An ancestral recombination graph of human, Neanderthal, and Denisovan genomes

Authors: Nathan K. Schaefer[1,2,4], Beth Shapiro[1,2,4], and Richard E. Green[3,4]

Affiliations:

[1]Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA.

[2]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA.

[3]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA 95064, USA.

[4]Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA.

*Correspondence to: ed@soe.ucsc.edu

Abstract: Many humans carry genes from Neanderthals, an important legacy of past admixture. Several methods have been described for detecting this archaic hominin ancestry within human genomes using patterns of linkage disequilibrium or direct comparison to Neanderthal genomes. Each of these methods is limited in sensitivity and scalability. We describe a new ancestral recombination graph inference algorithm that is scalable to large genome-wide data sets and demonstrate its accuracy on real and simulated data. We then generate the first genome-wide ancestral recombination graph of both human and archaic hominin genomes. From this, we generate a map within human genomes of archaic ancestry and of genomic regions devoid of genes shared with archaic hominins by either admixture or incomplete lineage sorting. We find that only 1.5-7% of the modern human genome is uniquely human. We also find evidence of at least two bursts of adaptive changes specific to modern humans within the last 600,000 years, consisting of both coding and regulatory changes, many of which may relate to brain development and function.

One Sentence Summary: A new method for mapping archaic hominin ancestry in human genomes reveals specific evolutionary changes unique to modern humans, including many involved in brain development and function.

Main Text:

Much of the current genetic variation within humans predates the split, estimated at 520-630 thousand years ago (kya) (*1*), between the populations that would become modern humans and Neanderthals. The shared genetic variation present in our common ancestral population is still largely present amongst humans today and was present in Neanderthals up until the time of their extinction. This phenomenon, which is known as incomplete lineage sorting (ILS), means that any particular human will share many alleles with a Neanderthal that are not shared with some other humans. Therefore, humans often share genetic variation with Neanderthals not by admixture but rather by shared inheritance from a population ancestral to us both. Because of this, any effort to map ancestry from archaic hominins in human genomes must disentangle admixture from ILS. Furthermore, a technique able to identify both admixture and ILS could

produce a catalog of uniquely human genomic regions that is free of both, and thereby shed light on the evolutionary processes that have been important in our origin as a unique species.

Ancestral recombination graph (ARG) inference (*2*) is a powerful starting point for such an analysis. An ARG can be conceptualized as a series of trees, mapped to individual sites, over phased haplotypes (chromosomes) in a panel of genomes. Ancestral recombination events, or sites at which chromosome segments with different histories were joined together by historical recombination, form boundaries between trees. Each ancestral recombination event manifests as a clade of haplotypes, all of which descend from the first ancestral haplotype to possess it, moving from one position in the tree upstream of the event to a new position in the downstream tree (*3*). ARGs are complete descriptions of phylogenomic data sets and present for recombining genomes what single trees present for nonrecombining genomes, *i.e.*, a complete description of their genetic relationships. As prior techniques for ancestry mapping can be thought of as summaries of the ancestral recombination graph, higher resolution ancestry maps could be produced if the ARG were known. Additionally, the ARG can be used to estimate the TMRCA between admixed and admixer haplotypes, providing additional information about historical admixture between humans and archaic hominins.

Given the utility of an ARG, it is not surprising that several methods have been devised for estimating ARGs from genetic data. These published approaches all have different strengths and limitations. BEAGLE (*3*), ArgWeaver (*4*), and Rent+ (*5*) were designed for small data sets and require substantial time and/or memory to be used with large sequencing panels. Margarita (*6*), randomly samples histories at ancestral recombination event boundaries and does not seek to produce parsimonious recombination histories (*6*). ArgWeaver (*4*), which is widely considered the gold standard in ARG inference, requires prior knowledge of demographic model parameters. Relate (*7*) is a relatively new method that scales well to large data sets and produces fully articulated trees with branch lengths, but in doing so necessarily describes relationships inferred from but not directly observed in the data, as do several other methods (*4*, *5*). The most computationally efficient approach, tsinfer (*8*), also scales to large data sets but does not infer branch lengths and assumes that frequency of an allele is correlated with its age. Since this assumption is violated at loci undergoing either admixture or selection, tsinfer is not well-suited for ARG inference using genetic data from Neanderthals, Denisovans, and modern humans.

Here, we present a heuristic, parsimonious ARG inference algorithm called SARGE (Speedy Ancestral Recombination Graph Estimator) and use it to build a genome-wide ARG of both modern human and archaic hominin genomes. SARGE can run on thousands of phased genomes, makes no prior assumptions other than parsimony, estimates branch lengths, and represents uncertainty due to missing mutations as polytomies in output trees. We validate SARGE using simulated data and demonstrate that it has high specificity compared to existing methods in reconstructing the topology of trees, but does so without assumptions about demographic history, making it more suitable for identifying archaic admixture segments. To achieve this high specificity, SARGE leaves uncertainty (polytomies) in output trees, resulting in lower sensitivity than existing methods.

We run SARGE on a panel of 279 modern human genomes, two high-coverage Neanderthal genomes, and one high-coverage Denisovan genome. Using the resulting ARG, we map Neanderthal and Denisovan ancestry, ILS, and the absence of both across modern human genomes. We find evidence of at least one wave of Neanderthal admixture into the ancestors of all non-Africans. We also identify several long and deeply divergent Neanderthal haplotype

blocks that are specific to some human populations. We find support for the hypothesis that Denisovan-like ancestry is the result of multiple introgression events from different source populations (*9*, *10*). We also detect an excess of Neanderthal and Denisovan haplotype blocks unique to South Asian genomes. Finally, we pinpoint human-specific changes likely to have been affected by selection since the split with archaic hominins, many of which are involved in brain development.

ARG Algorithm

To build an ARG containing both modern human and archaic hominin genomes without the use of a demographic model or the need to infer ancestral haplotypes, we developed a parsimony-based ARG inference technique, SARGE. SARGE uses both shared derived alleles and inferred, shared ancestral recombination events to articulate trees (Supplementary methods, *Fig. 1A*). SARGE uses the four-gamete test (*11*) to determine regions of recombination and the affected haplotypes. The crux of SARGE is a fast algorithm for choosing the branch movement(s) capable of explaining the highest number of discordant clades across a genomic segment that fails the four-gamete test. Once the branch movements, i.e.,inferred ancestral recombinations, are determined, further definition of clades is possible. Thus, the trees are articulated by both shared alleles and shared ancestral recombination events (*Fig. S1*, *Fig. S2*, Supplementary methods).

In the interest of parsimony, our method attempts to infer a set of ancestral recombination events that each explain as many four gamete test failures as possible. Because the four gamete test is known to underestimate the true number of ancestral recombination events (*12*, *13*), SARGE will systematically underestimate the true number of ancestral recombination events in a data set by design. Because of this, SARGE is not well-suited to certain tasks, such as the creation of fine-grained recombination maps. We have attempted to mitigate cases where a clade in the ARG should be broken by an unobserved ancestral recombination event, however, by introducing a propagation distance parameter that limits the genomic distance over which each observed clade is allowed to persist (Supplementary Methods, *Fig. 1A*).

SARGE is scalable to large data sets and achieves higher specificity than many other methods at the cost of lower sensitivity, by leaving uncertainty (polytomies) in the output data. Using simulated data, we find that SARGE runs quickly (*Fig. S5D*, *Fig. S 7*), requires little memory, and has 78.93% specificity (95% C.I. 78.09-78.95%) on average across a range of simulated data sets that include between 50-500 haplotypes (Supplementary Methods). SARGE is at least as specific as alternative techniques (*Fig. S6A,C*). Conversely, SARGE's sensitivity (25.36%; 95% C.I. 25.32-25.40%) is lower than that of other methods *(Fig. S6B,D*), as SARGE leaves an increasingly large number of polytomies in output trees as the number of input haplotypes increases (*Fig. 1B,C*).

We also find that the sensitivity of SARGE can be increased by increasing the propagation distance parameter (Supplementary Text, *Fig. S 8*), that missing clades are likely to be small clades that are close to the leaves of trees (Supplementary Text, *Fig. S 9*), and that incorrectly-inferred clades tend to be within a few kb of sites at which those clades exist in truth (Supplementary Text, *Fig. S 10*). We also find, using simulated data, that SARGE's branch lengths do not appear to be systematically biased upward or downward (Supplementary Text, *Fig. S 11*).

We ran SARGE on 279 phased human genomes from the Simons Genome Diversity Project (SGDP) (*14*), together with two high-coverage Neanderthal genomes (*1*, *15*) and one high-coverage Denisovan genome (*16*). In our analyses, we relied on modern human population labels defined by the SGDP for many analyses, but we split sub-Saharan Africans into one population containing only the most deeply-diverged lineages (Biaka, Mbuti, and Khomani-San), which we call "Africa-MBK," and the remaining genomes ("Africa"). Using these data, we find that the completeness of trees in the ARG is positively correlated with the local mutation rate to recombination rate ratio (*Fig. S13A*; Spearman's rho = 0.40; $p < 2.2 \times 10^{-16}$), and that the number of inferred ancestral recombination events per genomic window agrees with a previously published population recombination map (*17*) (*Fig. S13B*; Spearman's rho = 0.46; $p < 2.2 \times 10^{-16}$), as expected. Estimates of the mean time to most recent common ancestor (TMRCA) of groups, taken across all trees, were also concordant with prior knowledge (*Fig. 2A*). We note, however, that these TMRCA estimates are different from both pairwise coalescent times and population split times. Because we have included hundreds of modern human genomes, and because incomplete lineage sorting between modern humans and archaic hominins is widespread, the mean TMRCA of all humans in the SGDP panel is close to the mean TMRCA of all hominins in our data set (*Fig. 2A*). Our reported TMRCA values computed are also influenced by the demographic parameters implemented in our models (Supplementary Text).

Using these data, we found SARGE's inferences of ancestral recombination events to be accurate. Because SARGE articulates tree clades using either shared allelic variation or shared inferred ancestral recombination, it is possible to test the concordance of trees made from each source. On average, 13.2% of clades in the ARG are known only from inference of shared ancestral recombination events and not by the presence of a shared, derived allele. We created a similarity score between every pair of phased human genome haplotypes in our data set based on how often the haplotypes share ancestral recombination events. This score recapitulates relationships among humans known from SNP data alone (*Fig. 2B,C*; Pearson's r^2 with scores from SNP data = 0.989; $p < 2.2 \times 10^{-16}$). We note that genomes with the poorest correlation between SNP-based and recombination-based similarity scores to other genomes are those most likely to contain phasing errors (*Table S 1*).

Archaic hominin admixture

We used our ARG to find regions of each phased human genome that derived from admixture with archaic hominins (Supplementary Methods, *Fig. S16*). If humans and the archaic hominins in our panel were in populations that had sorted their lineages, this exercise would be simple with a complete and correct ARG. However, since human genome regions are often within a clade that includes hominin haplotypes due to incomplete lineage sorting, finding admixed segments requires analysis beyond simply finding clades that unite some human and archaic hominin haplotypes.

We started by selecting clades from ARG trees that united some modern humans with archaic hominins to the exclusion of some other modern humans. We then assigned each human genome haplotype in each such clade as putative Neanderthal, Denisovan, or ambiguous ancestry, depending on whether the clade contained Neanderthal, Denisovan, or both types of haplotypes. We then performed several filtering steps to remove such clades likely to result from ILS. First, we removed any clades that included more than 10% of the Africa-MBK haplotypes from the most basal human lineages, which are unlikely to be admixed. We then discarded clades that persisted for a short distance along the chromosome, or in which the TMRCA between

modern humans and archaic hominins was high (Supplementary Methods, *Fig. S16*). Because our method relies on both the haplotype block length and the TMRCA between admixed and introgressor haplotypes to identify admixed segments, we were able to identify some haplotypes that resemble archaic admixture in modern humans but that have relatively high sequence divergence to published archaic genomes (manifesting as high TMRCAs between archaic and modern genomes within these segments).

Using the resulting maps, we calculated genome-wide percent admixture estimates across populations and compared them to estimates based on the population-wide D-statistic (*18*, *19*) using basal Africa-MBK lineages as an outgroup. ARG-based estimates are similar to, but lower than, D-statistic based estimates in all non-African genomes. We detected slightly more admixture in sub-Saharan Africans (excluding Africa-MBK) than using the D-statistic (*Fig. 3A*), even when considering the lower end of 95% confidence weighted block jackknife intervals (*Table S2*). We note that a recent study that used an outgroup-free method to detect Neanderthal ancestry blocks in human genomes also found a higher average amount of Neanderthal ancestry in African genomes than has been previously reported (*20*).

We next looked for population-specific differences in archaic hominin ancestry in modern humans. Lengths of archaic haplotype segments and the TMRCA to admixer across those segments are both affected by the time of admixture and the divergence between the true admixers and available archaic hominin genomes. We therefore computed these two values for each ancestry type and compared them across individuals from different populations to look for evidence of distinctive, population-specific admixture events. This analysis revealed distinctive population-specific patterns for Neanderthal and Denisovan ancestry. Segments of ambiguous ancestry produce a pattern resembling a mixture of Neanderthal and Denisovan ancestry, as expected (*Fig. 3B-D, Fig. S 17, Fig. S 19, Fig. S 28, Fig. S 29*). We caution, however, that our approach can artificially shorten haplotype block lengths (Supplementary Methods, *Fig. S 14*), especially for populations such as Papuans and Australians that were absent from the 1000 Genomes Project panel (*21*) that was used for phasing (*14*). Nonetheless, Neanderthal haplotype block lengths in Oceania are not significantly shorter than in other populations (*Fig. 3B*), and incorrect phasing in archaic genomes does not appear to negatively affect results of admixture scans using simulated data (Supplementary text).

As expected, the ARG classifies a smaller fraction (0.10-0.46%) of sub-Saharan African genomes (excluding Mozabite and Saharawi individuals) as resulting from Neanderthal admixture compared to non-African genomes (0.73-1.3%). The haplotype segments of African genomes that are grouped together in clades with Neanderthal haplotypes are distinctive from the haplotype segments found in the genomes of people with non-African ancestry (*Fig. 3B, Fig. S 17A*). Namely, the African haplotypes are more dissimilar to the Neanderthal haplotypes with which they are grouped and tend to be shorter. These observations are qualitatively consistent with the model wherein genetic drift may group Neanderthal and African haplotypes, independent of a specific admixture event. It is also possible that these haplotypes are the result of true introgression events from unknown archaic hominins distantly related to the Neanderthal/Denisovan lineage (*22*). Another recent study using an inferred ARG also found mysterious, divergent haplotypes within sub-Saharan Africans that resembled unknown archaic introgression (*7*).

Unexpectedly, however, two of the SGDP African populations, Masai and Somali, are intermediate between non-African and African genomes when measuring lengths of archaic

haplotype segments and TMRCA to admixers within them (*Fig. 3B*). These Neanderthal haplotype blocks may have originated in ancient European migrants to eastern Africa (*23*) and spread beyond eastern Africa through gene flow, which is known to have affected even the basal Africa-MBK lineages (*24*).

To test this hypothesis, we re-computed the mean length and TMRCA of admixer genomes within archaic-introgressed haplotype segments across all individuals, using only geographically restricted segments. We defined these as any archaic haplotype segments found only in genomes that were sampled within a 3,000 km radius of each other (using geodesic distance between sampling coordinates). This analysis showed Masai and Somali genomes to possess fewer geographically restricted Neanderthal haplotypes than most other African genomes (*Fig. 3C*), concordant with the idea that they originated in Eurasian migrants.

Outside of Africa, our Neanderthal introgression maps largely agree with prior studies. We detect a mean TMRCA to Neanderthal of about 54 kya across all Neanderthal haplotype blocks in non-African populations, using published corrections for branch shortening in the archaic genomes (*1*). Remarkably, the mean TMRCA between genomic segments detected as Neanderthal admixture segments and the Neanderthal itself is consistent within several thousand years for all populations outside of Africa (*Fig. 3B*). We see slightly more Neanderthal ancestry in Central Asia, East Asia, and the Americas than in Europe, South Asia, and Southwest Asia (*Fig. 3A*). We also find more geographically restricted Neanderthal haplotype blocks in South Asia than elsewhere in mainland Eurasia, and the fewest geographically restricted Neanderthal haplotype blocks in the Americas (*Fig. 3C, Fig. S 26*).

Humans in Central and East Asia are known to have elevated Neanderthal ancestry compared to other populations (*25*). However, there is debate over whether this elevated Neanderthal ancestry is due to smaller past population size relative to other groups and the resulting stronger effect of genetic drift (*25*) or to additional pulses of Neanderthal admixture specific to these populations (*9*, *26*). Although we detect more Neanderthal ancestry in Central and East Asians than in West Eurasians, we detect a similar number of geographically restricted haplotype blocks (unique to a 3,000 km radius) in both groups (*Fig. 3C*). Further, Neanderthal haplotype blocks are shorter on average and therefore older in Central and East Asians than in West Eurasians (*Fig. 3B*). This implies that the excess Neanderthal ancestry in Central and East Asians mostly comprises broadly shared haplotype blocks from introgression common to all non-Africans, consistent with the drift scenario. Aside from these broadly shared haplotype blocks, we also observe geographically restricted Neanderthal haplotype blocks in each non-African population in our panel. These population-specific haplotype blocks tend to be longer than the shared haplotype blocks and to have an older TMRCA to the Neanderthal genome than the broadly shared haplotype blocks (*Fig. 3D*). These observations suggest that the population-specific haplotype blocks may be the result of more recent population-specific Neanderthal admixture, as has recently been suggested (*26, 27*, *28*).

We next investigated population-specific patterns within Denisovan ancestry segments and found that such segments probably originate from admixture with multiple, divergent individuals that were distantly related to the Denisovan genome. This implies that the Denisovan genome is not a good model for the actual population that admixed with humans with "Denisovan" ancestry. Prior studies have suggested that Denisovan-like haplotype blocks in humans have two or three distinct sources with different levels of divergence to the Denisovan genome, with the best-matching haplotype blocks in East Asia (*9*, *10*). We uncover the same

signal: geographically restricted Denisovan haplotype blocks have the lowest TMRCA to the Denisovan genome in East Asian genomes (mean TMRCA to Denisovan of 90.3 kya) (*Fig. S 17*, *Fig. S 18*).

Unexpectedly, we detected many Neanderthal and Denisovan-like haplotype blocks that are unique to South Asia (*Fig. 3C*, *Fig. S 17C*, *Fig. S36, Fig. S 26, Fig. S 27*), and many Neanderthal haplotype blocks that are unique to Oceania (*Fig. 3C*, *Fig. S35, Fig. S 26*). These geographically restricted Neanderthal haplotype blocks are no more divergent to the Neanderthal genome than those specific to other populations (*Fig. 3D*), complicating any interpretation of these regions.

Genomic regions free of admixture and ILS

Our ARG strategy allows us to bin the human genome into regions containing archaic admixture in at least some humans, regions of ILS, and regions free of both archaic admixture and ILS in all humans (hereafter archaic "deserts"). We find that approximately 7% of the human autosomal genome is human-unique and free of both admixture and ILS. Roughly 50% of the human genome contains regions where one or more humans has archaic ancestry obtained through admixture. If deserts are further restricted to regions that contain a high-frequency, human-specific derived allele, i.e., a substitution that can be assigned to the human lineage (hereafter "human-specific regions"), these comprise only 1.5% of the assayed genome (*Fig. 4A*). Despite comprising very little of the genome, however, human-specific regions are significantly enriched for genes, exons, and regulatory element binding sites, while deserts are enriched for both genes and regulatory element binding sites (*Table S3*). In line with previous studies (*29*, *30*), we find admixed regions to be depleted of genes. Regions of ILS are enriched for overlap with genes but significantly depleted of exons (*Table S3*).

To obtain an expectation of the extent of these different types of genomic regions, we ran a series of coalescent simulations with different amounts of archaic hominin admixture occurring in two pulses, as well as with no admixture (Supplementary Methods, Supplementary Text). Our observation in the real data – that only 7% of the autosomal genome is free of both archaic admixture and incomplete lineage sorting – is inconsistent with the results of these simulations, which suggest instead that this proportion should be larger (Supplementary Text, *Fig. 4A, Fig. S 38*). Two, non-mutually exclusive explanation for this difference are (1) the existence of more, geographically limited, archaic hominin admixture events than the two we modeled (Supplementary Text), and (2) selection acting on archaic admixed segments.

The power to detect deserts, i.e. regions in which no human carries a haplotype shared with an archaic hominin by ILS or admixture, can be expected to be affected by the number of human genomes available for analysis. To be certain we have found the true extent of archaic deserts, we inferred ARGs over random subsamples of the human panel, computing the extent of deserts and human-specific regions for each (*Fig. 4B*). We were able to recover the full extent of deserts using a subsample of 100 haplotypes, less than half the size of the full panel, suggesting that the panel is sufficiently large.

Timing of human-specific mutations

Given a clade of interest, mutations shared by all members of the clade must have arisen between that clade's TMRCA and its parent clade's TMRCA. Using this logic, and calibrating dates by using the chimpanzee genome as an outgroup and assuming 6.5 Mya human-chimp

divergence (*31*), we estimated ages of all human-specific mutations within deserts. Because the order of mutations along any given branch is unknowable, we took the midpoint of each branch, in years, to be the approximate age of each mutation. Combining these dates with a catalog of high-frequency, human-specific mutations as well as other annotation data (Supplementary Methods) allowed us to construct a picture of human-specific evolutionary changes through time.

We first examined whether there were one or multiple bursts of human-specific adaptive changes since divergence with Neanderthals and Denisovans. We compiled the ages of all fixed or nearly-fixed human-specific derived mutations within archaic hominin deserts that either were annotated as nonsynonymous substitutions (*32*) or fell within annotated regulatory element binding sites. The age distribution of these mutations is unimodal, with a peak around 300 kya (*Fig. 5A*).

We then compared the ages of mutations affecting pairs of genes that interact, according to the STRING database (*33*), to see if any clustered around specific time points (Supplementary Methods). We find two distinct bursts of such mutations, one concentrated around 300-350 kya and another around 100-150 kya (*Fig. 5B*). We note that, because many of our human-specific genes are likely functionally important and purifying selection can decrease genetic diversity, some of the time estimates for these mutations may be biased downward.

Estimating how and when the modern human lineage arose remains controversial. Dating the oldest population split within modern humans using genetic data has suggested times as recent as 200-100 kya (*34*, *35*). Archaeological evidence paints a more complex and older story, however: a recent study reported human remains with many modern features but archaic cranial morphology dated to about 315 kya (*36*), suggesting that not all human-specific traits arose at the same time. Other studies have found that accumulation of derived morphological features in humans occurred in approximately three periods, whose boundaries correspond roughly to the timing of mutational bursts we found, along with the 600-700 kya human/archaic hominin TMRCA (*37*).

Functional consequences of human-specific mutations

Comparison of the human and extinct hominin genomes could reveal instances of positive selection that are undetectable via allele frequency or haplotype-based analyses within modern humans or through comparative genomics between humans and other primate genomes (*38*). The ARG framework is an attractive approach as it pinpoints truly human-specific genomic regions unaffected by either admixture of ILS.

We performed a Gene Ontology (GO) (*39*) enrichment analysis on the human-specific haplotype regions that accounted for the lengths of candidate genes (Supplementary Methods) and found these regions to be heavily enriched for genes related to neuron growth, synapse assembly and function, and cell adhesion (*Table S4*). We note that this may occur because of positive selection in the ancestors of all humans, strong purifying selection that reduces the TMRCA within humans so that it excludes extinct hominins, or a combination of both. Further, this test does not suggest specific functional consequences of specific mutations.

We ordered human-specific derived mutations within desert regions found in our ARG by the strength of evidence that they were targeted by selection. To this end, we first sought to limit analysis to potentially functionally relevant mutations, defined as mutations that either caused a nonsynonymous substitution relative to archaic hominin genomes or fell within annotated

binding sites for regulatory elements known to affect specific genes (Supplementary Methods). We developed a simple score for each mutation based on its inferred age, where available, and the length of the surrounding desert region (*Fig. 5C*). The rationale for this approach is that older human-specific substitutions should be in shorter haplotypes as they would have undergone more generations of recombination. Mutations in haplotype regions that run counter to this expectation are *a priori* more likely to have been affected by positive selection. This approach is similar to the recently-described extended lineage sorting (ELS) scan (*40*), which prioritized long genomic intervals where modern human and archaic hominin lineages are completely sorted. Unlike the ELS method, however, our method only considers alleles that are fixed in modern humans, Our model also does not use a hidden Markov model to smooth transitions between sorted and unsorted haplotypes.

Several patterns emerge when considering genes with high-scoring human-specific mutations, and we highlight some of these key findings. Gene Ontology terms related to mRNA splicing, processing, and export are enriched in genes with high-scoring mutations (*Table S 6*). Of these, we find a regulatory mutation affecting one – LUC7L3 – that is somewhat tissue specific (tau, a measure of tissue specificity scaled from 0 to 1 = 0.713), most highly expressed in cerebellar tissue, and annotated to be involved in splice site selection. Its paralog LUC7L and the gene KHDC4, both of which also possess high-scoring (top 50th percentile) regulatory mutations, are involved in the same process. Additionally, the gene NOVA1, which harbors a nonsynonymous mutation in the top 95th percentile of our score distribution (*Fig. 5C*), is a neuronal splicing factor that regulates splicing of genes involved in synapse formation within the brain (*41*).

Other types of genes, largely related to brain function and development, appear to be affected by high-scoring mutations. Many genes localized to the centrosome and mitotic spindle are involved in maintaining the polarity of dividing neuroblasts, and some mutations affecting such genes are thought to be critical for the development of the human neocortex (*42*). We find the term “asymmetric neuroblast division” to be enriched in high-scoring genes (*Table S 6*). Among individual genes, we find a high-scoring nonsynonymous mutation affecting the centrosomal protein RABL6, which is highly expressed in cerebellar tissue and overexpressed in cancer (*43*) and a high-scoring regulatory mutation affecting INCENP, a protein crucial for cytokinesis that localizes to the mitotic spindle and centromere (*44*). Axon pathfinding is another process suggested to have been targeted by human-specific changes; the gene PIEZO1 is involved in this process (*45*) and harbors a high-scoring nonsynonymous mutation. Additionally, the protocadherin PCDHGB7, which contains a nonsynonymous substitution within a long desert region but which we could not date, is a member of a gene family that generates neuronal cell surface identity and is thought to help guide growing neurites (*46*). In addition to these, we find a number of other mutations potentially affecting genes involved in histone acetylation, neural cell migration, and the clearing of toxic substances from the brain (Supplementary Text).

Discussion

We implemented a new ancestral recombination graph inference approach, SARGE, and used it to build the first genome-wide ARG of both human and archaic hominin genomes. Analysis of the topology of these ARG trees confirms prior findings about archaic hominin admixture, but with important new biological insights. For one, we find that a surprisingly low fraction – 1.5-7% – of the human genome is uniquely human, with the remainder comprising lineages shared with archaic hominins from either incomplete lineage sorting or admixture. This

small human-specific fraction of the genome is enriched for genes related to neural development and function. We also find evidence for multiple waves of human-specific mutations that occurred through time, suggesting that the modern human phenotype may have developed in stages.

In addition to Neanderthal admixture into the ancestors of all modern non-African populations, we find evidence for other, population-specific episodes of admixture throughout Eurasia. The TMRCA to these population-specific Neanderthal haplotype blocks is deeper than the TMRCA to the Neanderthal haplotype blocks shared by all non-African populations. This deeper TMRCA suggests that Neanderthals contributing population-specific ancestry were less closely related to published (Altai and Vindija) Neanderthal genomes than were the Neanderthals that contributed the broadly shared Neanderthal haplotype blocks. We also find that Neanderthal ancestry is present to a smaller extent in some African genomes due to back-migration, consistent with other recent reports (*20*).

We note that our estimated TMRCA to Neanderthal within Neanderthal-introgressed segments in all non-African populations is recent – ~54 kya – and implies therefore that little genetic drift separates admixed humans from sequenced Neanderthals in these segments. This recent TMRCA suggests that the majority of Neanderthal ancestry in modern humans originated from Neanderthal gene flow into the ancestors of all non-Africans before populations diversified. It also suggests that at least one of the Neanderthal genomes used here is closely related to the Neanderthal(s) involved in this admixture event. The slightly elevated Neanderthal ancestry that others have described in Central and East Asian populations also appears to have originated in this first pulse, as Central and East Asian Neanderthal haplotypes are mostly shared with other, geographically distant populations. This observation favors the hypothesis that the increased Neanderthal ancestry in these populations relative to others is due to weaker selection against alleles that may be mildly deleterious (*47*), made possible because of smaller historical population sizes in this part of Eurasia, rather than to additional admixture events (*25*).

Finally, the genomes of some Oceanian and other populations harbor genes from a population most closely related to the archaic Denisovan genome. Importantly, the available Denisovan genome is less genetically similar to the admixing genome than the available Neanderthal genomes are to the admixing Neanderthals. While we are hopeful that future work may uncover a DNA-bearing fossil better representing the population involved in the Denisovan admixture, our approach allows identification of admixed regions that can be used to better describe the genome of the archaic hominin group involved in the admixture event. Larger panels of Denisovan admixed genomes may one day provide a nearly complete Denisovan genome scavenged in parts from the genomes of admixed human individuals.

The ARG also allows for prioritizing the selective importance of mutations specific to, and shared by, all modern humans by considering the TMRCAs of those mutations together with the lengths of their surrounding human-unique regions. Many of these selected human-specific mutations appear to affect genes involved in neural development and function, as well as RNA splicing. Using new tools for genome editing and brain organoid models for neural function, these mutations are obvious and important targets for experimental studies to determine what was selected in our human ancestors after divergence from our most closely related, extinct relatives.

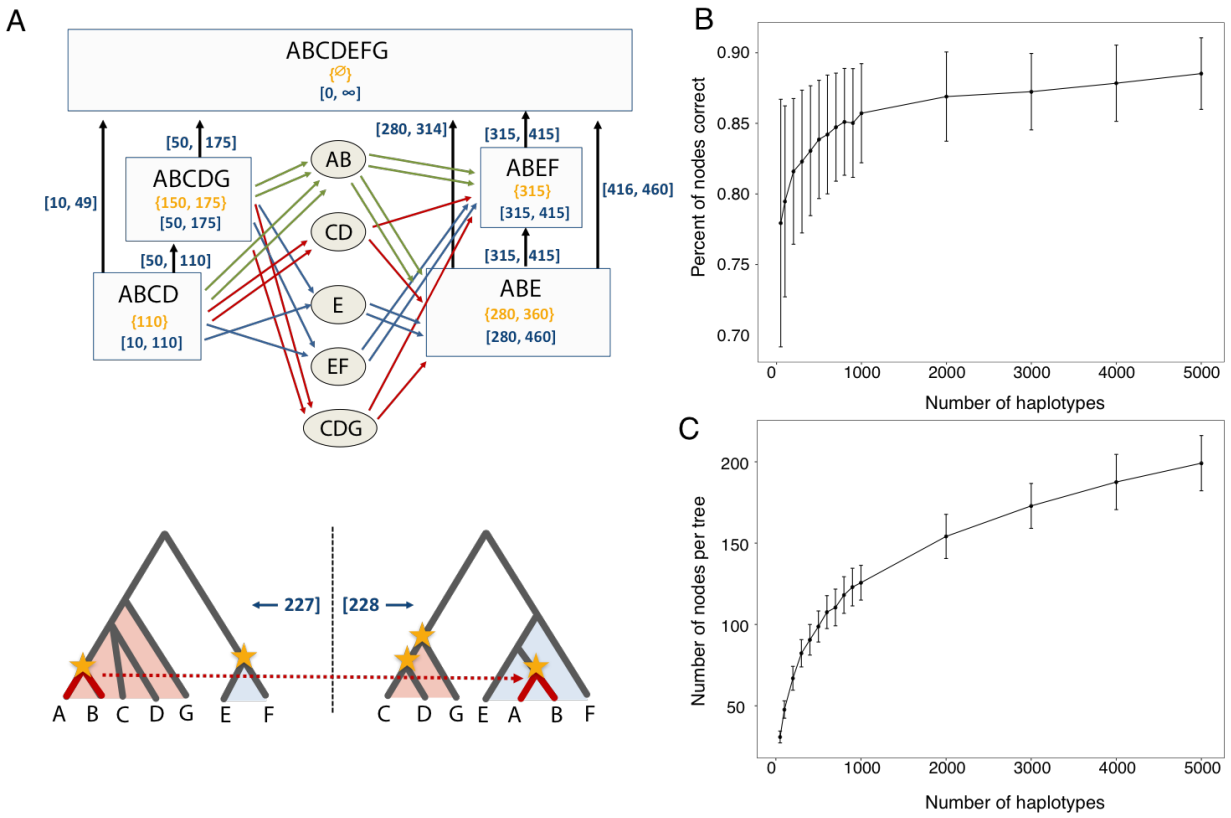

Fig. 1

A: schematic of data structure. Top: rectangles are "tree nodes" representing clades in trees. Each clade has member haplotypes (shown with letters A-G), and a start and end coordinate (blue numbers in brackets) determined by coordinates of SNP sites tagging the clade (yellow numbers in braces), along with a propagation distance parameter (100 in this example). Parent/child edges (vertical arrows) also have start and end coordinates determined by the nodes. Ovals are candidates for clades sharing an ancestral recombination event that can explain four gamete test failures; colored edges indicate potential paths between tree nodes through candidate nodes that could explain four gamete test failures (colors indicate types of paths). The candidate node with the most edges (here, AB) is eventually chosen as the most parsimonious branch movement, allowing for the inference of new nodes. The two trees at the bottom show the "solved" ancestral recombination event with the branch movement marked in red and all clades inferred without SNP data marked with yellow stars (haplotypes A and B share an ancestral recombination event; their ancestry is shared with haplotypes C,D, and G upstream of the recombination event and haplotype E downstream of it). The coordinates of the recombination event (blue numbers in brackets) are taken to be midway between the highest-coordinate upstream site (left side) and the lowest-coordinate downstream site (right side) involved in recombination. B: Accuracy of SARGE on simulated data (defined as percent of all clades correct according to the true ARG in the simulation), with increasing numbers of human-like haplotypes from an unstructured population. Error bars are one standard deviation across 5 replicates. C: Number of nodes per tree with increasing number of haplotypes in simulated data.

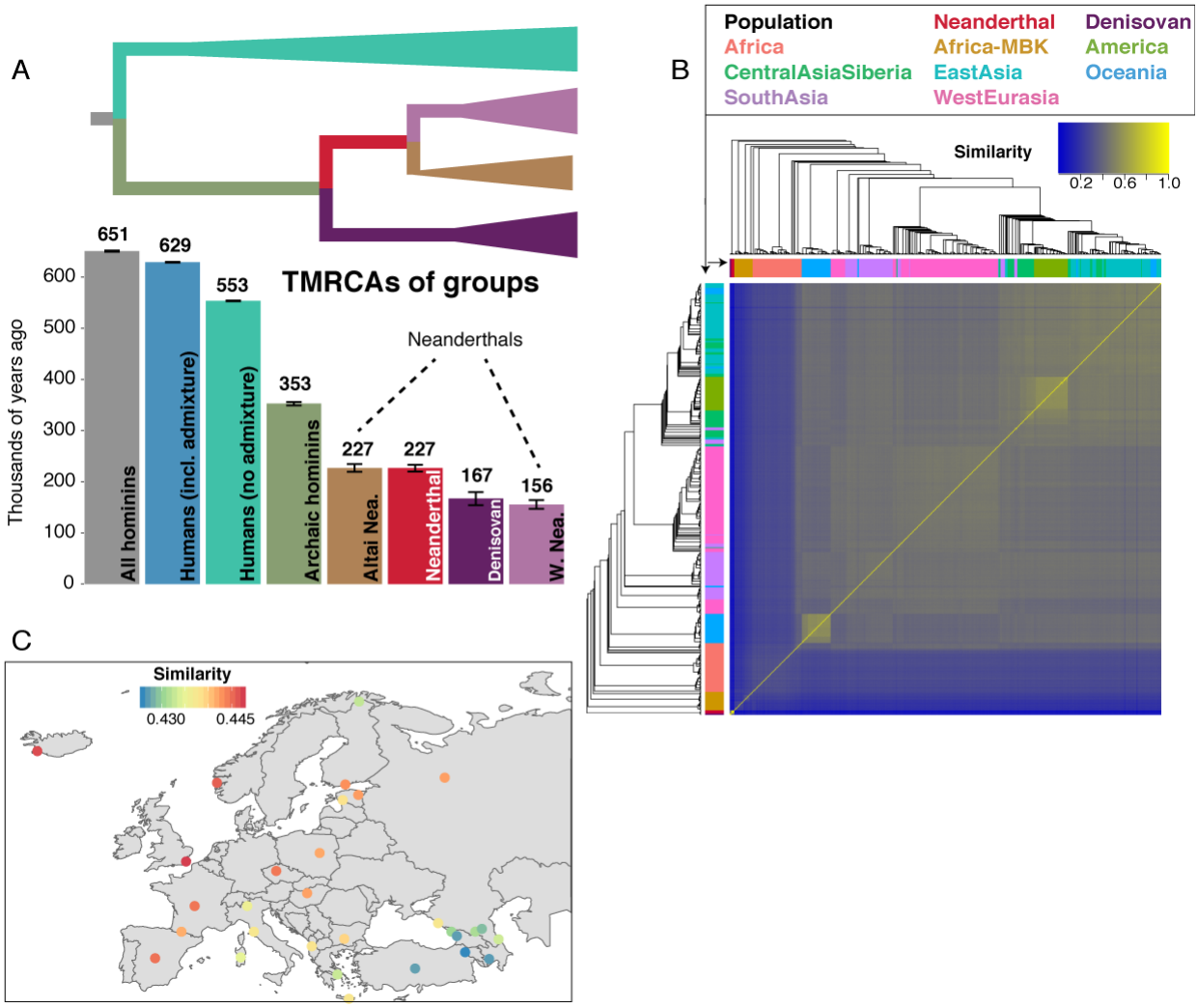

Fig. 2

A: Times to most recent common ancestor for various groups, averaged across all trees in the ARG. Branch shortening values for archaic samples were incorporated into calculations; error bars show the maximum and minimum value given using the maximum and minimum branch shortening values reported in (*1*). The lower value for humans comes from removing archaic-admixed clades from trees. B: UPGMA trees computed using nucleotide diversity from SNP data (top and left) against similarity matrix from shared recombination events inferred by SARGE. Light yellow boxes (similar groups) are Native Americans and Papuans. C: Average similarity between Orcadian haplotypes in the SGDP panel and other European haplotypes, calculated based on the number of shared ancestral recombination events. The best matches are in England, Iceland, and Norway, as expected.

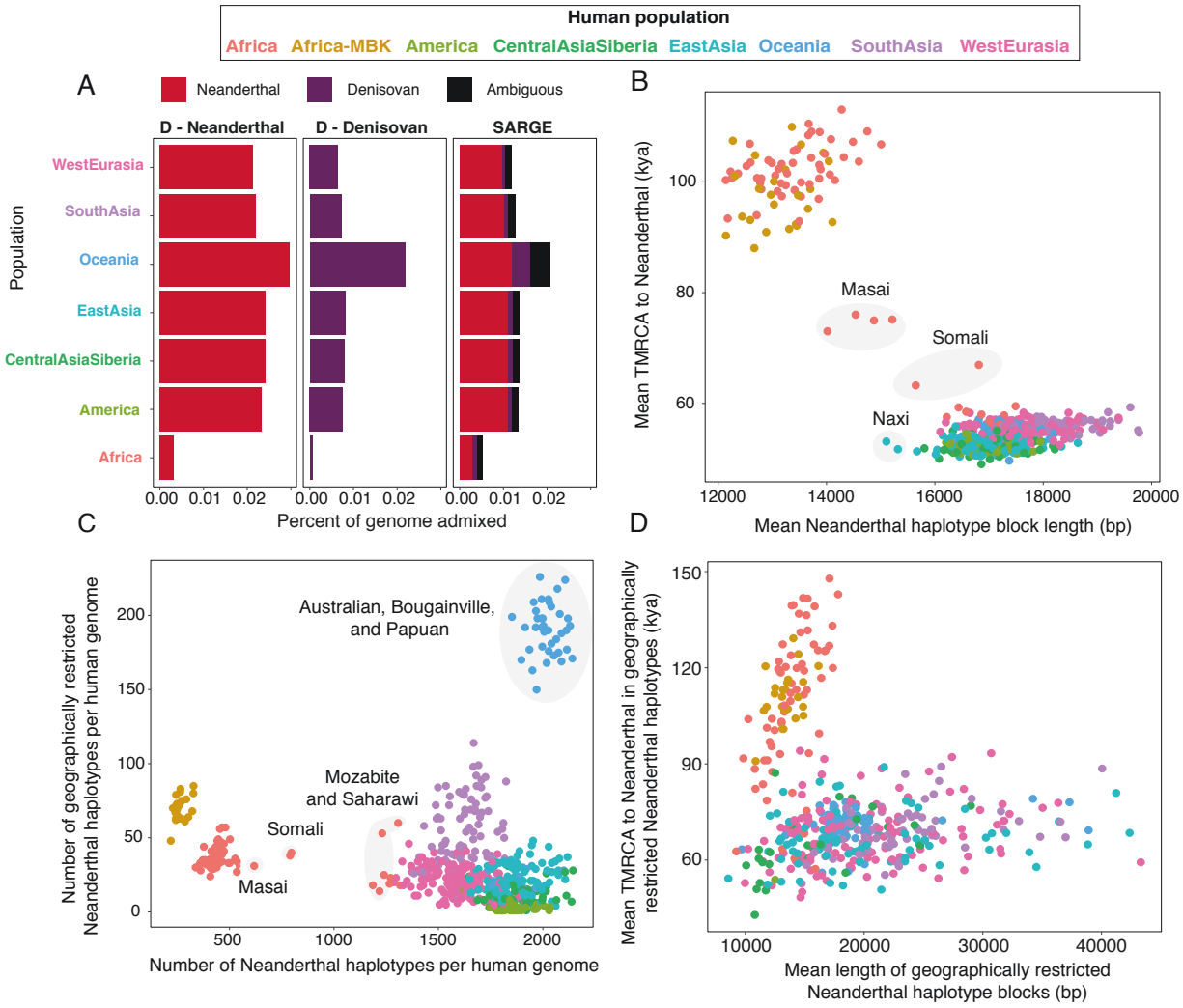

Fig. 3

A: Genome-wide percent Neanderthal, Denisovan, and ambiguous (either Neanderthal or Denisovan) across SGDP populations, using the ARG and an estimator based on the D-statistic. D-statistic calculations considered only one archaic population at a time as introgressor and thus does not detect ambiguous ancestry and also might count some Denisovan ancestry as Neanderthal, and vice-versa. B: For individual phased human genome haplotypes (points), mean time to most recent common ancestor (TMRCA) with Neanderthal in Neanderthal haplotype blocks (y-axis) and mean Neanderthal haplotype block length (x-axis). TMRCA calculations assume a total of 6.5 my human-chimp divergence and branch shortening values from (*1*), with a mutation rate of 1 x 10^{-9} per site per year.. C: Overall number Neanderthal haplotype blocks versus geographically restricted (unique to a 3,000 km radius) Neanderthal haplotype blocks. D: Same as B, but limited to geographically restricted (unique to a 3,000 km radius) Neanderthal haplotype blocks.

