www.sciencemag.org/cgi/content/full/science.aad9416/DC1

Supplementary Materials for

Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals

Benjamin Vernot, Serena Tucci, Janet Kelso, Joshua G. Schraiber, Aaron B. Wolf, Rachel M. Gittelman, Michael Dannemann, Steffi Grote, Rajiv C. McCoy, Heather Norton, Laura B. Scheinfeldt, David A. Merriwether, George Koki, Jonathan S. Friedlaender, Jon Wakefield, Svante Pääbo,* Joshua M. Akey*

*Corresponding author. E-mail: paabo@eva.mpg.de (S.P.); akeyj@uw.edu (J.M.A.)

Published 17 March 2016 on *Science* First Release
DOI: 10.1126/science.aad9416

This PDF file includes:

Supplementary Materials

Table of Contents

Description of Island Melanesian samples

Sampling description

The 35 samples analyzed are a subset of individuals covered in fieldwork led by J Friedlaender and G Koki during 1998, 2000, and 2003 (13). The study was performed in collaboration with the Institute for Medical Research of Papua New Guinea, and its protocol was approved by the Medical Research Advisory Committee of Papua New Guinea (their IRB) as well as the Institutional Review Boards of Temple University and the University of Michigan. Informed consent was obtained from each participant. DNA was extracted from each 10 ml blood sample with the Puregene Genomic DNA Isolation Kit (Gentra Systems, Minneapolis, Minnesota).

The original study focused on the islands immediately to the east of New Guinea, the region known as Northern Island Melanesia, The 35 samples analyzed here are from the Bismarck Archepelago: primarily New Britain and New Ireland, along with two nearby smaller islands (New Hanover/Lavongai and Saint Mattias/Mussau). A genealogy and residency questionnaire was taken, including parent and grandparent names, residence, and primary language. Table S1 provides details on locations and language affiliations of the samples.

The sampling strategy presumed that contemporary genetic variation in the region is closely related to population history, as manifested in the archaeological and linguistic record, as well as to geographical proximity. For this reason, the focus was on populations speaking different non-Austronesian languages and their immediate Austronesian-speaking neighbors in different islands.

Archaeological and linguistic background

Archaeological evidence suggests that modern humans reached parts of Island Melanesia as early as 50,000 to 30,000 years ago (*24, 25, 26*). These small groups remained relatively isolated until approximately 3,300 years ago (*25*), when populations with more complex agriculture and seafaring abilities arrived in the Bismarck Archipelago from the northern coast of New Guinea. This was associated with the spread of Austronesian languages. Proto Austronesian most likely originated in Taiwan 4,000 to 5,000 years ago (*27, 28*), and Austronesian now has approximately 1,200 member languages, nearly 1/5th of the human total. Almost all Austronesian languages spoken in the Pacific belong to its Oceanic branch. Its ancestor, Proto Oceanic, developed in the Bismarck Archipelago along the north shore of New Britain (*29*), associated with an early phase of the Lapita Cultural Complex.

The non-Austronesian languages spoken in New Guinea and Island Melanesia are thought to be remote descendants of languages spoken by the earlier migrants. Unlike the Austronesian languages for which lexical methods for reconstructing proto-languages are applicable, the relationships of these more diverse languages have been more difficult to reconstruct, since they are an extremely diverse set and include a number of unclassifiable isolates (*30*). Nevertheless, a recent application of cladistics to certain grammatical features and sound systems suggests that the set of non-Austronesian languages spoken in Island Melanesia are related to one another (*31*), in what has been called the East Papuan Phylum (*32*).

The earlier STR analysis of the larger dataset showed that patterns of genetic diversity in Island Melanesia is more closely related to geographic proximity and island size than to patterns of language affiliation (*13*). The Austronesian-speaking groups are genetically indistinguishable from their immediate Papuan-speaking neighbors (for example in New Britain, the Mamusi and

the Nakanai Loso cluster closely with their Papuan-speaking Ata neighbors). These results suggest that the Austronesian languages that were brought to these islands within the last 3,300 years were adopted by many formerly Papuan speaking groups without commensurate rates of genetic admixture. Austronesian/Lapita influences were lightly inscribed on the palimpsest of pre-existing population genetic diversity (*33*).

