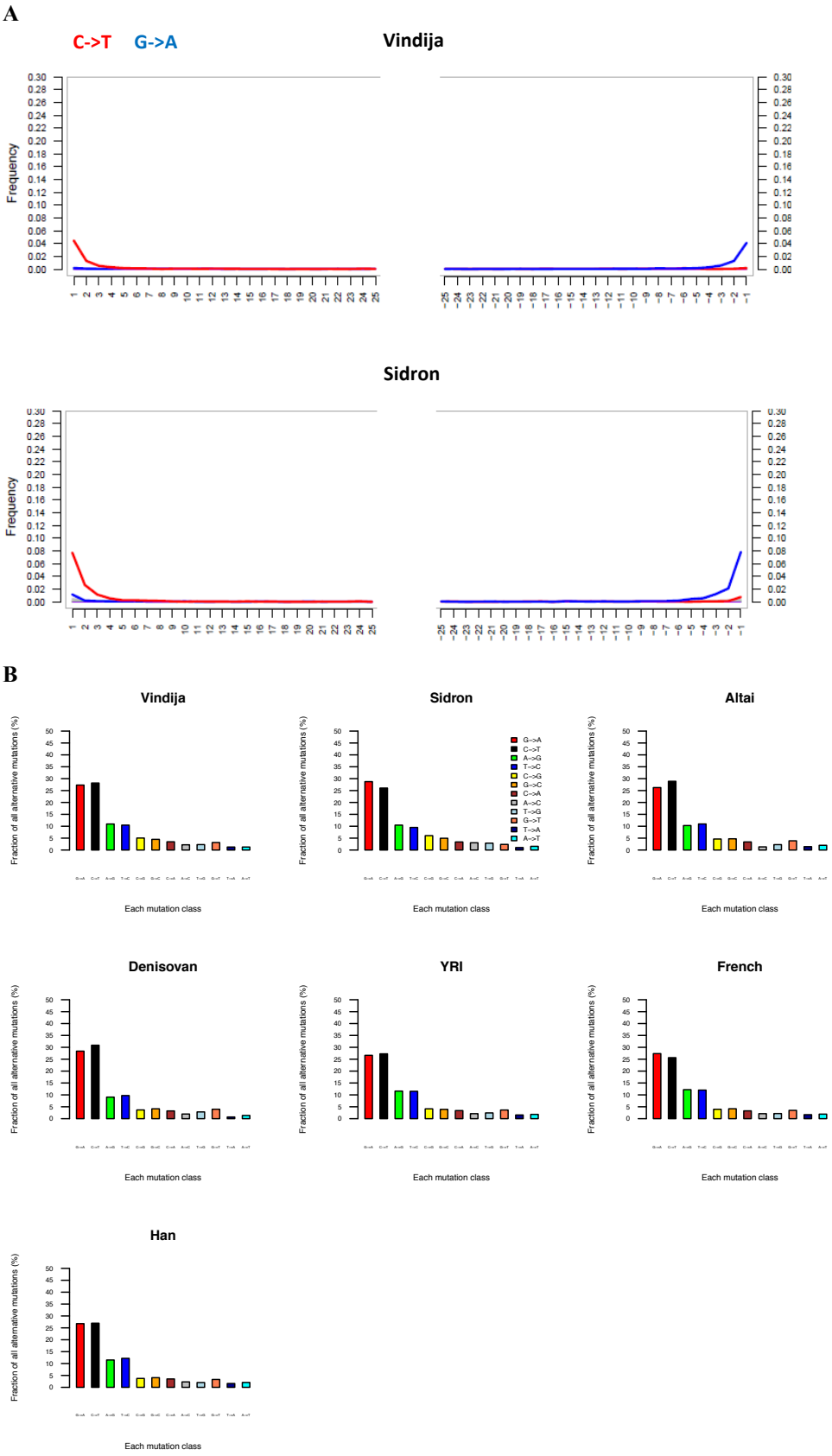


Figure S2. **A.** Residual deaminations, after the UDG treatment (3), in the forward (left) and reverse (right) strand as a function of their distance to the 5' or 3' end of the sequence, respectively. **B.** Proportion of each mutation class (ancestral to derived mutations) in the heterozygous positions of the four archaic (El Sidrón, Vindija, Altai and Denisovan), after computational mitigation of the residual deaminations, and of three present-day individuals.