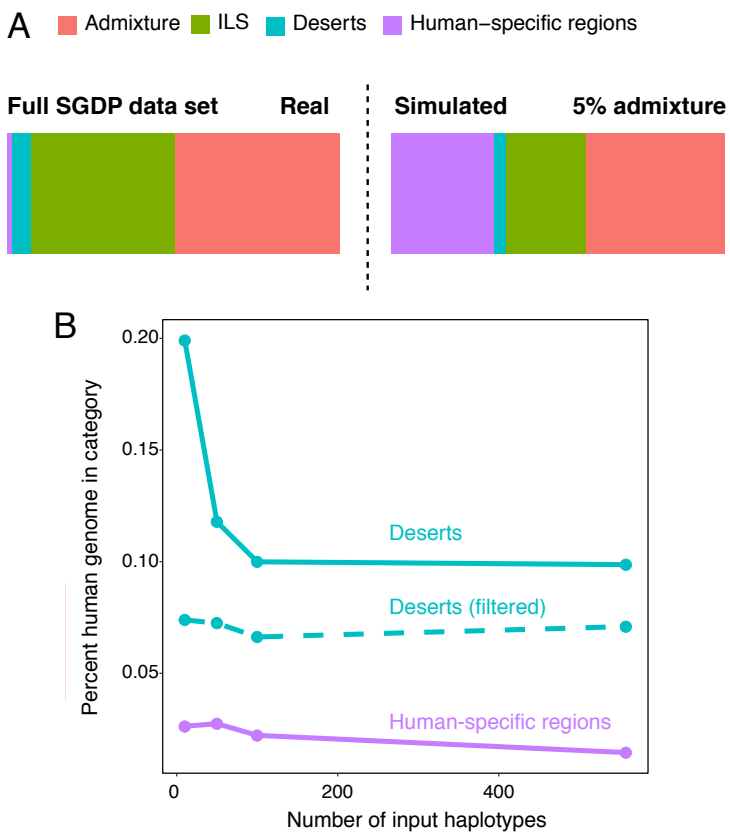

Fig. 4

A: Left panel: in the SGDP data set, fractions of the genome where any individual has archaic ancestry (Admixture), where any individual is incompletely sorted with archaic hominin lineages but where there is no archaic admixture (ILS), where there is no evidence of either admixture or ILS with archaic hominins (Deserts), and where there is a fixed derived allele private to and shared by all humans (Human-specific regions). Right panel: the same values from a simulated data set with single pulses of Neanderthal and Denisovan admixture, both with an admixture proportion (5%) that produced reasonable amounts of archaic ancestry per individual genome (*Fig. S 38C*) (right). B: For random subsamples of the SGDP data set, along with the full data set, fractions of the genome comprising deserts, deserts filtered for candidate archaic alleles using another data set, and human-specific regions are shown. The points on the far right (full data set) correspond to the desert and ILS bars in the left pane of A.

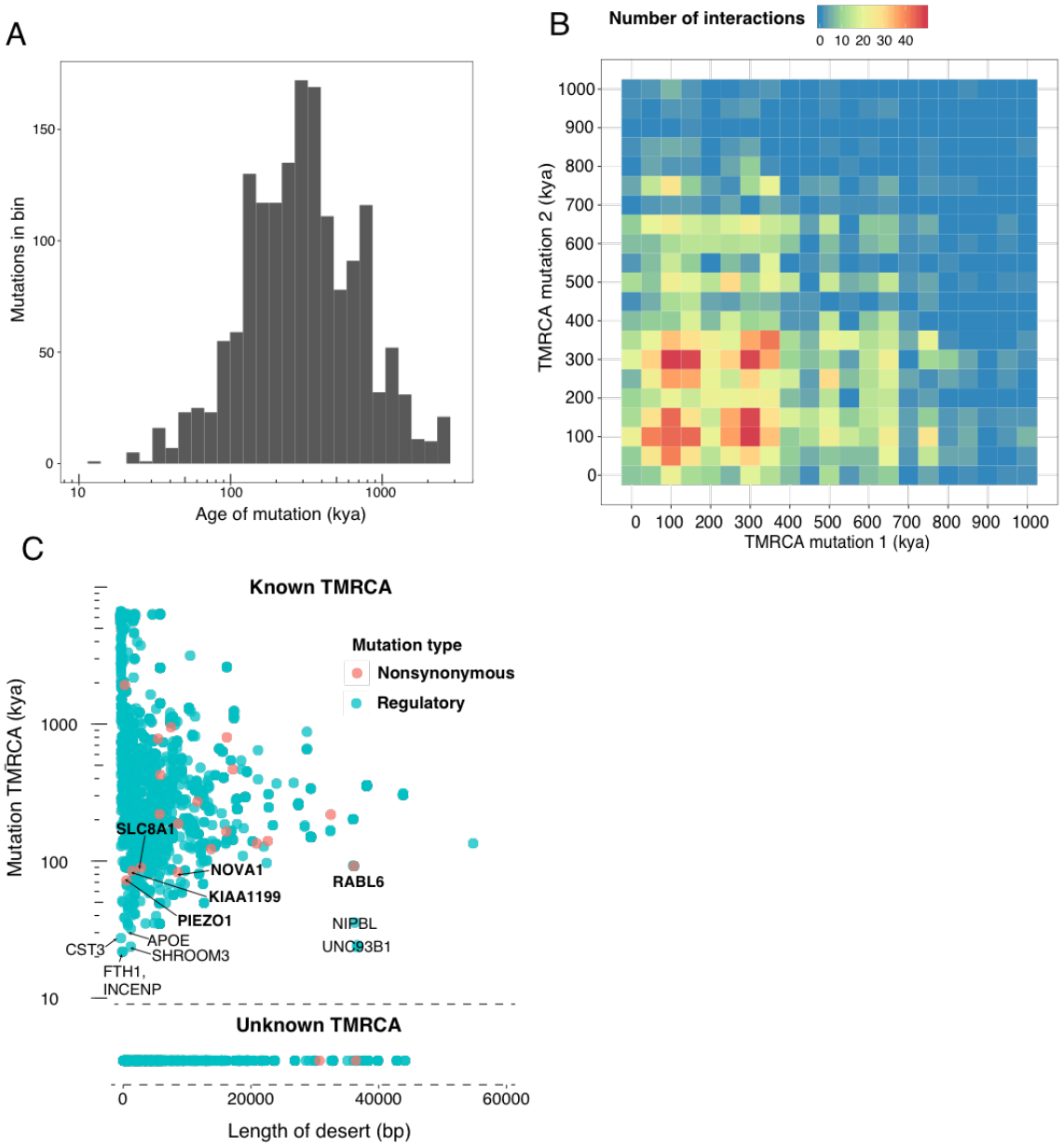

Fig. 5

A: Ages of candidate human-specific functional mutations (nonsynonymous substitutions and mutations within regulatory element binding sites) inferred by SARGE within desert regions (free of ILS and admixture between archaic hominins and modern humans). B: For interacting pairs of genes in the STRING database (*33*) for which candidate human-specific functional mutations affect both genes, the ages of the two mutations are shown. C: For each candidate human-specific functional mutation, the length of the surrounding desert region (x-axis) and inferred mutation age (y-axis) are shown. Mutations for which SARGE did not infer a date (mutations within CpG sites or for which the ancestral allele was unknown) are shown in the bottom panel. Mutations were scored based on length of desert and age; genes with regulatory mutations in the top 99.5th percentile of this distribution, or nonsynonymous mutations in the top 95th percentile, are shown, with nonsynonymous changes in bold.

References

1. K. Prüfer *et al.*, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*. **358**, 655–658 (2017).
2. R. C. Griffiths, P. Marjoram, Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**, 479–502 (1996).
3. Y. S. Song, J. Hein, Constructing minimal ancestral recombination graphs. *J. Comput. Biol.* **12**, 147–169 (2005).
4. M. D. Rasmussen, M. J. Hubisz, I. Gronau, A. Siepel, Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genet.* **10**, e1004342 (2014).
5. S. Mirzaei, Y. Wu, RENT+: An improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics*. **33**, 1021–1030 (2017).
6. M. J. Minichiello, R. Durbin, Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.* **79**, 910–22 (2006).
7. L. Speidel, M. Forest, S. Shi, S. R. Myers, A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
8. J. Kelleher *et al.*, Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330–1338 (2019).
9. S. R. Browning, B. L. Browning, Y. Zhou, S. Tucci, J. M. Akey, Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell*. **173**, 53–61.e9 (2018).
10. G. S. Jacobs *et al.*, Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell*. **177**, 1010–1021.e32 (2019).
11. R. R. Hudson, N. L. Kaplan, Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*. **111**, 147–64 (1985).
12. S. R. Myers, R. C. Griffiths, Bounds on the minimum number of recombination events in a sample history. *Genetics*. **163**, 375–94 (2003).
13. V. Bafna, V. Bansal, The number of recombination events in a sample history: conflict graph and lower bounds. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **1**, 78–90.
14. S. Mallick *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. **538**, 201–206 (2016).
15. K. Prüfer *et al.*, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*. **505**, 43–9 (2014).
16. M. Meyer *et al.*, A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science (80-.).* **338**, 222–226 (2012).
17. International HapMap Consortium *et al.*, A second generation human haplotype map of over 3.1 million SNPs. *Nature*. **449**, 851–61 (2007).
18. R. E. Green *et al.*, A draft sequence of the Neandertal genome. *Science*. **328**, 710–722 (2010).
19. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
20. L. Chen, A. B. Wolf, W. Fu, L. Li, J. M. Akey, Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell*. **180**, 677–687.e16 (2020).
21. 1000 Genomes Project Consortium *et al.*, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).
22. M. F. Hammer, A. E. Woerner, F. L. Mendez, J. C. Watkins, J. D. Wall, Genetic evidence

for archaic admixture in Africa. *Proc. Natl. Acad. Sci.* **108**, 15123–15128 (2011).
23. J. K. Pickrell *et al.*, Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad. Sci.* **111**, 2632–2637 (2014).
24. C. M. Schlebusch *et al.*, Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*. **358**, 652–655 (2017).
25. S. Sankararaman *et al.*, The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*. **507**, 354–7 (2014).
26. B. Vernot, J. M. Akey, Complex history of admixture between modern humans and Neandertals. *Am. J. Hum. Genet.* **96**, 448–53 (2015).
27. F. A. Villanea, J. G. Schraiber, Multiple episodes of interbreeding between Neanderthal and modern humans. *Nat. Ecol. Evol.* **3**, 39–44 (2019).
28. B. Y. Kim, K. E. Lohmueller, Selection and Reduced Population Size Cannot Explain Higher Amounts of Neandertal Ancestry in East Asian than in European Human Populations. *Am. J. Hum. Genet.* **96**, 454–461 (2015).
29. S. Sankararaman, S. Mallick, N. Patterson, D. Reich, The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Curr. Biol.* **26**, 1241–1247 (2016).
30. B. Vernot, J. M. Akey, Resurrecting surviving Neandertal lineages from modern human genomes. *Science*. **343**, 1017–21 (2014).
31. R. E. Green *et al.*, A draft sequence of the Neandertal genome. *Science*. **328**, 710–22 (2010).
32. M. Kuhlwilm, C. Boeckx, A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Sci. Rep.* **9**, 8463 (2019).
33. D. Szklarczyk *et al.*, The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
34. S. Fan *et al.*, African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol.* **20**, 82 (2019).
35. E. K. F. Chan *et al.*, Human origins in a southern African palaeo-wetland and first migrations. *Nature*. **575**, 185–189 (2019).
36. J.-J. Hublin *et al.*, New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*. **546**, 289–292 (2017).
37. G. Bräuer, The origin of modern anatomy: By speciation or intraspecific evolution? *Evol. Anthropol. Issues, News, Rev.* **17**, 22–37 (2008).
38. J. J. Vitti, S. R. Grossman, P. C. Sabeti, Detecting natural selection in genomic data. *Annu. Rev. Genet.* **47**, 97–120 (2013).
39. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
40. S. Peyrégne, M. J. Boyle, M. Dannemann, K. Prüfer, Detecting ancient positive selection in humans using extended lineage sorting. *Genome Res.* **27**, 1563–1572 (2017).
41. J. Ule *et al.*, Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet.* **37**, 844–52 (2005).
42. B.-I. Bae, D. Jayaraman, C. A. Walsh, Genetic changes shaping the human brain. *Dev. Cell*. **32**, 423–34 (2015).
43. J. Montalbano, W. Jin, M. S. Sheikh, Y. Huang, RBEL1 is a novel gene that encodes a nucleocytoplasmic Ras superfamily GTP-binding protein and is overexpressed in breast cancer. *J. Biol. Chem.* **282**, 37640–9 (2007).

44. S. M. Cutts *et al.*, Defective chromosome segregation, microtubule bundling and nuclear bridging in inner centromere protein gene (Incenp)-disrupted mice. *Hum. Mol. Genet.* **8**, 1145–55 (1999).
45. D. E. Koser *et al.*, Mechanosensing is critical for axon growth in the developing brain. *Nat. Neurosci.* **19**, 1592–1598 (2016).
46. J. Brasch *et al.*, Visualization of clustered protocadherin neuronal self-recognition complexes. *Nature*. **569**, 280–283 (2019).
47. K. Harris, R. Nielsen, The Genetic Cost of Neanderthal Introgression. *Genetics*. **203**, 881–91 (2016).
48. O. Delaneau, J. Marchini, 1000 Genomes Project Consortium, 1000 Genomes Project Consortium, Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat. Commun.* **5**, 3934 (2014).
49. K. Prüfer, snpAD: An ancient DNA genotype caller. *Bioinformatics* (2018), doi:10.1093/bioinformatics/bty507.
50. M. Patterson *et al.*, WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol.* **22**, 498–509 (2015).
51. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. **27**, 2987–93 (2011).
52. P.-R. Loh *et al.*, Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
53. Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. **437**, 69–87 (2005).
54. J. Casper *et al.*, The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* **46**, D762–D769 (2018).
55. R. M. Durbin *et al.*, A map of human genome variation from population-scale sequencing. *Nature*. **467**, 1061–1073 (2010).
56. M. Ehrlich, R. Y. Wang, 5-Methylcytosine in eukaryotic DNA. *Science*. **212**, 1350–7 (1981).
57. J. Kelleher, A. M. Etheridge, G. McVean, Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).
58. P. R. Staab, S. Zhu, D. Metzler, G. Lunter, scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*. **31**, 1680–2 (2015).
59. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet.* **5**, e1000695 (2009).
60. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).
61. K. J. Karczewski *et al.*, Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 531210 (2019).
62. A. Favorov *et al.*, Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Comput. Biol.* **8**, e1002529 (2012).
63. A. Frankish *et al.*, GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

64. S. Fishilevich *et al.*, GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford).* **2017** (2017), doi:10.1093/database/bax028.
65. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
66. GTEx Consortium *et al.*, Genetic effects on gene expression across human tissues. *Nature*. **550**, 204–213 (2017).
67. N. Kryuchkova-Mostacci, M. Robinson-Rechavi, A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).
68. K. Prüfer *et al.*, FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics*. **8**, 41 (2007).
69. V. M. Narasimhan *et al.*, Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat. Commun.* **8**, 303 (2017).
70. K. Harris, Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 3439–44 (2015).
71. K. E. Langergraber *et al.*, Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 15716–21 (2012).
72. G. C. Gibb, S. F. K. Hills, Intergenerational mutation rate does not equal long-term evolutionary substitution rate. *Proc. Natl. Acad. Sci.* **110**, E611–E611 (2013).
73. P. Moorjani, C. E. G. Amorim, P. F. Arndt, M. Przeworski, Variation in the molecular clock of primates. *Proc. Natl. Acad. Sci.* **113**, 10607–10612 (2016).
74. N. Patterson, D. J. Richter, S. Gnerre, E. S. Lander, D. Reich, Genetic evidence for complex speciation of humans and chimpanzees. *Nature*. **441**, 1103–1108 (2006).
75. T. Rito *et al.*, A dispersal of Homo sapiens from southern to eastern Africa immediately preceded the out-of-Africa migration. *Sci. Rep.* **9**, 4728 (2019).
76. R. Tobler *et al.*, Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. *Nature*. **544**, 180–184 (2017).
77. S. Ramachandran *et al.*, Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15942–7 (2005).
78. A. Durvasula, S. Sankararaman, Recovering signals of ghost archaic admixture in the genomes of present-day Africans. *bioRxiv*, 285734 (2018).
79. S. L. McIntire, R. J. Reimer, K. Schuske, R. H. Edwards, E. M. Jorgensen, Identification and characterization of the vesicular GABA transporter. *Nature*. **389**, 870–6 (1997).
80. E. Ravindran *et al.*, Homozygous ARHGEF2 mutation causes intellectual disability and midbrain-hindbrain malformation. *PLoS Genet.* **13**, e1006746 (2017).
81. C. Hashizume *et al.*, Nucleoporin Nup62 maintains centrosome homeostasis. *Cell Cycle*. **12**, 3804–16 (2013).
82. K. I. Swenson, K. E. Winkler, A. R. Means, A new identity for MLK3 as an NIMA-related, cell cycle-regulated kinase that is localized near centrosomes and influences microtubule organization. *Mol. Biol. Cell*. **14**, 156–72 (2003).
83. T. K. Sears, J. M. Angelastro, The transcription factor ATF5: role in cellular differentiation, stress responses, and cancer. *Oncotarget*. **8**, 84595–84609 (2017).
84. O. D. Gil, G. Zanazzi, A. F. Struyk, J. L. Salzer, Neurotrimin mediates bifunctional effects on neurite outgrowth via homophilic and heterophilic interactions. *J. Neurosci.* **18**, 9312–

25 (1998).
85. T. Maes, A. Barceló, C. Buesa, Neuron navigator: a human gene family with homology to unc-53, a cell guidance gene from Caenorhabditis elegans. *Genomics*. **80**, 21–30 (2002).
86. D. L. C. van den Berg *et al.*, Nipbl Interacts with Zfp609 and the Integrator Complex to Regulate Cortical Neuron Migration. *Neuron*. **93**, 348–361 (2017).
87. B. B. Muhoberac, R. Vidal, Abnormal iron homeostasis and neurodegeneration. *Front. Aging Neurosci.* **5**, 32 (2013).
88. E. Shokri-Kojori *et al.*, β-Amyloid accumulation in the human brain after one night of sleep deprivation. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4483–4488 (2018).
89. S. A. Kaeser *et al.*, Cystatin C modulates cerebral beta-amyloidosis. *Nat. Genet.* **39**, 1437–9 (2007).
90. D. M. Holtzman, J. Herz, G. Bu, Apolipoprotein E and apolipoprotein E receptors: normal biology and roles in Alzheimer disease. *Cold Spring Harb. Perspect. Med.* **2**, a006312 (2012).
91. H. A. Pearson, C. Peers, Physiological roles for amyloid beta peptides. *J. Physiol.* **575**, 5–10 (2006).
92. H. B. Nygaard, C. H. van Dyck, S. M. Strittmatter, Fyn kinase inhibition as a novel therapy for Alzheimer's disease. *Alzheimers. Res. Ther.* **6**, 8 (2014).
93. I. Letunic, P. Bork, Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242-5 (2016).

Acknowledgments: The authors thanks Sam Vohr and Russ Corbett-Detig for their helpful input. **Funding:** This work was funded by the National Institutes of Health (NIH) T32 grant HG00834, National Science Foundation (NSF) grant DEB-1754451, and by the Howard Hughes Medical Institute. **Author contributions:** All authors conceived of and helped direct the study. NKS performed the analyses. All authors contributed to writing this paper. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** Data used in this study were downloaded from public databases. Code is available at https://github.com/nkschaefer/sarge.

Supplementary Materials:

Materials and Methods

Figures S1-S31

Tables S1-S6

Supplementary Materials for

An ancestral recombination graph of human, Neanderthal, and Denisovan genomes

Nathan K. Schaefer, Beth Shapiro, and Richard E. Green

Correspondence to: ed@soe.ucsc.edu

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S31

Tables S1 to S6

Materials and Methods

Data processing

We downloaded data from the Simons Genome Diversity Project (SGDP) panel (*14*), along with two Neanderthal (*1*, *15*) genomes and one Denisovan (*16*) genome. The Simons data were downloaded in pre-phased form from https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data/; phasing was done using SHAPEIT2 (*48*). We note that the hosts of the data state that the genotypes they provide at sites lacking a homologous chimpanzee are unreliable; we discarded all such sites from analysis.

Existing variant call sets for the ancient samples were either created using a genotype caller that did not account for ancient DNA damage (*15*, *16*) or were subjected to a mapability filter that discarded many sites in the genome (*1*). Because our method is sensitive to genotype errors and seeks to make inferences at every possible site in the genome, we chose to re-call variants in these three genomes using the ancient DNA-aware genotype caller snpAD version 0.3.0 (*49*). For all snpAD runs, we required a minimum base quality of 25 and treated different types of libraries separately, separating UDG-treated and non-UDG treated libraries in the case of the Vindija Neanderthal, and separating single-stranded and double-stranded library data for the Altai Neanderthal and Denisovan.

Although the SGDP data were already phased, phasing posed a challenge for the Neanderthal and Denisovan data, for which there is no reference panel and for which DNA is fragmented into short segments. Fortunately, the comparatively low nucleotide diversity in these archaic hominins results in the presence of long runs of homozygosity, which are phased by definition. As an unbiased first step, we performed read-backed phasing using WhatsHap version 0.16 (*50*) (with default parameters, plus –ignore-read-groups). Before filtering SNPs for quality and coverage, this phased 722,828 of 11,746,838 heterozygous sites (6.2%) in the Altai Neanderthal, 346,992 of 48,083,469 heterozygous sites (0.7%) in the Vindija Neanderthal, and 514,575 of 33,951,346 heterozygous sites (1.5%) in the Denisovan. Many of the remaining, unconfidently phased heterozygous sites were excluded by other, later filtering steps, however: in our final, filtered data set, we were left with only 1,677,774 of 49,876,210 total SNPs (3.4%) for which at least one archaic hominin individual was heterozygous and not phased by read-backed phasing.

Following read-backed phasing, we merged archaic hominin VCF files (using bcftools merge from bcftools version 1.8 (*51*)) and then phased the merged files using Eagle2.4 (*52*), with the 1000 Genomes Project data (*21*) as a reference panel. We used Eagle2's default parameters, but specified that it should not impute missing data (--noImpMissing) and that it should output alleles that it could not phase (--outputUnphased). After this, we randomly assigned both alleles at every unphased heterozygous site to one or the other haplotype. Although this decision, along with the use of a modern human reference panel, undoubtedly introduced haplotype switch errors, we deemed this preferable to excluding sites that were not confidently phased (which would require us to exclude data from all of the Simons Genome Diversity Project individuals at the same sites). To mitigate problems arising from this decision, we avoided performing any haplotype-specific analyses on the archaic hominin genomes. When creating maps of archaic hominin ancestry in modern humans, for example, we track only whether a modern human haplotype is in a clade with one or more archaic hominin haplotypes at each site, but not which specific archaic hominin haplotype is in the clade. After running the ARG, we computed the discordance between similarity scores per genome haplotype computed using SNP data and

computed using shared ancestral recombination events; this discordance should be largely driven by phasing switch errors, which can cause the inference of erroneous ancestral recombination events. We found that this discordance ranged from about 8x higher (in Denisovan) to 2-3x higher (in Neanderthals) in archaic hominins than in the most discordant human genomes (*Table S 1*).

We merged the phased archaic hominin files with the SGDP data, using bcftools merge with the –missing-to-ref option, and then used bcftools norm to remove duplicate alleles (-d). To avoid mis-identifying all SGDP samples as homozygous reference at sites that were originally excluded from the SGDP data set, we limited the variant call set for each chromosome to the sites between the first and last site in the SGDP data on that chromosome. To mitigate the same problem, we also removed any site for which all non-reference alleles in our SGDP data were private to archaic hominins, but for which non-reference alleles were present in modern humans within the 1000 Genomes data set (*21*). We then discarded all sites for which any individual had a missing genotype or genotype quality below 25 or for which any archaic sample fell within the upper or lower tail of its genome-wide coverage distribution (extracted from the VCF file). The allowed coverage ranges (determined by eye) were 23-70X for the Altai Neanderthal, 10-43X for the Denisovan, and 10-47X for the Vindija33.19 Neanderthal.

Finally, we polarized our variant call set into ancestral and derived alleles, using the chimpanzee reference genome panTro4 (*53*) (mapped to hg19 by the UCSC Genome Browser team (*54*) and downloaded in AXT format) as an ancestral sequence, discarding any variant that was an indel, had more than two alleles, or lacked a known chimpanzee homolog. We chose panTro4 as an ancestral sequence rather than a composite ancestral sequence as some other studies have done (e.g. (*55*)) because it allowed us to more easily estimate branch lengths, at the cost of discarding some sites. Additionally, because our approach assumes the infinite sites model of mutation, we excluded all CpG dinucleotide sites from analysis, as methylated cytosines in CpG dinucleotides are highly mutable and are thus more likely than other nucleotides to undergo repeated mutations (*56*).

Ancestral recombination graph inference

We developed an ancestral recombination graph inference program called SARGE (available at https://github.com/nkschaefer/sarge), which is optimized for speed and low memory usage, in addition to making minimal model assumptions. SARGE assumes parsimony and the infinite sites model and uses the four gamete test (*11*) as a central insight. SARGE avoids using statistical techniques to smooth branch lengths or infer clades, opting instead to describe only that which can be inferred directly from the input data. The result is a set of trees that contain polytomies and have relatively low-resolution branch lengths.

Our algorithm centers on the observation that a single tree cannot contain two clades that share members unless one is a superset of the other. We assume that every shared derived allele in our data set defines a clade. It has been shown that, under this assumption, pairs of sites for which the inferred clades share members, but for which neither is a superset of the other, mark ancestral recombination events, or breakpoints between different trees. This is referred to as the "four haplotype test" or "four gamete test" (*3*, *11*). One could use this technique to map ancestral recombination events, which mark boundaries between trees, articulate trees using the sites within these boundaries. In practice, however, this can only produce minimally articulated trees. In the case of organisms with low nucleotide diversity, this is because there will not often be

enough polymorphic sites between ancestral recombination breakpoints to observe many of the possible clades per tree. In the case of organisms with high nucleotide diversity, however, it will be possible to detect far more ancestral recombination events, thus making the size of "bins" between ancestral recombination breakpoints smaller and leading to the same problem.

Our algorithm therefore seeks to infer all relevant information about each ancestral recombination event. An ancestral recombination event can be conceptualized as a branch movement (*3*), and so each consists of a set of haplotypes moving from one clade in an upstream tree into a new clade in a downstream tree. Given two clades that share members, but for which neither is a superset of the other (henceforth described as a failure of the four haplotype test), and assuming that this four haplotype test failure describes only one ancestral recombination event, there are then three possible branch movements than can explain it (Fig. S1). We refer to the clade in the upstream tree from which a subclade moved as α, the clade in the downstream tree into which a subclade moved as β, and the subclade that moved positions as γ. Four haplotype test failures are possible between the following sets of clades (with the clade in the upstream tree listed first and the clade in the downstream tree listed second): α/ α, α/ β, and β/ β. In the case of an upward branch movement, all four haplotype test failures are α/ α, and all four haplotype test failures are of the type β/ β in the case of downward branch movements. The members of the moving clade γ can then be inferred once the type of four haplotype test failure is known. Denoting the members of the upstream clade as U and the members of the downstream clade as D, γ contains U ∖ D in the α/ α case, U ∩ D in the α/ β case, or D ∖ U in the β/ β case.

Inferring branch movements between two trees

With this insight, we developed a simple algorithm to infer the most parsimonious ancestral recombination event (branch movement) between two trees, if the trees are known *a priori* and fully articulated. Consider a clade to be a set of haplotypes, and take U to be the set of nodes in the first (upstream) tree and D to be the set of nodes in the second (downstream tree). Then,

```
DECLARE set A1 ← ∅
DECLARE set A2 ← ∅
DECLARE set B1 ← ∅
DECLARE set B2 ← ∅
DECLARE set Γ ← ∅
DECLARE bool finished ← FALSE
DECLARE set F ← ∅
DECLARE set G ← ∅
DECLARE map C ← {}
FOR u in U:
      FOR d in D:
            IF | u ∩ d | > 0 and u ⊈ d and u ⊉ d:
                  F ← F ∪ (u, d)
                  FOR g in [(u ∩ d), (u \ d), (d \ u)]:
                        IF g in G:
                              C[g]++
                        ELSE:
                              G ← G ∪ g
                              C[g] ← 1
DECLARE F'← F
```

```
WHILE not finished:
      DECLARE γ ← argmax(C)

      Γ ← Γ∩γ
      G ← ∅
C ← {}
      FOR (u, d) in F':
            IF γ in [(u ∩ d), (u \ d), (d \ u)]:
                  F' ← F' \ (u, d)
            ELSE:
                  (u,d) ← (u \γ, d \γ)
                  FOR g in [(u ∩ d), (d \ u), (u \ d)]:
                        IF g in G:
                              C[g]++
                        ELSE:
                              G ← G ∪ g
                              C[g] ← 1
      IF | F' | == 0:
            finished ← TRUE
FOR γ in Γ:
      FOR (u,d) in F:
            IF γ ⊂ u:
                  A1 ← A1 ∪ u
            ELSE:
                  B1 ← B1 ∪ u
            IF γ ⊂ d:
                  B2 ← B2 ∪ d
            ELSE:
                  A2 ← A2 ∪ d
```

After this, the sets A1 and A2 contain clades that lost members by recombination, before and after the recombination event, respectively. Likewise, B1 and B2 contain clades that gained members by recombination, before and after the recombination event, respectively. Γ contains all clades defined by shared ancestral recombination events in the genomic interval between the two trees. This is also illustrated in Fig. S2.

If the B1 and B2 sets are empty, then the recombination event was an upward movement (the clade that moved, defined by recombination, left a clade to join that clade's parent). If the A1 and A2 sets are empty, then the recombination event was a downward movement (the clade that moved, defined by recombination, left a clade to join that clade's child). Otherwise, the recombination event was a lateral movement (the clade that was left and the clade that was joined do not share members).

At the step where a "moving clade" (γ) representing a single ancestral recombination event is chosen, it is possible for ties (candidate moving clades that explain an equal number of four haplotype test failures) to exist. This means that there were multiple, equivalent ways to describe the ancestral recombination event.

General case algorithm

Extrapolating this approach to ARG inference poses several problems. First, it cannot be known *a priori* which clades belong together in trees. Grouping clades together into upstream and downstream sets is therefore a difficult problem that we solve by exploring many possible groupings and bound using heuristic assumptions (Supplemental text). Second, many of the clades that could inform ancestral recombination events will be unobserved, if they are not tagged by mutations at sites in the data set.

Knowing this, we infer ancestral recombination events using the available mutations and then use these inferred ancestral recombination events to infer clades that they imply (Fig. S5B). Namely, we assume that γ clades should exist as clades in the ARG, whether or not they are tagged by mutations, because the haplotypes in γ share at least one ancestral recombination event. All subclades within the upstream α clade, with the γ clade haplotypes removed, must also exist as clades in the downstream tree. Likewise, all subclades within the downstream β clade, with the addition of γ haplotypes, must also exist in the upstream tree. Similarly, all subclades within the downstream α clade must exist in the upstream tree, with γ clade haplotypes added, and all subclades within the upstream β clade must exist in the downstream tree, with γ clade haplotypes added (Fig. 1A, bottom panel). Finally, in the case of an upward or downward branch movement (inferred by the absence of any β clades or α clades in the four haplotype test failures, respectively), the union of all clades failing the four haplotype test should exist as a clade in the ARG

The other key component of our algorithm is a "propagation distance" parameter, p. This parameter describes how far upstream and downstream (in physical distance) each site's clade is allowed to communicate its existence. Because the all-versus-all clade comparisons required by our algorithm would become very computationally expensive without knowing *a priori* which clades belong to adjacent trees, this parameter helps bound the number of comparisons and thus the execution time. It also allows us to avoid storing an entire ARG over a chromosome in memory at once. As we read new sites into memory, we can identify nodes sufficiently far away upstream to be unaffected by the new data. We can then "solve" ancestral recombination events for those upstream nodes, and other nodes even further upstream, whose ranges leave them unaffected by the newly-solved recombination events, can be written to disk and erased. Because errors and violations of the infinite sites model (such as back-mutations) invariably exist, this parameter has the extra benefit of limiting how far along a chromosome erroneous data can propagate (although a cascade of incorrect clades inferred by recombination could hypothetically propagate errors outside of the range of the original, erroneous node).

We create a graph containing two types of nodes: "tree nodes," which are part of the ARG, and "recombination nodes," which represent candidate γ clades for unsolved ancestral recombination events. Each tree node represents a given clade over a contiguous genomic span and has a start and end coordinate, a set of positions of SNPs that tag it, and a set of other sites at which it was inferred to exist as part of a recombination event. Tree nodes have parent/child edges, also with start and end coordinates, and there is a single root node that spans the entire chromosome. Node range coordinates are initially set to the furthest upstream (lowest coordinate) site owned by the node minus the propagation distance, up to the furthest downstream (highest coordinate) site owned by the node, plus the propagation distance. When a node encounters another node with which it fails the four haplotype test, however, its coordinates are adjusted – either its end coordinate is set to the furthest-downstream site at which it is known to exist, or its start coordinate is set to the furthest-upstream site at which it is known to exist. Nodes also can have recombination edges, which point to nodes with which they fail the four

haplotype test, with paths through recombination nodes (Fig. 1A). These edges are analogous to the edges described in the two-trees algorithm (Fig. S2). When a recombination event is solved, all nodes implied by the recombination event are created as tree nodes in the ARG (Fig. 1A), with "solved" recombination edges describing the inferred recombination event, to avoid creating redundant recombination events in the future. Furthermore, when no possible γ node explaining a four haplotype test failure can exist (i.e. all three possible clades fail the four haplotype test with existing ARG nodes within the ranges over which they must exist), we add "unsolvable" recombination edges connecting the two nodes that fail the four haplotype test. These edges allow us to adjust start and end coordinates of the nodes without inferring the branch movement that separates them.

The propagation distance parameter p allows us to bin the ARG into regions 2*p bases wide, each of which undergoes a different process simultaneously. Because 2*p is the maximum number of bases within two sites can affect each other in the ARG, any node tied to a site more than 2*p bases upstream of the most recently-observed site is already informed by all available input data. This means we can solve ancestral recombination events affecting nodes more than 2*p upstream of the most recently-observed sites. Likewise, nodes far enough upstream to be unaffected by these ancestral recombination events being solved can be written to disk, and nodes far enough upstream as to not affect those being written to disk can be erased from memory. For a cartoon of the different ARG operations allowed in different genomic bins, see Fig. S 3.

We determine branch lengths when writing trees. Since each tree is defined only at a single site, we determine a node's branch length by counting the number of mutations it owns within the range defined by the edge to its parent at the current site. If this parent/child edge expands beyond the range $[s - p, s + p]$, where s is the current site and p is the propagation distance, we limit to mutations found only within that range. In the case of the root node, because this node consists of all haplotypes in the data set and cannot be affected by ancestral recombination events, the branch length will always equal the total number of fixed differences between haplotypes in the data set and the outgroup, divided by two times the propagation distance. We then divide the number of mutations by the number of bases in the range over which they were collected. In the case where a parent/child edge is valid only at a single site, this will lead to the extremely large branch length of 1. To help compensate for this, when we load trees from an output file, we scale each branch length by dividing it by the total height of the tree, both above and below that branch length (Fig. S 4). This puts all branch lengths on a scale between 0 and 1. When all fixed differences between the ancestral sequence and the reference genome are included as sites that can contribute to the root branch length in the ARG (as in this study), these branch lengths can then be multiplied by two times the divergence time between the ancestral and reference genomes to get approximate (low resolution) branch lengths. We note that many clades in our ARG have branch lengths of zero, meaning that no mutations were observed on those lineages. We also note that the number of times a given node serves as a γ clade in an ancestral recombination event also provides a measure of age. Although we store these values, we do not use them when computing branch lengths in this study, since it is difficult to reconcile time measured using two different types of units (mutations and shared recombination events). Thus, clades inferred solely from ancestral recombination events will have branch lengths of zero.

QC Simulations

For the sake of assessing our and other ARG inference programs, we simulated sampling an increasing number of haplotypes from a single panmictic population with no history of growth or bottlenecks (QC simulations).

Our QC simulations were done using msprime (*57*). We chose a recombination rate of 1 centimorgan per megabase and a mutation rate of $1*10^{-9}$ per year with a 25-year generation time, giving a per-generation mutation rate of $2.5*10^{-8}$. Additionally, we chose a heterozygosity value of 10.1 per 10,000 bases, comparable to the rate in modern sub-Saharan Africans (*15*). We simulated 1 megabase of sequence per run, running 5 replicates each of simulations with 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, and 5000 haplotypes. The complete command used was `mspms` **`X`** `1 -t 1010.0 -r 404.0 1000000 --precision 6 -T`, where X is the number of haplotypes. Whenever there were duplicate base positions in a simulated data set, we ignored the allele data at all but the first occurrence of each position. We also repeated these simulations with a lower mutation rate of $1*10^{-8}$ per site per generation, to see how this affected SARGE's ability to correctly infer clades. The mean sensitivity and specificity given in the main text are computed on both of these sets of simulations, with numbers of haplotypes ranging from 50 to 500.

We then ran SARGE on each data set with a propagation distance of 25,000 bases, along with tsinfer (*8*) (converting its output to a sequence of trees linked to specific variable sites), Relate (*7*) with the mutation rate set to that known from the simulation and with the haploid effective population size set to two times the simulation's effective population size, and Rent+ (*5*) with the –t option to infer branch lengths. For each inferred tree, we loaded the tree output by msprime for the same variable site and defined the inferred ARG's specificity as the fraction of all clades in all inferred trees that existed as clades in the msprime tree at the same sites. We also computed the sensitivity, defined as the fraction of all clades in true trees that existed as clades in the inferred ARGs. For the two methods, Relate and Rent+, that produced fully articulated trees, these two values were the same. Results of QC simulations are shown in Fig. S5, Fig. S6, and Fig. S *7*.

Demographic simulations

Several of our other analyses and tests required simulated data more similar to our real data (a global panel of human genomes, plus one eastern Neanderthal, one western Neanderthal, and one Denisovan genome). We therefore created a second type of simulated data set using a popular human demographic model.

We used scrm (*58*) for these simulations because it allows users to sample haplotypes from time points in the past, mimicking the branch shortening due to "missing evolution" when analyzing ancient genomes. We combined a popular, three population demographic model for modern humans (*59*) with populations meant to approximate the Altai (*15*) and Vindija (*1*) Neanderthals and the Altai Denisovan (*16*). We again assumed a 1 centimorgan per megabase recombination rate and a $1*10^{-9}$ per year mutation rate, along with a 25-year generation time, giving a per-generation mutation rate of $2.5*10^{-8}$. In addition to the demographic model parameters listed in (*59*), we modeled a Neanderthal/human split time of 575kya (*15*), an Altai/Vindija split time of 137.5kya, a Neanderthal/Denisovan split of 420kya, and modeled the heterozygosity in all Neanderthals as $1.6*10^{-4}$ and in Denisovans as $1.8*10^{-5}$(*1*). We chose 100kya as the divergence time between the Vindija and introgressing Neanderthal, no split between the introgressing and sampled Denisovan populations, and we modeled

human/Neanderthal admixture as a single pulse 50kya, in the population ancestral to both Europeans and Asians, and human/Denisovan admixture as a single pulse 20kya, in the Asian population. We assigned 57.kya of missing evolution to the Vindija haplotypes, 123kya of missing evolution to the Altai haplotypes, and 80kya of missing evolution to the Denisovan haplotypes (*1*). Our simulated chromosome was 25 megabases long, and we sampled 2 haplotypes from each Neanderthal and Denisovan (but not the introgressing Neanderthal), as well as 20 haplotypes from each modern human population (African, Asian, and European) population. For the sake of simplicity ascertaining archaic introgressed haplotypes, we modeled no archaic introgression into Africans and set the gene flow parameters between Africans and other populations to zero.

We then ran five different simulations, varying the human/archaic admixture proportion in each (the admixture proportion was the same for both the Neanderthal and Densisovan admixture events). The admixture proportions we used were 5%, 10%, 25%, 50%, and 75%. The full command for each run was `scrm 66 1 -t 17253.7128713 -r 7342.00547714 25000000 -T -I 7 20 20 20 0 0 0 0 -eI 0.0772268560362 0 0 0 0 0 2 0 -eI 0.167529158597 0 0 0 2 0 0 0 -eI 0.108962054372 0 0 0 0 0 0 2 -n 1 1.68 -n 2 3.74 -n 3 7.29 -n 4 0.231834158238 -n 5 0.231834158238 -n 6 0.231834158238 -n 7 0.0260813428018 -eg 0 2 116.010723 -eg 0 3 160.246047 -m 2 3 2.797460 -m 3 2 2.797460 -ej 0.028985 3 2 -en 0.028985 2 0.287184 -em 0.028985 1 2 7.293140 -em 0.028985 2 1 7.293140 -es 0.027240513593 3` **`[1-x]`** `-ej 0.027240513593 8 7 -es 0.0681012839825 2` **`[1-x]`** `-ej 0.0681012839825 9 5 -ej 0.136202567965 6 5 -ej 0.197963 2 1 -en 0.303501 1 1 -ej 0.187278530952 5 4 -ej 0.572050785453 7 4 -ej 0.783164765799 4 1`, where [1-x] means one minus the admixture proportion in a given run. We again discarded all but the first instance of every unique base position in the output file, and we converted the "true" trees into SARGE format for running analyses.

We used the simulation with what we deemed was the most realistic admixture proportion (0.05) for a variety of other QC assessments of SARGE, including testing the propagation distance parameter, comparing sensitivity and specificity to Relate (*7*) and tsinfer (*8*), checking accuracy of inferred branch lengths, and ascertaining whether incorrectly inferred clades were inferred to exist near sites where they could be correctly inferred to exist. We also ran one more iteration of this simulation, but with 150 haplotypes sampled per human population, for a total of 456 simulated genomes. This simulation was used to test the accuracy of inferred branch lengths, in addition to the simulation with 50 modern human haplotypes. For more information see Supplementary Text.

Comparing tree articulation to mutation/recombination rate ratio

Tree articulation refers to the extent to which all branches in a tree are defined. A fully articulated tree containing n lineages has n-1 total nodes, whereas a completely un-articulated tree would consist of a single large polytomy. Because SARGE leaves polytomies in output trees, we sought to determine whether polytomies (which manifest as a lower amount of articulation) tend to occur in regions with a low mutation to recombination rate ratio.

We binned the genome into 50kb blocks and counted the mean number of nodes per tree, a measure of articulation, across all trees within each window. We then measured the mutation

rate by sampling the branch length of the root node of each ARG tree (this is the number of mutations separating all hominin lineages from the chimpanzee genome, collected over 2*(propagation distance) bases and reported in units of mutations per base). Assuming 6.5 Mya for the hominin-chimpanzee split and a 25-year generation time, we transformed numbers of mutations into a per-site, per-generation mutation rate by dividing by 13,000,000 divided by 25 and taking means across windows. Finally, we took the mean recombination rate in cM/Mb from the sex-averaged Oxford map contained within Eagle2 (*52*) and converted it to Morgans per base, to get a value in the same units.