Whole genome sequencing and data filtering

Whole-genome sequencing (WGS) was performed on approximately 1μg of genomic DNA at the New York Genome Center. Sequencing libraries were prepared using TruSeqDNA Nano 350bp kits and 150 bp paired-end reads were obtained on a HiSeq XTen sequencing platform. FASTQ files for each individual are available at dbGap. Individuals were sequenced to a median depth of 40x (range 33x-47x; Table S2).

Duplicates reads were removed using Picard (http://broadinstitute.github.io/picard/). Local realignment around indels and base quality score recalibration were performed using GATK *(34)* to generate the final bam files. High quality variant calls were obtained following GATK3.2-2 pipeline (see GATK Best Practices documentation for more details). Variants were called on each sample using GATK HaplotypeCaller and per-sample gVCFs were generated. Subsequently, joint genotyping was performed using GATK GenotypeGVCFs. Variant calls were annotated with SnpEff *(35)* and VCFtools *(36)*. After obtaining genotype calls for all 35 samples, the following filters were applied:

- Minimum base quality of 20
- Minimum mapping quality of 30
- Segmental duplications (*37*) were removed and downloaded from: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/genomicSuperDups.txt.gz

- Mappable regions were determined by examining all 35 base long "reads" that overlap a each site. A site is mappable if the majority of overlapping reads are mapped uniquely or without 1-mismatch hits to hg19 (*38*).
- CpGs were masked as in *(3)*.
- Sites within 5bp of indels were removed.
- As most analyses were done in the context of some portion of the 1000 Genomes dataset, the 1000 Genomes accessibility mask was applied, downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20141020.pilot_mask.whole_genome.bed
- We also applied the Altai and Denisovan minimal filter mask *(3)*, downloaded from: https://bioinf.eva.mpg.de/altai_minimal_filters/
- For each sample we required a minimum depth of 8 reads, and a maximum of the 99.5th percentile of autosomal depth for each sample (corresponding on average to a cutoff of 79x). These masks were merged for the 35 samples, and in the final mask a site was only considered if it passed all 35 masks.

SNP genotyping was performed on each sample using an Illumina HumanExome-12 v1.2 BeadChip, which facilitated sample tracking and validating genotypes inferred from WGS. Overall, concordance between SNP chip and WGS genotypes was >99% for each individual.

Kinship and haplotype inference

Kinship analysis was performed with KING version 1.4 *(39)* on unlinked 244,770 SNP genotypes. We identified and removed all first and second degree relatives. In deciding on which individuals to remove, we preferentially retained individuals with the highest sequence coverage.

For most of the analyses described in the main text, we used a subset of 27 unrelated individuals (Table S2).

We computationally phased the 35 PNG samples using Beagle version 4.0 *(40)*. 2504 computationally phased genomes from the 1000 Genomes Project were also used as a reference panel. Specifically, we first phased PNG sites that were also present in the 1000 Genomes dataset, using 1000 Genomes sites as a reference panel. We then merged these with the remaining Melanesian sites, which were phased using the phased Melanesian sites as a reference panel. Related individuals, including one trio, were included in this process as they share a significant number of haplotypes, and this information facilitates phasing unrelated individuals from the same population. We then merged the fully phased Melanesian VCFs with the 1000 Genomes VCFs. Any sites that were a) unmasked in either the Melanesian or 1000 Genomes datasets, and b) present in one dataset but absent in the other, were assumed to be homozygous reference.