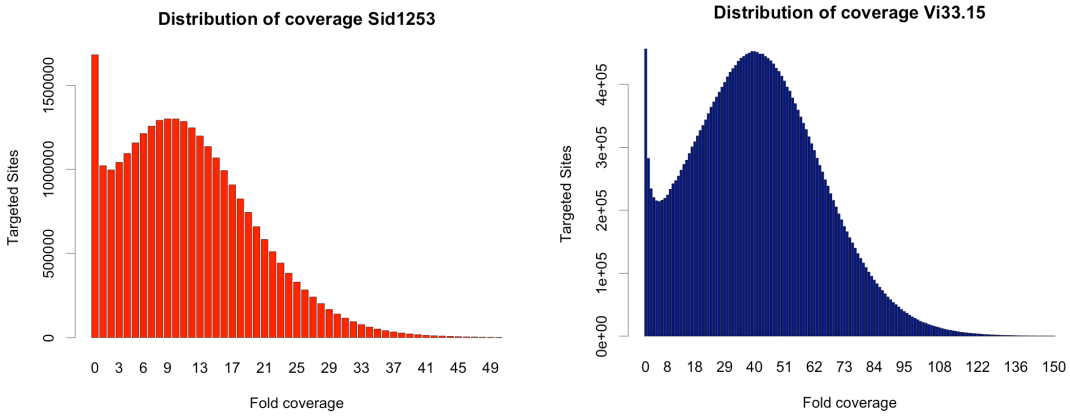


Figure S3. Coverage distribution for El Sidrón (Sid1253) and Vindija (Vi33.15) exome captures. Targeted sites refers to the number of bases at a given fold coverage.

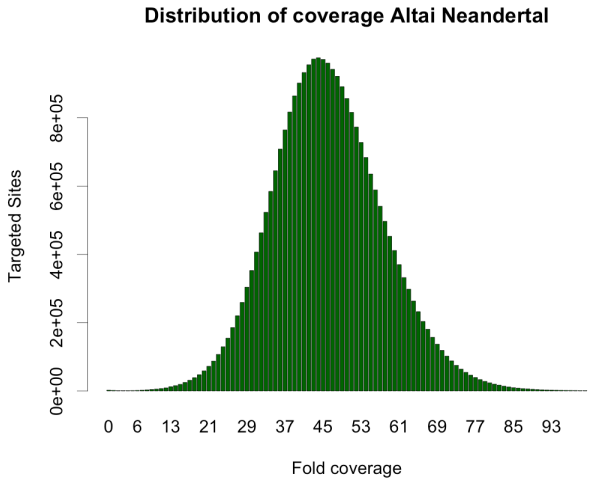


Figure S4. Coverage distribution for the Altai exome. Targeted sites refers to the number of bases at a given fold coverage.

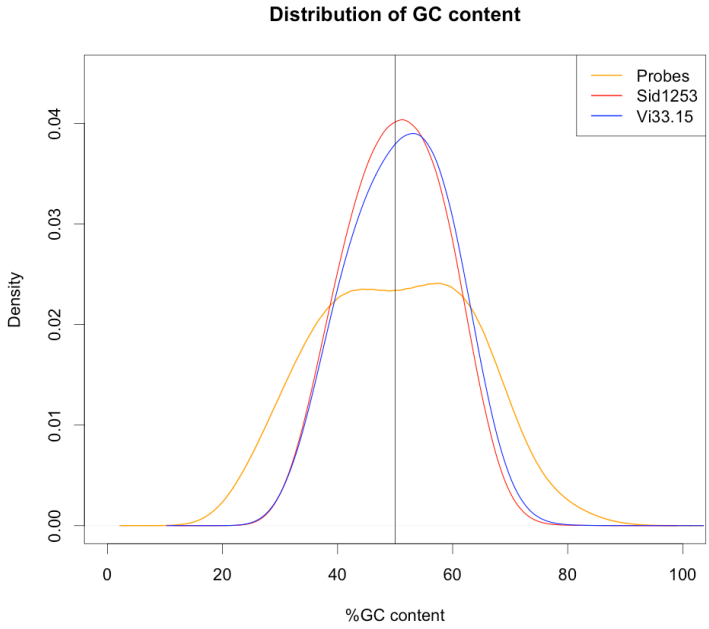


Figure S5. Coverage density of probes in El Sidrón (Sid1253) and Vindija (Vi33.15) reads at different levels of GC content.