Comparing number of ancestral recombination events to recombination rate

We binned the genome into 50kb blocks and counted the number of ancestral recombination events occurring completely within each bin. We then computed the mean recombination rate in cM/Mb in the same windows, using the sex-averaged Oxford map contained in Eagle2 (*52*).

Admixture scans

The central challenge of creating admixture maps is to disentangle incomplete lineage sorting (ILS) from admixture. Both processes create local trees in the genome that group candidate admixed haplotypes with admixer haplotypes. Clades resulting from ILS are older than those resulting from admixture, however; they should therefore persist for shorter stretches along the genome and have older times to most recent common ancestor (TMRCAs). Clades resulting from ILS are separable from those resulting from admixture using these two metrics, although the low resolution of branch lengths in the inferred ARG makes this problem more difficult than when true branch lengths are known (Fig. S 12). Because of this, we established several filtering steps to distinguish ILS from admixed clades in our data set.

To map Neanderthal and Denisovan ancestry, we first scanned through ARG output for all clades that grouped some modern human haplotypes with one or more admixer haplotypes (Neanderthal and/or Denisovan) to the exclusion of some other modern human haplotypes. Since SARGE produces many polytomies, this carries the risk of observing a parent of one or more true admixed clades, but not the true admixed clade. This would manifest as a clade containing many modern humans, in addition to one or more archaic hominins, and would falsely be interpreted as a very high-frequency archaic-introgressed haplotype. To mitigate this problem, we defined the Mbuti, Biaka, and Khomani-San genomes as an outgroup population (Africa-MBK) and discarded any clade that contained more than 10% of the outgroup members. We set this 10% threshold, rather than excluding all clades containing outgroup members, because of later filtering steps also designed to eliminate ILS clades. Next, we discarded any clades that persisted for less than 5kb along the chromosome, as well as one extremely long haplotype that fell within a centromere, as annotated by the UCSC Genome Browser (*54*). For each clade that passed our selection criteria, we visited each non-archaic hominin member and determined whether that member possessed candidate Neanderthal, Denisovan, or undetermined ancestry by assessing whether it was closer (by tree topology, ignoring branch lengths) to a Neanderthal or Denisovan haplotype, or equidistant to both.

We then computed the mean time to most recent common ancestor (TMRCA) between each human member and the candidate archaic introgressor across each haplotype. SARGE

reports each branch length as a percent of the total height of the tree both above and below that branch (Fig. S *4*), which in our data set includes all fixed differences between the genomes in our data set and the reference chimpanzee genome. The units in which these TMRCAs are expressed are therefore a percent of two times the total divergence time between humans and chimpanzees. These values were then corrected for branch shortening, according to the mean branch shortening or "missing evolution" values published in a prior study (*1*), converted to branch units by dividing each number by 13 million years. To correct for branch shortening, we added the amount of missing evolution reported for a given archaic genome (in the same units as ARG branch lengths) to the archaic genome's branch when computing TMRCAs. In other words, if the TMRCA between a modern and archaic haplotype is x and the branch shortening value for the archaic haplotype is y, then the corrected TMRCA between them is (2x + y)/2. We note that the values published in (*1*) were produced assuming a 13 my human/chimpanzee split time and a mutation rate of 0.5 x 10^{-9} per base per year. Our TMRCA estimates produced assuming a 6.5 my human/chimp divergence time were in line with prior estimates, however (Fig. 2A). We therefore chose to use a 6.5 Mya human/chimp divergence time to scale all of our branch lengths. Using this value along with the branch shortening values in (*1*) is equivalent to assuming a 1 x 10^{-9} mutation rate per base per year. We are aware that this value is controversial, and that mutation rates estimated using human pedigrees and the fossil record are sometimes in conflict (*15*) (Supplementary Text).

Our ascertainment strategy allowed some candidate admixed haplotypes to intersect. If a human haplotype was inferred to have Neanderthal ancestry at one site and Denisovan ancestry at the next variable site, for instance, it is likely that the two haplotypes actually originated from the same source. Because of this possibility, we merged all such haplotypes together, averaging the TMRCA to admixer in each, weighting by the relative lengths of the two haplotypes being merged. We then repeated this process until each haplotype was unable to merge with any others. Because each such haplotype resulted from the combination of haplotypes with different types of ancestry, all merged haplotypes were labeled as ambiguous, rather than Neanderthal or Densiovan, origin.

At this stage, the set of haplotypes likely still contained haplotypes resulting from incomplete lineage sorting rather than admixture. To help mitigate this, we assigned each a score designed to be high in cases of admixture and low in cases of ILS. We chose a date that predates the human/archaic hominin separation – 700 kya – and computed the p-value of each haplotype having originated at that time or earlier according to its length and according to its TMRCA to the candidate admixer. For this, we assumed both TMRCAs to admixer and haplotype lengths to be exponentially distributed, and we assumed neutral evolution, a standard recombination rate of 1 cM/Mb, a 25-year generation time, and 6.5 Mya human-chimp divergence. We assigned each haplotype a probability of resulting from ILS, based on its TMRCA: if the TMRCA (as a percent of the height of the tree) is y, then $p_{TMRCA} = 1 - e^{-\left(\frac{13e6}{700000}\right)(y)}$. We then assigned each a probability of resulting from ILS, based on its length: if x is the length of a haplotype, in bases, then $p_{len} = e^{-\frac{700000}{1e8*25}x}$. We then computed a score, using a pseudocount of 0.001 to avoid taking the logarithm of 0: $S_{adm} = -(\log_{10}(p_{TMRCA} + 0.001) + \log_{10}(p_{len} + 0.001))$.

Because our prior filtering strategies already removed many ILS haplotypes, we sought to find the smallest threshold for S_{adm} that gave reasonable results. We therefore tested S_{adm} cutoffs at every fifth percentile of the distribution across all haplotypes, and computed the coefficient of variation (standard deviation divided by mean) for TMRCAs to admixer in each class of archaic-introgressed haplotypes (Neanderthal, Denisovan, and ambiguous) within each

human population defined within the Simons Genome Diversity Project (*14*). Because we expect members of these populations to have somewhat shared histories, we expected the coefficient of variation to the admixer to decrease and level out as an appropriate cutoff was reached, reflecting the removal of highly variable segments with old TMRCAs (ILS). We found such a cutoff at the 30th percentile of the score distribution and excluded haplotypes with scores below this threshold (*Fig. S16*). Because our resulting archaic introgression maps still underestimated archaic ancestry per genome relative to the D-statistic (*Fig. 3A*) and because our real data set overestimated the extent of ILS relative to every demographic simulation we ran and underestimated admixture relative to the size of inferred deserts (*Fig. 4*), we believe that this cutoff was reasonable, if not overly conservative.

We also quantified uncertainty in our percent admixture estimates using the weighted block jackknife technique (*31*) with 10 megabase blocks. We generated windows using BEDTools, omitted each from our admixture maps in turn, and re-calculated the percent of each type of archaic ancestry in each SGDP genome from our admixture maps. We then used this distribution of archaic ancestry estimates for each individual to obtain a 95% confidence interval. Mean percent archaic ancestry, along with the minimum lower and maximum upper bound of the 95% confidence interval for each type of ancestry across all genomes in each population are given in Table S2.

Because our method relies on phased data, improper phasing could pose problems for this by breaking haplotypes where they should not be broken. Improper phasing is most likely to be a problem in the archaic genomes, for which reads were short and there is no phasing reference panel, so many sites were phased randomly, after phasing the others using a (likely inadequate) modern human reference panel (see Data Processing section above). We worked around this issue by only conducting admixture scans on modern human genomes – those more likely to be properly phased – and tracking only whether each candidate admixed clade contained Neanderthal or Denisovan haplotypes, without regard to which specific Neanderthal or Denisovan haplotypes they contained. In other words, a clade containing one modern human genome haplotype and one Neanderthal haplotype would still be considered the same clade (potentially from the same admixture event) if at the next site the Neanderthal haplotype contained within was replaced by another. To ensure that this did not negatively affect our results, we repeated some admixture scans on simulated data after intentionally introducing phasing errors into the archaic genomes; we did not see evidence that this negatively affected the results (Supplementary Text).

Analysis of introgressed haplotypes

We obtained geographic coordinates for each human genome from the Simons Genome Diversity Project data (*14*). For each type of archaic hominin ancestry, we then found segments of consensus across all our haplotypes using BEDTools multiinter (*60*). In addition to computing the frequency of each introgressed haplotype across our human genome panel ("global frequency") this way, we also used this data set to obtain the set of geographic coordinates for each human sharing each archaic hominin haplotype. Using geopy (https://github.com/geopy/geopy), we then computed the geodesic distance in kilometers between each pair of humans and selected the maximum such value as the geographic radius covered by the haplotype. To obtain geographically restricted haplotypes, we then discarded any

archaic hominin haplotypes for which any part intersected a consensus haplotype with a geographic radius of 3,000 km or more (using bedtools subtract –A).

Creating the four-part map (admixture, deserts, human-specific regions, and ILS)

Once our individual admixed haplotypes were compiled (see Admixture Scans section), we merged all admixed haplotypes passing filters (using BEDTools merge (*60*)) to create a full set of all regions containing admixture.

Next, we defined a desert region as a region lacking any clade that fails the four gamete test with a clade comprising all modern human genomes in our data set. We required desert regions to begin and end with such trees. The possibility exists, however, that alleles indicative of archaic hominin admixture and/or ILS could exist within these regions, but at sites excluded from the Simons Genome Diversity Project data set. To help mitigate this possibility, we downloaded a de-identified set of genome-wide variant calls from up to tens of thousands of human genomes from the gnomAD database (*61*). We then polarized gnomAD variants into ancestral and derived alleles using the chimpanzee genome panTro4 (*53*) and compiled a catalog of non-fixed derived alleles in modern humans, defined as non-chimpanzee alleles existing at greater than 1% and less than 99% frequency in the gnomAD database. We chose these cutoffs because the large size of the database introduces the possibility of back-mutations and false variant calls due to sequencing error. We then identified all such variants matching any archaic hominin variants in our call set, at positions passing our quality filters. These variants were treated as indicators of archaic hominin admixture and/or ILS.

To filter our desert regions according to this set of admixture and/or ILS-indicative alleles, we first compiled the set of all such alleles falling within each desert region, using BEDTools intersect. We then split our desert regions containing these alleles into two new regions, one with the original start coordinate and an end coordinate halfway between the start coordinate and the first admixture and/or ILS-indicative allele, by genome coordinate (rounded down). The second region was defined with a start coordinate halfway between the last such admixture and/or ILS-indicative allele (rounded up) and the original end coordinate. All desert regions not containing any such alleles were kept unchanged.

To find human-specific regions, we downloaded a published catalog of human-specific derived mutations, based on an aggregate of multiple mutation databases (*32*). We filtered this catalog to variants where the archaic hominin frequency was 0 and the human-specific frequency was above 99%; we chose this cutoff rather than 100% due to the large numbers of samples in databases introducing the possibility for back-mutations and false variant calls due to sequencing errors. We found that, of 118,519 sites passing these criteria, 51,162 had matching archaic variants in our call set and were therefore discarded. This is likely due to the fact that the catalog was generated using mapability-masked variant calls for the archaic hominins, while our archaic allele call set spanned the entire genome. Another 46,206 variants from the catalog were missing from our data set, either because they were missing from the Simons Genome Diversity Project data set, lacked a chimpanzee allele, were not biallelic, failed one or more quality filters, or fell within a CpG site.

Ultimately, 17,888 alleles from the catalog fell within our filtered set of deserts, resulting in 10,503 “human-specific” desert regions containing these alleles.

With deserts, human-specific, and admixed regions thus defined, we created a BED file of ILS regions by subtracting these other regions from the autosomal genome (using BEDTools complement (*60*)).

We tested our four types of genomic regions for enrichment or depletion of intersection with various genomic features using the GenometriCorr R package (*62*). We used the Gencode gene annotation (*63*), using Ensembl version 94 on human genome version GRCh38 lifted over to GRCh37 coordinates, limited to protein coding genes, for locations of both genes and exons. For both gene and exon annotations, we merged together the locations of all annotations on the autosomes. Our regulatory element binding sites are from the filtered "double-elite" set within the GeneHancer database (*64*), downloaded from the UCSC Genome Browser ("interactions" table on hg19) (*54*), which we also limited to autosomes and merged. Our overlap p-values were computed using the projection test in GenometriCorr, and our distance p-values are from the Kolmogorov-Smirnov distance correlation test in the same package. For each test, we limited background regions to 50kb genomic windows that contained polymorphic sites in the Simons Genome Diversity Project data.

Comparing desert sizes to coalescent simulations

For the sake of having a model against which to compare the extent of genomic regions we found to contain admixture, ILS, the absence of both (archaic hominin deserts), and the absence of both plus a fixed derived allele specific to modern humans (human-specific regions), we ran a coalescent simulation based on a simple three-population demographic model for modern humans (*59*), with the addition of two Neanderthals from different populations, both differently related to the population that admixed with modern humans (*1*, *15*) and one Denisovan (see "Demographic simulations section" above).

After obtaining our set of deserts and human-specific regions using real data, we scanned our inferred ARG on data from each demographic simulation for deserts, human-specific regions, and admixture. Due to the small size of the simulated chromosome, the small number of simulated samples, the presence of an unadmixed outgroup unlike in real data, we did not use the same technique as in real data to detect admixture. Instead, we first scanned for all clades that grouped one or more modern human haplotypes with archaic hominin haplotypes to the exclusion of other modern human haplotypes, discarding any haplotype containing African genomes. Since we had prior knowledge of the admixture times in our simulations, we then required all admixed haplotypes to have a TMRCA to the admixer of no greater than four times the time of introgression (80 kya for Denisovan ancestry, or 200 kya for Neanderthal or ambiguous ancestry). This worked well, producing admixture proportions in agreement with expectation (*Fig. 4C*) and comparable to the D-statistic, owing to the presence of an unadmixed outgroup, as well as the uniform mutation and recombination rates producing more uniform haplotype block lengths. We then computed deserts the same way as in real data, but without the need to filter out regions missing from our input data. We then identified human-specific regions by searching for deserts that contained a derived allele specific to, and fixed in, our simulated human genomes. With maps of admixture, deserts, and human-specific regions produced this way, we then determined ILS regions by subtracting these regions from the rest of the genome (using BEDTools complement (*60*)).

For each simulation, we plotted the different metrics of interest (individual percent archaic ancestry and extent of regions of admixture, desert regions, and human-specific regions across all genomes) and modeled the space between points as straight lines. We then compared the same metrics computed using real data and found the intersections between our real data values and the lines between simulated data points; the x-value of each point of intersection was taken to be the inferred admixture proportion using each statistic (*Fig. 4*).

Haplotype block lengths in SARGE-inferred versus true ARGs

Using the results of our coalescent simulations (see Comparing desert sizes to coalescent simulations above), we sought to assess whether haplotype block lengths inferred by SARGE were reliable compared to true haplotype block lengths given by the ARG output by the simulation software. Haplotype block lengths manifest in the ARG as the number of bases for which the clade sharing the haplotype persists along the chromosome. We therefore randomly sampled the distance of persistence of 100,000 clades from the simulation's true ARG and that inferred by running SARGE with a 25kb propagation distance. Although the median haplotype block length in the inferred ARG is close to the median haplotype block length in the true ARG, the mean true haplotype block length is 2.46 times the mean inferred haplotype block length (*Fig. S 14*). This suggests that SARGE artificially breaks some haplotypes, possibly due to mistakes made inferring ancestral recombination events.

We note that this problem is almost certainly exacerbated in real data, for several reasons. For one, the clustering of SNPs and the existence recombination hotspots in real data, in contrast to the uniform mutation and recombination rates used in our simulation, will require SARGE to infer more ancestral recombination events, and the proximity of these ancestral recombination events to each other will increase the likelihood of making mistakes when inferring them. Second, phasing errors will create false ancestral recombination events and artificially shorten haplotype block lengths. We see evidence that incorrect phasing due to the absence of suitable reference data likely caused the inference of incorrect ancestral recombination events, which defined clades that disagreed with those learned from SNP data (*Table S 1*). Wherever this happens in the genome, it will incorrectly break haplotypes.

Because of this potential limitation, we do not seek to biologically interpret haplotype block lengths, *i.e.* to infer times of admixture directly from lengths of haplotypes resulting from admixture. Instead, we only compare haplotype block lengths across admixed individuals and populations to gain an idea of relative times of admixture.

Determining true extent of deserts

To determine if the deserts and human-specific regions we detected represent the full extent of those regions across all humans, or whether they are a superset that would decrease with the examination of more genomes, we randomly sampled (without regard to population or phylogenetic position) sets of 10, 50, and 100 human haplotypes from the SGDP data set and added all archaic hominin haplotypes (Altai, Vindija33.19, and Denisovan) to each set. We re-ran SARGE on each of these data sets, with the same parameters as the full run (excluding CpG sites, and with 25kb propagation distance) and scanned the results of each for deserts and human-specific regions. For each desert region, we report both the full extent of the desert region and the extent after removing regions around alleles shared between modern and archaic hominins and present at between 1% and 99% in modern humans, according to the gnomAD database (*61*).

Functional analysis of human-specific alleles

Our first test for functional significance of human-specific regions was a position-based Gene Ontology enrichment analysis. Using the October 29, 2018 version of the Gene Ontology (GO) database (*65*) together with the Gencode gene annotation (*63*), using Ensembl version 94

on human genome version GRCh38 lifted over to GRCh37 coordinates, limited to whole protein coding genes (introns included), we created a merged BED file of all genome regions mapped to each GO term. We then performed an overlap enrichment test between human-specific regions and each GO term's regioons using the GenometriCorr R package (*62*), with the merged set of all 50 kb windows in the genome containing variant sites in the Simons Genome Diversity Project data as background regions. We took the right-tailed projection test p-value, multiplied by the number of tests, as the p-value for each term.

We then sought to look at specific mutations hypothesized to have a functional impact. After obtaining a catalog of human-specific derived alleles in deserts (see Creating the four-part map section), we screened these alleles for potential functional impact in two ways. First, we selected any of these mutations reported as nonsynonymous variants ("HHMCs") by the catalog's authors. We then intersected all mutations with a heavily-filtered "double elite" set of regulatory element binding sites, mapped to the genes they regulate, by the GeneHancer database (*64*). We used the resulting set of 2,686 candidate functionally-significant human-specific mutations within deserts, mapped to their affected genes, in all downstream functional analyses.

Dates were obtained, where available, for these mutations by searching ARG trees for clades tagged with their SNP positions. Since each of these lineages was human-specific, all haplotypes within each clade were modern humans and no adjustments needed to be made for branch shortening. Because it is impossible to determine the order of mutations along a branch of a tree, each mutation's age was taken to be the height of the midpoint of the branch on which it occurred. In other words, for each clade tagged with a SNP of interest, we calculated the height (distance from the present time) of the clade and the height of its parent; the mean of these two values was used as the age of the mutation.

We compared the ages of mutations affecting interacting sets of genes using data from the STRING database (*33*), limited to interactions with a score of greater than 700. Each pair of candidate functional mutations (nonsynonymous or regulatory) with dates inferable from the ARG were searched for interactions of any kind among the genes they affect.

We searched genes affected by our candidate functional mutations for tissue specificity using data from the GTEx database (*66*). We downloaded a data set containing the median expression value for each gene across a set of tissues in a wide variety of samples. We then discarded all expression values for cell lines, which may exhibit unusual expression patterns that do not correspond to healthy living tissues. In case a gene still had data for multiple tissues reported in this data set, we re-computed the mean across all tissues reported. For each gene, we then computed tau, a measure of tissue specificity, due to its robustness when compared to several other tissue-specificity metrics (*67*). For analyses we performed that required tissue-specific genes, we chose a tau cutoff of 0.9 to determine specificity.

When analyzing tissue-specific expression of genes affected by human-specific mutations through time, simply reporting the number of genes affecting each tissue would likely only reflect the fact that many more genes are specifically expressed in certain tissues than others (*66*). To highlight tissues affected by mutations at specific time points, we used the genome-wide tissue specific expression pattern as background. For each time bin we chose, we compiled a list of all genes affected by candidate regulatory and nonsynonymous human-specific mutations within that time bin. We then limited these genes to those showing tissue-specific expression (tau > 0.9) and counted the number of affected tissue-specific genes each non-cell line tissue in the GTEx database had in each time point. We normalized these counts by the total number of affected tissue-specific genes per time point. Next, we obtained a background distribution by

counting the number of total tissue-specific genes per tissue genome wide, normalized by the total count of tissue-specific genes genome-wide. In each time bin, we reported our normalized counts per tissue by the background normalized count to obtain a measure of enrichment.

For determining the types of biological processes affected by human-specific mutations through time, we first discarded all mutations inferred to be older than 1 my and placed all other mutations into bins 100 ky wide. For each time bin, we then gathered the set of all genes falling within that time bin and compiled all Gene Ontology terms with which each such gene is annotated. For each GO term, we then searched up the GO term hierarchy for parent terms "developmental process" (GO:0032502), "immune system process" (GO: 0002376), "metabolic process" (GO: 0008152), and "reproductive process" (GO: 0022414). We counted the number of times each of these terms was the parent of one or more terms in a given time bin, allowing each term to count toward more than one parent category.

We created scores for each human-specific mutation based on the age of the mutation as well as the length of the desert region surrounding it. Taking each variant's TMRCA to be t and each variant's surrounding desert length to be l, and the maximum TMRCA for all variants t_{max} and the maximum desert length for all variants l_{max}, then each variant's score was calculated as $s_v = -\log_{10}\left(1 - \frac{l}{l_{max}+1}\right) - \log_{10}\left(\frac{t}{t_{max}}\right)$ for variants with known TMRCA and $s_v = -\log_{10}\left(1 - \frac{l}{l_{max}+1}\right)$ for variants without a known TMRCA. As this score was intended to be higher for selected variants (with long desert length and recent TMRCA), we performed Gene Ontology enrichment analysis using the Wilcoxon rank-sum test implemented in FUNC (*68*), running the refinement routine with p-value cutoffs of 0.01 and reporting genes enriched in high scores.

Supplementary Text

Adding nodes to the ARG

Every node in the ARG must be "anchored" at one or more genomic positions. This is because each node's start and end coordinates depend on these positions, along with the propagation distance p. When a node's range is interrupted by a new node with which it fails the four haplotype test being created in the middle of its range, for example, the set of genomic sites the node "owns" are used to determine its new range. Additionally, to make it easier to look through the ARG, we store a mapping of sites to ARG nodes with those sites. Since it is set up this way, we do not allow any node to be created if the site at which it is originally anchored already is tied to an existing ARG node with which it fails the four haplotype test.

When a new node is to be created, we first check to see if it can be merged with an existing node. If two nodes have the same clade, if the ranges implied by their sites and the propagation distance overlap, and if they do not fail the four gamete test with a node between them, then they can be merged. If the new node matches two existing nodes that it overlaps both upstream and downstream, then all three nodes are merged.

If a new node does not merge with an existing node, then we compile all four haplotype test failures it has with other nearby nodes (within propagation distance). If the new node is in the middle of an existing node's range and the new node fails the four haplotype test with that node, then that existing node is split into two nodes. Otherwise, start and end coordinates are adjusted: if node A and B fail the four haplotype test and node A is upstream of (has lower coordinates than) B, the end coordinate of A is set to the highest coordinate site that it owns and the start coordinate of B is set to the lowest coordinate site that it owns (Fig. S40).

Once all node ranges have been adjusted according to four haplotype test failures, then all parent/child relationships are created. To do this, a depth-first search is performed across the ARG down from the root, across the entire range of the new node. The new node may have different parents and children across its range: each parent/child edge has start and end coordinates (Fig. 1A).

After all parent/child edges are added, recombination edges are added to the graph. Upstream and downstream nodes failing the four gamete test with the new node are sorted by distance from the new node. If any pair of upstream or downstream nodes in these sets fail the four gamete test with each other, the one further away from the new node is removed from the set. Recombination edges are then added between the new node and all remaining nodes, except for four gamete test failures that can be explained by a previously-solved recombination event.

Recombination edges include paths through candidate moving clades, which are not yet part of the ARG (Fig. 1A). Normally, if an upstream clade u and a downstream clade d fail the four haplotype test, then clades with the members of $u \cap d$, $u \setminus d$, and $d \setminus u$ are added as candidate moving clades. If any of these candidate moving clades fails the four gamete test with a node already in the ARG between u and d, however, it will not be created.

A special case for adding nodes exists for clades where every haplotype shares the derived allele. These sites can only contribute to the branch length of the root node. Therefore, we store a single root node whose start and end coordinates span the entire chromosome. If a mutation is observed for which every haplotype shares the derived allele, it is added to the root node. Inferred (non-mutation) sites with this clade are ignored and not added to the root node.

Similarly, clades where every haplotype shares the ancestral allele are not informative for the ARG and are skipped altogether.

Solving ancestral recombination events

The process of "solving" ancestral recombination events consists of finding a node with unsolved recombination edges connecting it to one or more nodes downstream, finding a subgraph of the ARG containing other nodes involved in this or possibly other recombination events, filtering the subgraph so that it only describes a single recombination event, and then choosing the most likely γ node that could explain the recombination event (similar to the "two-trees" algorithm, Fig. S1). Finally, the chosen γ node is added to the ARG as a standard tree node, the start and end coordinates of all nodes involved are adjusted to account for the inferred recombination event, and any nodes that do not exist in the ARG but whose existence is implied by the recombination event are created (Fig. S2B). This process is the core of the ARG inference algorithm, as it allows for the creation of nodes not directly observed in the input SNP data.

One concept used by several stages of this algorithm is that of tree-compatibility between two nodes (Fig. S41). Two nodes are tree-compatible if, according to their clades, genomic positions, and genomic positions of their four haplotype test-failing partner nodes, they can both exist in the same tree. At this stage in ARG building, start and end coordinates have not yet been finalized, so we cannot define compatibility based on coordinates alone. However, if two nodes already have overlapping start and end coordinates, then they must be compatible. Additionally, two nodes that fail the four haplotype test cannot be compatible. Otherwise, we must rely, for upstream nodes, on the lowest start coordinate of all downstream tree nodes connected to the node via recombination edges. Likewise, for downstream nodes, we consider the highest end coordinate of all upstream tree nodes connected to the node via recombination edges. We refer to this value, in both cases, as the "closest recombination partner" of the node; this determines how

far the node's end coordinate (if upstream of a recombination event) or start coordinate (if downstream of a recombination event) could be extended in the ARG. Whether or not any two tree nodes are tree-compatible depends on the location of both nodes' closest recombination partners. If a node A is upstream of node B, then in order for nodes A and B to be tree-compatible, node B must be upstream of node A's closest downstream recombination partner and node A must be downstream of node B's closest upstream recombination partner (Fig. S41).

Before solving an ancestral recombination event, SARGE must find a subgraph of the ARG containing a set of tree-compatible upstream nodes *U*, a set of tree-compatible downstream nodes *D*, and a set of candidate γ nodes *L* that connect together nodes in *U* and *D*. Ideally, *U*, *D*, and *L* should correspond to a single ancestral recombination event; however, in practice, there are situations in which a single event is difficult or impossible to distinguish from multiple events (Fig. S43). We will hereafter refer to this ARG subgraph used to infer ancestral recombination events as a "recombination graph."

Collecting a recombination graph begins with a "key" node *k*, which is a tree node in the ARG with unsolved recombination edges to downstream nodes. To begin, we visit each candidate γ node downstream of *k* and add it to *L*. Next, we visit all upstream tree nodes connected to every node in *L* and add them to *U*, if they are tree-compatible with *k*. We then follow all recombination edges from nodes in *U*, through candidate γ nodes, to tree nodes with start coordinates higher than the end coordinate of *k*. These nodes are added to *D*, and all candidate γ nodes along their paths to nodes in *U* are added to *L*. We then revisit nodes in *U*; any that are not connected via recombination edges to nodes in *D* are removed from *U*.

At this point, the highest end coordinate of nodes in U and the lowest start coordinate in D give boundaries between which the ancestral recombination event must have happened. Therefore, any node in U or D whose closest recombination partner falls within, rather than outside, these boundaries, is removed (Fig. S 42).

The next step is to filter *U*, *D*, and *L* to a set of nodes describing only a single recombination event. This is the most intensive part of the algorithm, as it must explore a large set of choices. At this stage, the recombination graph is likely to represent several different recombination events, which must be pared down to one before a branch movement can be inferred. The goal of this step is to obtain a set of tree-compatible upstream nodes U and a set of tree-compatible downstream nodes D, where all nodes in U and D are tree-incompatible with each other. Additionally, the closest downstream recombination partner of each node in U must be present in D, and the closest upstream recombination partner of each node in D must be present in U.

Because the nodes in U must all be tree-incompatible with the nodes in D, then if there exists a pair (u, d) of candidate upstream and downstream nodes that are tree-compatible, either u or d must be excluded. We therefore define a set C containing pairs of sets of upstream and downstream nodes (u_s, d_s). For each such pair, including the nodes in u_s in the recombination graph requires excluding the nodes in d_s, and vice versa.

```
DECLARE set C ← ∅
FOR u in U:
      DECLARE set u_s ←[u]
      DECLARE set d_s ← ∅
      FOR d in D:
            IF tree-compatible(u, d):
                  d_s ← d_s U d
      DECLARE bool found ← FALSE
```

```
FOR (us', ds') in C:
    IF ds' ⊆ ds:
        ds'  ← ds' ∪ ds
        us'  ← us' ∪ us
        found ← TRUE
        break
IF not found:
    C ← C ∪ (us, ds)
```

Next, because of the rule that the closest downstream recombination partner of each node in U must exist in D and vice versa, we store a collection of pairs of "partner" sets P. Each member of P is a pair of sets of upstream and downstream nodes (u_s, d_s), where including the nodes in us requires also including the nodes in ds.

```
DECLARE set P ← ∅
FOR u in U:
    DECLARE set us ←[u]
    DECLARE set ds ←[closest_recombination_partner(u)]
    FOR (us', ds') in P:
        IF |us ∩ us'| > 0 or |ds ∩ ds'| > 0:
            us'  ← us' ∪ us
            ds'  ← ds' ∪ ds
            found ← TRUE
            break
    IF not found:
        P ← P ∪ (us, ds)
```

Given all choices described by the node sets in *C* and *P*, we now build a set *S*, where each entry is a set of upstream and downstream nodes (U, D) that could describe a single recombination event. To populate *S*, we first enumerate all possible choices in *C*, then filter according to the constraints imposed by the pairs in *P*.

```
DECLARE pair of sets (ufirst, dfirst) ← first set pair in C
C ← C \ (ufirst, dfirst)
DECLARE set S ← [(ufirst, D \ dfirst), (U \ ufirst, dfirst)]
FOR (u, d) in C:
    DECLARE set Snew ← ∅
    FOR (u', d') in S:
        IF | u ∩ u' | > 0 and | d ∩ d' | > 0:
            IF | u' \ u | > 0 and | d' \ d | > 0:
                Snew ← Snew ∪ (u' \ u, d')
                Snew ← Snew ∪ (u', d' \ d)
    S ← Snew
FOR (u, d) in S:
    FOR (u', d') in P:
        IF | u ∩ u' | > 0 or | d ∩ d' | > 0:
            IF not ( u ⊇ u' and d ⊇ d'):
                u ← u \ u'
                d ← d \ d'
                IF u == ∅ or d == ∅:
                    S ← S \ (u, d)
```

We now have in *S* a set of choices of full recombination graphs. In the spirit of parsimony, we choose the set with the highest total node count (the recombination graph

containing the most possible four gamete test failures). If there is a tie, we choose the set of nodes covering the smallest genomic span (the lowest start coordinate in D minus the highest end coordinate in U). The reasoning behind this choice is that sets of nodes covering greater genomic distances are more likely to be affected by multiple ancestral recombination events than sets spanning smaller genomic distances.

Before solving the recombination event, we check pairs of upstream and downstream nodes in the recombination graph (members of U and D) that fail the four gamete test, to see whether their four gamete test failure could be explained by a previously-solved ancestral recombination event. If so, both nodes in the pair are removed from the recombination graph.

At this stage, there are still boundary cases in which it is impossible to determine if a given recombination graph describes one or more recombination events (Fig. S43). Because of this, we employ a heuristic check to see whether the graph might describe multiple recombination events. If the last (highest-coordinate) node in U, U_L, does not fail the four gamete test with the first (lowest-coordinate) node in D, D_F, then we gather two alternative recombination graphs. One excludes U_L and includes any additional downstream nodes in D made possible by this exclusion. The other excludes D_F and includes any additional upstream nodes in U made possible by this exclusion. If either of these alternative recombination graphs covers a smaller genomic distance than the main graph being considered, we take this as evidence that the main recombination graph might describe multiple ancestral recombination events. If this is the case, we defer solving it until neighboring recombination events have been solved. If a given "key" node is visited a second time, and thus the same recombination graph is revisited, the recombination graph is solved regardless of the outcome of these checks.

At this stage, nodes in U should belong to a single upstream tree, nodes in D should belong to a single downstream tree, and nodes in L represent candidate ancestral recombination clades. Similar to the "two trees" algorithm explained in Materials and Methods and Fig. S2, we choose the L clade that explains the most four gamete test failures between nodes in U and D. If there is a tie, we choose the node in L that is the most compatible with the existing ARG topology in the surrounding region. We determine this by checking how many bases the L clade – and each other new clade it implies – can exist along the chromosome. If p is the propagation distance parameter, and a candidate L clade is γ, then:

```
DEFINE function fourhap_test(set x, set y):
      IF | x ∩ y | > 0 and x ⊄ y and x ⊉ y:
            RETURN TRUE
      ELSE:
            RETURN FALSE
DEFINE clade_type(set x, set γ):
      RETURN clade type stored in recombination edge connecting node x
            to node γ
DEFINE set dists ← ∅
DEFINE recomb_start ← max(end coordinate in U)
DEFINE recomb_end ← min(start coordinate in D)
DEFINE int ldist1 ← p
      FOR int pos = recomb_start-1; pos >= recomb_start– p; --pos:
      FOR ARG clade at pos c:
            IF fourhap_test(c, γ):
                  ldist1 ← recomb_start – pos
                  break
DEFINE int ldist2 ← p
```

```
FOR int pos = recomb_end + 1; pos <= recomb_end + p; ++pos:
      FOR ARG clade at pos c:
            IF fourhap_test(c, γ):
                  ldist2 ← pos – recomb_end
                  break
dists ← dists ∪ (ldist1 + ldist2)
FOR u in U:
      u′ ← u
      IF clade_type(u, γ) == α:
            u′ ← u / γ
      ELSE IF clade_type(u, γ) == β:
            u′ ← u ∪ γ
      IF | u′ | > 0:
            DEFINE udist ← p
            FOR int pos = recomb_end + 1; pos <= recomb_end + p; ++pos:
                  FOR ARG clade at pos c:
                        IF fourhap_test(c, u′):
                              udist ← pos – recomb_end
                              break
            dists ← dists ∪ udist
FOR d in D:
      d′ ← d
      IF clade_type(d, γ) == α:
            d′ ← d ∪ γ
      ELSE IF clade_type(d, γ) == β:
            d′ ← d \ γ
      IF | d′ | > 0:
            DEFINE ddist ← p
            FOR int pos = recomb_start - 1; pos <= recomb_start - p;
                  --pos:
                  FOR ARG clade at pos c:
                        IF fourhap_test(c, d′):
                              ddist ← recomb_start - pos
                              break
            dists ← dists ∪ ddist
meandist = 1/|dists| ∑_{i=1}^{|dists|} dists_i
```

Then, the clade in L tied for the most four haplotype test failures in U and D, with the highest meandist is chosen as the correct γ clade.

The recombination event must have happened between the highest end coordinate in U and the lowest start coordinate in D. We choose two adjacent sites within this interval, as close to the center of it as possible, as the boundaries of the inferred recombination event. Naming these site coordinates x and y, where x is lower than y, we expand the end coordinate of each node in U to x and the start coordinate of each node in D to y.

Finally, we create new nodes implied by the recombination event. If the chosen moving clade from L is denoted γ, x and y are the chosen coordinates between which the recombination event happened, and p is the propagation distance, then

```
DEFINE set U_new ← ∅
DEFINE set D_new ← ∅
DEFINE set union_all ← ∅
```

```
DEFINE bool alpha_exists ← FALSE
DEFINE bool beta_exists ← FALSE
FOR u in U:
      union_all ← union_all ∪ u
      IF clade_type(u, γ) == α:
            alpha_exists ← TRUE
            D_new ← D_new ∪ (u \ γ)
      ELSE IF clade_type(u, γ) == β:
            beta_exists ← TRUE
            D_new ← D_new ∪ (u ∪γ)
FOR d in D:
      union_all ← union_all ∪ d
      IF clade_type(d, γ) == α:
            alpha_exists ← TRUE
            U_new ← U_new ∪ (d ∪ γ)
      ELSE IF clade_type(d, γ) == β:
            beta_exists ← TRUE
            U_new ← U_new ∪ (d \ γ)
```

The γ clade is added to the ARG, anchored at x and y, with range [x-p, y+p]. All nodes in U_new are added, anchored at x and with range [x-p,x]. All nodes in D_new are added, anchored at y and with range [y, y+p]. Finally, if there were no α nodes, or if there were no β nodes (determined by the values of alpha_exists and beta_exists), we create an ARG node with the members of union_all. If there were no α nodes (alpha_exists is false), this node is anchored at x with range [x-p, x+p]; if no β nodes (beta_exists is false), it is anchored at y with range [y-p, y+p].

Information about the solved recombination event is stored in another type of edge, in order to distinguish four gamete test failures belonging to solved recombination events from unsolved ones.

Finalizing ARG node ranges

Because of the heuristic nature of our method, some ancestral recombination events go unsolved. Additionally, some may be unsolvable (for example, if all three candidate γ nodes for a four haplotype test failure fail the four haplotype test with existing ARG nodes in their range). When this is the case, we seek to expand the ranges of all nodes involved in recombination to their fullest extent. In other words, for every pair of nodes that fail the four haplotype test with each other, we want to ensure that the upstream node's end coordinate and the downstream node's start coordinate are set to sites approximately in the center of the genomic interval between the two nodes. If this is not done, there will be additional polytomies in the ARG. Therefore, when we are about to write a tree at site index s to disk, we seek to ensure that site index $s + 1$ will be covered either by a node in the tree covering s or by a downstream node that fails the four haplotype test with a node in the tree covering s. In this case, s + 1 is not the very next genomic position after s, but the next genomic position with a SNP in the ARG:

```
DEFINE set U ← ∅
DEFINE set D ← ∅
FOR ARG node u at site s:
      IF u end coordinate < s + 1 and u end coordinate + p >= s+1:
```

```
            DEFINE bool u_pass ← TRUE
            FOR downstream recombination partner d of u:
                IF d start coordinate <= s + 1:
                    u_pass ← FALSE
                    break
            IF u_pass:
                U ← U U u
                FOR downstream recombination partner d of u:
                    D ← D U d
FOR d in D:
    FOR upstream recombination partner u of d:
        IF u end coordinate <= s:
            U ← U U u
FOR site z = s + 1; z <= lowest site among nodes in D; ++z:
    DEFINE set U_z ← ∅
    DEFINE set D_z ← ∅
    FOR u in U:
        IF u end coordinate > z and u end coordinate + p >= z:
            DEFINE bool u_pass ← TRUE
            FOR downstream recombination partner d of u:
                IF d start coordinate <= z:
                    u_pass ← FALSE
                    break
            IF u_pass:
                U_z ← U_z U u
    FOR d in D:
        IF d start coordinate > z and d start coordinate – p <= z:
            DEFINE bool d_pass ← TRUE
            FOR upstream recombination partner u of d:
                IF u end coordinate >= z:
                    d_pass ← FALSE
            IF d_pass:
                D_z ← D_z U d
    IF | U_z | == 0:
        FOR d in D_z:
            Expand d start coordinate to z
    ELSE IF | D_z | == 0:
        FOR u in U_z:
            Expand u end coordinate to z
    ELSE:
        DEFINE int udist ← z – highest end coordinate in U_z
        DEFINE int ddist ← lowest start coordinate in D_z – z
        DEFINE float r ← random decimal in [0,1]
        IF udist < ddist or (udist == ddist and r < 0.5):
            FOR u in U_z:
                Expand u end coordinate to z
        ELSE IF ddist < udist or (udist == ddist and r >= 0.5):
            FOR d in D_z:
                Expand d start coordinate to z
```

Collapsing to trees

To avoid making it necessary to hold the ARG over an entire chromosome in memory at once, or to load the entire ARG for all analyses, we represent the ARG on disk as a series of trees. At every site, the ARG collapses to a tree, so we write out each tree independently to disk, along with its chromosome and base position, in a custom serial binary format. We find that our files compress well with GZIP, and we provide utilities for indexing and retrieving specific genomic regions from files, and for converting our trees to Newick format.

Testing the propagation distance parameter

Using data from the demographic simulation with the plausible admixture proportion of 0.05 (see "Demographic simulations" section in Supplementary Methods), we sought to assess the impact of the choice of propagation distance on SARGE's ability to correctly define clades. We ran SARGE on this data set using a variety of propagation distances: 5 kb, 10 kb, 25 kb, 50 kb, 100 kb, and 500 kb.

For each simulation, we measured the specificity (defined as the percent of clades in each tree inferred by SARGE that were present in the true tree from the simulation) and sensitivity (defined as the percent of clades in the true tree from the simulation that were correctly recovered by SARGE). For comparison, we also ran tsinfer, a recently described ARG inference program that scales well to large data sets and also leaves polytomies in output trees (*8*) and Relate, another recently described program that does not produce polytomies (*7*). For the Relate run, we used the mutation rate from the simulation and set the haploid effective population size to two times the effective population size in Africans, according to simulation parameters. We note that this simulation only contained 66 haplotypes and thus SARGE likely achieved higher sensitivity on this data set than it would on one with more haplotypes, as its specificity falls on large data sets, due to increasing numbers of polytomies (*Fig. S6*).

Using SARGE over increasingly large propagation distanves, specificity converged to 0.70 and sensitivity to 0.49 (Fig. S *8*). Using tsinfer, sensitivity was the same as what SARGE achieved with the 500 kb propagation distance (0.49) but specificity was 10% lower (0.60). Relate performed similarly to tsinfer, but with higher sensitivity (0.55) and lower specificity (0.55). Interestingly, although Relate and tsinfer recovered more true clades than SARGE, due to both methods producing fewer polytomies, this difference was minimal (0% for tsinfer and 5% for Relate) when using a large propagation distance on this data set.

A propagation distance as large as 500 kb is impractical on large data sets because it drastically increases the number of comparisons between sites and therefore the execution time. We used a propagation distance of 25 kb for all our analyses of real data, which in this case allowed for reasonably fast computation, as well as high specificity (0.74) and reasonable sensitivity (0.40, compared to the maximum of 0.49).