Integration with SNP data from worldwide populations

We obtained a published population dataset genotyped on the Affymetrix Human Origin SNP Array *(15)*. The fully public curated dataset was downloaded from http://genetics.med.harvard.edu/reich/Reich_Lab/Datasets_files/EuropeFullyPublic.tar.gz. The files in PACKEDANCESTRYMAP formats, were converted to PLINK formats using the "CONVERTF" utility of ADMIXTOOLS software version 3.0 *(15)*.

The Human Origin dataset is released in a publicly curated version from which all individuals and all markers loci failing QC were previously removed, as described in Lazaridis et al. *(16)*, leaving 1,935 modern human individuals genotyped for 600,841 SNPs. The public

dataset also includes sequencing data from 5 archaic human samples (Vindija Neandertal, high coverage Altai Neandertal, low and high coverage Denisovan and the Mezmaiskaya Neandertal), 11 ancient humans, the human reference genome (hg19) and 5 Primates samples (see (*16*) for a detailed description of the Human Origin Dataset).

We next merged the Island Melanesian sequencing data with the genotyping dataset. We started by extracting positions from our multi-VCF that overlap with loci genotyped in the Affymetrix Human Origin SNP Array, whereas positions not called in the Melanesian multi-VCF were set as homozygous for the reference allele, with one exception: only variant and homozygous sites that passed our quality filters were used (see Whole Genome Sequencing and Filtering).

The so obtained Melanesian multi-VCF, filtered for Human Origins positions with high confidence, was converted to PLINK formats using v*cftools (36)*. The Island Melanesian genotypes were then merged with the Human Origin array data. After discarding sex-linked, mitochondrial and multiallelic sites and SNPs with ambiguous strand identification in the Human Origins dataset, 593,269 autosomal SNPs are left for the analysis.

We next removed populations with less than three individuals from the published dataset, and related individuals among the Melanesian samples; our final analysis data consisted of 1,964 present day humans from 170 worldwide populations, 5 archaic humans, and the chimpanzee sequencing data. A list of the samples included in the dataset is provided in Table S3, together with geographic coordinates that were used for the map in Figure 1.

This merged filtered dataset (we will refer to as "HO DATASET"), was used for subsequent analysis including Principal Component Analysis (PCA) and estimates of ancestry proportions using *f4* statistics.

PCA, Admixture, and *f4*-ratio analyses

To explore the Melanesian genomes diversity in the context of worldwide variation, we performed a Principal Component Analysis (PCA) with the merged HO DATASET (above). The original merged dataset was pruned using in PLINK (*41*) in order to avoid the effect of variants in high linkage disequilibrium, employing a window of 50 SNPs advanced by 5 SNPs and a r^2 threshold of 0.4 (--*indep-pairwise* 50 5 0.4). After removing SNPs with a missing genotype rate >= 0.05 and SNPs with MAF <= 0.05, totally 164,642 SNPs were left for the analysis.

The PCA was carried out using the R package SNPRelate (*42*) on 1,964 modern human individuals (Figure 1B main text). As shown in Figure 1B, the first component, which captures 5.79% of total variation, separates Africa from Eurasia and America (for color codes see Figure 1A and Figure S1). Populations in Middle East, Arabian Peninsula and East Africa are intermediate between the African cluster and the Eurasian cluster. The second component, which accounts for 4.49% of the variation, separates West Eurasia and South Asia from East Asia and America. Finally, our Melanesian individuals form a cluster with other Oceanian populations (Papuans from highland New Guinea, Australians and Bougainville from Nasioi) included in the reference panel.

To focus on Oceanian population affinities, we performed a PCA on 54 individuals that include Papuans (from highland New Guinea), samples from Bougainville (from Nasioi), Australians and our Melanesians (Figure S3C). We used 133,867 pruned SNPs filtered using the same approach as above. This PCA clearly resemble a geographic pattern where the first component separates the eastermost (Bougainville) and the westermost population (Papuan from New Guinea), wheres the second component differentiate our Melanesian samples from the other populations. The Melanesian cluster, that harbor the greatest internal differentiation, span across

the upper part of the plot, and it is likely to reflect the heterogeneity of the sampling locations. Among the Melanesian individuals, "Papuan" speaking groups appear to be indistinguishable from their immediate Austronesian-speaking neighbors confirming previous work based on microsatellite data (*13*).