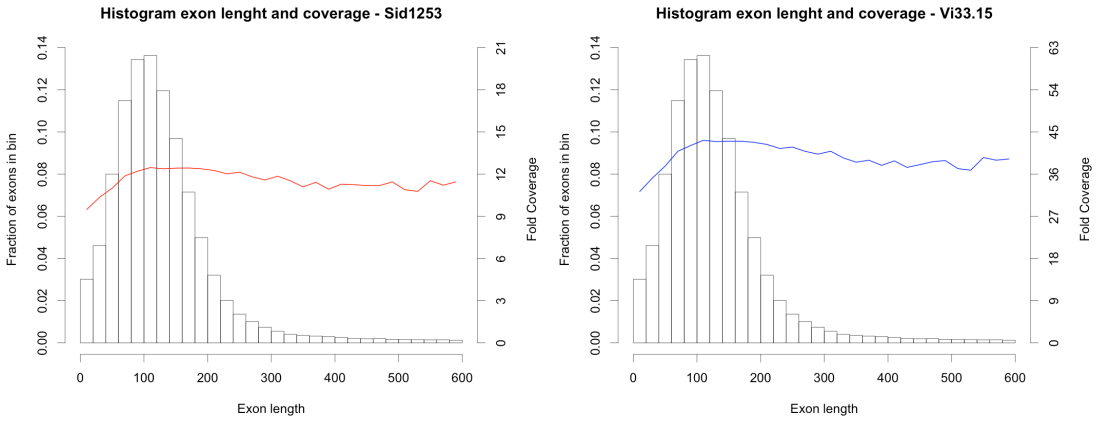


Figure S6. Distribution of coverage for exons of different length in El Sidrón (Sid1253) (red line) and Vindija (Vi33.15) (blue line) exomes. Only exons ranging from 1 to 600 bases long are included (97% of exons in our array design are within this range).

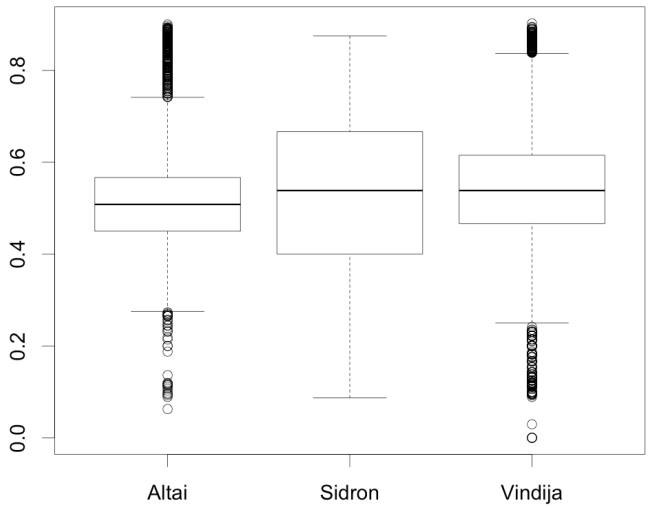
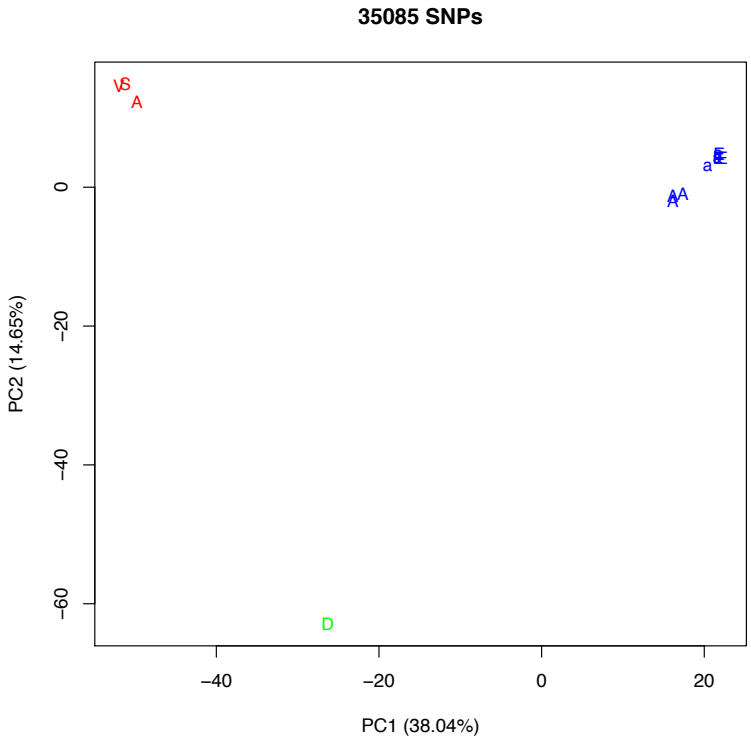