Properties of missing and incorrectly-inferred clades

Some clades inferred by SARGE are inaccurate. For each such clade, we sought to determine how problematic it might be for downstream inferences. To do this, we used the results of our demographic simulation with a 0.05 admixture proportion (see Supplementary Methods) and found all clades inferred by SARGE with a 25 kb propagation distance that were absent from the true simulation ARG at those sites. We then looked along the chromosome, both upstream and downstream, for the nearest true ARG tree in which the incorrect clade *did* exist (if the clade existed at all elsewhere in the ARG). If this number was very large, this would suggest

that these clades often do not exist anywhere near where they are inferred, possibly because they are the result of incorrectly solving ancestral recombination events. If small, however, then many incorrectly inferred clades are the result of getting the boundaries between ancestral recombination events slightly wrong, and downstream inferences will suffer less, as wrong clades will be in close proximity to loci where they are correct.

In addition to performing this analysis for SARGE, we also ran Relate (*7*) and tsinfer (*8*) on the same data set, using the mutation rate known from the simulation and two times the effective population size in Africans in the simulation as the haploid effective population size for Relate. We also computed the distance of each incorrectly-inferred clade to the nearest site where it was correct in both of these simulations.

Overall, we found that about 14% of clades incorrectly inferred by SARGE did not exist on the chromosome. Of the 86% that did exist, the median distance to a locus where the clade was correct is 3.5 kb (mean distance 92 kb). This is a lower percent of completely missing clades, and a lower distance to positions where clades are correct than that obtained using both Relate and tsinfer (Table S 7, Fig. S *10*).

We also sought to learn where in ARG trees incorrectly-inferred and missing (due to polytomy) clades tended to occur. Using this same demographic simulation, we labeled every clade in every true ARG tree as either correctly identified (present at the same site in SARGE results), incorrectly identified (failing the four gamete test with another clade present at the same site in SARGE results), or missing due to polytomy (not present in the SARGE tree at the same site, but passing the four gamete test with all present clades). We then examined the size (number of member haplotypes) distribution of clades falling into each category. We find that missing clades skew smaller (closer to the leaves of trees) than correctly and incorrectly identified clades. Incorrectly identified clades tend to be larger (closer to the root), and correctly identified clades are intermediate in size between the other two categories. Repeating this analysis with a larger simulated data set (an unstructured population of 500 haplotypes from our QC simulations described in Supplementary Methods), the pattern becomes more visible, although there are many more missing clades (Fig. S *9*).

We conclude from these analyses that SARGE's inaccurate inferences are less problematic (and more likely to occur close in the genome to where they are accurate) than those produced by both Relate and tsinfer. SARGE does leave far more polytomies than either of the other two programs, however, although these polytomies are often concentrated near the leaves of trees, where clades are less useful for making broad phylogenetic inferences.

Testing the accuracy of inferred branch lengths

We sought to test the accuracy of SARGE's inferred branch lengths, compare to those inferred by the recently-described ARG inference program Relate (*7*), and uncover any systematic biases in branch length estimates, using simulated data. For this, we used data from our demographic simulation with a 0.05 admixture proportion (see "Demographic simulations" section in Supplementary Methods). We ran SARGE with a 25 kb propagation distance (the same as was used on our real data set) on the simulated data. We also ran Relate, using two times the simulation's effective population size of Africans as the haplotype N parameter, and using the same per-generation mutation rate as in the simulation. Using both the SARGE and Relate output, we then scanned for clades that were correct according to the true ARG. We then extracted the branch lengths from these clades in both the true and inferred ARG, converting the simulated branch lengths into years (by multiplying by 4*base N*generation time) and SARGE's

branch lengths into years (by multiplying by the simulation's TRMCA of all groups). This strategy is slightly different from our real data, in which we collected all fixed differences between sample haplotypes an outgroup (chimpanzee) genome and multiplied by two times the human/chimp divergence time. In the case of this simulated data, ancestral and derived alleles were known *a priori* and we therefore did not need to use an outgroup. Similarly, Relate requires model parameters for estimating branch lengths, and these were directly known from running the simulation.

We found that SARGE branch lengths were less tightly correlated to true branch lengths than Relate branch lengths were to true branch lengths (SARGE $r^2 = 0.60$; Relate $r^2 = 0.79$). However, we also found that Relate systematically underestimated branch lengths (Fig. S 11). The median difference between SARGE's inferred branch lengths and true branch lengths was approximately -15 ky, while the same value for Relate data was -26 ky. We then scaled these same values by the true branch lengths to obtain a percent error of each inferred branch length estimate: $\frac{|inferred-true|}{true}$. These percent error estimates were very similar for both programs: SARGE median = 1; Relate median = 0.97, which suggests that many inferred branch lengths were either close to zero or double their true value. One cause of this could be the inherent difficulty of inferring lengths of branches with zero mutations – SARGE sets such branch lengths to zero, guaranteeing a percent error of one; Relate's randomly sampled estimates are apparently not much more reliable.

Because SARGE produces more polytomies than Relate as the size of data set increases, we repeated this test using a larger data set: we re-ran the same simulation, but with 450 instead of 50 modern human haplotypes. In this case, we found that SARGE's performance suffered some: r^2 between true and inferred branch lengths SARGE = 0.44; Relate = 0.79. The median difference between inferred and true branch lengths decreased, however: SARGE = -4.6 kb and Relate = -6.0 kb. Median percent error stayed roughly the same: Relate = 0.98 and SARGE = 1. SARGE's branch length estimates also remained unbiased, unlike those from Relate (Fig. S 11).

We take from this exercise that SARGE's branch length estimates, which are based purely on counting mutations and are not smoothed by a model, are imperfect but relatively unbiased estimates of true branch lengths. Relate, on the contrary, is accurate more of the time, but systematically biased toward underestimation.

Converting branch lengths into years in real data

Because branch lengths in SARGE are divided by the total height of the tree both above and below each branch (Fig. S 4), they are reported in units of the total divergence time between the genomes in the data set and the outgroup genome used to determine ancestral and derived states. If the divergence time between the genomes in the data set and the outgroup species is known, then these branch lengths can be converted to years by multiplying by two times this divergence time. In the previous exercise where we assessed accuracy of branch lengths on simulated data, the coalescence time of all lineages under study, as well as the ancestral or derived state of each allele, was already known. We therefore did not simulate an outgroup genome, and the total height of the tree was converted to years by multiplying branch lengths by the coalescence time of all lineages. In real data, where these parameters are unknown, this is not possible.

The mutation rate in humans has been the subject of controversy. Mutation rates estimated by comparing parents to offspring are about half as fast (~1×10^{-8} per base per generation) as mutation rates estimated by calibrating with dated fossils (~2×10^{-8} per base per generation) (*15*); this discrepancy has led at least one recent study to set the human-chimpanzee split time at 13

Mya rather than 6.5 Mya, so as to account for a slower assumed mutation rate (*1*). Other methods for estimating the mutation rate have been developed that use population genetic techniques to estimate the mutation to recombination rate ratio. These methods, as well as a recent approach that used the rate of heterozygosity within identical-by-descent sequences that individuals inherited from a recent common ancestor as a proxy for the *de novo* mutation rate, have produced rate estimates that are intermediate between these two values (~1.5-1.7x10^{-8} per base per generation) (*69*). Interestingly, both this recent study and prior work (*70*) suggest that mutation rates, as well as their sequence contexts, can differ among human populations.

The divergence time between humans and chimpanzees is also controversial. Early studies using genetic divergence to estimate this time placed the human/chimpanzee split time too recently to reconcile with some paleontological data. An approach that combined human and chimpanzee generation time estimates with estimated per-generation mutation rates (to avoid fossil date calibration) placed the split at 7-8 Mya, which is compatible with the fossil record (*71*). Some researchers took issue with this estimate, however, in part because of its use of slow single-generation mutation rates inferred from parent/offspring sequencing data (*72*). Another split time estimated using mutations accumulated at CpG sites placed the divergence around 12 Mya (*73*). One possible explanation for the variation in estimated divergence times is a complex speciation scenario (74), involving multiple splits interspersed with periods of interspecific hybridization. The study describing this scenario produced a divergence time estimate of less than 6.3 Mya (*74*).

Our method is mostly agnostic about the human mutation rate: branch lengths can be converted into years by multiplying by two times the human-chimpanzee divergence time. We obtained TMRCA estimates that agree with previous knowledge including the timing of the out-of-Africa migration event using a 6.5 Mya split time (Fig. 2A), and so we chose this value for downstream analyses. In general, branch lengths can be rescaled to use a different chimpanzee divergence time T, by multiplying by $T/6.5\text{x}10^{6}$. For TMRCAs of clades that include archaic hominin genomes, however, we also incorporated branch shortening values, which quantify "missing evolution" due to sampling genomes from the past, which were reported in a prior study (*1*). This study reported branch shortening values in years that were calculated assuming the slow mutation rate reported in parent/offspring sequencing studies (~1x10^{-8} per base per generation) and a relatively old human/chimpanzee split time (13 Mya). By using these values in our study, along with a 6.5 Mya human/chimpanzee split time, we implicitly assumed a higher mutation rate (~2x10^{-8} per base per generation) more in line with estimates calibrated using fossil dates than per-generation estimates produced using parent/offspring sequencing data.

If we were to re-calculate TMRCAs between admixed modern humans and archaic hominin genomes within archaic-introgressed haplotype blocks using the older (13 Mya) human/chimp divergence time, TMRCAs between humans and the Neanderthal genome would predate the out-of-Africa migration event, which is estimated to have happened within the last 70 ky (*75*). Conversely, if we adjusted the archaic branch shortening values to use the low (~1x10^{-8} per base per generation) mutation rate along with the recent (6.5 Mya) human/chimp divergence, these branch lengths would be reduced by half and would therefore further lower our human/archaic hominin TMRCAs within these haplotype blocks. The estimated TMRCA would then be more recent than the timing of the settlement of Australia, which was estimated recently from mtDNA sequences to be around 50 kya (*76*). We therefore feel that our results comprise evidence in favor of using the faster (1x10^{-8} per site per generation) mutation rate estimated using fossil calibration, along with a 6.5 Mya human/chimpanzee divergence time.

Comparing nucleotide diversity to phylogenetic position in ARG trees

One simple model for human demographic history is the serial founder effect model wherein each human population diverged from a previous source population by undergoing a dispersal-related population bottleneck (*77*). This model has been invoked to explain the decreasing genetic and haplotype diversity seen in populations as a function of their distance from Africa, the geographic source of much of human genetic diversity. Under this model, each human population was founded by members of a preceding population and carries a subset of that population's genetic diversity. As a result, the degree to which a population is ancestral to others – measured as how often that population occupies a basal position in ARG trees – should correlate with the nucleotide diversity within that population. We calculated the probability of each genome haplotype belonging to the smaller of the two clades at the deepest split in well-articulated trees, and plotted this value against per-site nucleotide diversity (Fig. S 44). The residuals to the best-fit line computed excluding archaic hominins provide a measure of how well each haplotype's phylogenetic placement agrees with this model. Among modern humans, residuals are highest in basal sub-Saharan Africans (Africa-MBK) and Papuans. One possible explanation for this observation is that both groups have undergone bottlenecks subsequent to their formation. Additionally, the residual for the Denisovan is 9.5% higher than for the Neanderthals. Although the Denisovan lineage is thought to have separated from modern humans at the same time as the Neanderthal lineage (*1*), it is also thought to possess up to 8% ancestry from a more-diverged, "super archaic" source (*15*), which could help explain this observation.

Testing the effect of incorrect phasing on admixture mapping

SARGE requires phased input, and improper phasing causes SARGE to infer incorrect ancestral recombination events (Table S 1). This is especially a problem for archaic genomes, which certainly contain phasing errors because they lack representation in phasing reference panels and are too fragmented to be experimentally phased. Because of this issue, we do not perform any analyses on individual archaic genomes; rather, we use the archaic genomes to locate and track segments of archaic ancestry in modern human genomes. We infer the existence of blocks of archaic admixture and/or incomplete lineage sorting wherever a tree places a modern human closer to archaic hominins than to other modern humans; we allow the archaic hominin genome haplotypes to change places within these blocks without breaking them, provided they are of the same type (all Neanderthal genome haplotypes can trade places with each other within these blocks, and Denisovan genome haplotypes can also trade places with each other). Nonetheless, we sought to assess whether incorrectly phased archaic hominin genomes would introduce problems for admixture scans in modern humans.

We started with the output of our demographic simulation (see "Demographic simulations" section in Supplementary Methods) with a 0.05% admixture proportion. For each archaic hominin genome in the simulation (Altai Neanderthal, Vindija Neanderthal, and Denisovan), we then simulated phasing errors by randomly swapping the two haplotype's alleles at a randomly-chosen 50% of sites. We then ran SARGE on this data set with the same parameters as before (25 kb propagation distance) and scanned human haplotypes for archaic admixture. Because SARGE makes some decisions randomly and this can produce different output even in two runs with the same parameters on the same data set, we also re-ran SARGE on the original, properly phased data set to get a second replicate unaffected by phasing errors.

When computing the overall percent Neanderthal and Denisovan ancestry for each modern human haplotype using the different data sets, results are very similar. Comparing the two replicates of the properly phased data set, $r^2 = 0.9998$, and the mean difference in admixture proportion estimates is 0.017%. Comparing the properly phased with the improperly phased data set, $r^2 = 0.9990$ and the mean difference in admixture proportion estimates is 0.044%.

We also checked to see whether improper phasing impaired our ability to locate specific archaic ancestry blocks (in specific parts of the genome). To this end, for each modern human genome in the simulation, we computed the Jaccard statistic (using BEDTools) between its ancestry maps produced using the properly and improperly phased data sets, as well as between the two replicate ancestry maps using the properly phased data set. The Jaccard statistic is a ratio of set intersection to set union, where 0 indicates no overlap and 1 indicates complete overlap between two maps. Overlap was high in all cases: for the two replicate properly phased data sets, the mean Jaccard statistic between ancestry maps was 0.899 (full range 0.772-0.973) and the mean Jaccard between ancestry maps using properly and improperly phased data was 0.937 (full range 0.859-0.982).

We concluded from these experiments that the phasing errors present in archaic hominin genomes likely do not have a large effect on our admixture analyses, and that whatever effect they do have is likely smaller in magnitude than the effects of the random choices built into SARGE.

Setting a distance for geographically restricted introgressed haplotypes

In order to better understand how Neanderthal and Denisovan admixture might have affected different human populations differently, we sought to define whether each introgressed haplotype block was broadly shared by many different human populations or limited to specific geographic regions. Because the SGDP population labels we used (Africa, Africa-MBK, America, CentralAsiaSiberia, EastAsia, Oceania, SouthAsia, and WestEurasia) are very broad, we decided to use geodesic distances between reported sampling coordinates of each individual instead. For each introgressed haplotype, we computed a maximum pairwise distance (in km) between each pair of genomes that possessed that haplotype. We then chose 3,000 km as a cutoff below which all introgressed haplotypes were considered geographically restricted, and above which all introgressed haplotypes were considered geographically widespread.

To test whether 3,000 km was a reasonable cutoff, we also applied cutoffs of 1,000, 2,000, and 10,000 km (as well as no cutoff, treating all introgressed haplotypes as geographically restricted). For each cutoff, we then plotted the distribution of TMRCAs to admixers and haplotype block lengths across all modern human genomes in geographically restricted introgressed haplotypes. We also used the Wilcoxon rank-sum test (wilcox.test in R) to compare each of these distributions to that created using the 3,000 km cutoff. Considering TMRCAs to admixers, neither the 1,000 or 2,000 km cutoffs were significantly different from the 3,000 km cutoff, using a significance threshold of 0.001. 1,000, 2,000, and 3,000 km cutoffs were all significantly different from the 10,000 km cutoff and no cutoff, however (Fig. S 45). Considering lengths, every cutoff (and no cutoff) was significantly different from the 3,000 km cutoff for Neanderthal haplotype blocks. For Denisovan haplotype block lengths, however, neither the 1,000 nor 2,000 km cutoff significantly differed from the 3,000 km cutoff, while the 10,000 km cutoff and no cutoff both did (Fig. S 46).

From this exercise, we deduced that a 3,000 km cutoff produced results that did not significantly differ from those produced using 1,000 or 2,000 km cutoffs, but that did differ from

a 10,000 km cutoff. As 3,000 km is large – over 1/3 the width of Eurasia – we took it to be a cutoff that would still allow for some widespread sharing of introgressed haplotype blocks and avoid discarding too many archaic haplotypes and causing sampling error in some genomes.

Results of coalescent simulations

We ran a series of coalescent simulations against which to compare the amount of the human genome we found to contain admixture with archaic hominins, ILS with archaic hominins, regions free of both (deserts) and deserts containing fixed human-specific derived alleles (human-specific regions) (Supplementary Methods). We did not seek to model all structure within modern human populations or widely-hypothesized selection against weakly deleterious archaic hominin alleles (*47*). Rather, we used a simple, three-population model of human history (*59*), which included Africans, Europeans, and Asians, but without migration to and from Africa, in order to have an unadmixed outgroup for ascertaining archaic hominin admixture. Our goal was to obtain a null model for the relative extents of regions of admixture, ILS, deserts, and human-specific regions throughout the genome. We modeled one pulse of Neanderthal admixture into the ancestors of all non-Africans, followed by a later pulse of Denisovan admixture into Asians. We repeated our simulations with increasing admixture proportions, in order to find the best-fit admixture proportion for each observation in our real data. Ultimately, comparing these simulations with our real data suggests that we have underestimated the amount of admixture relative to ILS in our real data, and that there were probably numerous population-specific archaic admixture events aside from those we modeled.

In the absence of admixture, the entire genome is separable into regions of ILS and deserts. Selection aside, the extent of ILS in our simulation with no admixture depends only on fairly well-understood parameters, and is not affected by details omitted from the model such as population structure in modern humans. We find ILS to cover 37% of the genome and deserts to cover 63% in our simulation with no admixture (*Fig. S 38A*).

For the closest possible comparison with real data, we computed the extent of desert regions in our outgroup Africa-MBK population, in which admixture was minimal (covering 5% of the genome). In this population, ILS covers 64% of the genome and deserts cover only 31% of the genome (*Fig. S 38A*). Assuming that human speciation involved selection for uniquely human alleles and against alleles shared with archaic hominins, our simulation with no admixture should place more of the genome in regions of ILS and less of it in desert regions than in real data, meaning that the extent of deserts in this simulation should be considered a lower bound. The presence of fewer deserts in real data than in this simulation therefore suggests either that some regions in which we detect ILS haplotypes actually contain admixture from archaic hominins, that selection actually worked to decrease rather than increase the extent of desert regions, or that another more fundamental model parameter, such as the split time between modern humans and archaic hominins, was incorrect.

Because SARGE produces slightly shorter haplotypes on average than the true ARG, using simulated data (*Fig. S 14*) and because this problem is likely exacerbated in real data due to uneven mutation and recombination rates across chromosomes, it is likely that we have underestimated admixture and incorrectly labeled some of it as ILS. This is supported by our finding lower population-average archaic ancestry proportions than the D-statistic-based estimator (*Fig. 3A*). Another reason we may have under-estimated archaic ancestry in some populations is that we require admixed individuals to remain in a clade with at least one admixer

haplotype for the full extent of the admixed haplotypes. In the event that there is admixture but the introgressor is highly diverged from the archaic hominin genomes in our panel, we would be likely to break such admixed haplotypes into erroneously small pieces. This, in turn, would make them more likely to be labeled as ILS rather than admixture (Supplementary Methods, *Fig. S16*). The fact that we discover more ILS (43%) in our full panel than in any simulation, including our admixture-free simulation (37%) supports this.

As we increased the amount of admixture in our simulations, we expected admixture to overwrite ILS and desert regions equally often, due to the absence of selection. In real data, however, we expect selection to remove admixed alleles from desert regions more often than from ILS regions, resulting in more extensive deserts and less extensive ILS relative to simulated data. Working from this assumption, the amount of admixture included in the simulation that most resembles our data (in terms of the extent of deserts versus admixture and ILS across the genome) should be lower than the amount of admixture that occurred in reality.

We plotted the extent of admixture + ILS versus deserts + human-specific regions across all simulations, along with those values computed from real data and found that the real data values correspond to simulations with 18.2% admixture proportions (*Fig. S 38B*). Such simulations produced unrealistically high individual percent archaic ancestry estimates, however (*Fig. S 38C*). In real data, the existence of population structure can help explain this discrepancy in two ways.

First, population structure could increase the power of drift to randomly eliminate some, and increase the frequency of other, archaic hominin haplotypes in individual human populations. This could result in individual populations each maintaining small numbers of archaic haplotypes from a shared, ancestral admixture event, each covering different parts of the genome.

Second, later admixture events involving small, isolated populations would increase the total amount of the human genome containing admixture without contributing to the overall percent archaic hominin ancestry in individuals that do not belong to those populations. If this happened, migrants from these populations could later contribute archaic hominin haplotypes to other populations, which would then be widely shared within those populations and exist at relatively high frequency in our panel; this is what we see, for example, in the case of Denisovan haplotypes in West Eurasians (*Fig. S23*, *Fig. S 28B*). We also find further evidence of population-specific admixture events in the presence of geographically restricted Neanderthal and Denisovan haplotype blocks in our real data set, which have different distributions of haplotype block lengths and TMRCAs to admixer than geographically widespread archaic hominin haplotype blocks (*Fig. 3B,D*; *Fig. S 17A,D*).

If we also allow for the possibility of more admixture events than the two we modeled, then each could have had a lower admixture proportion than those in our model. This seems likely, given the presence of many geographically restricted archaic hominin haplotype blocks (*Fig. 3C*, *Fig. S 17B-C*, *Fig. S 18B-C*), the existence of mysterious Neanderthal and Denisovan-like haplotype blocks detected in sub-Saharan Africans but unlike those detected in non-Africans (*Fig. 3B*, *Fig. S 17B*), findings of prior studies (*9*, *22*, *78*), and the evidence that we may have mis-labeled some ILS as admixture, even in the deeply-divergent Africa-MBK lineages thought to be free of Neanderthal and Denisovan ancestry.

Timing and functional consequences of human-specific mutations

After identifying regions of the human genome free of incomplete lineage sorting and admixture with archaic hominins in all modern human genomes sampled, we identified fixed or nearly-fixed human-specific derived mutations within these regions (Supplementary Methods). Using the chimpanzee genome as an outgroup and assuming 6.5 mya human-chimpanzee divergence (*31*), we computed the TMRCA in years of each human-specific clade and its parent. We then took the mean of these two numbers (the midpoint of the branch containing all derived alleles specific to and shared by all modern humans) to be the age of each human-specific mutation. We then compiled all such mutations that either created nonsynonymous substitutions relative to the Neanderthal and Denisovan genomes (*32*) or fell within an annotated binding site for a regulatory element known to target specific genes (*64*) (Supplementary Methods). We then performed several analyses to determine whether particular biological processes or tissues were predominately affected by mutations that occurred at distinct points in time.

With our list of affected genes and approximate ages of mutations affecting them, we compared ages of mutations affecting interacting pairs of genes, according to the STRING database (*33*) (Supplementary Methods) and found two distinct bursts of mutations affecting interacting sets of genes, at approximately 100 kya and 300 kya (*Fig. 5B*). We then performed Gene Ontology (GO) enrichment analyses on the sets of genes affected by mutations 300-350 ky old, and affected by mutations 100-150 ky old, using FUNC (*68*) to determine whether different biological processes were affected by mutations at the different time points. We found a variety of biological process terms to be enriched in the gene sets at both time points (*Table S 5*); we were therefore unable to identify any specific biological process as the main target of selection at either time point.

We next sought to identify whether any tissues were predominantly affected by mutations clustered together in time, or whether the same tissues tended to be acted on repeatedly by different mutations over time. To this end, we computed the tissue specificity of each gene in our set by calculating tau (*67*) from median tissue-specific expression values across many samples in the GTEx database (*66*), excluding cell line data. We then compared the ages of all mutations affecting high-tau (> 0.9) genes specific to the same tissues. After normalizing counts of mutations affecting specific tissues at specific time points to account for the overall number of genes specific to each tissue (Supplementary Methods), we found that most tissues were acted on repeatedly by mutations over time (*Fig. S 39A*).

There are several exceptions to this pattern, however. For example, the two tissues acted on most recently but not affected by older mutations are the frontal cortex and basal ganglia, with mutations 100-200 ky old (*Fig. S 39A*). The genes affected by these mutations are CREG2 and SLC32A1. CREG2 has little known about its function. SLC32A1 codes for a transmembrane protein that transports the inhibitory neurotransmitter GABA into synaptic vesicles (*79*), with a nonsynonymous substitution in one of its intra-vesicle, lumenal domains (*32*).

In addition, most mutations affecting brain-specific genes happened after 300 kya (*Fig. S 39A*), coincident with a peak in changes to developmentally-relevant genes (*Fig. S 39B*) and postdating the age of human remains discovered to have some modern features coupled with archaic cranial morphology (*36*).

We then sought to identify whether broad functional categories of genes (genes involved in developmental, immune system, metabolic, and/or reproductive processes as annotated in the Gene Ontology database (*39*)) were affected by mutations occurring at specific points in time

(Supplementary Methods). When considering relative numbers of mutations affecting development, immunity, metabolism, and reproduction, we find an uptick in metabolic changes beginning 700 kya; the rate of accumulation of such changes was consistent until 400 kya, when it accelerated (*Fig. S 39B*). In contrast, developmental changes took an extra 100 kya to begin to rapidly accumulate (*Fig. S 39B*). Although our data are low-resolution, this could imply that changes in diet or energy usage were important in the very early development of our species.

Prioritizing selected human-specific derived mutations

Starting with our list of human-specific derived mutations within deserts, limited to those that either caused a nonsynonymous substitution relative to archaic hominins or fell within an annotated binding site for a regulatory element believed to affect specific genes (Supplementary Methods), we sought to prioritize these mutations by the strength of evidence that they were acted on by selection. To this end, we created a score for each mutation based on its age (where available) and the length of the surrounding desert region (Supplementary Methods). We expected mutations targeted by selection to have a recent age (both because they arose recently in time and should have reduced haplotype diversity in the event of either positive or purifying selection) and a long surrounding desert region (because recombination has not yet had time to break down such haplotypes into smaller pieces). We ranked genes by scores of mutations affecting them (*Fig. 5C*), in order to identify genes and biological processes acted upon by selection since the split between modern humans and archaic hominins. According to a Wilcoxon rank-sum test using these scores, many biological processes appear to have been affected by such selection (*Table S 6*).

In addition to genes involved in mRNA splicing and brain development (main text), we find a number of high-scoring regulatory mutations affecting genes involved in histone acetylation (*Table S 6*), which suggests another way that a small number of mutations could lead to large-scale changes in gene expression that could in turn cause phenotypic differences. These include four mutations affecting KAT7 and KAT8, three affecting ASH1L, three affecting ING4, and one affecting SETD2.

We find several mutations affecting centrosomal and/or mitotic spindle or kinetochore-associated genes that may be involved in neural cell proliferation, beyond those mentioned in the main text. We find three high-scoring regulatory mutations affecting ARHGEF2, which is a Rho GTPase activator involved in cell division and cell migration localized to the mitotic spindle, whose mutation has been linked to microcephaly and other brain development disorders (*80*). All mutations affecting genes localized to the centrosome or kinetochore are in the top 50^{th} percentile of our score distribution; these include nonsynonymous mutations affecting ALMS1, KATNA1, KIF18A, RABL6, and SPAG5. We also find centrosomal genes affected by multiple candidate regulatory mutations, including 10 mutations affecting the nucleoporin NUP62, which maintains centrosomes and is required for successful mitotic division (*81*), 9 affecting MAP3K11, which influences microtubule organization (*82*), and 8 affecting ATF5, which is a cancer drug target, suggesting it can affect cell proliferation (*83*).

In addition to mutations affecting axon pathfinding-related genes mentioned in the main text, we find a moderately high-scoring potential regulatory mutation affecting NTM (neurotrimin), a brain-specific gene also thought to contribute to cell-surface protein diversity involved in axon guidance (*84*), and two regulatory mutations affecting NAV1, a gene believed to be involved in axon guidance due to homology with a similar *C. elegans* gene (*85*).

Control of the migration of neural cells might also play a factor in human-specific developmental changes. We find a regulatory element mutation affecting the transcription factor NIPBL to be in the top 99.5th percentile of the score distribution; aside from many other functions, this gene is involved in brain development and required for proper cortical neuron migration (*86*).

Some other top-scoring mutations may be involved in control of toxic substances in the brain. One top-scoring regulatory mutation affects the ferritin heavy chain gene FTH1, which can sequester iron, an element whose over-accumulation in brain tissue can cause neurodegenerative disease (*87*). Amyloid-beta, which aggregates to form pathogenic plaques involved in Alzheimer's disease, is another regulated substance that is cleared from the brain during sleep (*88*). Two of the highest-scoring regulatory mutations affect amyloid-beta binding proteins CST3, which can inhibit amyloid-beta aggregation (*89*), and APOE, which also influences amyloid-beta accumulation in Alzheimer's disease and may play neurodevelopmental roles as well (*90*).

There is evidence that amyloid-beta, when properly regulated, helps control synaptic activity by regulating other genes (*91*); the Gene Ontology term "response to amyloid-beta" is enriched in interacting sets of genes affected by mutations that arose in the burst 300-350 kya (*Table S 5*). This is due to five regulatory mutations affecting FYN, a tyrosine kinase activated by amyloid-beta that influences synaptic plasticity through NMDA receptor phosphorylation, among other neurological functions (*92*); its interaction partner GRIN1, an NMDA receptor subunit, also has four regulatory mutations, the two of which we could date arose around the same time.

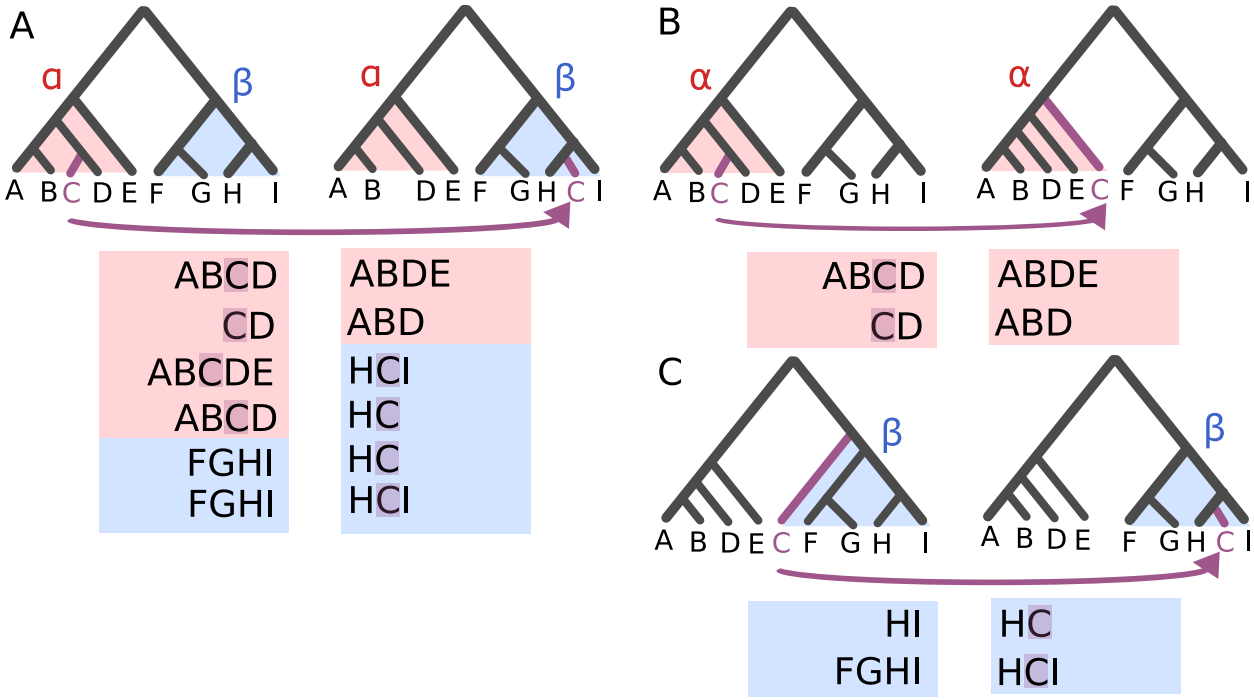

Fig. S1.

Different types of four haplotype test failures. In each, the γ clade is highlighted in purple, α in red, and β in blue. A: Lateral branch movement. Four haplotype test failures of type α/ α, α/ β, and β/ β are observed. B: Upward branch movement. Only α/ α four haplotype test failures are observed. C: Downward branch movement. Only β/ β four haplotype test failures are observed.

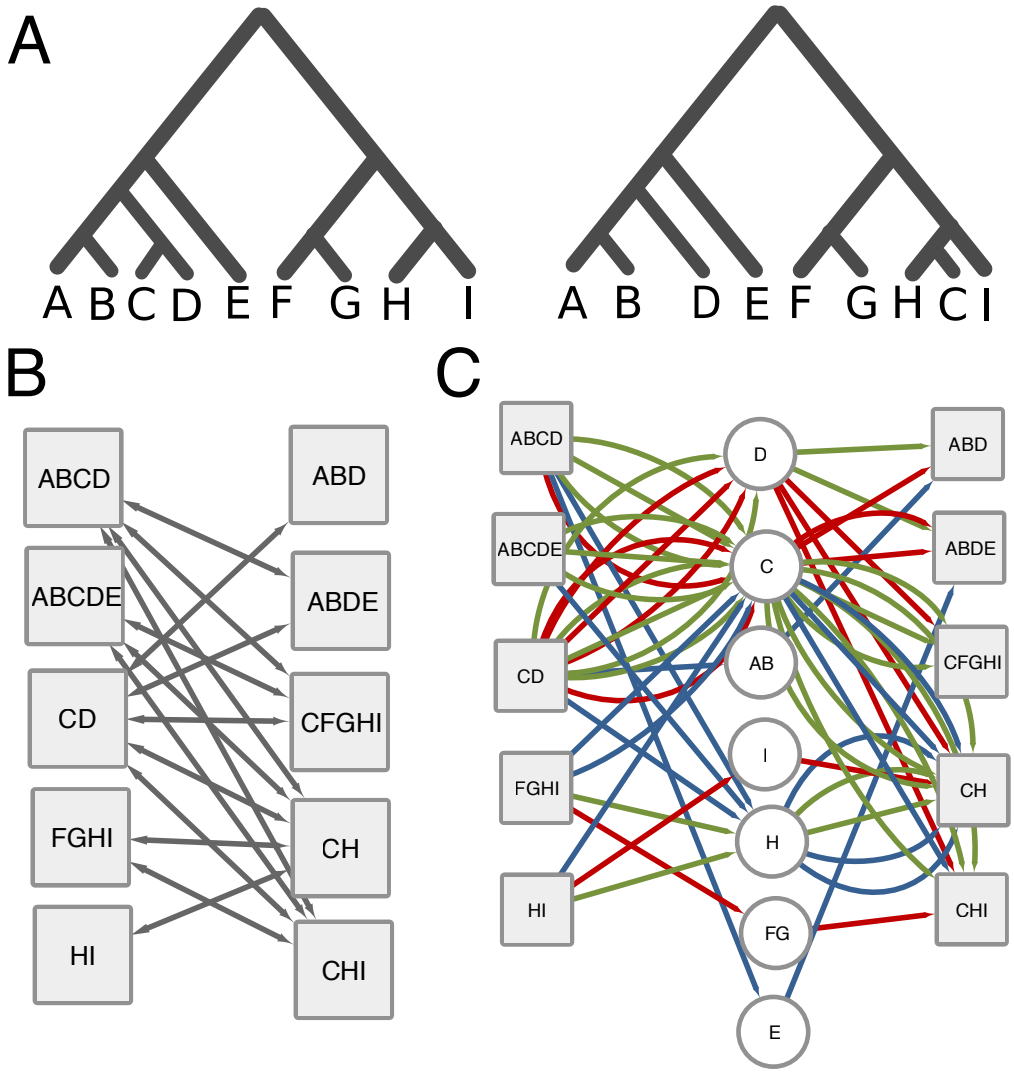

Fig. S2.

Example of algorithm for inferring branch movements between to trees known *a priori*. A: Two trees, which differ by one branch movement. B: Clades from the two trees that fail the four haplotype test. Left column shows clades from the first (upstream) tree and right column shows clades from the second (downstream) tree; arrows indicate four haplotype test failures. C: Graph showing all possible branch movements that could explain the four haplotype test failures shown in B. The left and right columns are "tree" nodes, while the center column lists candidate γ clades. Colors indicate types of four haplotype test failures: red paths are conditional on a failure being the α/ α type, green on it being α/ β, and blue on it being β/ β. In this case, a single candidate γ clade (C) has the most edges and can explain all four haplotype test failures. This is interpreted as the clade C moving from the smallest observed α clade in the first tree (CD) to the smallest observed β clade in the second tree (CH). If no β clades from the second tree are observed, the branch movement goes upward to a clade containing the union of all clades failing the four haplotype test. If no α clades from the first tree are observed, the branch movement goes downward from a clade containing the union of all clades failing the four haplotype test.

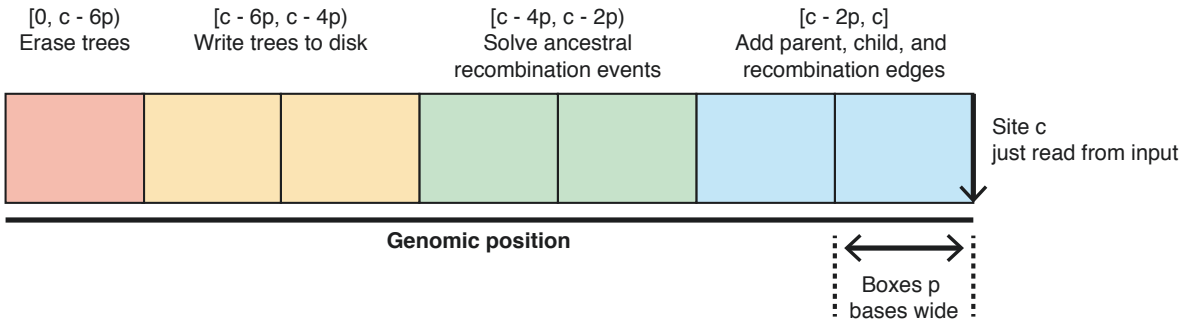

Fig. S 3

How the propagation distance parameter affects the operations SARGE performs on ARG nodes (clades). SARGE does not store all nodes across an entire chromosome in memory at once; once sites that define nodes are sufficiently far away from the most recently-observed site as to not be affected by it, they can be written to disk and erased. In this figure, c is the position (in reference genome coordinates) of the most recently-read site, and p is the propagation distance (in base pairs). Clades within 2p bases of one another can "communicate" with one another and form parent/child and recombination edges.

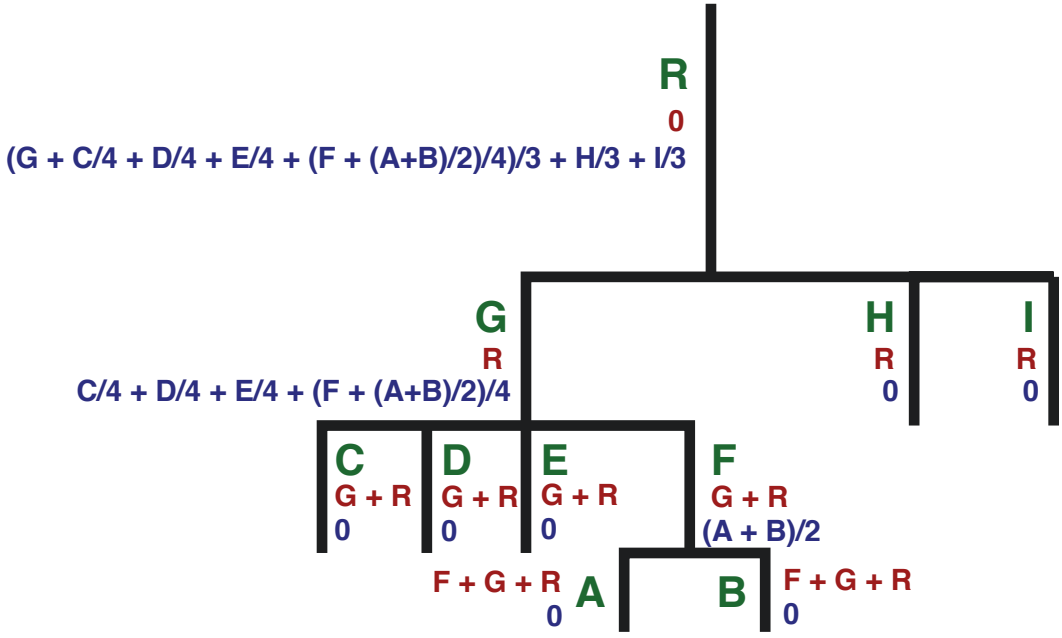

Fig. S 4

Calculation of branch lengths. Each non-root branch (A-I) stores the number of mutations that tag its child clade, as well as the genomic interval over which it exists. Each green letter represents a branch length, calculated as the number of mutations on the branch divided by the genomic interval over which those mutations were observed. In the case of the root branch (R), this genomic interval will always be two times the propagation distance, since the root clade (consisting of all haplotypes in the data set) cannot be affected by ancestral recombination events. The red values below the green values are the sum of all branch lengths above each branch, and the blue values are the sum of all branch lengths below each branch. The scaled branch lengths used in this study are then the individual branch lengths divided by the height of the tree at those branches, or the green values divided by the sum of the red and blue values.

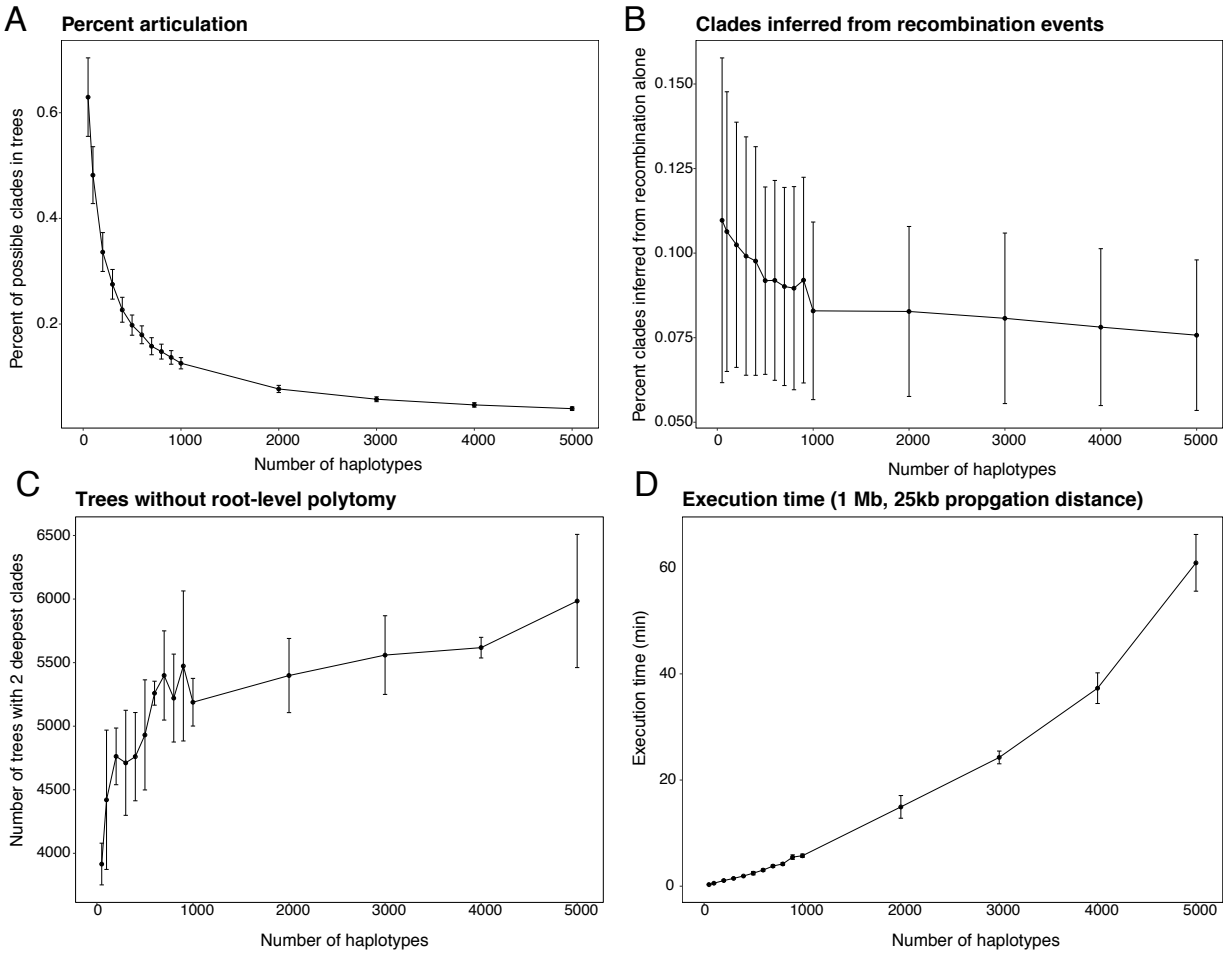

Fig. S5.

Properties of SARGE performance on simulated data with a sub-Saharan African-like level of heterozygosity, constant population size history, and no structure. Points are means; error bars show one standard deviation. A: Tree articulation as a percent of all nodes possible (given the number of haplotypes), with increasing number of haplotypes. B: Percent of all clades (across all trees) inferred from solving recombination events (rather than shared mutations). C: Number of trees across the chromosome with two children of the root node (no root-level polytomies). D: Execution time as a function of the number of input haplotypes. Real data, where SNPs and recombination events cluster in the genome, is likely to increase execution time.

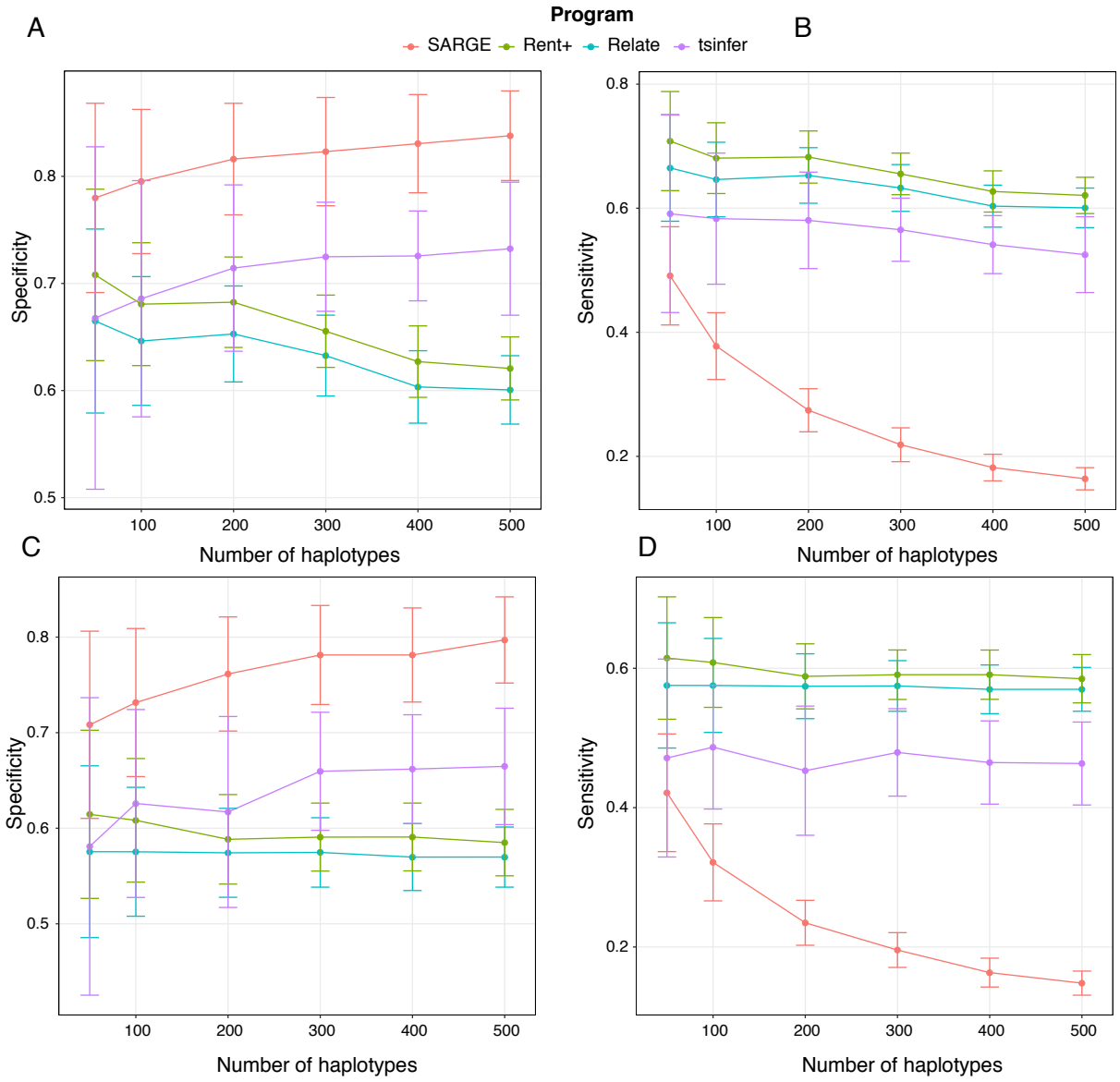

Fig. S6

Comparisons of SARGE to three other ARG inference programs (*5*, *7*, *8*), using a simulated data with sub-Saharan African-like heterozosity, constant population size, and no structure. In each comparison, error bars represent one standard deviation across 5 replicates and ARGs were inferred across data sets with increasing numbers of haplotypes. A and B used a 2.5 x 10^{-8} mutation rate per site per generation; C and D used a 1.0 x 10^{-8} mutation rate per site per generation. A and C show specificity, defined as the percent of nodes in an inferred ARG that were correct according to the true ARG. B and D show sensitivity, defined as the percent of nodes in the true ARG that were present in the inferred ARG. Sensitivity and specificity are equal for the two methods (Rent+ and Relate) that produce fully articulated trees (without polytomies).

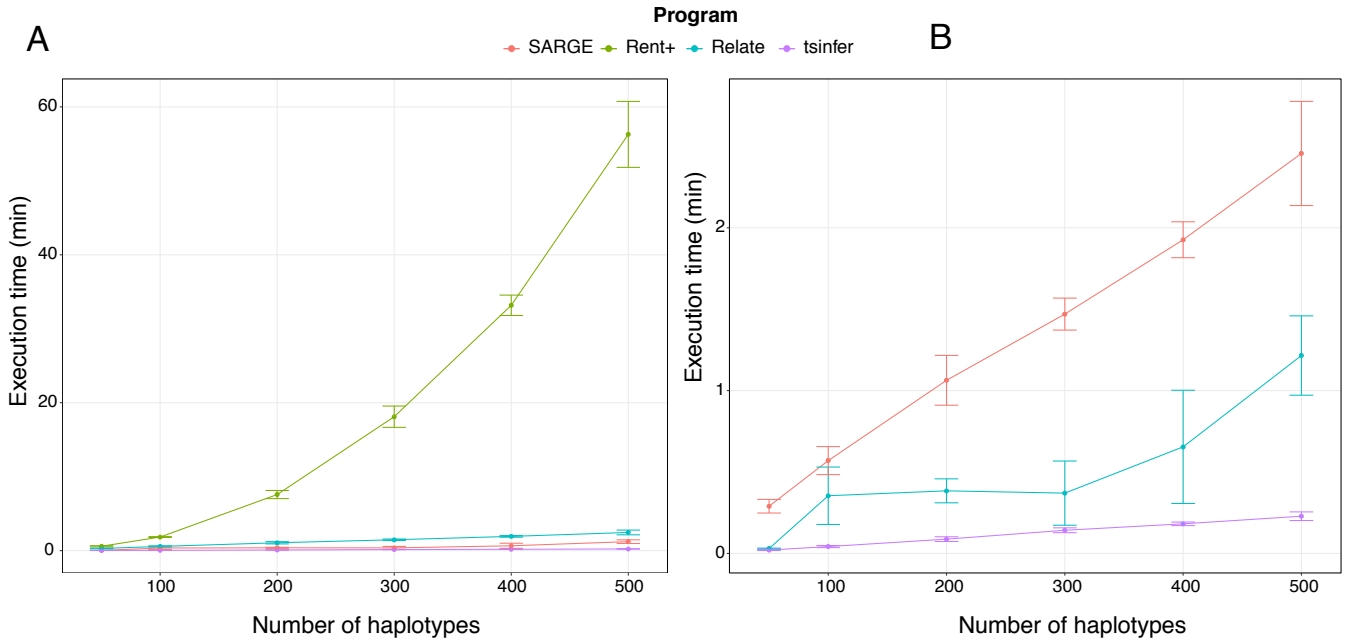

Fig. S 7

Execution time for SARGE and three other tested ARG inference programs (*5*, *7*, *8*), using a simulated data set with sub-Saharan African-like heterozygosity, constant population size, and no structure, with a 2.5 x 10^{-8} mutation rate per site per generation. Because tsinfer is a Python module, its execution time does not include file I/O. A: Showing all programs. B: Omitting Rent+.

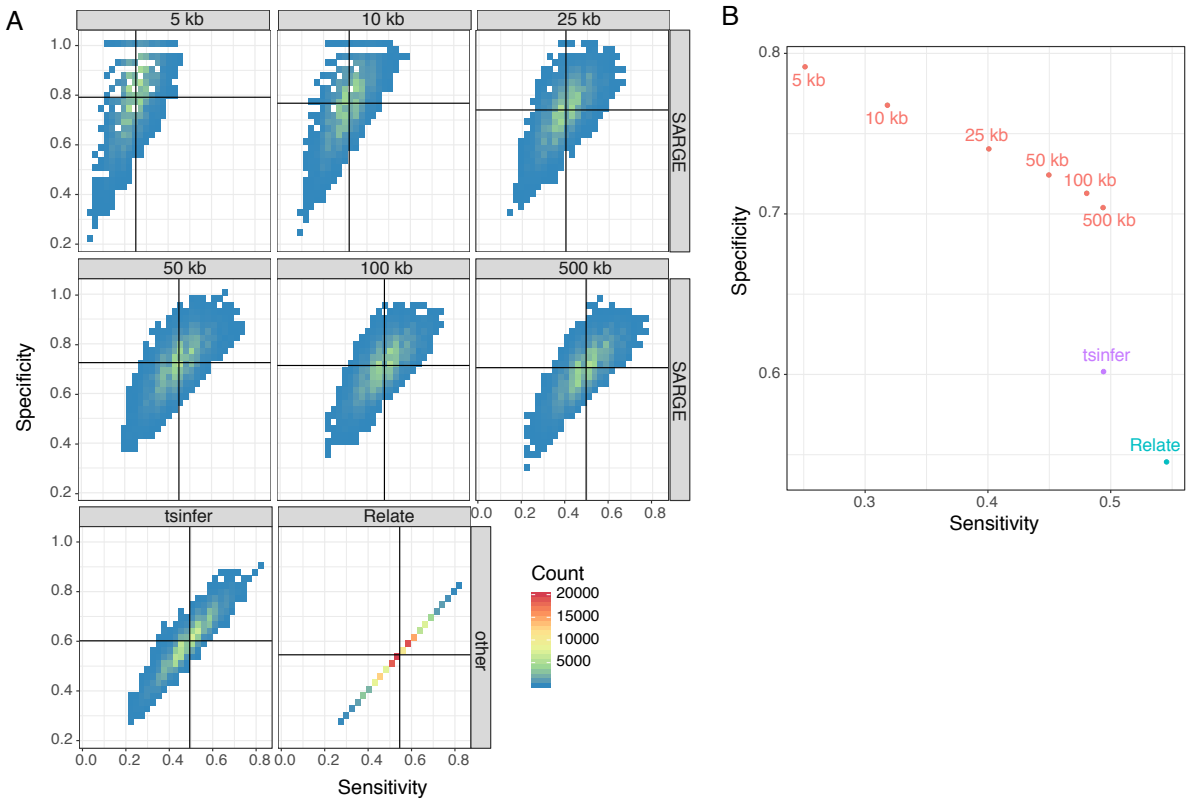

Fig. S 8

Effect of propagation distance parameter on SARGE sensitivity and specificity on simulated data. A: Using a simulated data set of human and archaic genomes (see Supplementary Methods, "Demographic simulations" section), specificity (percent clades from inferred ARG present in true trees) on y-axis versus sensitivity (percent true trees recovered correctly by inferred ARG) on x-axis. Numbers at the top of boxes are SARGE propagation distances (in kb) or other programs (tsinfer or Relate) used to infer ARGs (*8*) instead of SARGE. Relate shows a unique pattern because it, like the simulation, produces fully articulated trees (without polytomies); sensitivity and specificity are both therefore always equal (the denominators – the number of inferred clades and the number of true clades – are identical). Horizontal lines are mean specificity across the entire ARG and vertical lines are mean sensitivity across the entire ARG. B: Mean sensitivity and specificity values (same as horizontal and vertical lines in A) across SARGE runs with different propagation distances, compared to tsinfer and Relate. Colors show programs used (green = SARGE, blue = tsinfer, red = Relate).

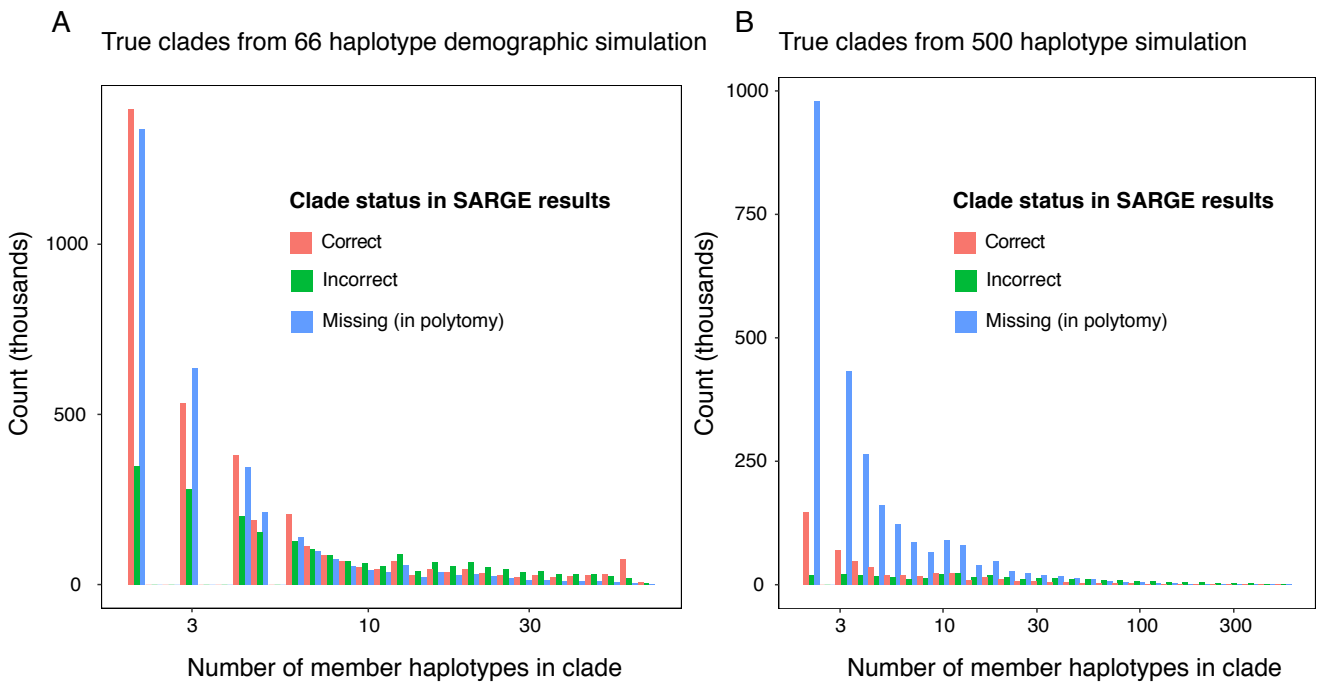

Fig. S 9

A: SARGE was run on a simulated data set of 66 human and archaic hominin genomes (Supplementary Methods), and each clade in each tree in the true (simulation) ARG was searched for in the SARGE-inferred ARG. Clades present in the SARGE results (red), clades incompatible with (failing the four gamete test with) clades at the same site in the SARGE results (green), and clades missing due to a polytomy in SARGE results (blue) are shown. The x-axis shows the size of the clades. B: The same as A, but using a larger (500-haplotype) simulation with no population structure ("QC simulations" in Supplementary Methods).

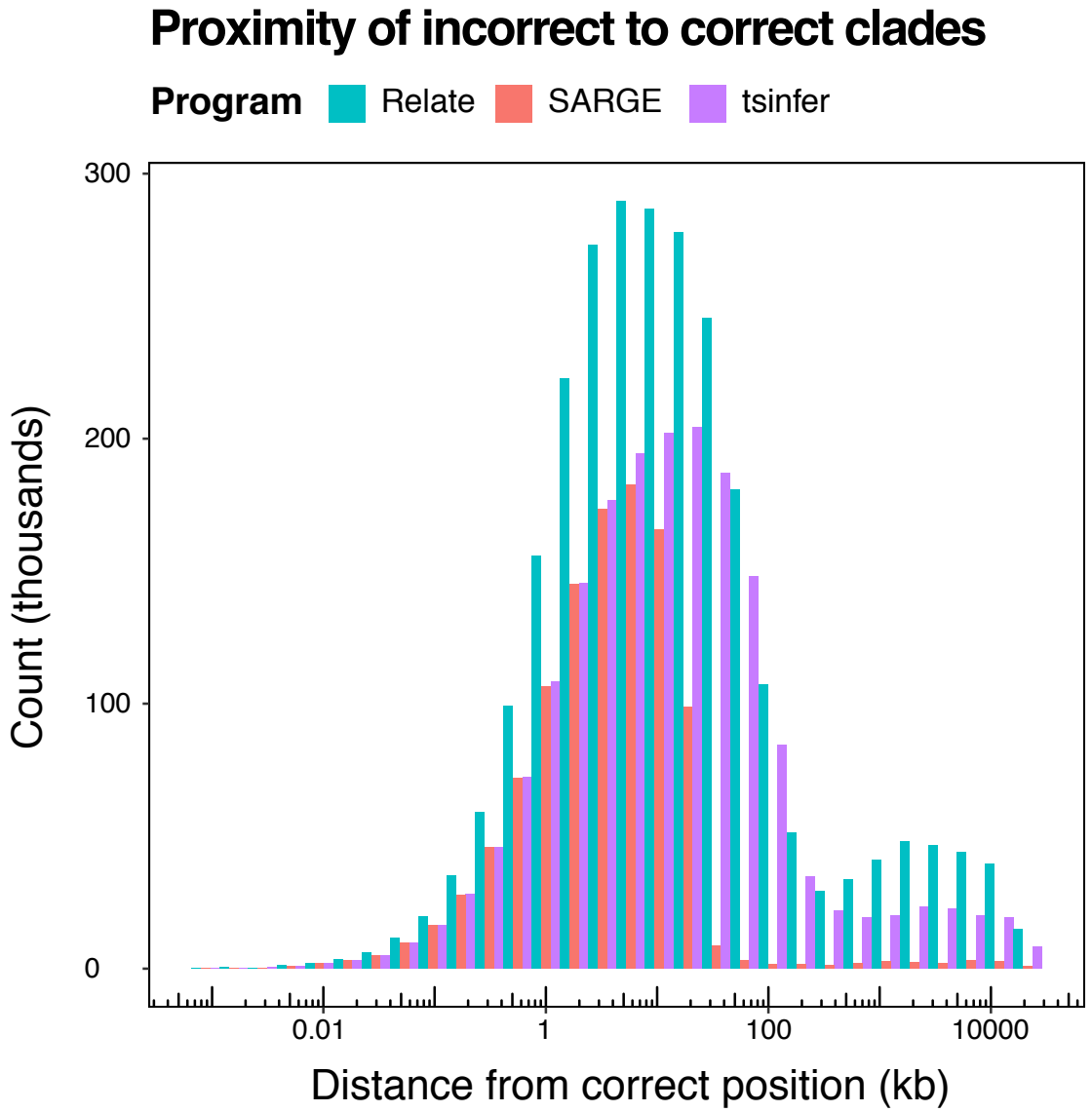

Fig. S 10

Using a simulated data set approximating humans and archaic hominins (see Supplementary Methods), ARGs were inferred using three different programs, and incorrect clades (clades in inferred ARG trees that do not exist in the true simulation trees) were considered. For each incorrect clade, the minimum distance on the chromosome to a position where that clade is correct (in the true simulation ARG) is shown.

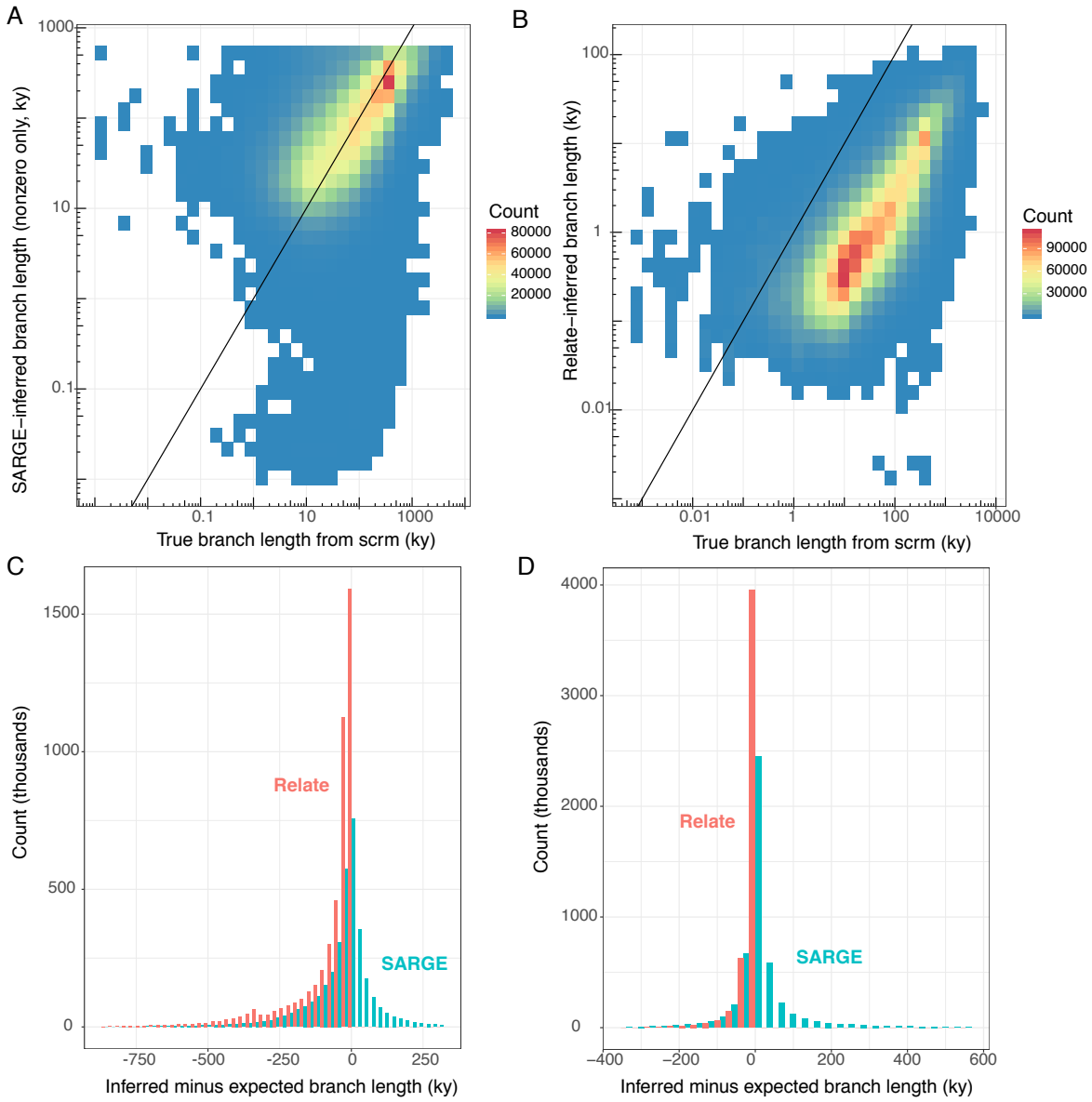

Fig. S 11

Accuracy of inferred branch lengths, using simulated data. Data are from a 25 Mb simulated data set using a demographic model of human and archaic hominin evolution, including archaic hominin admixture (see Supplementary Methods). A: For each correct clade inferred by SARGE, its true branch length (x axis) is compared to its inferred branch length (y axis). The line shows the expectation if inferred branch lengths always matched true branch lengths. B: Same as A, but for the ARG inferred by Relate (*7*). C: Inferred minus true branch lengths for both SARGE and Relate. To help with readability, only the 1st to 99th percentile of both distributions is shown. D: Same as C, but using a larger simulated data set with the same parameters, but 450 modern human haplotypes and 6 archaic hominin haplotypes. Because Relate trees contain fewer polytomies and thus have many more branch lengths, only a sample of 5 million branch length differences was selected from each ARG to plot.

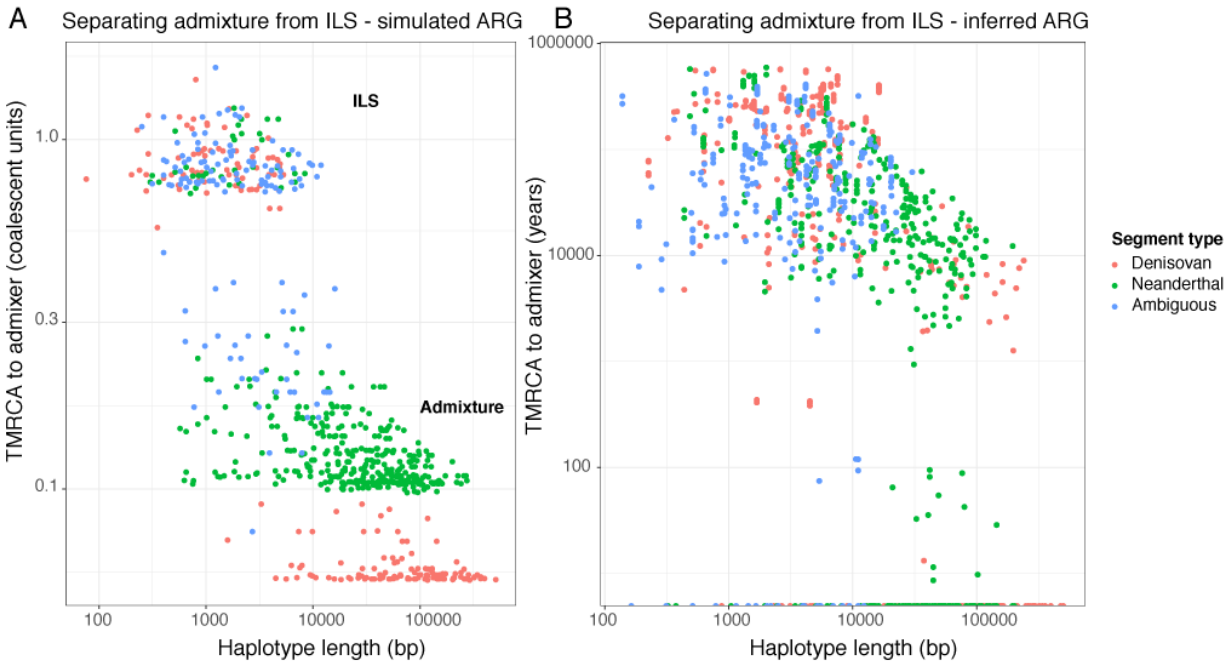

Fig. S 12

Using a simulated data set in which humans received one pulse of Neanderthal admixture 50 kya and one pulse of Deniosvan admixture 20 kya (Supplementary Methods), all clades grouping some human haplotypes with archaic hominin haplotypes, to the exclusion of other human haplotypes, were selected from the true simulation ARG (A) and the ARG inferred from SARGE (B). Each point is one such clade, colored by the type of archaic hominin genomes it contains (ambiguous means that human haplotypes are equally related to Neanderthal and Denisovan genomes within the clade). The persistence of each clade along the genome (x-axis) and mean TMRCA between archaic and human haplotypes within the clade (y-axis) are shown. In B, many such clades had a TMRCA to admixer of 0 years and are shown along the bottom of the panel.

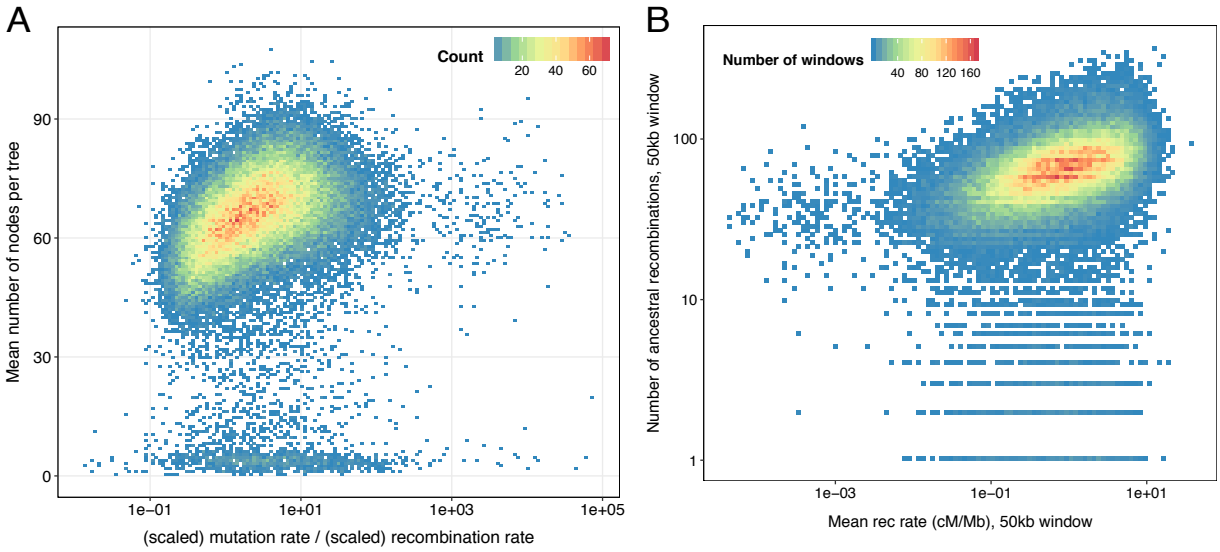

Fig. S13

How mutation and recombination rate variation affects ARG trees. The ARG was inferred on Simons Genome Diversity Project data (*14*) with three archaic hominin genomes (*1*, *15*, *16*) included. A: In 50kb genomic windows, mean tree articulation (number of nodes per tree y-axis) versus mutation rate to recombination rate ratio within the window (x-axis). B: Number of inferred ancestral recombination events per 50kb genomic window (y-axis) vs. mean population recombination rate (cM/Mb; x-axis). Data used Simons Genome Diversity Project data (*14*) with three archaic hominin genomes included. The two numbers are correlated (Spearman's rho = 0.46; $p < 2.2e\text{-}16$).

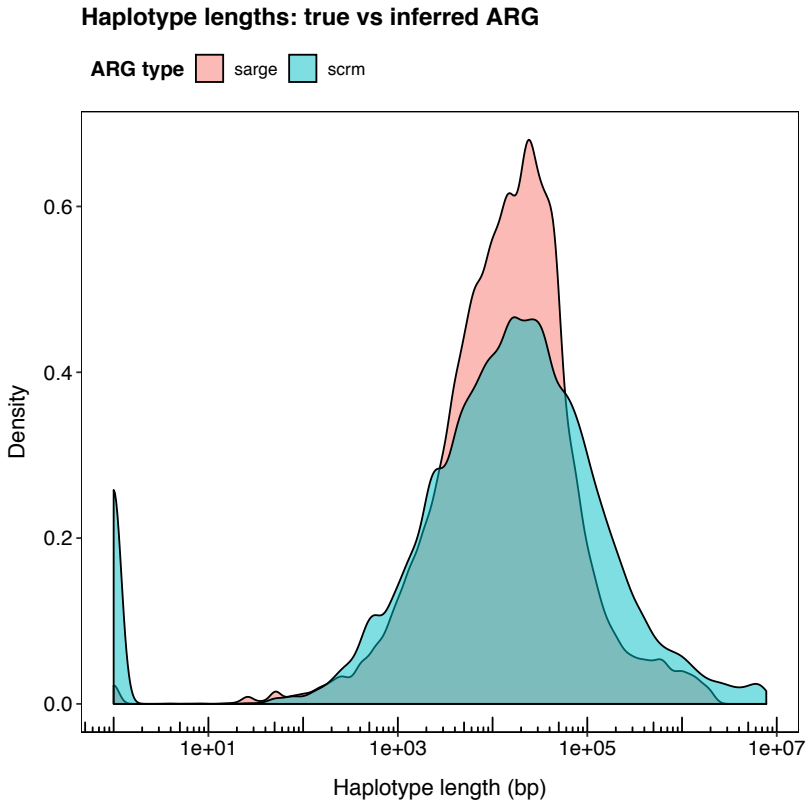

Fig. S 14

Persistence of clades (haplotype block lengths) in a true vs. inferred ARG. The true ARG and input data for the inferred ARG are from a scrm (*58*) simulation with Neanderthal and Denisovan admixture proportions of 0.05 (Supplementary Methods). Shown are haplotype block lengths for 100,000 clades randomly sampled from each ARG. The difference in haplotype block length distribution is owed to SARGE artificially breaking long haplotypes: while the mean true haplotype block length is 2.46x the mean inferred haplotype block length, the median true haplotype block length is only 1.04x the median inferred haplotype block length. In real data, haplotype block lengths can be shorter still due to variation in mutation and recombination rates across chromosomes, as well as phasing errors in input data, which will introduce artificial ancestral recombination events that can break haplotypes.

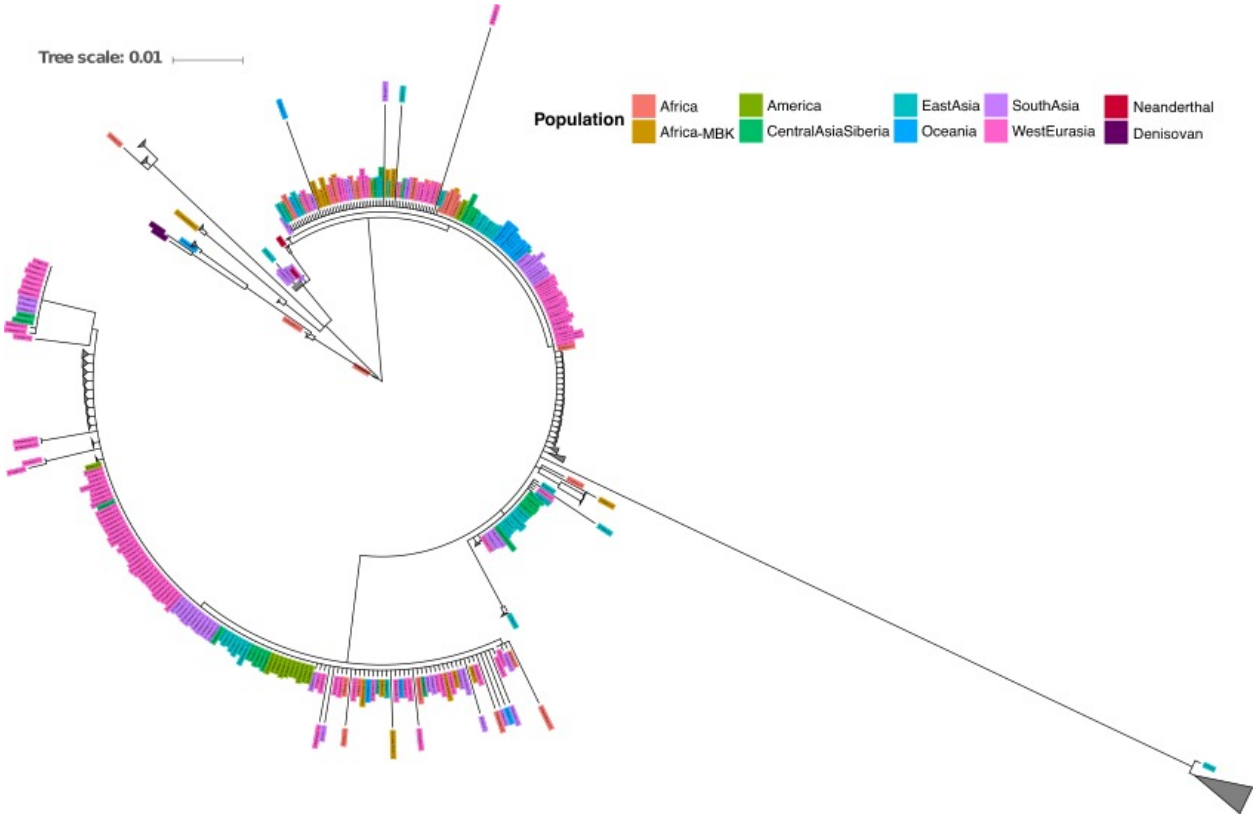

Fig. S15

Example of a typical tree in our human data set. Shown is the tree for chromosome 1, position 4870939 (hg19 coordinates), which contains 90 nodes (the average for all trees genome-wide). Branch lengths are measured as a percent of human-chimpanzee divergence (assumed to be 13 million years). Since clades inferred from recombination alone do not receive a branch length, a pseudocount of 0.0008 (corresponding to about 10 thousand years) was added to each branch length. All clades whose average distance to leaves is less than 0.005 (approx. 65 ky) are collapsed, shown as triangles with size proportional to the number of leaves contained within. The tree was generated using ITOL (*93*); all populations shown are from the Simons Genome Diversity Panel (*14*) plus Neanderthal (*1*, *15*) and Denisovan (*16*)genomes. Africa-MBK consists of the most basal African lineages (Khomani-San, Mbuti, and Biaka).

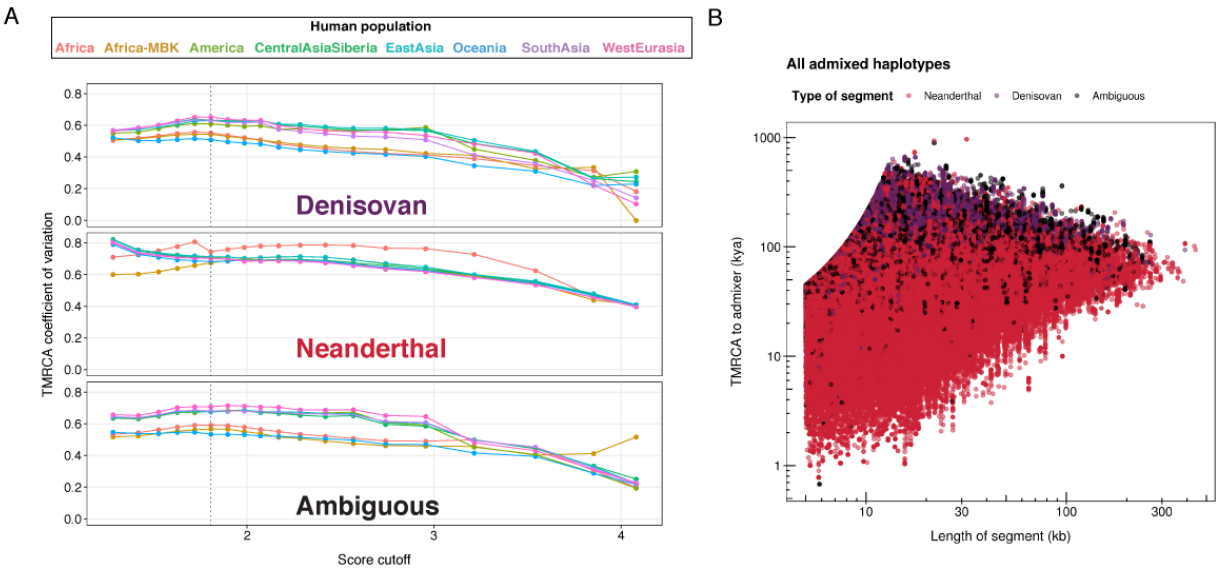

Fig. S16

Separating admixed haplotypes from incomplete lineage sorting (ILS). A: After selecting candidate admixed haplotypes and excluding any that contained more than 10% outgroup haplotypes (from the most basal sub-Saharan lineages, here referred to as Africa-MBK), we assigned each haplotype a score, which increased with both haplotype block length and low time to most recent common ancestor (TMRCA) with the admixer. For different score cutoffs, we calculated the coefficient of variation (standard deviation divided by mean) of the TMRCA to admixer within each population and chose the score at which this value began to stabilize for each type of admixture as the cutoff (vertical line). B: TMRCA to admixer and haplotype block length of all admixed haplotypes passing the cutoff.