Figure S7. Reference allele frequencies in El Sidrón and Vindija captured exomes and in the Altai exome (protein-coding regions from the Altai genome).

A



B

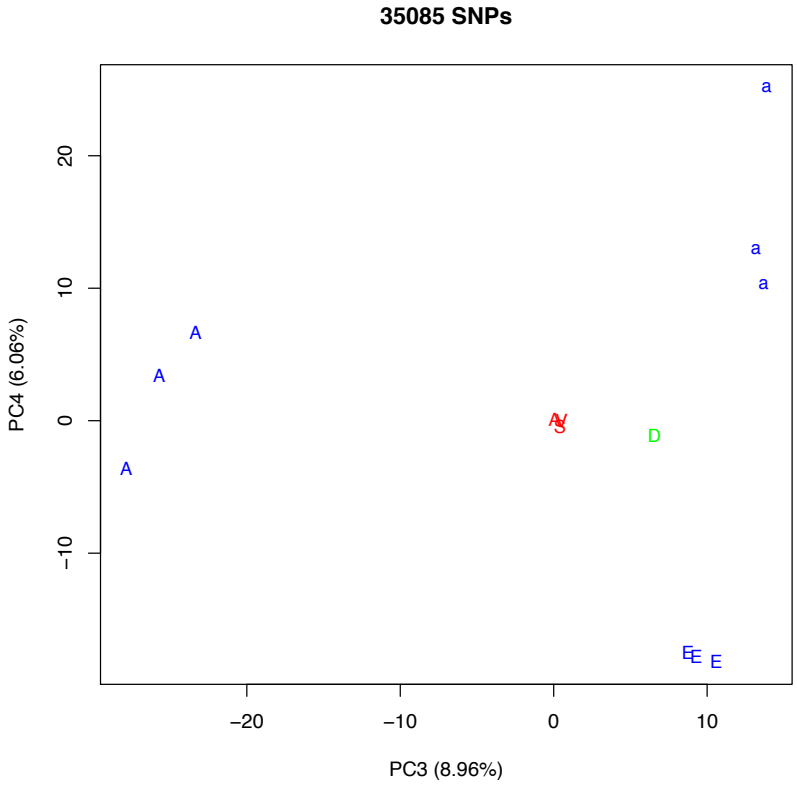


Figure S8. Principal components (PC) analysis of Neandertals, the Denisovan and nine present-day human exomes from 35,085 coding SNPs. The percentage of the variance explained by each PC is given in parenthesis. Legend: El Sidrón (S, in red), Vindija (V, in red), Altai (A, in red), Denisovan (D, in brown), Africans (A, in blue), European (E, in blue), Asian (a, in blue).