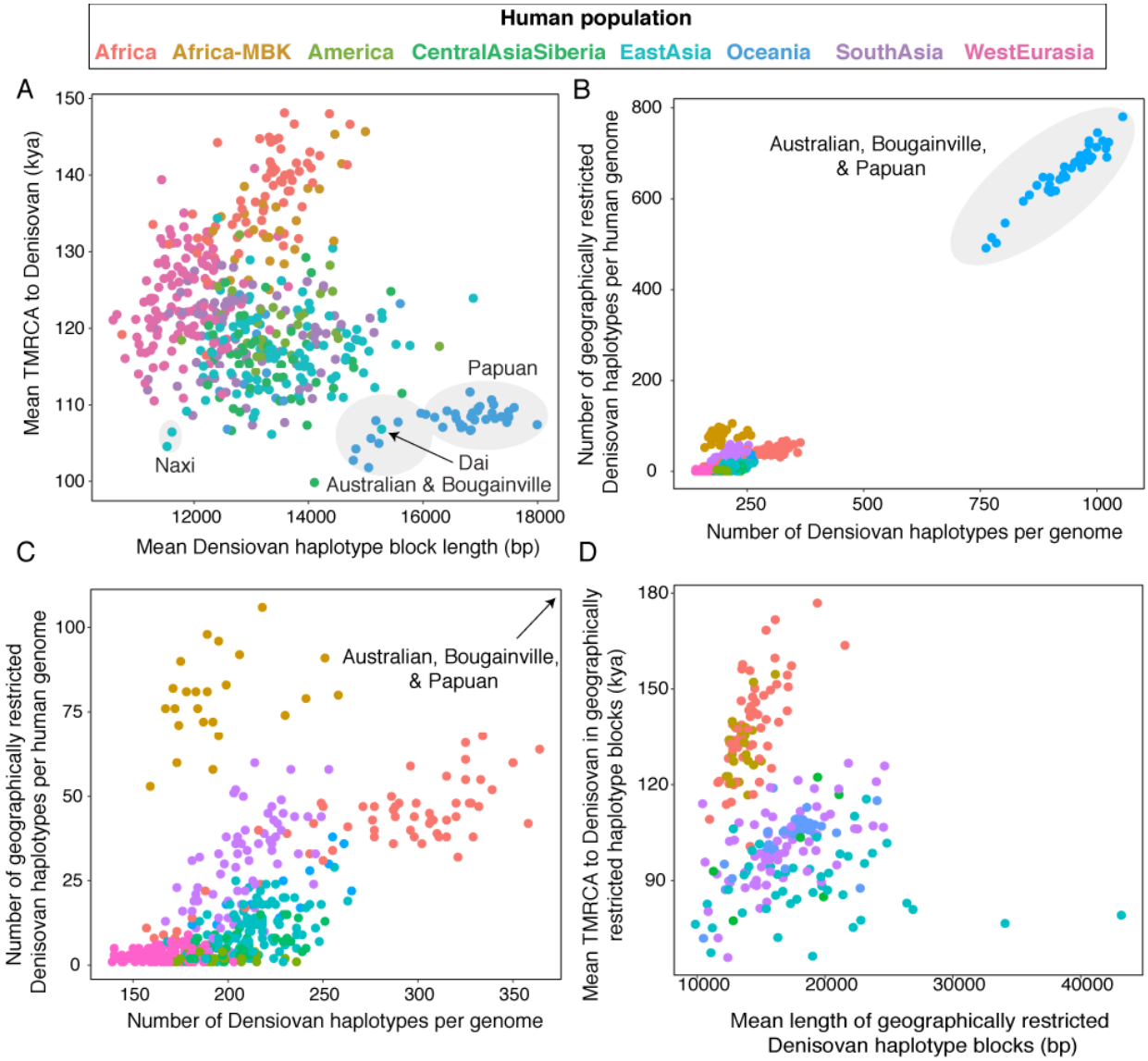

Fig. S 17

Properties of Denisovan-introgressed haplotype blocks. A: For each human genome haplotype (dots), mean haplotype block length for Denisovan haplotypes (x-axis) and mean TMRCA to Denisovan within Denisovan haplotyeps (y-axis) are reported. One outlier for short haplotypes (S_Naxi-2) appears to have phasing errors (*Table S 1*). B: Total number of unique Denisovan haplotype blocks per genome (x-axis) and number of geographically restricted (unique to a 3,000 km radius) Denisovan haplotype blocks per genome (y-axis). C: Same as B, but with the Australian, Bougainville, and Papuan cluster removed for readability. D: Same as A, but for only geographically restricted haplotype blocks. Only genome haplotypes with more than 10 unique Denisovan haplotype blocks are included.

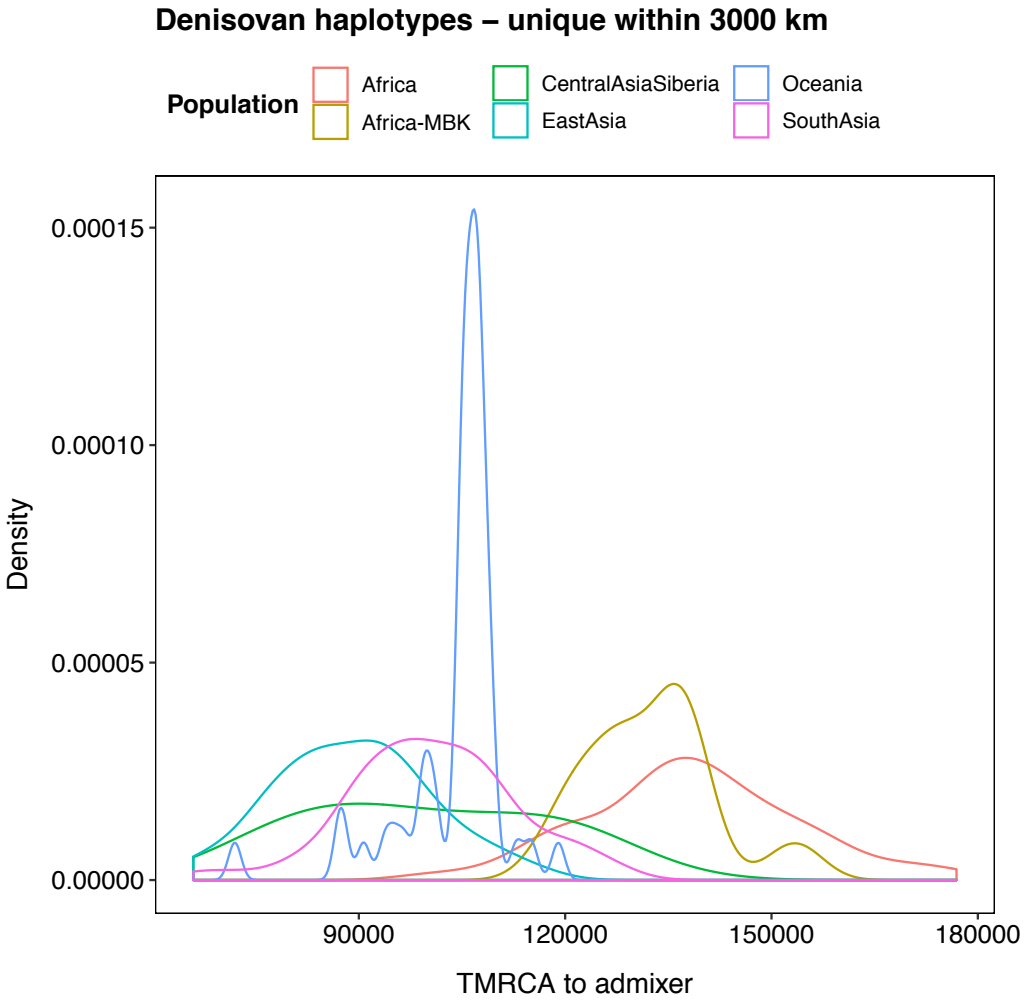

Fig. S 18

TMRCAs to the Denisovan genome in Denisovan-introgressed segments restricted to genomes sampled within 3,000 km of each other.

5

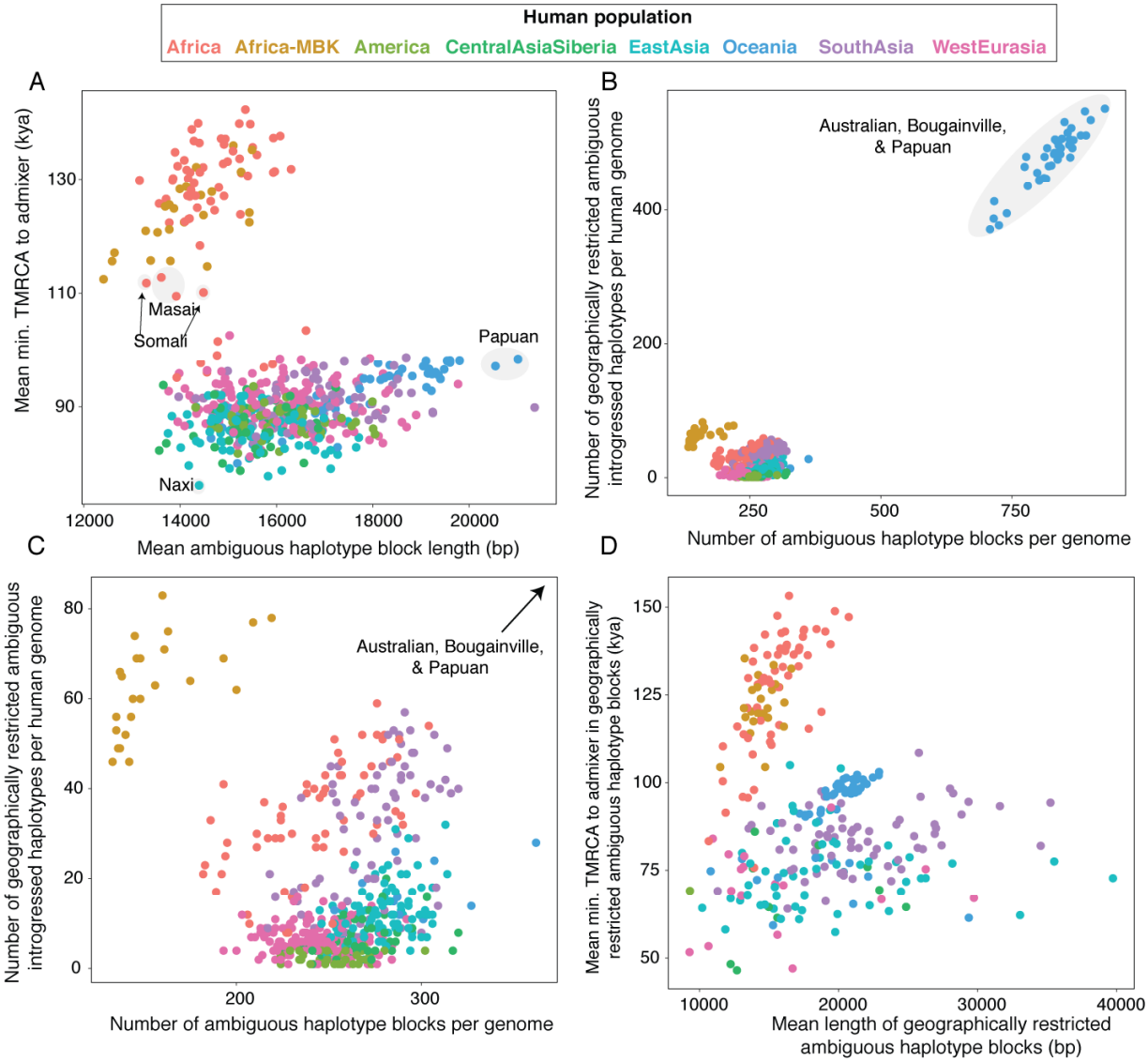

Fig. S 19

Properties of ambiguous (Neanderthal or Denisovan origin) archaic-introgressed haplotype blocks. A: For each human genome haplotype (dots), mean haplotype block length for ambiguous haplotypes (x-axis) and mean minimum TMRCA to an admixer haplotype (Neanderthal or Denisovan) within ambiguous haplotyeps (y-axis). One outlier for short haplotypes (S_Naxi-2) appears to have phasing errors (*Table S 1*). B: Total number of unique ambiguous haplotype blocks per genome (x-axis) and number of geographically restricted (unique to a 3,000 km radius) ambiguous haplotype blocks per genome (y-axis). C: Same as B, but with the Australian, Bougainville, and Papuan cluster removed for readability. D: Same as A, but for only geographically restricted haplotype blocks. Only genome haplotypes with more than 10 unique ambiguous haplotype blocks are included.

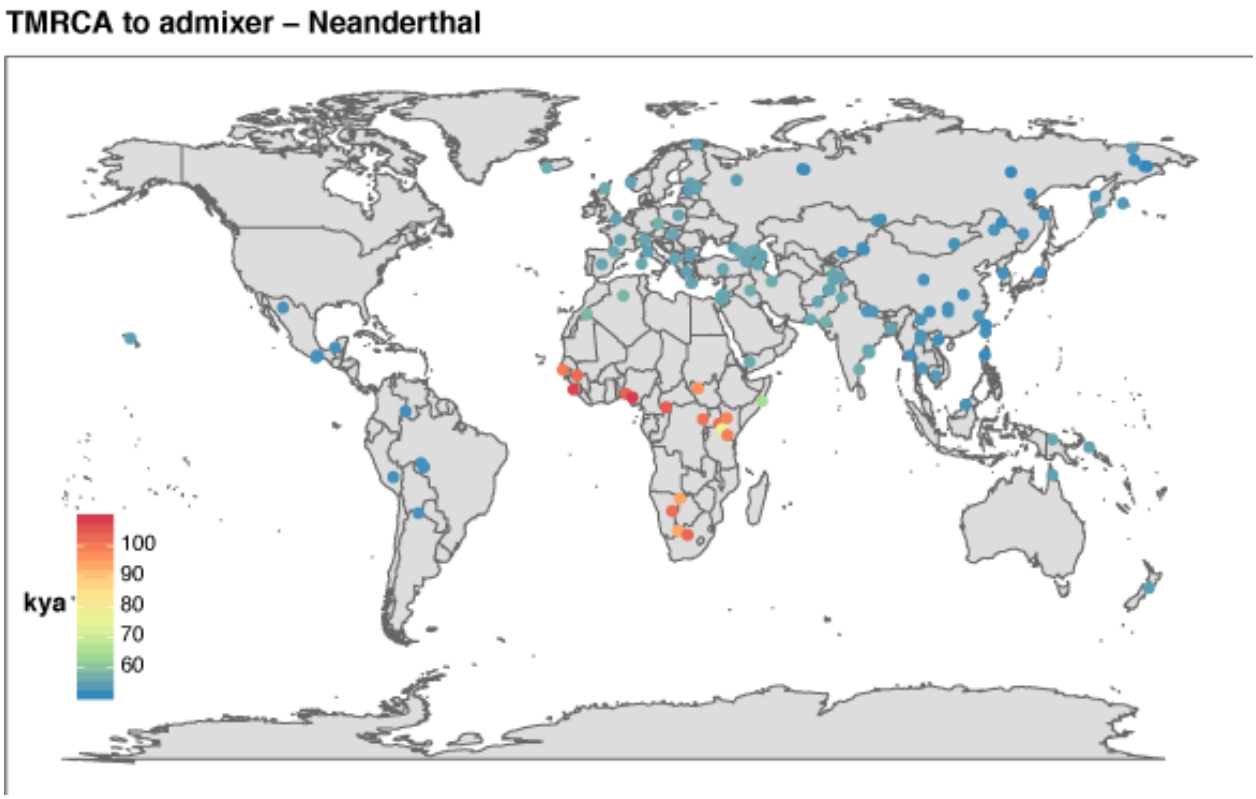

Fig. S20
Worldwide distribution of times to most recent common ancestor (TMRCA) to the closest Neanderthal haplotype of Neanderthal-like haplotypes in modern humans. Points are averages across all haplotypes within all genomes from each location. Numbers are corrected for branch shortening, using the values given for the two Neanderthal genomes in (*1*).

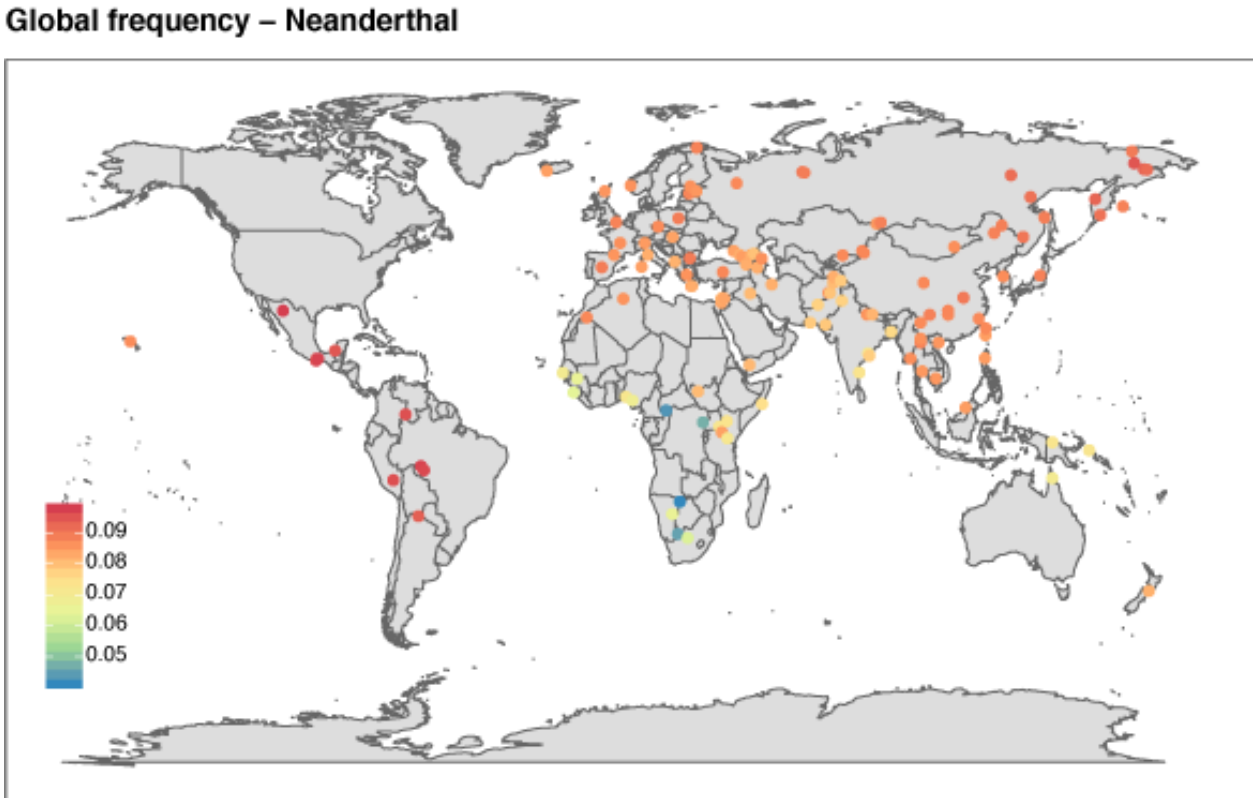

Fig. S21
Worldwide distribution of frequencies of individual Neanderthal-like haplotypes in modern humans. For each introgressed haplotype, its frequency in all humans worldwide was computed, and these values were averaged across all haplotypes within all human genomes from each geographic location.

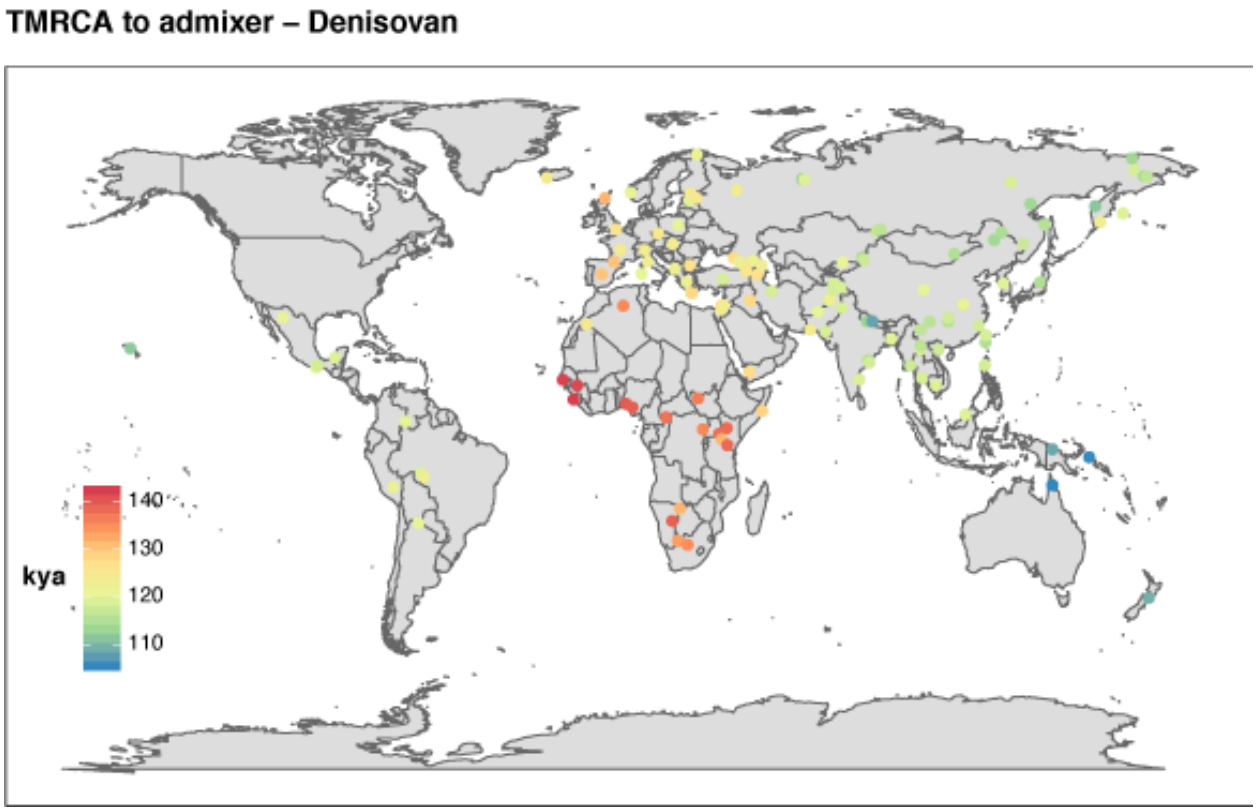

Fig. S22
Worldwide distribution of times to most recent common ancestor (TMRCA) to the closest Denisovan haplotype of Denisovan-like haplotypes introgressed in modern humans. Points are averages across all haplotypes within all genomes from each location. Numbers are corrected for branch shortening, using the value given for the Denisovan genome in (*1*).

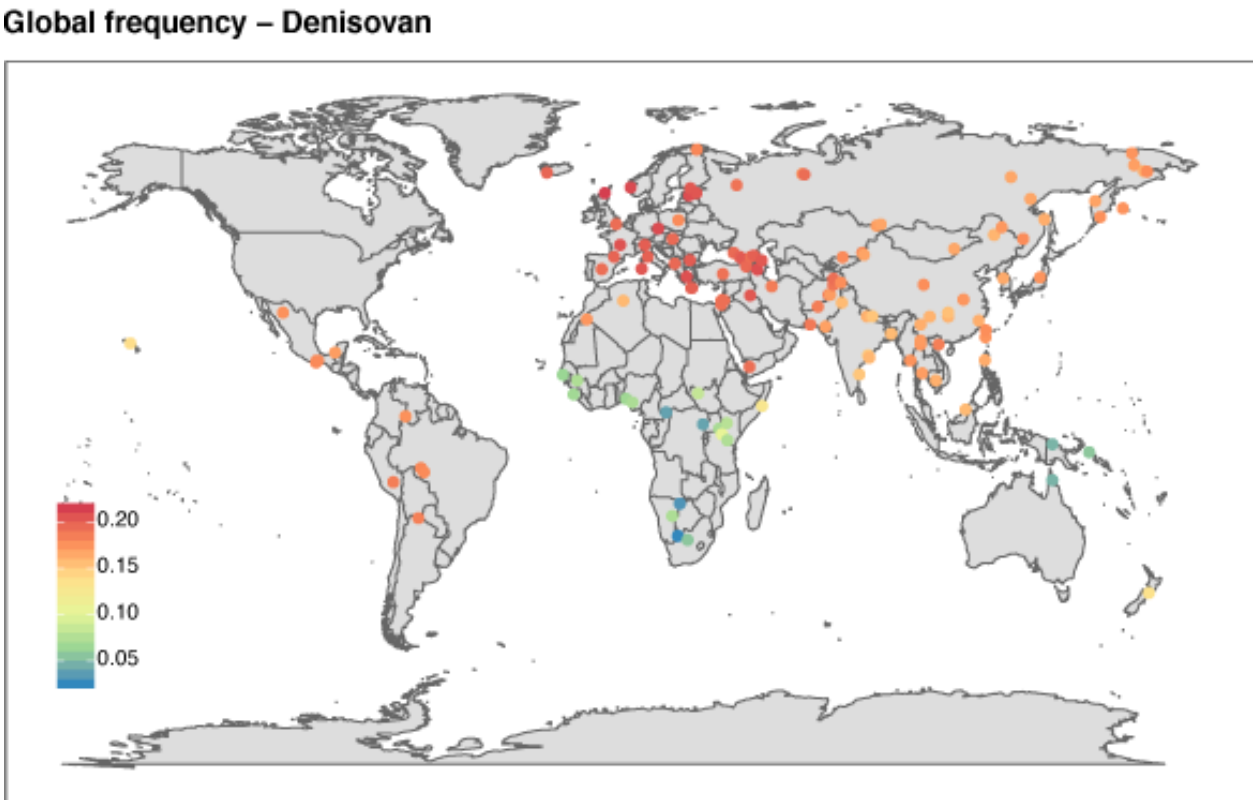

Fig. S23
Worldwide distribution of frequencies of individual Denisovan-like haplotypes in modern humans. For each introgressed haplotype, its frequency in all humans worldwide was computed, and these values were averaged across all haplotypes within all human genomes from each geographic location.

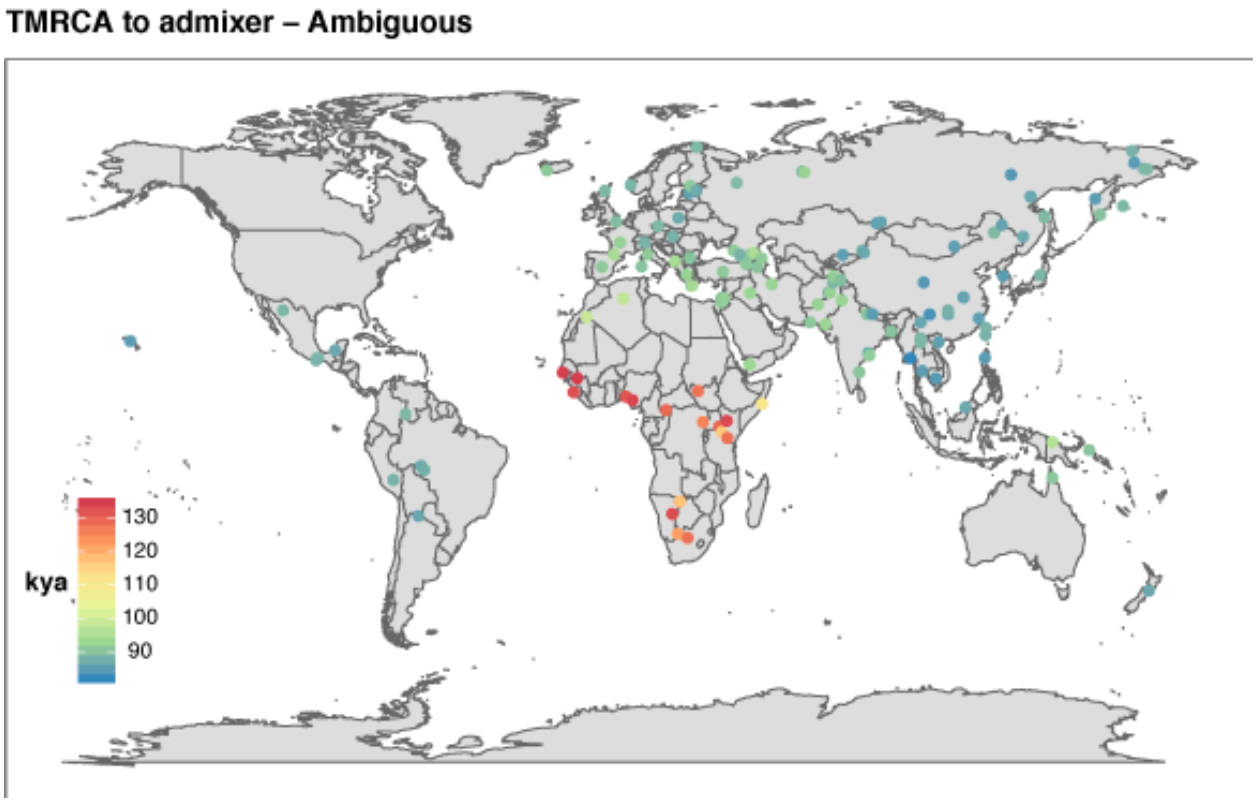

Fig. S24

Worldwide distribution of times to most recent common ancestor (TMRCA) to the closest admixer haplotype (Neanderthal or Denisovan) of introgressed haplotypes of ambiguous origin in modern humans. Points are averages across all haplotypes within all genomes from each location. Numbers are corrected for branch shortening, using the values given in (*1*). Some ambiguous haplotypes are the result of merging Neanderthal and Denisovan haplotypes together; in those cases, TMRCAs are averages of those in the two original haplotypes, weighted by the number of bases in each.

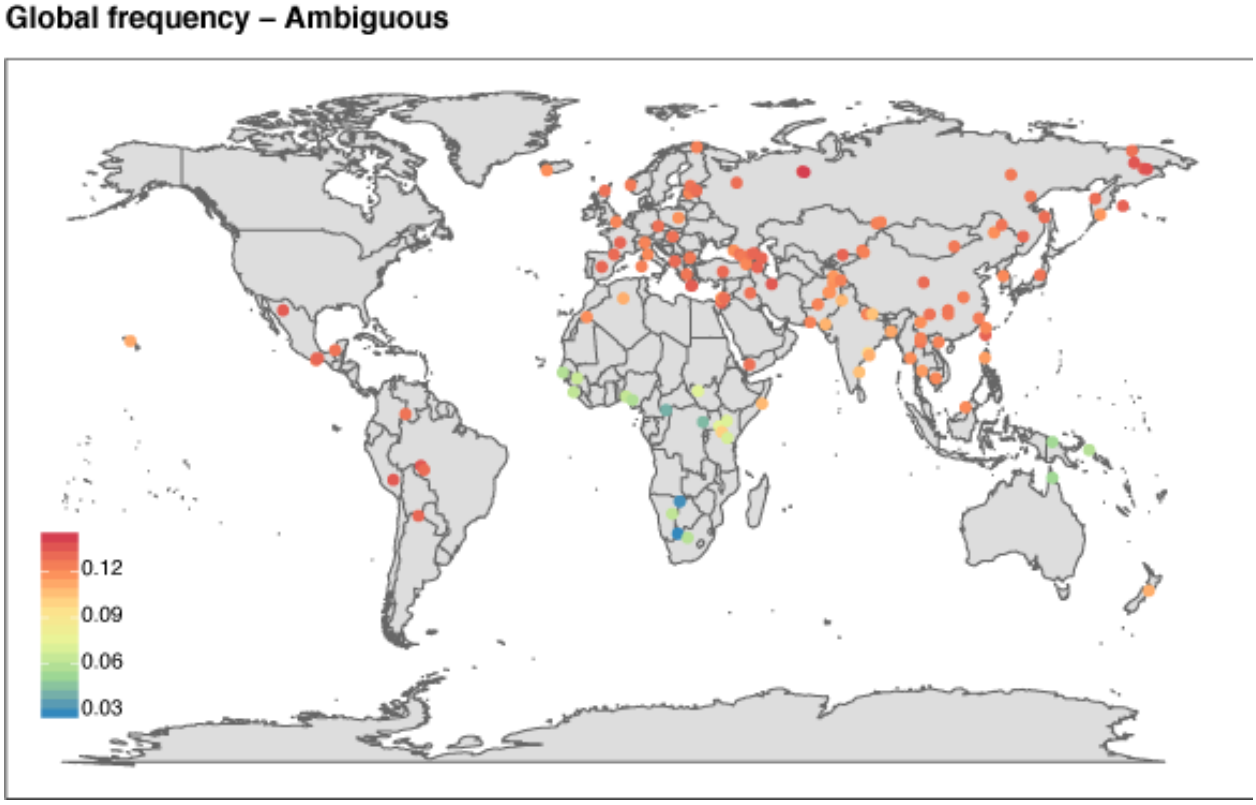

Fig. S 25

Worldwide distribution of frequencies of individual ambiguous origin introgressed haplotypes in modern humans. For each introgressed haplotype, its frequency in all humans worldwide was computed, and these values were averaged across all haplotypes within all human genomes from each geographic location.

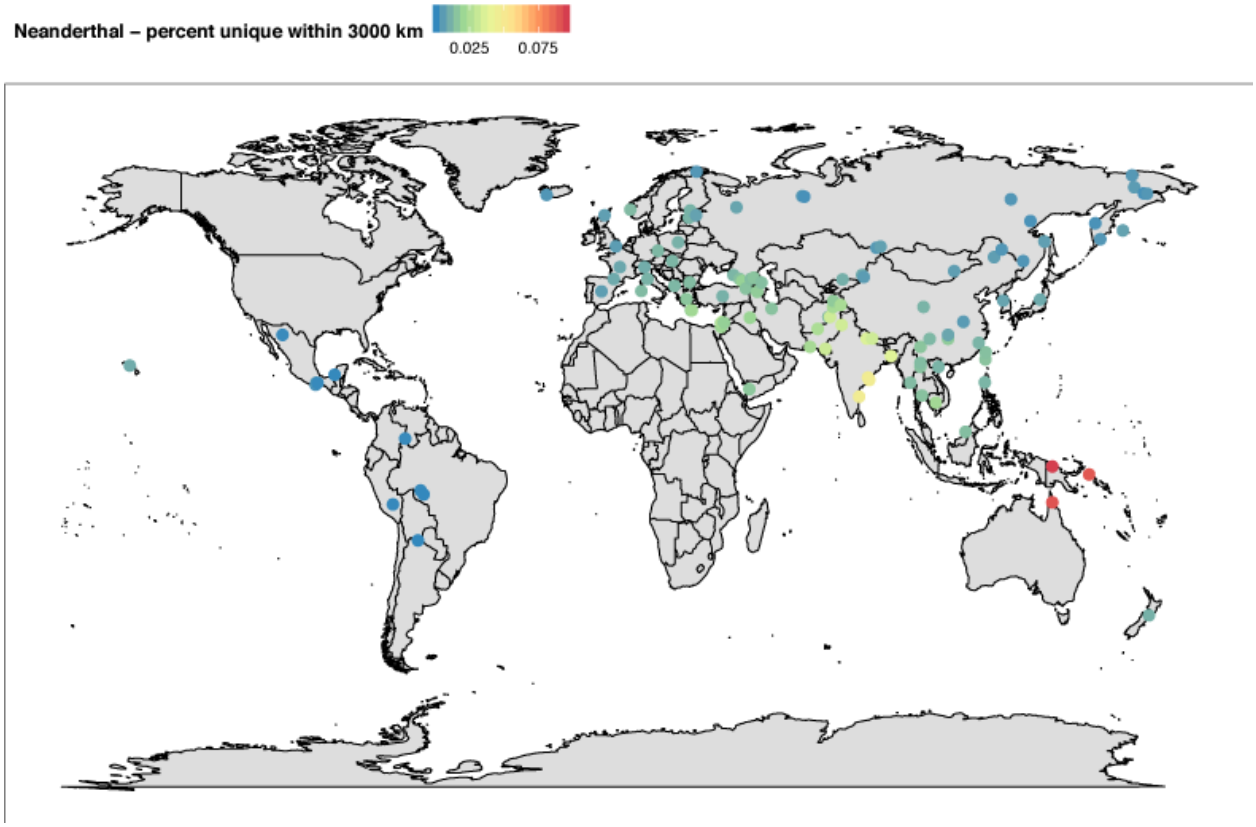

Fig. S 26

For all non-African haplotypes, the percent of geographically restricted (limited to genomes sampled within 3,000 km of each other) Neanderthal-introgressed segments, of all Neanderthal-introgressed segments.

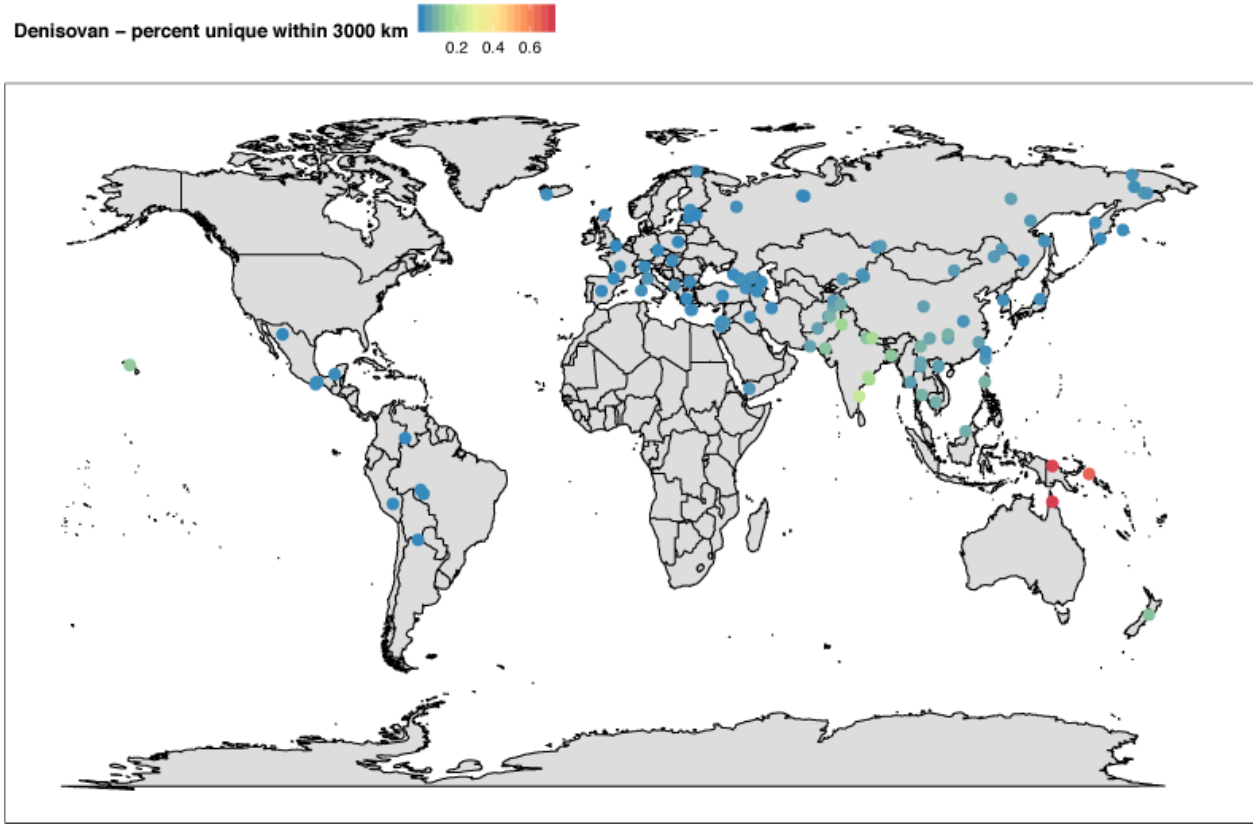

Fig. S 27

For all non-African haplotypes, the percent of geographically restricted (limited to genomes sampled within 3,000 km of each other) Denisovan-introgressed segments, as a percent of all Denisovan-introgressed segments.

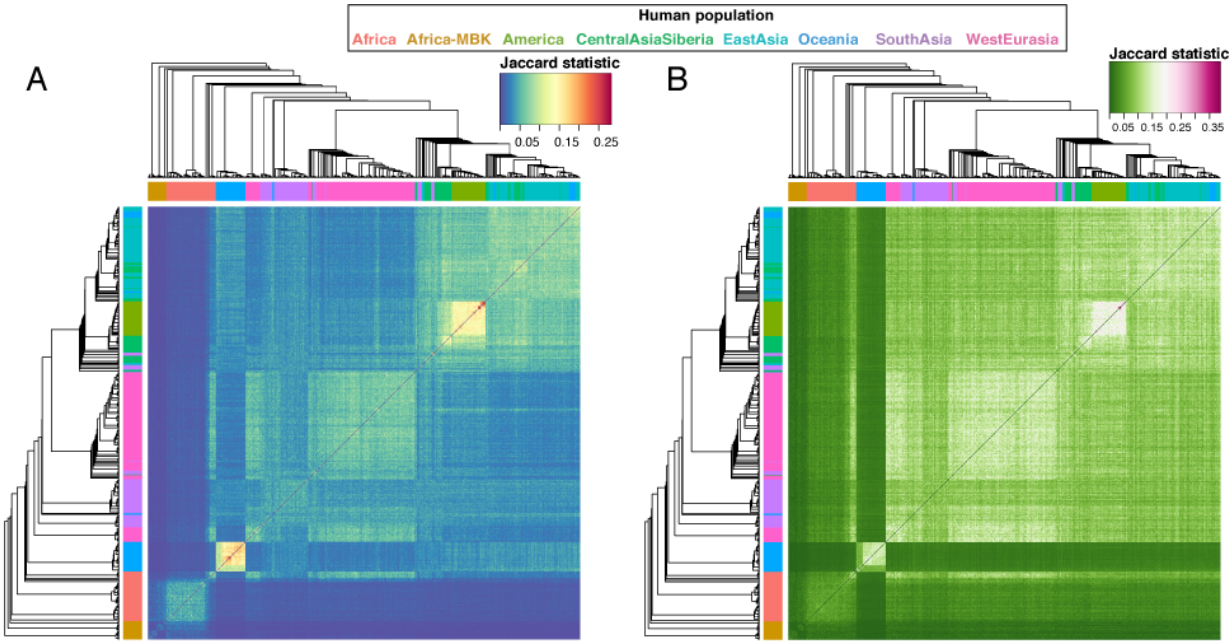

Fig. S 28

Sharing of archaic hominin haplotypes between human genome haplotypes. Genomes are arranged (rows and columns) according to a tree inferred via UPGMA on genome-wide SNPs from the input data set (top and bottom of matrices); colors below the trees correspond to SGDP population identifiers (top). A: Sharing of Neanderthal-introgressed haplotypes, as measured by the Jaccard statistic. B: Sharing of Denisovan-introgressed haplotypes, as measured by the Jaccard statistic.

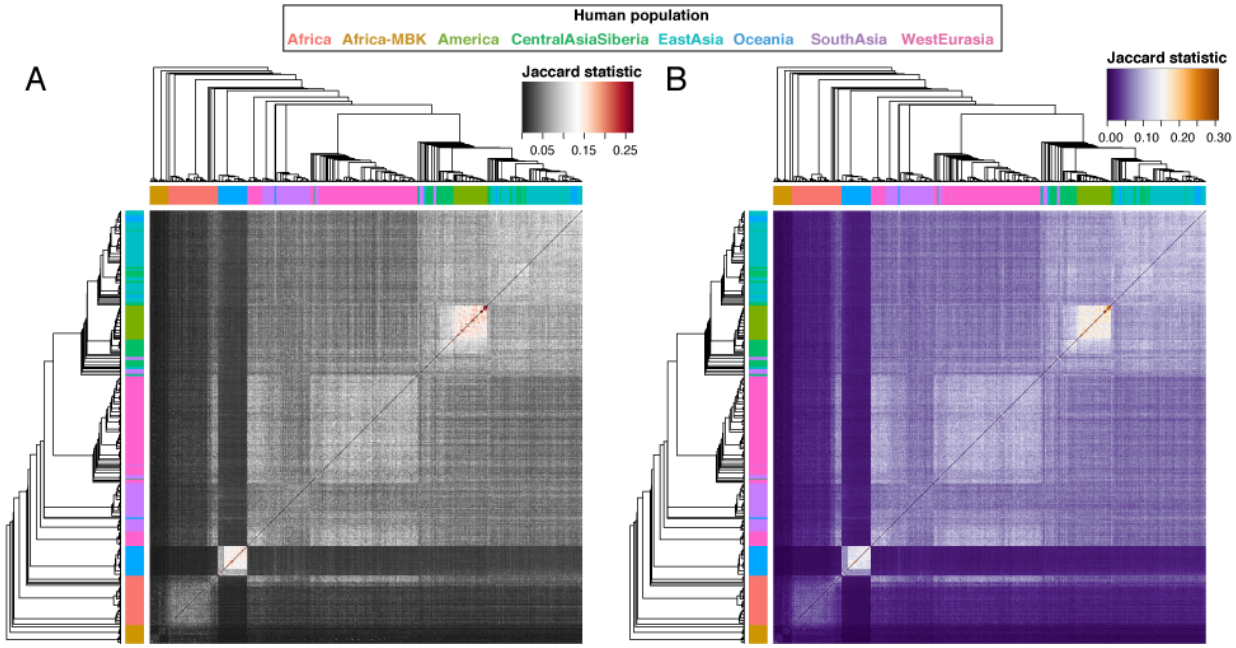

Fig. S 29

Sharing of archaic hominin haplotypes between human genome haplotypes. Genomes are arranged (rows and columns) according to a tree inferred via UPGMA on genome-wide SNPs from the input data set (top and bottom of matrices); colors below the trees correspond to SGDP population identifiers (top). A: Sharing of ambiguous (Neanderthal or Denisovan) introgressed haplotypes, as measured by the Jaccard statistic. B: Sharing of ambiguous introgressed haplotypes, combined with Denisovan introgressed haplotypes, as measured by the Jaccard statistic.

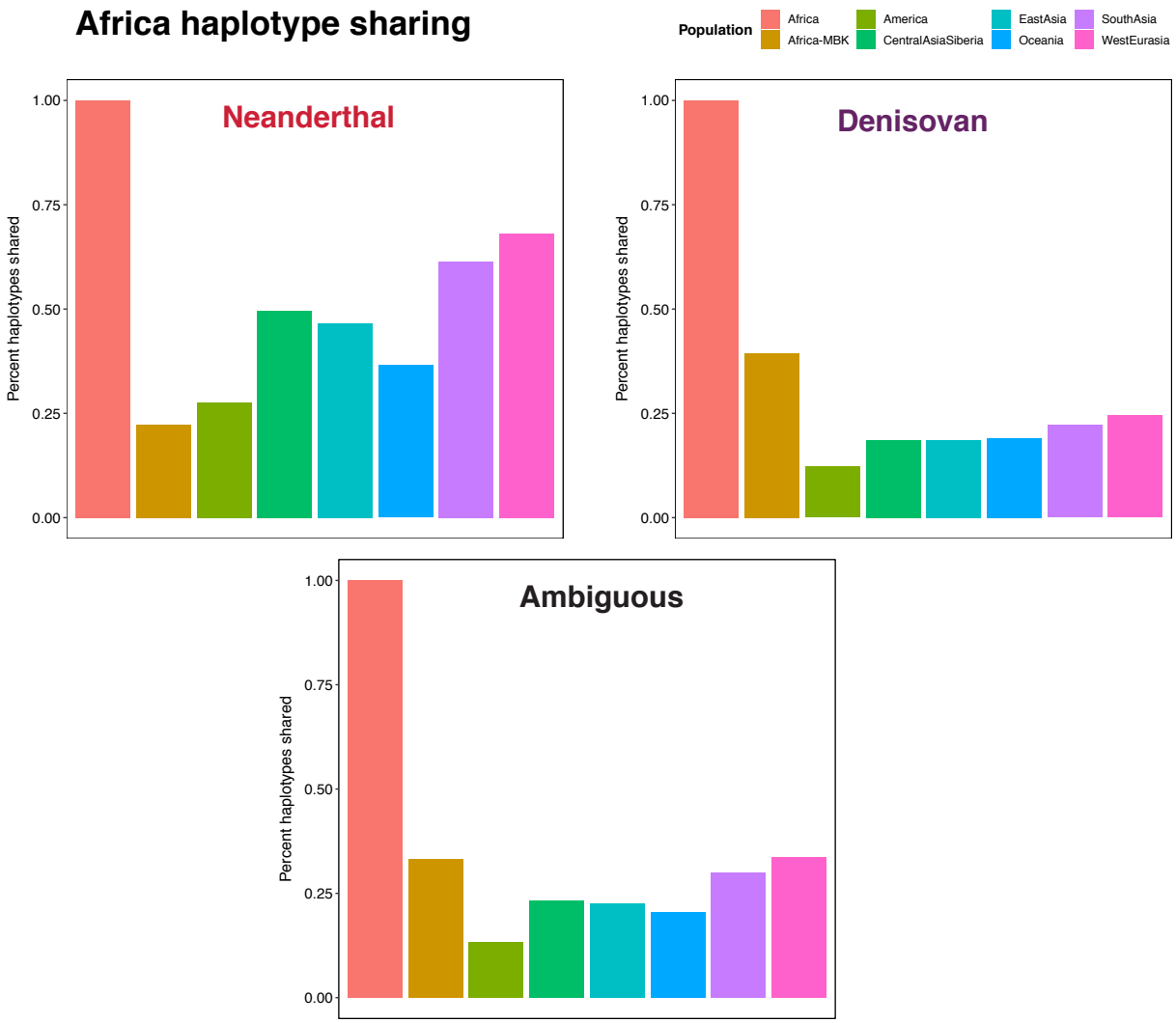

Fig. S30

Percent of archaic-introgressed haplotypes in Africa (excluding Biaka, Mbuti, and Khomani-San) shared with other SGDP populations. Africa-MBK consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be < 10% frequency.

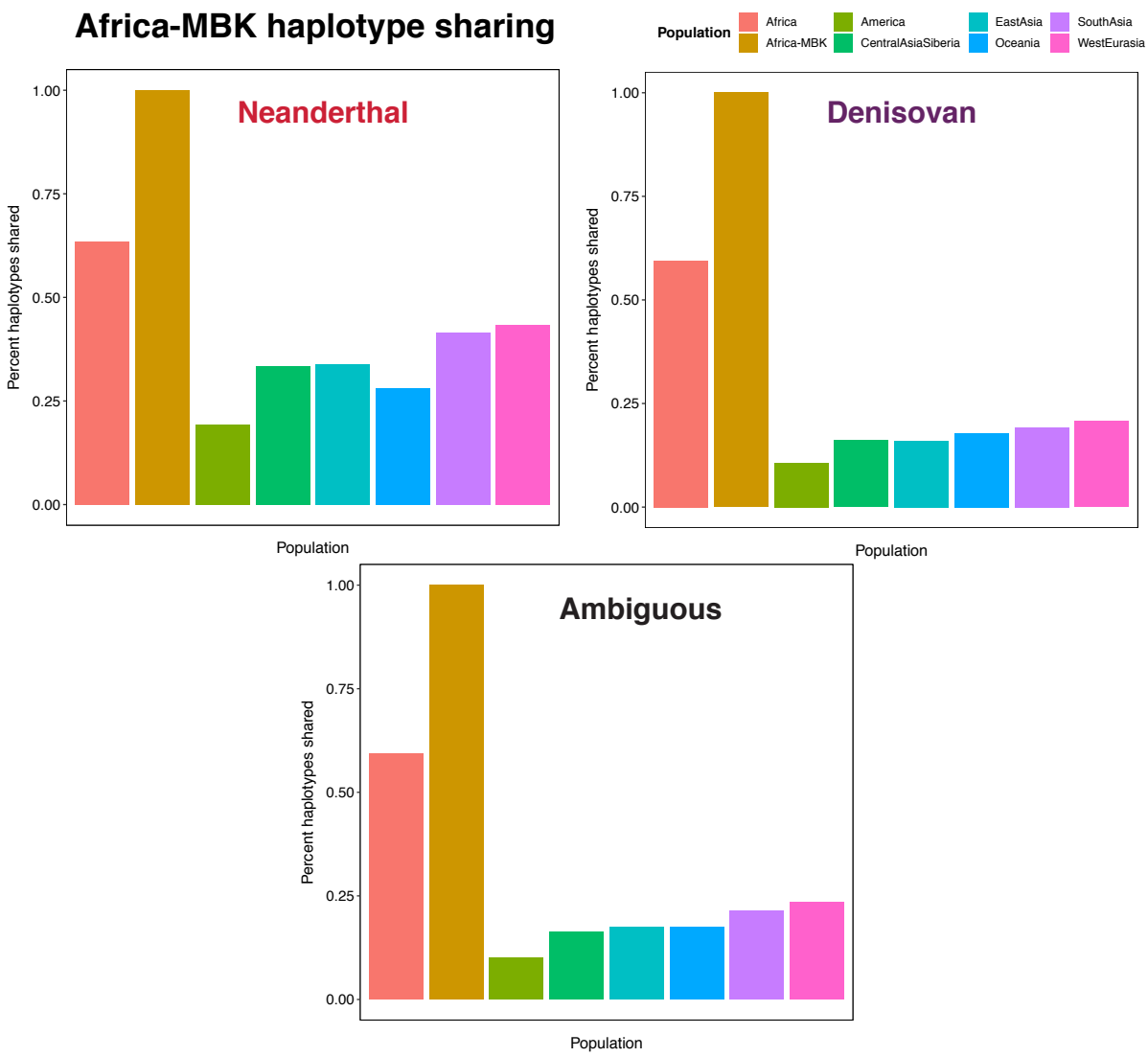

Fig. S31
Percent of archaic-introgressed haplotypes in basal African lineages used as an outgroup (Biaka, Mbuti, and Khomani-San) shared with other SGDP populations.

5

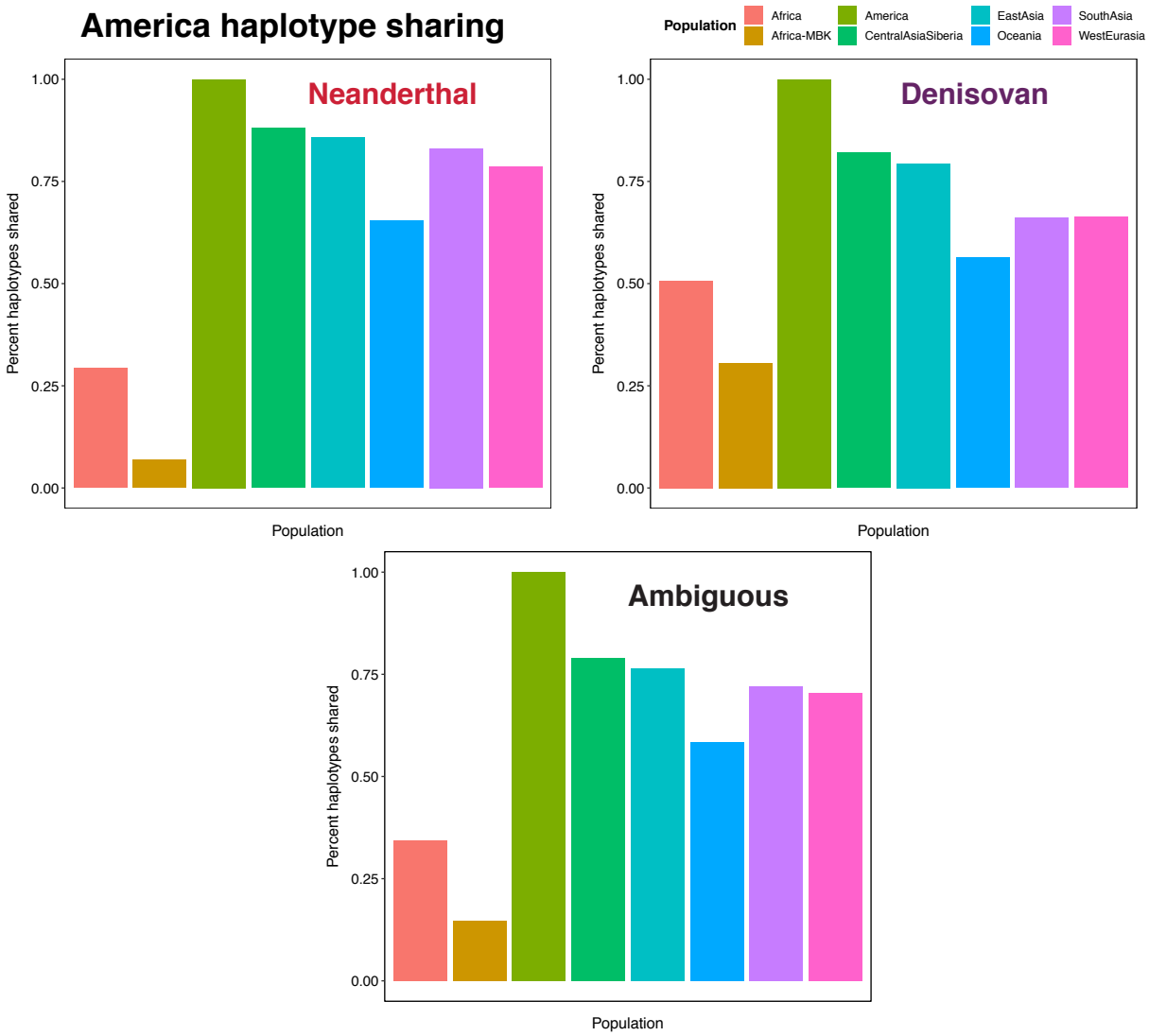

Fig. S32
Percent of archaic-introgressed haplotypes in America shared with other SGDP populations. Africa-MBK consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be < 10% frequency.

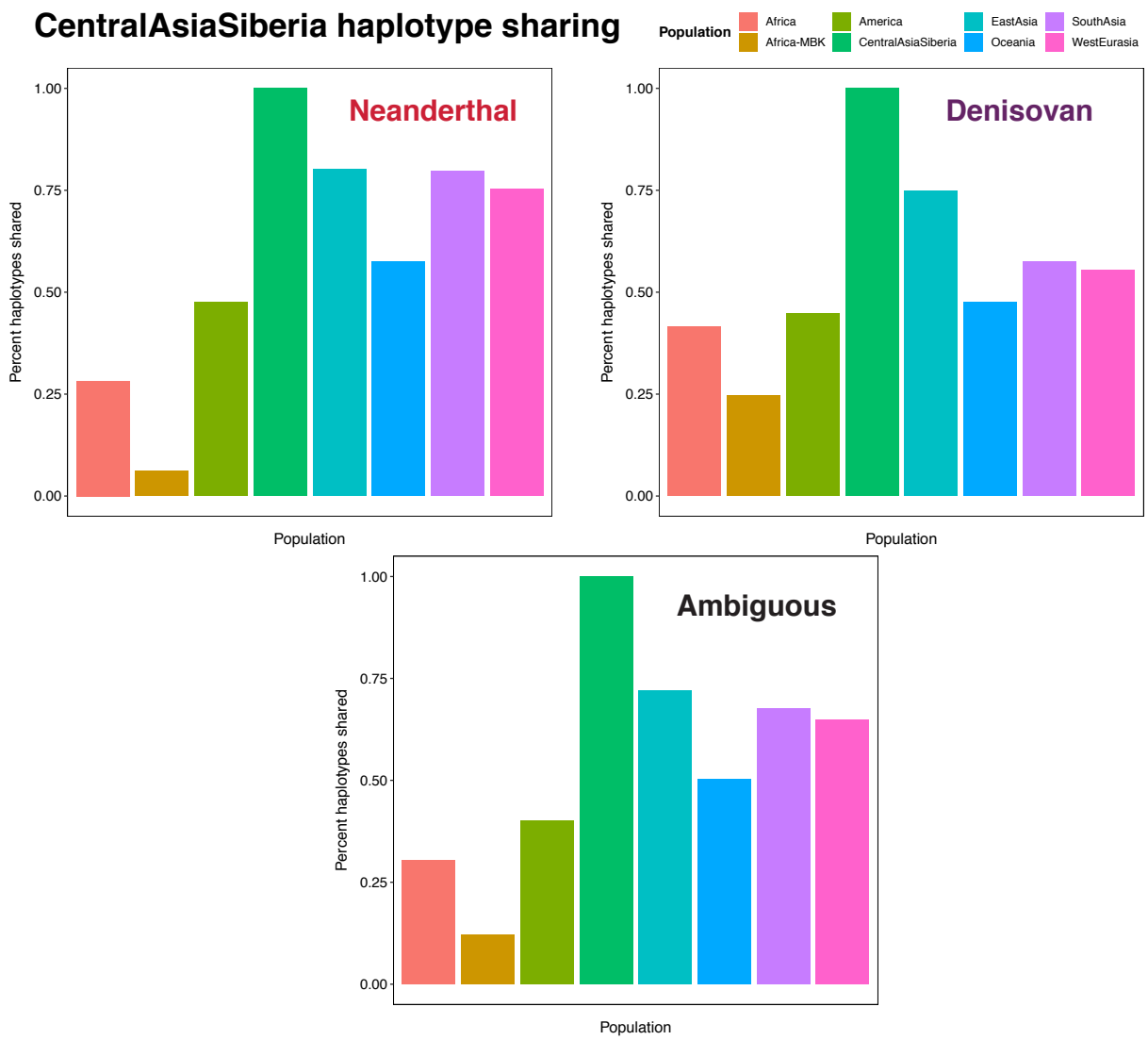

Fig. S33
Percent of archaic-introgressed haplotypes in CentralAsiaSiberia shared with other SGDP populations. Africa-MBK consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be < 10% frequency.

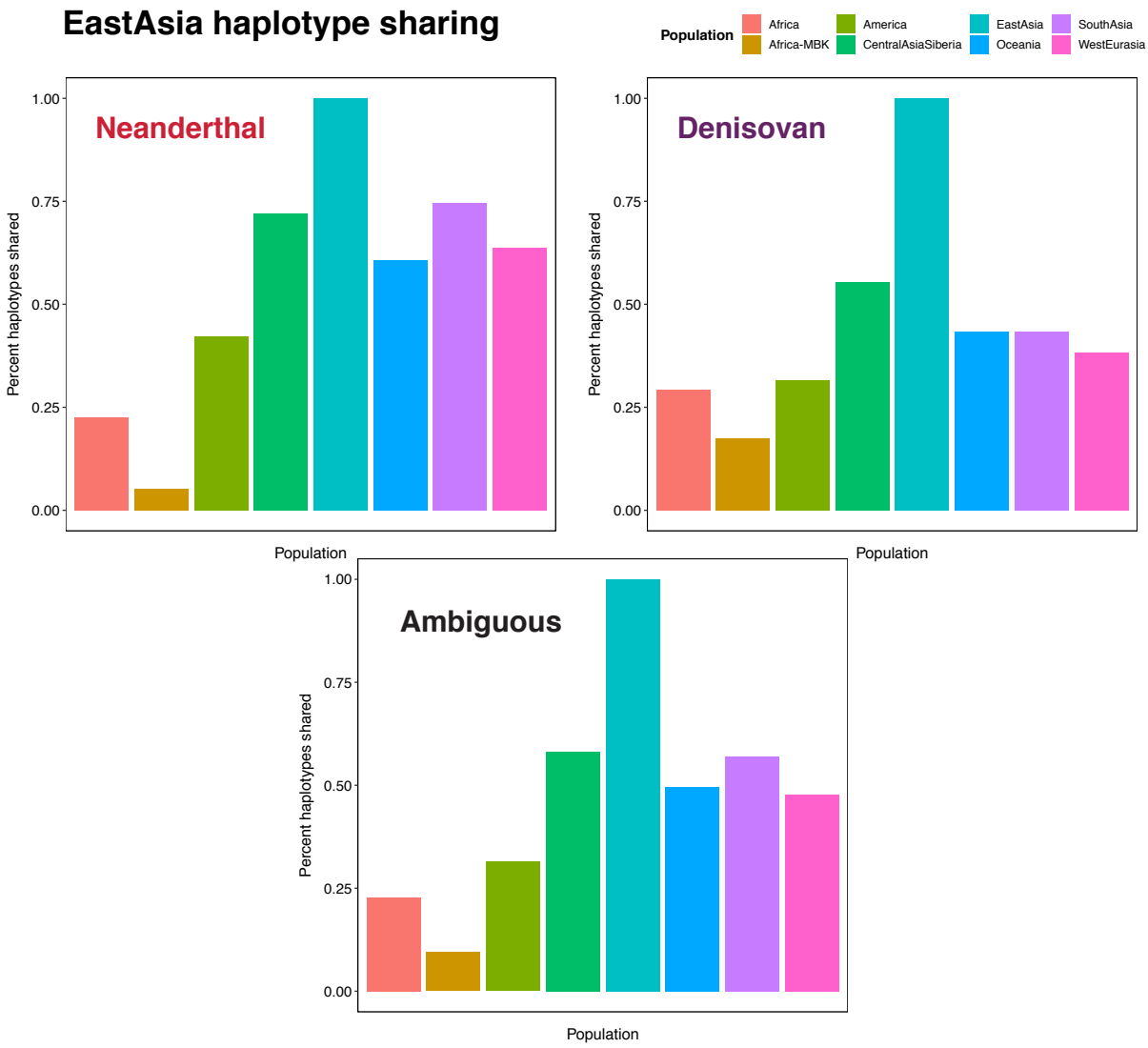

Fig. S34
Percent of archaic-introgressed haplotypes in EastAsia shared with other SGDP populations. Africa-MBK consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be < 10% frequency.

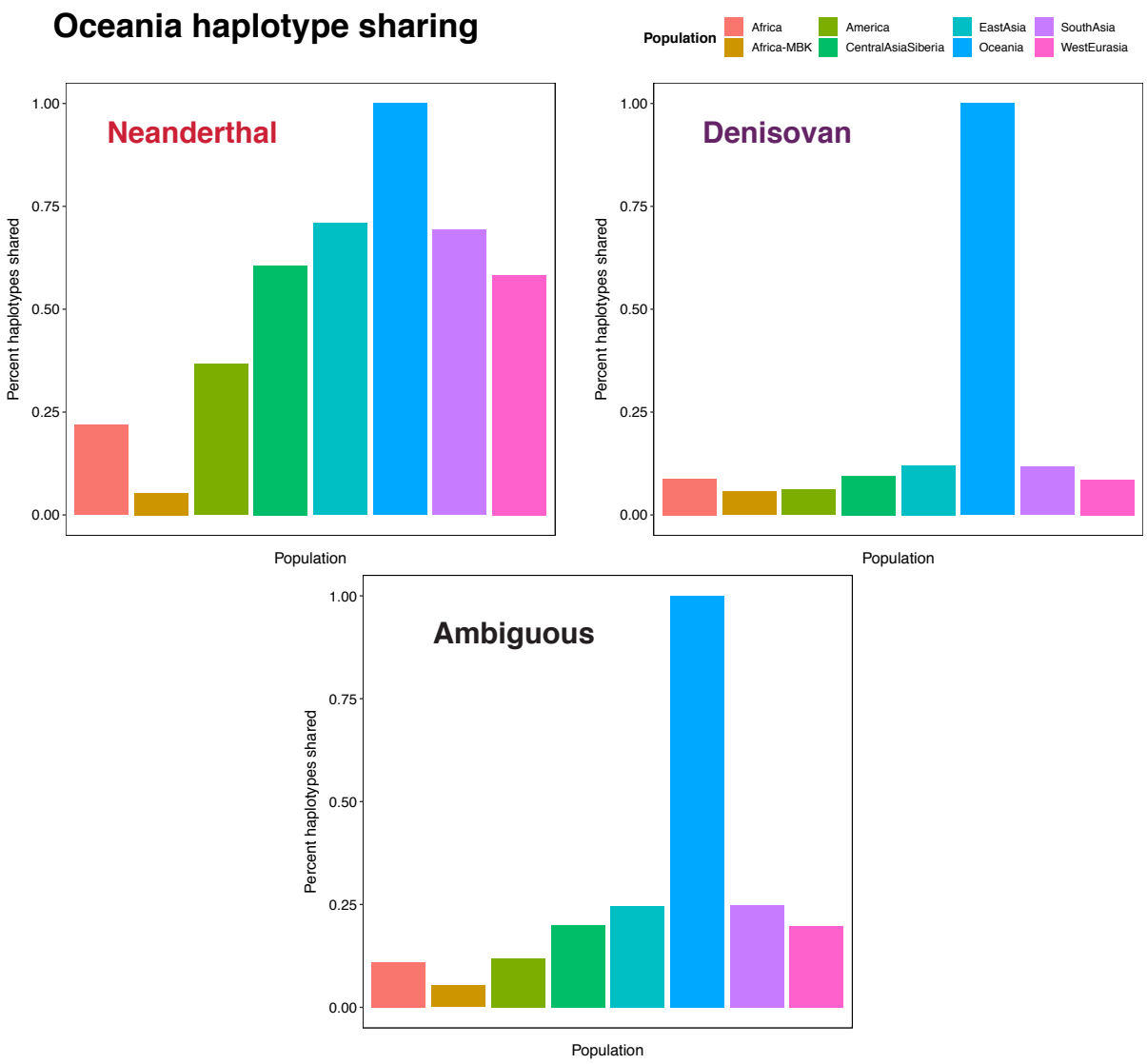

Fig. S35
Percent of archaic-introgressed haplotypes in Oceania shared with other SGDP populations. Africa-MBK consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be < 10% frequency.

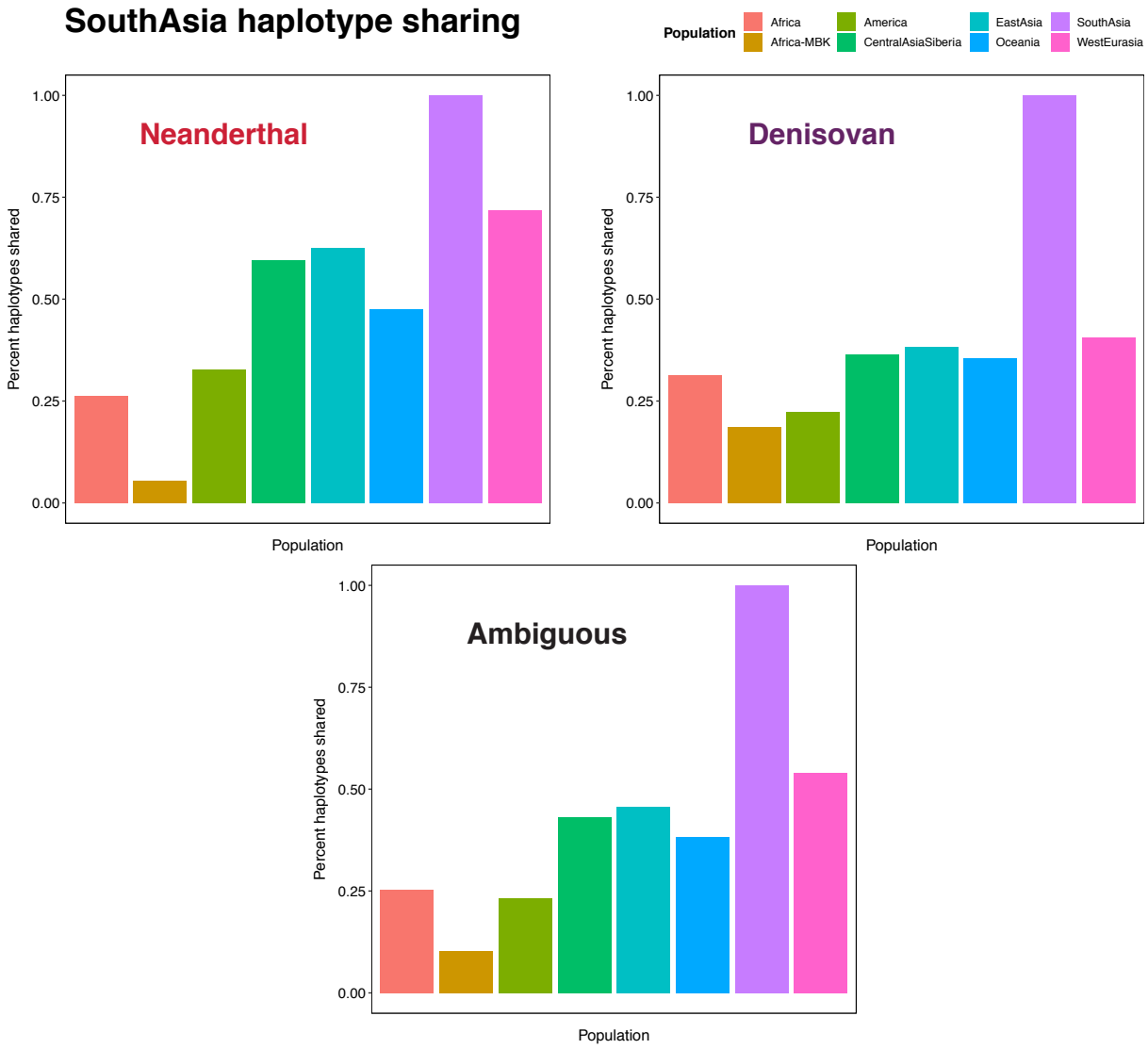

Fig. S36
Percent of archaic-introgressed haplotypes in SouthAsia shared with other SGDP populations. Africa-MBK consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be < 10% frequency.

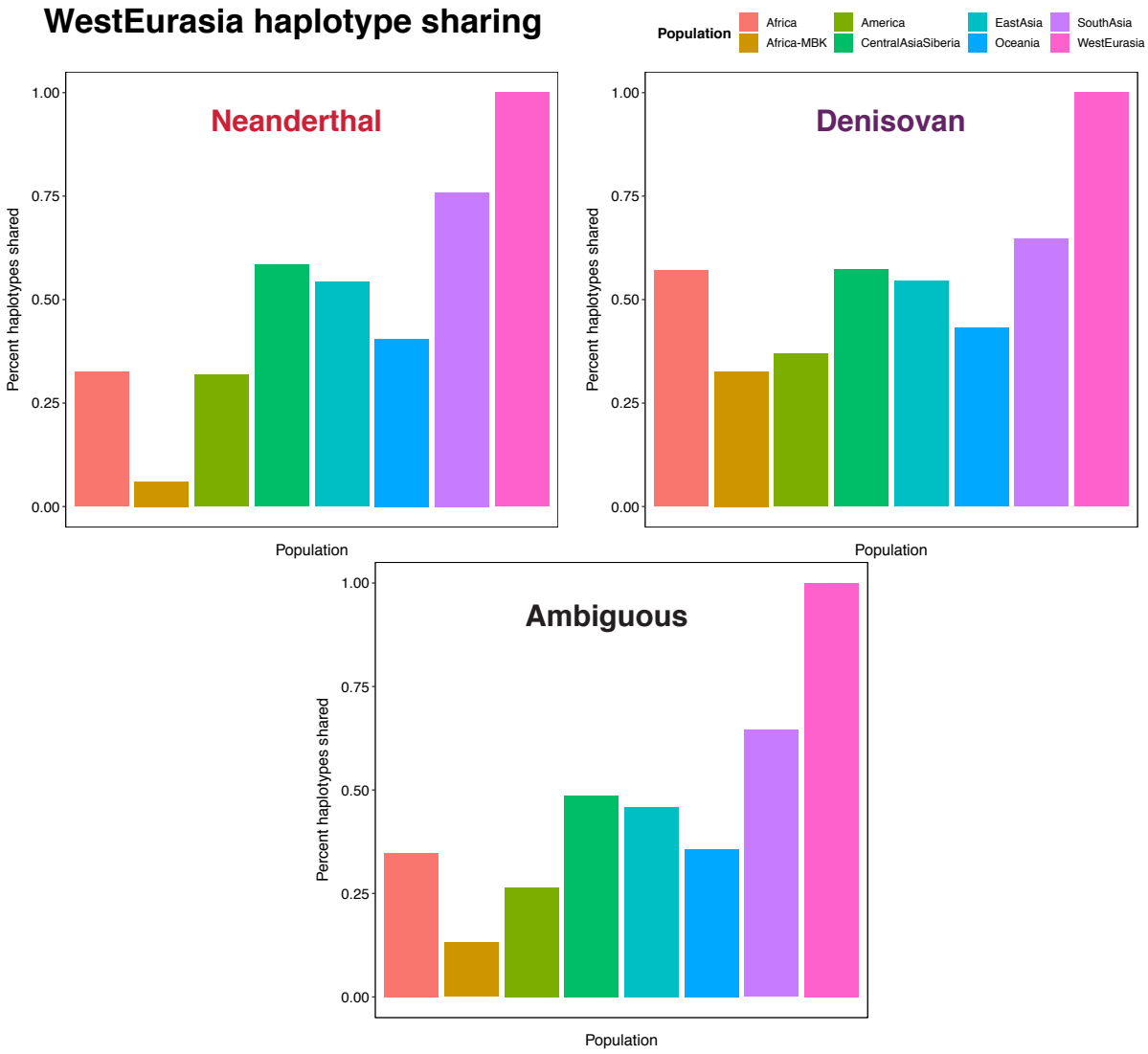

Fig. S37
Percent of archaic-introgressed haplotypes in WestEurasia shared with other SGDP populations. Africa-MBK consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be < 10% frequency.

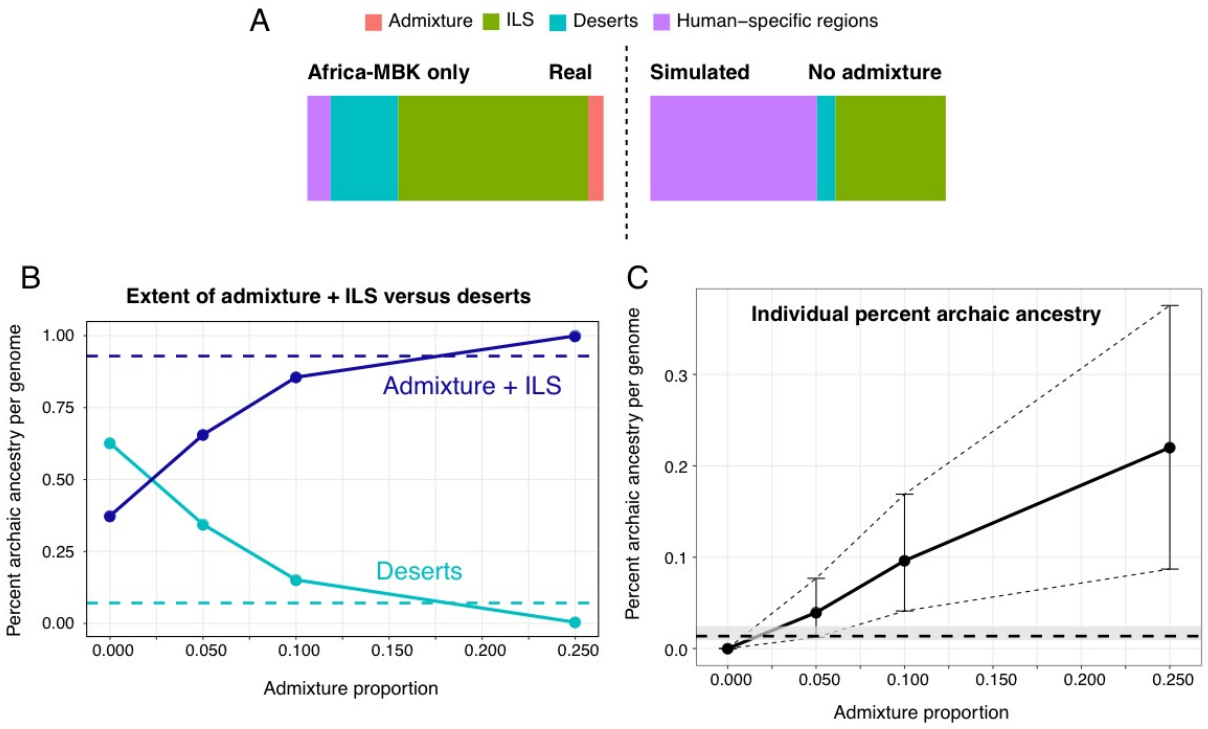

Fig. S 38

Coalescent simulations with one Neanderthal and one Denisovan admixture event were run, with increasing admixture proportions (Supplementary Methods). An ARG was inferred over each, and several values were computed and compared to those from real data. A: Percent of the genome in regions containing archaic hominin admixture in any individual, incomplete lineage sorting in any individual, neither (deserts), or neither plus a derived allele specific to and fixed in all humans (human-specific regions). Left: Using the ARG inferred on our full real data set, but considering only the most basal (Africa-MBK) human lineages, thought to be relatively free of admixture. Human-specific regions still require alleles to be fixed across all humans, not only Africa-MBK individuals. Right: results using a simulation that did not include any archaic hominin admixture. B: The full extent of regions containing admixture or ILS with any archaic hominin across all sampled humans, and the extent of regions free of both admixture and ILS (deserts). Horizontal lines show values computed from real data. C: The range of percent archaic ancestry per individual (*f*) (points are mean values; error bars show maximum and minimum value). The horizontal dotted line shows this (mean) value for real data, and the shaded rectangle shows maximum and minimum values.

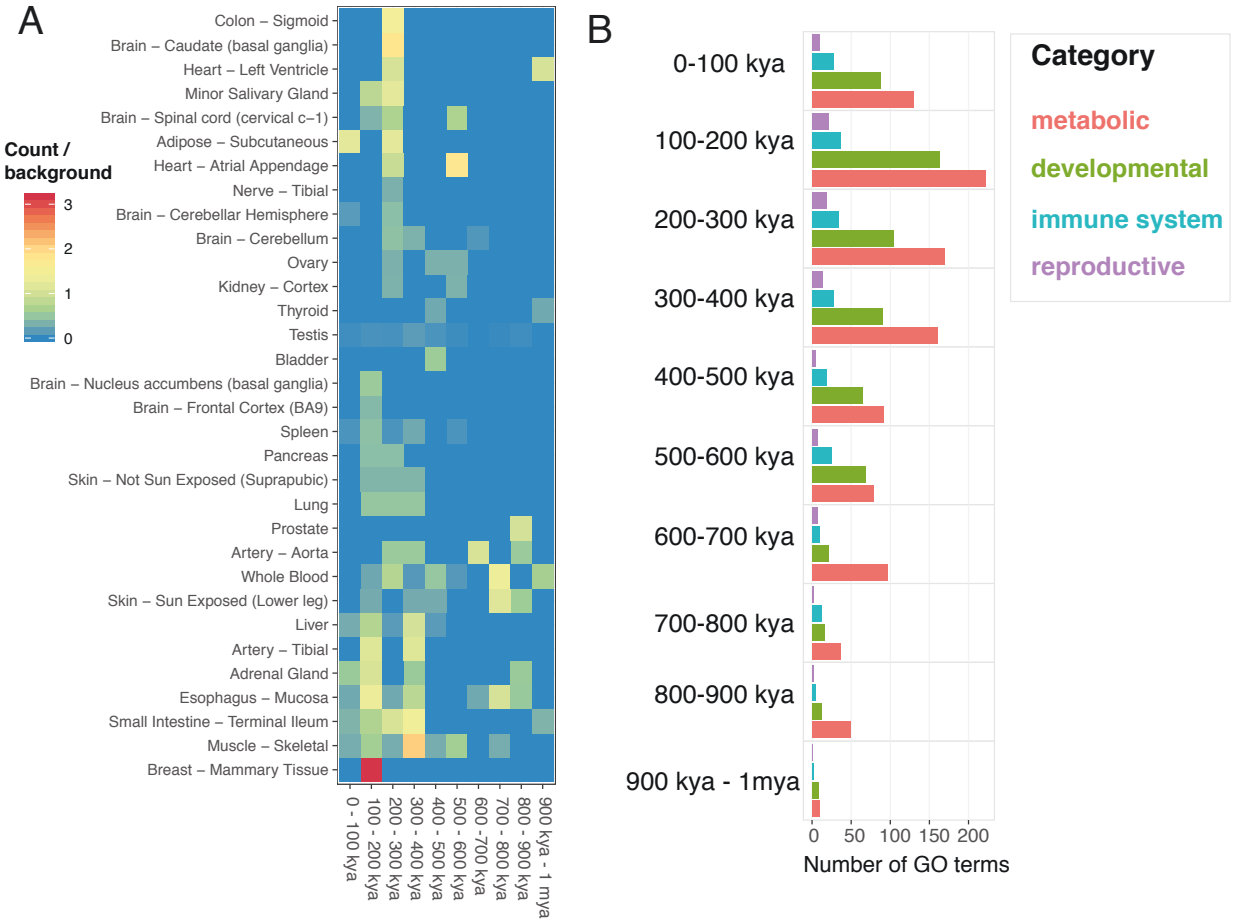

Fig. S 39

Human-specific derived changes through time. A: For human-specific derived mutations affecting tissue-specific genes (tau > 0.9) in time bins dating back 1 my, enrichment of affected tissues in each time bin. Expression data came from the per-gene median tissue expression values across many samples from the GTEx database (*66*), with cell line "tissues" excluded. Values shown are the percent of all mutations in a time bin affecting a given tissue, divided by the percent of all total tissue-specific genes affecting that tissue. B: For human-specific derived mutations affecting all genes in time bins dating back 1 my, the relative number of Gene Ontology (*65*) terms per time bin below the terms "metabolic process," "developmental process," "immune system process," and "reproductive process" are shown.

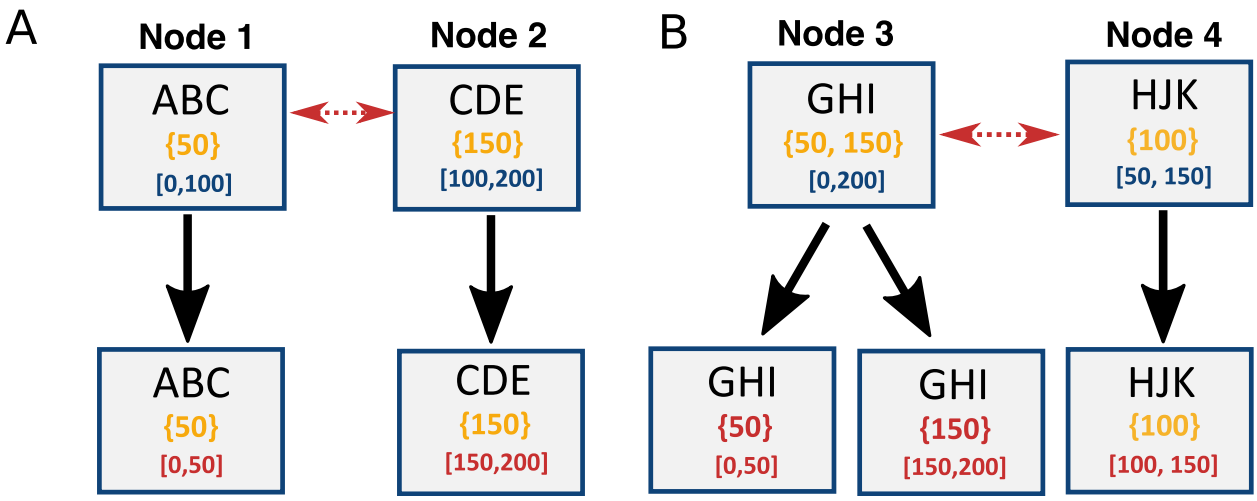

Fig. S40.