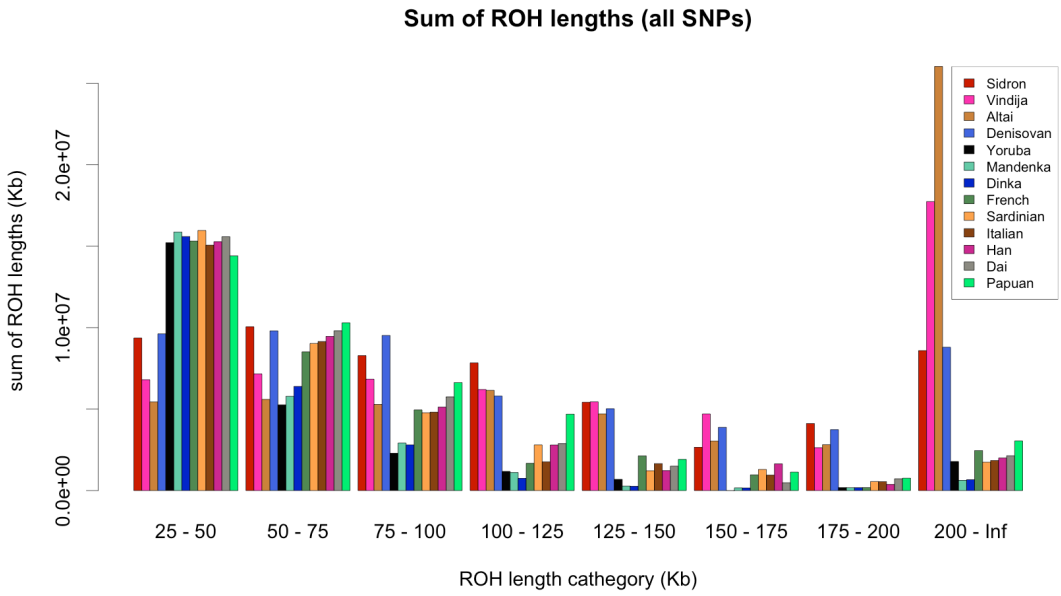


Figure S9. Sum of the length of the runs of homozygosity (ROH) in Neandertals, the Denisovan and nine present-day humans. Similar results are obtained when using the number of SNPs in the individual with the lowest heterozygosity (Altai) in all individuals.