How node indices are adjusted when four haplotype test failures are encountered. Black letters represent clades, yellow numbers in curly braces represent site indices, and blue numbers in brackets represent start and end coordinates (inclusive). Red text indicates an adjusted value. Red arrows show four haplotype test failures, and black arrows represent changes made to nodes. A: a simple case where the furthest donwnstream site owned by node 1 (50) is upstream of the furthest upstream site owned by node 2 (150). In this case, node 1's end coordinate is set to its furthest downstream site, and node 2's start coordinate is set to its furthest upstream site. B: Node 4 interrupts the range of node 3. Node 3 must be split into two nodes, and all three resulting nodes must have their ranges adjusted.

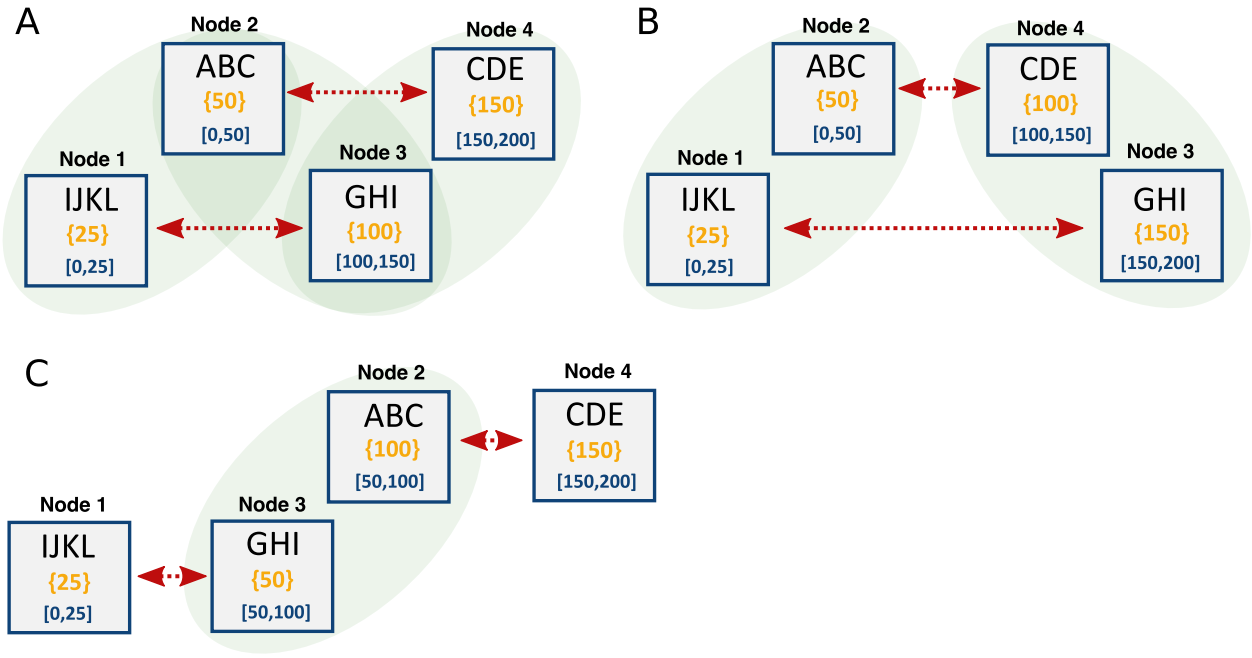

Fig. S41.

Tree-compatibility in three different situations. Gray squares are nodes, black letters are clades, yellow numbers in curly braces are genomic positions, and blue numbers in brackets are start/end coordinates. Red arrows indicate four haplotype test failures, and green ovals denote tree compatibility. A: three pairs of nodes are compatible (can belong to the same trees as each other). B: only two pairs of nodes are tree-compatible. C: Only one pair of nodes is tree-compatible.

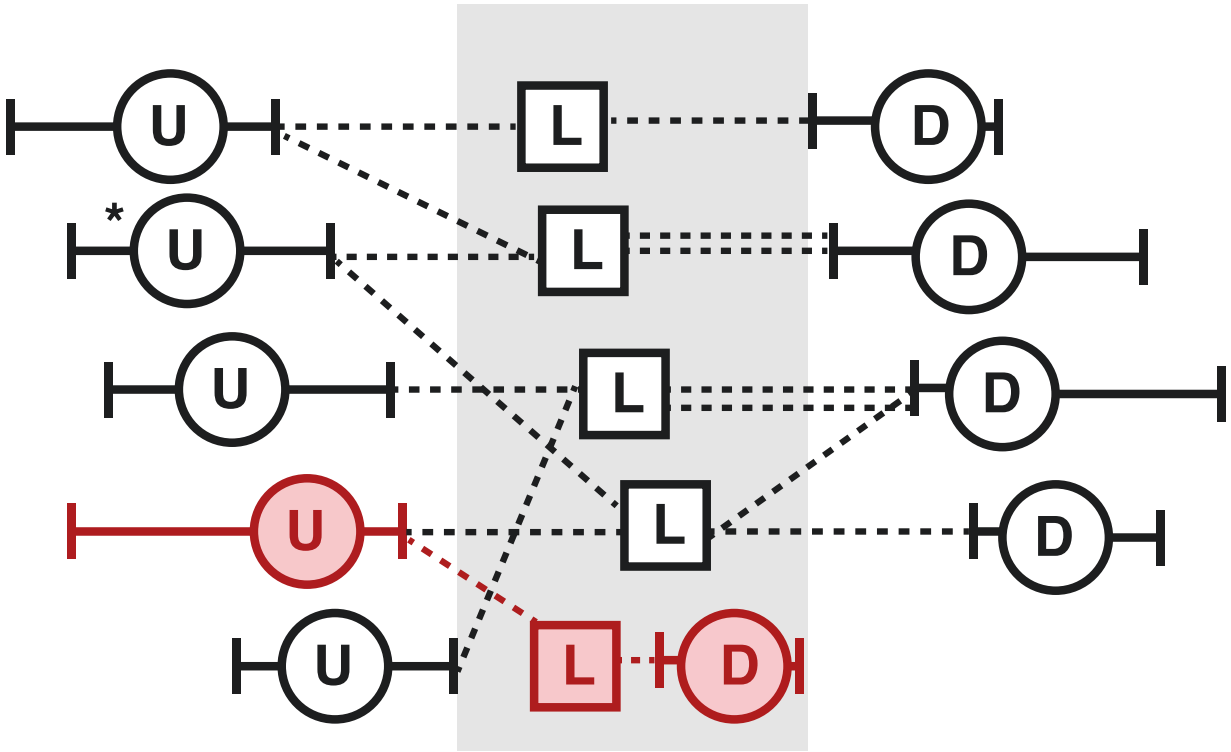

Fig. S 42

Filtering a recombination graph after finding the initial set of nodes. All upstream nodes in the set are marked with U, downstream nodes marked with D, and candidate "leaving" nodes are squares marked with L. Genomic intervals over which each node is known to exist are marked with solid horizontal lines, and dashed lines represent recombination edges (each connects an upstream node with a downstream node, through a candidate leaving node). The initial "key" node is the upstream node marked with an asterisk. After the recombination graph is gathered, the genomic interval in which the recombination event must have happened is shaded gray. One of the upstream nodes in the set has a closest downstream recombination partner node within this interval, which is not part of the downstream set (all three nodes shown in red). This upstream node therefore does not help describe the same ancestral recombination event as the other nodes and will be removed.

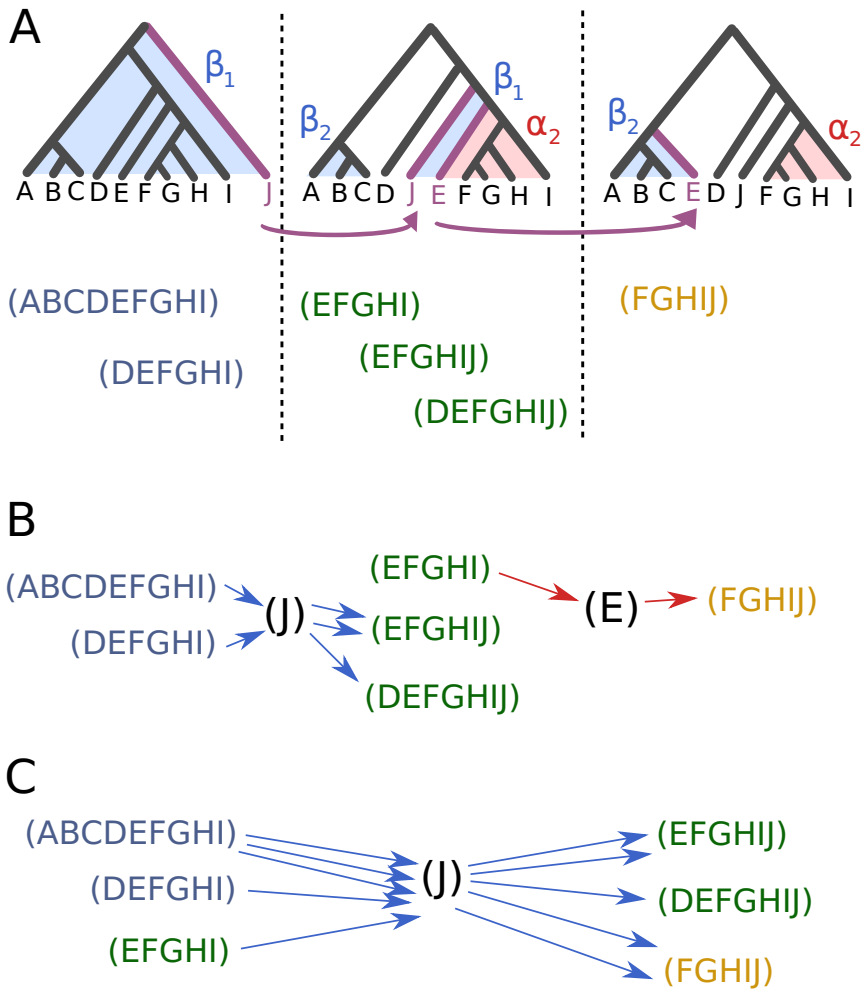

Fig. S43.

An example case in which multiple ancestral recombination events may be considered as one. A: The true ARG across three adjacent genomic regions. Clades involved in recombination are marked α and β; subscripts denote the recombination event (first or second) to which they correspond. Clades observed in SNP data appear below each tree in the order in which they are observed; colors mark the true tree to which each clade belongs. Purple branches are true γ clades, and purple arrows show ancestral recombination events. B: The correct grouping of nodes path through them in a recombination graph. First, (J) moved downward from (ABCDEFGHIJ) to (EFGHIJ). Then, (E) moved from (EFGHI) to (ABCE). C: A likely incorrect inference made, if nodes are not grouped correctly into trees. It appears most parsimonious to say that (J) moved down from (ABCDEFGHIJ) in the first tree to (FGHIJ) in the third tree, skipping the middle tree altogether. If this choice is made, genomic positions for the ancestral recombination event will also be wrong, as it chooses the narrowest possible interval, which would place it between the first and second tree. Note that observing the clade (ABCE) in the third tree might help avoid this problem.

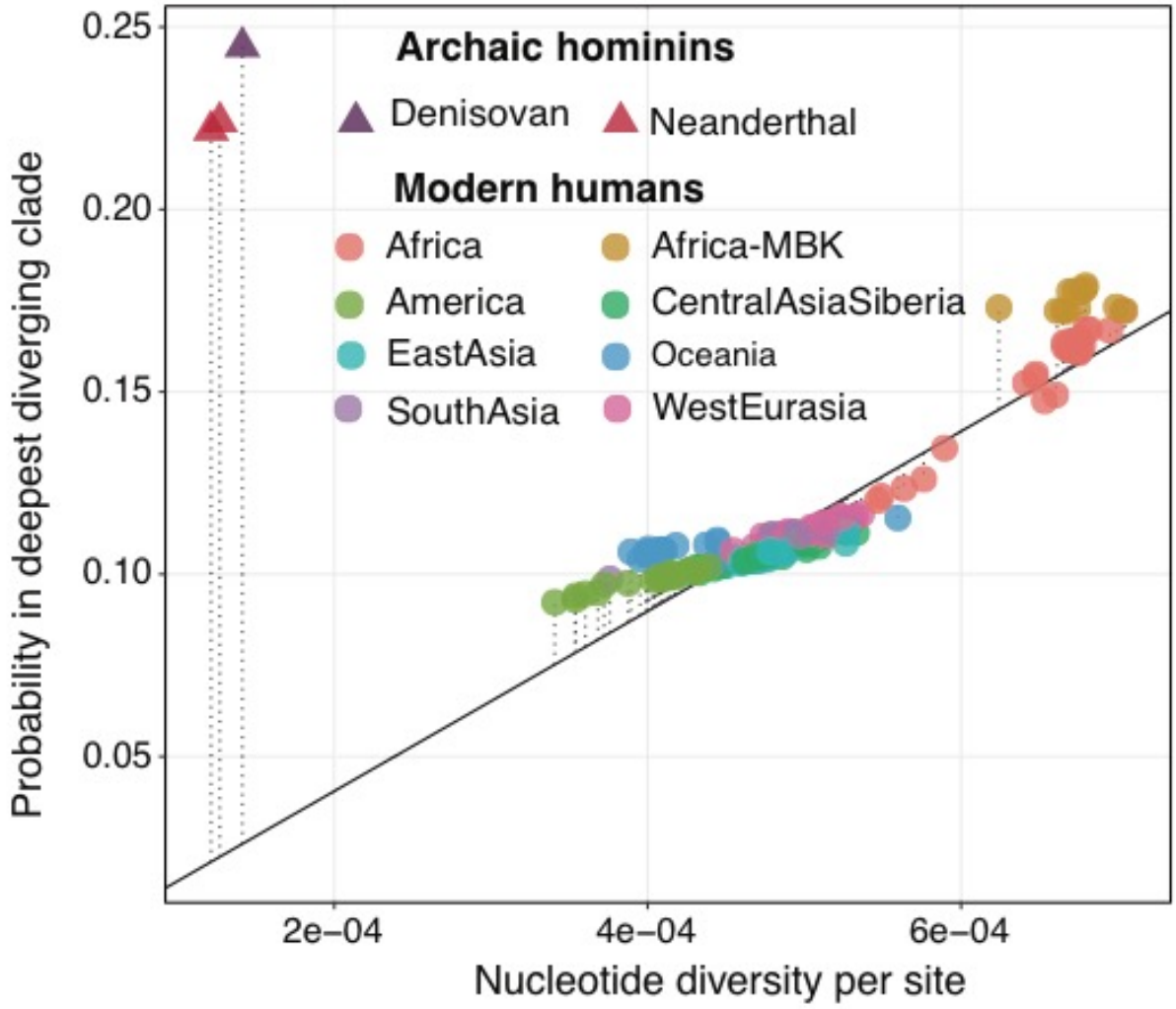

Fig. S 44

Probability of a haplotype belonging to the deepest diverging clade in a tree (defined as the smaller of the two children of the root, only when the root node is bifurcating and the two children have unequal numbers of leaves) against per-site nucleotide diversity, computed using only sites used to build the ARG (biallelic SNPs passing quality filters, where the chimpanzee allele is known, excluding CpG sites). Dotted lines are residuals to the best fit line computed excluding archaic hominins (solid line).

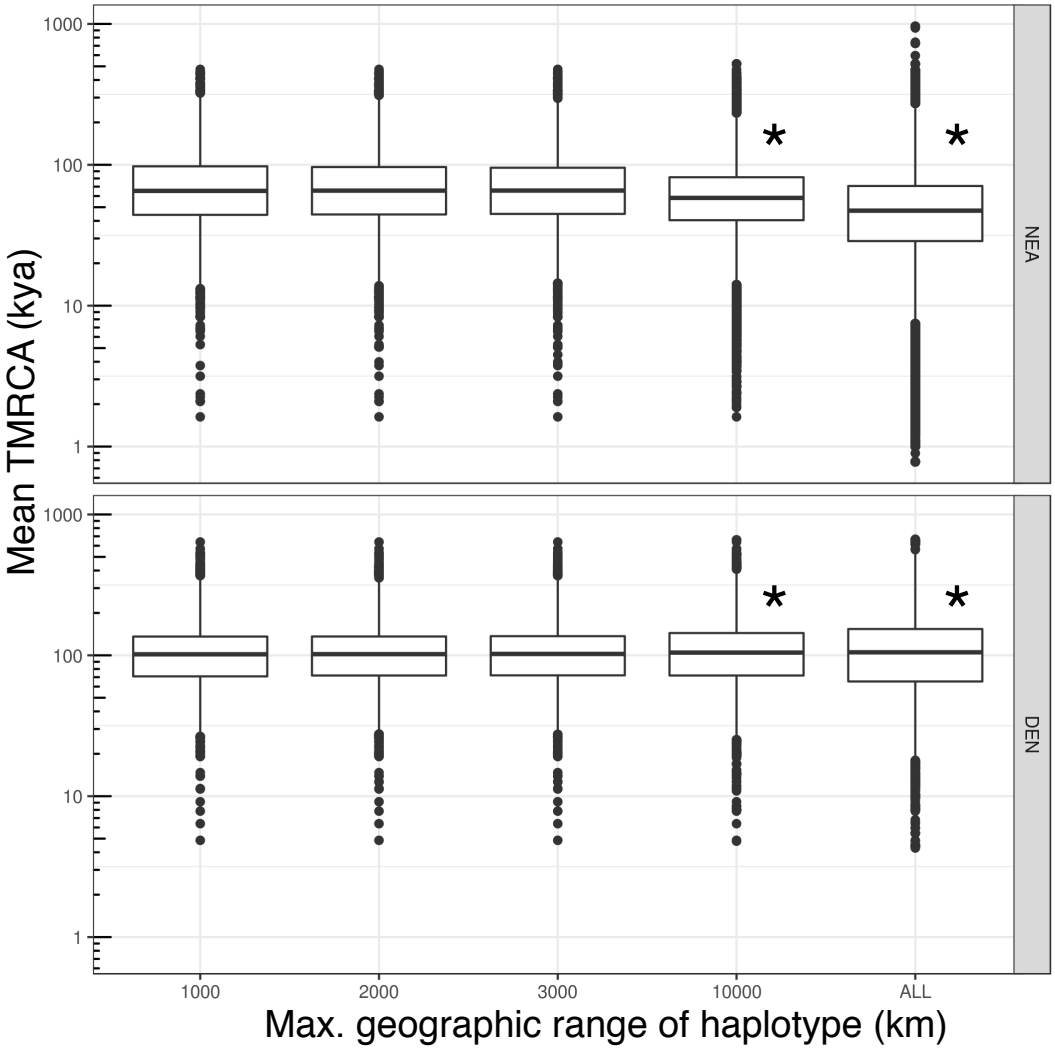

Fig. S 45

Effects of choosing different geographic range cutoffs on the distributions of TMRCAs between admixed and admixing individuals within geographically restricted Neanderthal and Denisovan introgressed haplotype blocks in modern humans. NEA = Neanderthal-introgressed haplotypes; DEN = Denisovan-introgressed haplotypes. Asterisks denote distributions significantly different (via Wilcoxon rank-sum test) from that produced using a 3,000 km cutoff (the choice made in this study).

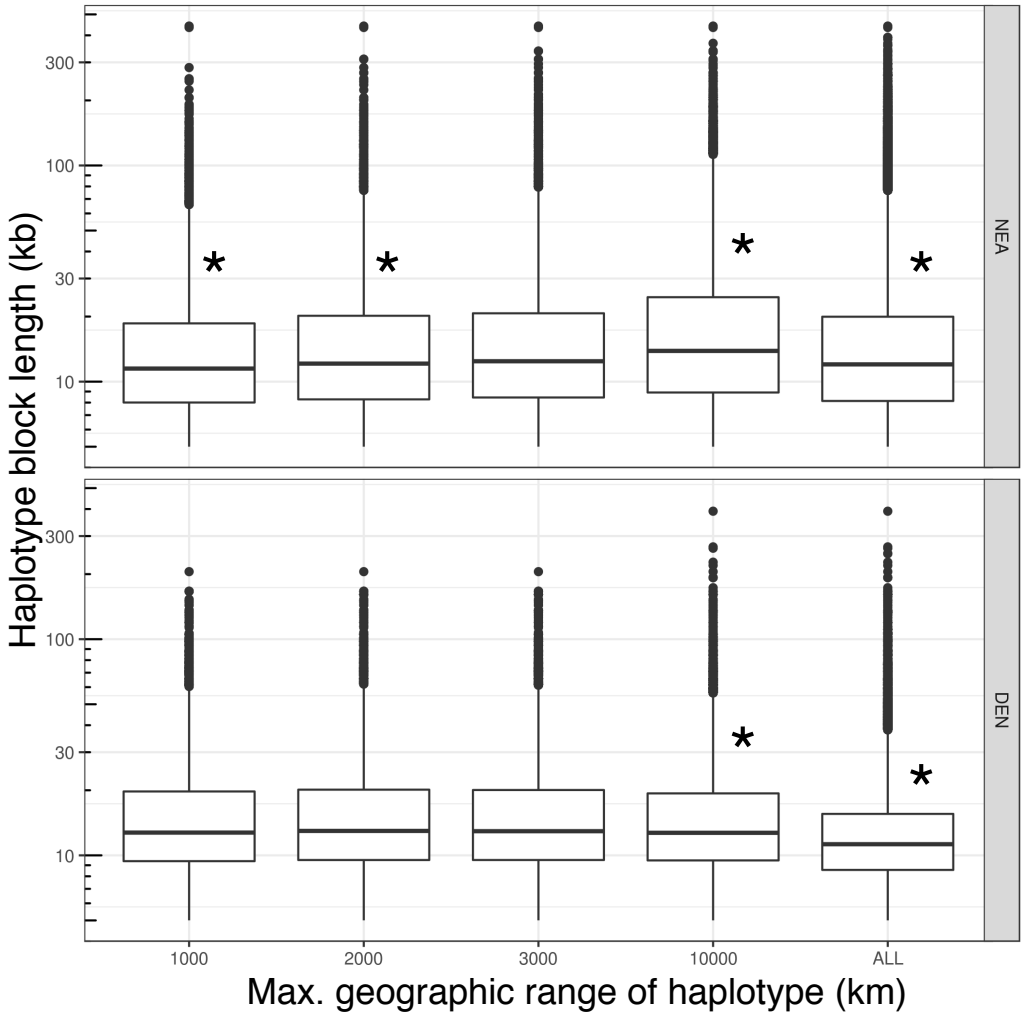

Fig. S 46

Effects of choosing different geographic range cutoffs on the distributions of lengths of geographically restricted Neanderthal and Denisovan introgressed haplotype blocks in modern humans. NEA = Neanderthal-introgressed haplotypes; DEN = Denisovan-introgressed haplotypes. Asterisks denote distributions significantly different (via Wilcoxon rank-sum test) from that produced using a 3,000 km cutoff (the choice made in this study).

Genome haplotype	Mean % error	Genome haplotype (continued)	Mean % error (continued)
Denisova-1	42.27	S_Khomani_San-1-1	3.36
Denisova-2	40.68	S_Ju_hoan_North-2-2	3.31
Altai-2	15.72	S_Biaka-1-2	3.14
Altai-1	14.95	S_Mbuti-2-1	3.10
Vindija33.19-2	12.58	S_BantuHerero-2-2	2.93
Vindija33.19-1	8.79	S_Khomani_San-2-2	2.78
S_Khomani_San-1-2	5.22	S_Mbuti-1-2	2.59
B_Ju_hoan_North-4-2	5.15	S_Ju_hoan_North-3-1	2.55
S_Naxi-2-2	4.56	S_BantuTswana-1-2	2.47
S_Ju_hoan_North-1-2	3.98	S_Biaka-2-2	2.46
S_Khomani_San-2-1	3.94	S_Mbuti-3-1	2.33
B_Mbuti-4-2	3.85	S_Biaka-2-1	2.17
B_Mbuti-4-1	3.82	S_Ju_hoan_North-1-1	2.13
S_Naxi-2-1	3.71	S_BantuHerero-2-1	2.00

Table S 1

Genome haplotypes with poor correlation between SNP-based and inferred ancestral recombination event-based similarity scores to other genomes. We created a SNP-based distance matrix by counting the number of clades that include one but not the other of each pair of genome haplotypes; we then repeated this process using clades defined by shared ancestral recombination events to build a recombination-based distance matrix. After dividing both matrices by their maximum values and subtracting them from 1 to transform them into similarity matrices, we computed the percent error between the similarity scores in both matrices. This value was $|s_{SNP} - s_{recomb}|/(0.5 * (s_{SNP} + s_{recomb}))$, where $s_{SNP}$is a similarity score from the SNP matrix and s_{recomb} is the corresponding similarity score from the recombination matrix. We report here all genome haplotypes for which the mean percent error across all columns was greater than 2%. All genomes reported here are either archaic hominins (for which there are only short read fragments and no phasing reference panels), sub-Saharan African genomes, including some from the most basal lineages, for which there is little published data and for which reference panels are poor, and the S_Naxi-2 genome, which another study reported showed signs of improper phasing .

Population	Percent Neanderthal	Percent Denisovan	Percent ambiguous	Mean frequency (Neanderthal)	Mean frequency (Denisovan)	Mean frequency (Ambiguous)
Africa	0.28% (0.15-0.80%)	0.13% (0.06-0.19%)	0.12% (0.09-0.17%)	7.7 %	8.1%	7.5%
Africa-MBK (Mbuti, Biaka, Khomani-San)	0.12% (0.10-0.16%)	0.091% (0.07-0.12%)	0.08% (0.06-0.12%)	4.5%	3.5%	3.6%
America	1.1% (0.96-1.2%)	0.091% (0.07-0.12%)	0.14% (0.12-0.18%)	9.6 %	18%	13%
CentralAsiaSiberia	1.1% (0.92-1.3%)	0.095% (0.07-0.12%)	0.14% (0.10-0.19%)	9.0 %	17%	13%
EastAsia	1.1% (0.90-1.3%)	0.10% (0.07-0.14%)	0.15% (0.12-0.18%)	8.7%	17%	12%
Oceania	1.2% (1.0-1.3%)	0.44% (0.09-0.62%)	0.44% (0.13-0.65%)	7.5%	5.5%	6.1%
SouthAsia	1.0% (0.85-1.2%)	0.092% (0.06-0.13%)	0.16% (0.11-0.20%)	7.8%	17%	11%
WestEurasia	0.97% (0.73-1.2%)	0.068% (0.05-0.10%)	0.14% (0.11-0.19%)	8.5%	20%	13%

Table S2

Demographic parameters of Neanderthal and Denisovan admixture from ARG inference. Genome-wide percents given are the percent of the autosomal genome classified as Neanderthal or Denisovan origin (or equidistant from each; “ambiguous” column), using a score cutoff (Supplementary Methods, Fig. S16). Numbers in parenthesis are minimum lower end and maximum upper end of 95% block jackknife confidence intervals across all genomes in each population. Frequencies given are calculated using only confidently admixed haplotypes and are the frequencies across all human haplotypes in the Simons Genome Diversity Project Panel.

Map type	Feature type	Projection upper p	Distance p
Admixture	genes	0.995*	0.268
Admixture	exons	0.978	0.624
Admixture	regulatory elt. binding sites	5.95e-2	0.585
ILS	genes	4.10e-05*	0.575
ILS	exons	0.999*	1.27e-05*
ILS	regulatory elt. binding sites	0.00*	5.22e-09*
Deserts	genes	0.00*	0.107
Deserts	exons	0.410	2.73e-09*
Deserts	regulatory elt. binding sites	0.00*	1.90e-09*
Deserts with human mutation	genes	0.00*	2.53e-05*
Deserts with human mutation	exons	0.00*	7.95e-06*
Deserts with human mutation	regulatory elt. binding sites	0.00*	3.20e-03*

Table S3

Overlap of genomic regions with archaic admixture in any human genome (Admixture), incomplete lineage sorting with archaic hominins in any human genome (ILS), neither admixture nor ILS with archaic hominins in any human genome (Deserts), and deserts with a fixed derived allele specific to humans (Deserts with human mutation) with other genomic features. Genes are whole protein coding genes from Gencode (*63*), using Ensembl version 94 on human genome version GRCh38 lifted over to GRCh37 coordinates. Exons are for protein-coding genes from the same annotation. Regulatory elements are from the filtered "double-elite" set in the GeneHancer database (*64*), obtained from the UCSC Genome Browser's Table Browser utility (*54*). Distance-based p-values are from the "relative distance" Kolmogorov-Smirnov test and project p-vaues measure overlap, both implemented in the GenometricCorr R package (*62*). Significant ($p < 0.01$ or $p > 0.99$) values are marked with asterisks.

p	GO ID	term
0	GO:0050775	positive regulation of dendrite morphogenesis
0	GO:0099151	regulation of postsynaptic density assembly
0	GO:0099545	trans-synaptic signaling by trans-synaptic complex
0	GO:1905606	regulation of presynapse assembly
1.45E-11	GO:0051965	positive regulation of synapse assembly
1.86E-11	GO:0099560	synaptic membrane adhesion
1.18E-10	GO:0045944	positive regulation of transcription by RNA polymerase II
1.40E-10	GO:0007185	transmembrane receptor protein tyrosine phosphatase signaling pathway
3.28E-10	GO:0030182	neuron differentiation
1.76E-09	GO:0010828	positive regulation of glucose transmembrane transport
6.08E-09	GO:0007157	heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules
7.35E-09	GO:0070413	trehalose metabolism in response to stress
7.47E-09	GO:0097105	presynaptic membrane assembly
9.96E-09	GO:0032446	protein modification by small protein conjugation
1.99E-08	GO:0043367	CD4-positive, alpha-beta T cell differentiation
3.42E-08	GO:0006796	phosphate-containing compound metabolic process
3.48E-08	GO:0007059	chromosome segregation
3.85E-08	GO:0051463	negative regulation of cortisol secretion
3.85E-08	GO:0061582	intestinal epithelial cell migration
1.02E-07	GO:0016575	histone deacetylation
1.08E-07	GO:2000773	negative regulation of cellular senescence
1.44E-07	GO:0033277	abortive mitotic cell cycle
1.49E-07	GO:0043523	regulation of neuron apoptotic process
1.52E-07	GO:0000381	regulation of alternative mRNA splicing, via spliceosome
2.02E-07	GO:0010764	negative regulation of fibroblast migration
2.43E-07	GO:0033628	regulation of cell adhesion mediated by integrin
3.48E-07	GO:1901407	regulation of phosphorylation of RNA polymerase II C-terminal domain
5.62E-07	GO:0043369	CD4-positive or CD8-positive, alpha-beta T cell lineage commitment
6.15E-07	GO:0060134	prepulse inhibition
8.72E-07	GO:0006362	transcription elongation from RNA polymerase I promoter
8.90E-07	GO:2000301	negative regulation of synaptic vesicle exocytosis
9.16E-07	GO:0002318	myeloid progenitor cell differentiation
1.01E-06	GO:0048680	positive regulation of axon regeneration
1.19E-06	GO:0015728	mevalonate transport
1.19E-06	GO:0051780	behavioral response to nutrient
1.22E-06	GO:0021707	cerebellar granule cell differentiation
1.40E-06	GO:0000122	negative regulation of transcription by RNA polymerase II
1.66E-06	GO:0008380	RNA splicing
2.77E-06	GO:0046642	negative regulation of alpha-beta T cell proliferation
3.40E-06	GO:1901673	regulation of mitotic spindle assembly
3.43E-06	GO:0070475	rRNA base methylation
6.61E-06	GO:0032922	circadian regulation of gene expression
7.49E-06	GO:0031401	positive regulation of protein modification process
1.01E-05	GO:0032528	microvillus organization
1.10E-05	GO:0060325	face morphogenesis
1.18E-05	GO:0006805	xenobiotic metabolic process
1.20E-05	GO:0035019	somatic stem cell population maintenance
2.15E-05	GO:0042593	glucose homeostasis
2.15E-05	GO:0043153	entrainment of circadian clock by photoperiod
2.95E-05	GO:0002669	positive regulation of T cell anergy
3.06E-05	GO:0007049	cell cycle
4.18E-05	GO:0090085	regulation of protein deubiquitination
5.70E-05	GO:1900424	regulation of defense response to bacterium
6.24E-05	GO:0072619	interleukin-21 secretion
6.24E-05	GO:1901256	regulation of macrophage colony-stimulating factor production
6.24E-05	GO:2001182	regulation of interleukin-12 secretion

6.61E-05	GO:0061470	T follicular helper cell differentiation
7.01E-05	GO:0045580	regulation of T cell differentiation
8.00E-05	GO:1904861	excitatory synapse assembly
0.00010737	GO:1901509	regulation of endothelial tube morphogenesis
0.000107471	GO:0030520	intracellular estrogen receptor signaling pathway
0.00011123	GO:0030099	myeloid cell differentiation
0.000112339	GO:0090050	positive regulation of cell migration involved in sprouting angiogenesis
0.000122928	GO:0032480	negative regulation of type I interferon production
0.000128903	GO:0072757	cellular response to camptothecin
0.000132611	GO:0030220	platelet formation
0.000152959	GO:0042117	monocyte activation
0.000158838	GO:0031532	actin cytoskeleton reorganization
0.000167342	GO:1905377	response to D-galactose
0.000172017	GO:0035926	chemokine (C-C motif) ligand 2 secretion
0.000213716	GO:0045655	regulation of monocyte differentiation
0.000214274	GO:0006357	regulation of transcription by RNA polymerase II
0.000242844	GO:2000646	positive regulation of receptor catabolic process
0.000248198	GO:0048609	multicellular organismal reproductive process
0.000258532	GO:0090150	establishment of protein localization to membrane
0.000267499	GO:0072711	cellular response to hydroxyurea
0.000304723	GO:0050706	regulation of interleukin-1 beta secretion
0.000315893	GO:0000303	response to superoxide
0.000380935	GO:0001525	angiogenesis
0.000394229	GO:0042921	glucocorticoid receptor signaling pathway
0.000400655	GO:0050658	RNA transport
0.000403785	GO:1902476	chloride transmembrane transport
0.000409687	GO:1902605	heterotrimeric G-protein complex assembly
0.000413288	GO:0021697	cerebellar cortex formation
0.000425503	GO:0060323	head morphogenesis
0.000426453	GO:0009887	animal organ morphogenesis
0.00043196	GO:0032784	regulation of DNA-templated transcription, elongation
0.000747324	GO:0046887	positive regulation of hormone secretion

Table S4

Significantly enriched ($p < 0.001$) Gene Ontology (*39*) biological_process terms in desert regions containing fixed human-specific derived alleles. Enrichment was tested by fetching all regions of the genome annotated with each GO term, then testing for overlap with filtered desert regions containing human-specific differences, using the projection test implemented in the GenometriCorr R package (*62*), then applying the Bonferonni correction to p-values.

Time bin (kya)	P	GO ID	term
100	0.000104124	GO:0035878	nail development
100	0.000152424	GO:0009650	UV protection
100	0.000152424	GO:0031581	hemidesmosome assembly
100	0.000209759	GO:0071391	cellular response to estrogen stimulus
100	0.000331521	GO:0008283	cell proliferation
100	0.000435455	GO:0006293	nucleotide-excision repair, preincision complex stabilization
100	0.000630144	GO:0070911	global genome nucleotide-excision repair
100	0.000630144	GO:0097186	amelogenesis
100	0.000959005	GO:0048565	digestive tract development
100	0.000959005	GO:1901796	regulation of signal transduction by p53 class mediator
300	0.00044997	GO:1904645	response to amyloid-beta
300	0.000935046	GO:0000304	response to singlet oxygen
300	0.000935046	GO:0018964	propylene metabolic process
300	0.000935046	GO:1905429	response to glycine
300	0.000935046	GO:1905430	cellular response to glycine
300	0.000935046	GO:1990771	clathrin-dependent extracellular exosome endocytosis

Table S 5

Singificantly enriched ($p < 0.001$) biological_process Gene Ontology terms attached to interacting sets of genes affected by candidate regulatory element mutations or nonsynonymous substitutions, where all such mutations occurred within the listed time bin (100 = between 100-150 kya; 300 = between 300-350 kya).

p	GO ID	term
0.000111966	GO:0038063	collagen-activated tyrosine kinase receptor signaling pathway
0.00011562	GO:0010569	regulation of double-strand break repair via homologous recombination
0.000139583	GO:0048025	negative regulation of mRNA splicing, via spliceosome
0.000149125	GO:2000327	positive regulation of nuclear receptor transcription coactivator activity
0.000156365	GO:0055059	asymmetric neuroblast division
0.000167805	GO:1902412	regulation of mitotic cytokinesis
0.000181854	GO:1901842	negative regulation of high voltage-gated calcium channel activity
0.000218898	GO:0000462	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
0.00022847	GO:0007076	mitotic chromosome condensation
0.000236442	GO:0071044	histone mRNA catabolic process
0.000270717	GO:1905064	negative regulation of vascular smooth muscle cell differentiation
0.000305361	GO:0098789	pre-mRNA cleavage required for polyadenylation
0.000313394	GO:0043928	exonucleolytic nuclear-transcribed mRNA catabolic process involved in deadenylation-dependent decay
0.000360947	GO:1990414	replication-born double-strand break repair via sister chromatid exchange
0.000362148	GO:0003192	mitral valve formation
0.000418857	GO:0010792	DNA double-strand break processing involved in repair via single-strand annealing
0.000421671	GO:0010793	regulation of mRNA export from nucleus
0.000434558	GO:0010724	regulation of definitive erythrocyte differentiation
0.000442968	GO:0040016	embryonic cleavage
0.000484611	GO:0099527	postsynapse to nucleus signaling pathway
0.000488376	GO:0006376	mRNA splice site selection
0.000488585	GO:0043983	histone H4-K12 acetylation
0.000514264	GO:0043981	histone H4-K5 acetylation
0.000514264	GO:0043982	histone H4-K8 acetylation
0.000522681	GO:0010842	retina layer formation
0.000524426	GO:0097676	histone H3-K36 dimethylation
0.00053366	GO:0045577	regulation of B cell differentiation
0.000536417	GO:0035166	post-embryonic hemopoiesis
0.000595826	GO:0000027	ribosomal large subunit assembly
0.000632034	GO:0010603	regulation of cytoplasmic mRNA processing body assembly
0.000632994	GO:1904431	positive regulation of t-circle formation
0.000643467	GO:0031124	mRNA 3'-end processing
0.000649581	GO:0097155	fasciculation of sensory neuron axon
0.000658939	GO:0021747	cochlear nucleus development
0.000767928	GO:0055113	epiboly involved in gastrulation with mouth forming second
0.000773163	GO:0070550	rDNA condensation
0.000773163	GO:1905406	positive regulation of mitotic cohesin loading
0.000777374	GO:0010711	negative regulation of collagen catabolic process
0.000777374	GO:0060311	negative regulation of elastin catabolic process
0.000778431	GO:1901630	negative regulation of presynaptic membrane organization
0.000778431	GO:1903002	positive regulation of lipid transport across blood brain barrier
0.000778431	GO:1905855	positive regulation of heparan sulfate binding
0.000778431	GO:1905860	positive regulation of heparan sulfate proteoglycan binding
0.000778431	GO:1905890	regulation of cellular response to very-low-density lipoprotein particle stimulus
0.000801461	GO:0062030	negative regulation of stress granule assembly
0.000805269	GO:0035971	peptidyl-histidine dephosphorylation
0.000812273	GO:0048024	regulation of mRNA splicing, via spliceosome
0.000822891	GO:0018323	enzyme active site formation via L-cysteine sulfinic acid
0.000822891	GO:0036471	cellular response to glyoxal
0.000822891	GO:0036526	peptidyl-cysteine deglycation
0.000822891	GO:0036527	peptidyl-arginine deglycation
0.000822891	GO:0036528	peptidyl-lysine deglycation
0.000822891	GO:0036529	protein deglycation, glyoxal removal
0.000822891	GO:0036530	protein deglycation, methylglyoxal removal

0.000822891	GO:0036531	glutathione deglycation
0.000822891	GO:0045560	regulation of TRAIL receptor biosynthetic process
0.000822891	GO:0050787	detoxification of mercury ion
0.000822891	GO:0106045	guanine deglycation, methylglyoxal removal
0.000822891	GO:0106046	guanine deglycation, glyoxal removal
0.000822891	GO:1903073	negative regulation of death-inducing signaling complex assembly
0.000822891	GO:1903122	negative regulation of TRAIL-activated apoptotic signaling pathway
0.000822891	GO:1903168	positive regulation of pyrroline-5-carboxylate reductase activity
0.000822891	GO:1903178	positive regulation of tyrosine 3-monooxygenase activity
0.000822891	GO:1903197	positive regulation of L-dopa biosynthetic process
0.000822891	GO:1903200	positive regulation of L-dopa decarboxylase activity
0.000822891	GO:2000277	positive regulation of oxidative phosphorylation uncoupler activity
0.000829029	GO:1902889	protein localization to spindle microtubule
0.000829029	GO:1990280	RNA localization to chromatin
0.000851135	GO:0002380	immunoglobulin secretion involved in immune response
0.000867097	GO:0035855	megakaryocyte development
0.000870262	GO:2000795	negative regulation of epithelial cell proliferation involved in lung morphogenesis
0.000874578	GO:0031397	negative regulation of protein ubiquitination
0.00088442	GO:0009443	pyridoxal 5'-phosphate salvage
0.000898789	GO:0072334	UDP-galactose transmembrane transport
0.000905934	GO:0032049	cardiolipin biosynthetic process
0.000916438	GO:0120049	snRNA (adenine-N6)-methylation
0.000933051	GO:0110024	positive regulation of cardiac muscle myoblast proliferation
0.000937536	GO:0045403	negative regulation of interleukin-4 biosynthetic process
0.000937536	GO:0060377	negative regulation of mast cell differentiation
0.000944871	GO:0044794	positive regulation by host of viral process
0.000948251	GO:0071348	cellular response to interleukin-11
0.0009655	GO:0045951	positive regulation of mitotic recombination
0.000970954	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay

Table S 6

Significantly enriched ($p < 0.001$) Gene Ontology (*65*) terms describing genes affected by fixed human-specific derived mutations in candidate regulatory element binding sites or nonsynonymous mutations, ranked by size of surrounding desert region and age of mutation (longer deserts and more recent mutations were ranked higher). Low p-values may indicate continued strength of purifying selection on these mutations. Testing was done using the Wilcoxon rank-order test implemented in FUNC (*68*).

Program	% missing on chrom	Median dist (kb)	Mean dist (kb)
SARGE	13.7%	3.51	91.9
Relate	35.7%	10.2	565
tsinfer	37.7%	10.5	406

Table S 7

For a single demographic simulation of humans and archaic hominins (Supplementary Methods), ARGs were inferred using SARGE, and two published programs, Relate (*7*) and tsinfer (*8*). Each time an inferred ARG contained a clade that did not exist in the true tree at the same site, the physical distance along the chromosome to the nearest site at which that clade did exist in the true ARG was computed. Incorrectly-inferred clades which did not exist anywhere in the true ARG are also shown (% missing on chrom column).