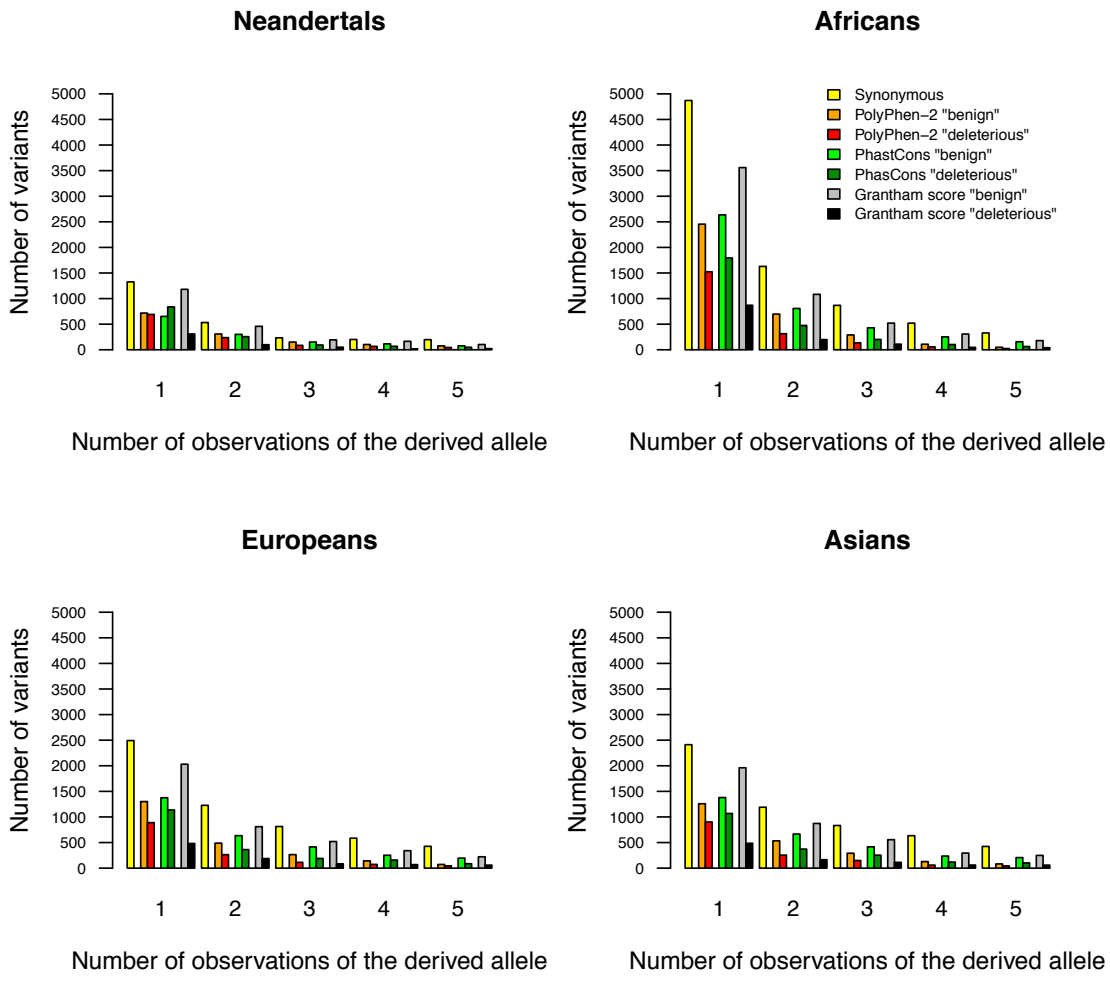


Figure S10. The number of synonymous and non-synonymous (either "benign" or "deleterious") derived alleles in each frequency class. Neandertal positions are homozygous ancestral in the three African individuals. The higher number of synonymous alleles in present-day humans than in Neandertals reflects the higher genetic diversity of humans today. The number of alleles inferred to be "deleterious" out of all non-synonymous alleles by PolyPhen-2 and PhastCons is noticeably larger in Neandertals than in present-day humans at the lowest frequency.

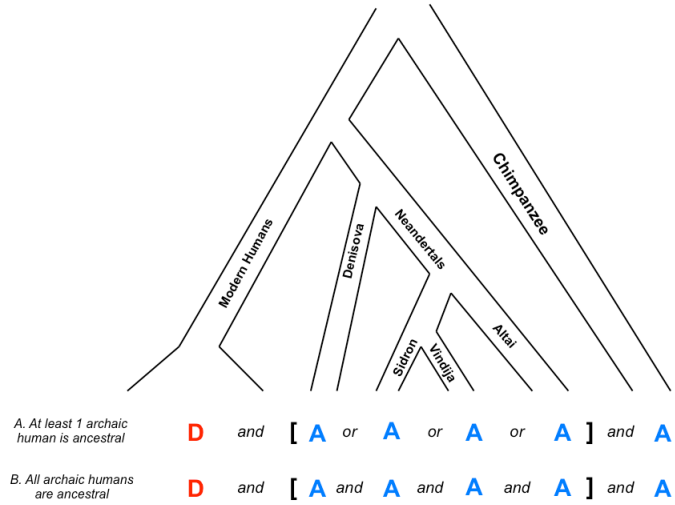


Figure S11. Present-day human-specific sites were selected by either **A**) requiring that at least 1 of the archaic humans have at least 1 ancestral allele or **B**) requiring that all archaic humans be homozygous ancestral. For simplification, this chart only shows 1 allele per lineage and does not reflect the true divergence distances between each group. A=ancestral. D=derived.

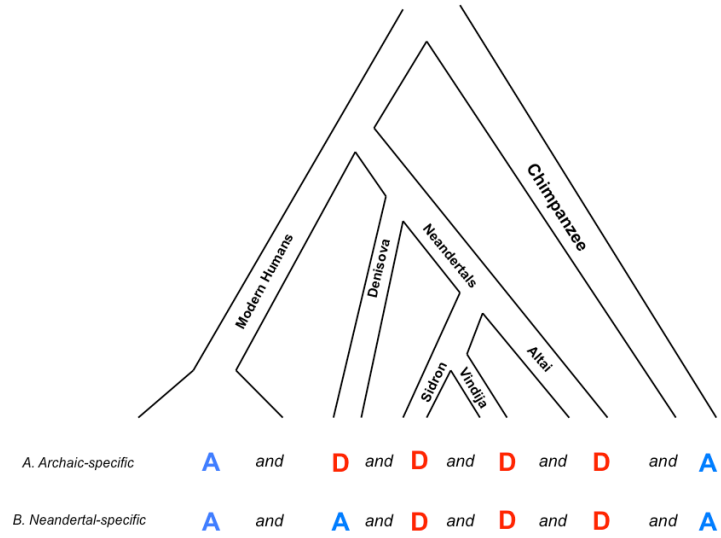


Figure S12. A) Archaic-specific sites were chosen by requiring that present-day humans carry the ancestral allele, while all the archaic human individuals carry the derived allele. **B**) Neandertal-specific sites were chosen by requiring that both Denisova and present-day humans carry the ancestral allele, while all the Neandertals carry the derived allele. For simplification, this chart only shows 1 allele per lineage and does not reflect the true divergence distances between each group. A=ancestral. D=derived.