To disentangle the archaic ancestry signal, we used a PCA projection analysis where we performed PCA on the Altai Neandertal, the Denisovan, and the chimpanzee genome, included in the merged HO DATASET using the R package SNPRelate (*42*), and projected 1,964 present-day humans onto the plane described by the top two principal components (Fig. S1). To allow a better visualization, we plotted the mean values for the top two components, for each of the 170 populations (Fig. 1C, Fig. S1). The first component describes the genetic similarity of modern humans to both archaic species, while the second component contrasts modern humans with respect to their similarity to Neanderthals and Denisovans. In particular, unlike other Eurasian populations, Oceanians appear to be closer to Denisovan, recapitulating previously reported observations (*4,9,10*).

In order to separate the signal of Denisova admixture from the signal deriving from Neandertal admixture, we computed a *f4*-ratio, a method to estimate ancestry proportions in an admixed population (*10,15*):

$$P_D(X) = \frac{f4(\text{Yoruba, Altai Neandertal; Han, X})}{f4(\text{ Yoruba, Altai Neandertal; Han, Denisovan})}$$

where *X* is a target population in the merged HO DATASET. Following previous work (*10*), we use this *f4*-ratio to estimate Denisova admixture proportions in Oceanians, using Han Chinese to correct for levels of Neandertal ancestry in Oceanians.

We converted PLINK formats to eigenstrat formats using the "CONVERTF" utility of

ADMIXTOOLS v.3 (*15*). We then obtained the *f4*-ratio estimates with the "qpF4ratio" software of ADMIXTOOLS, which computes a standard error using a weighted block jackknife for each estimated quantity (block size set to 5 cM) (*15*). $P_D(X)$ values plotted in Fig. 1D are for all Oceanic populations included in the merged HO DATASET. Estimates of $P_D(X)$, standard errors, and *Z*-scores are provided for populations in Oceania in Table S4.

All our Island Melanesians show large and significant *Z*-scores (*Z*-score ≥ 4; *15,17*) and overall ancestry proportions for Oceanian populations are consistent with previous estimates (*4,10*). Interestingly, although Baining - a "Papuan" speaking group from West New Britain - has the highest $P_D(X)$ in our samples (0.034 ± 0.004), the other "Papuan" speaking group (Ata) has a value of 0.025 ± 0.0035, less than the Austronesian average of 0.0264.

Note that we report $P_D(X)$ estimates only for Oceanian populations since not all Eurasian populations conform to the population phylogeny we assumed in our *f4*-ratio implementation. We did not compute *f4*-ratios using other African populations as outgroup, as it has been extensively shown that the use of different African groups does not produce noticeable effects on ancestry estimates in tests for archaic admixture (*4,8,10*).

Finally, we computed a *f4*-ratio in the form *f4(Yoruba, Papuan; Han, Melanesian)/(Yoruba, Papuan; Han, Papuan)* to estimate the amount of "Papuan" ancestry in our Melanesian individuals. Results are shown in Table S7. We use these estimates to estimate how much inter-individual variation in Denisovan ancestry is due to heterogeneity in Papuan ancestry among individuals.

ADMIXTURE analysis

We used the unsupervised clustering algorithm ADMIXTURE to infer ancestral clusters

in our Melanesian samples and 159 worldwide population of the Human Origin dataset. In order to avoid the effect of variants in high linkage disequilibrium, we pruned our dataset using PLINK (*41*) employing a window of 50 SNPs advanced by 5 SNPs and a r^2 threshold of 0.4 (*--indep-pairwise* 50 5 0.4). After the LD pruning, 282,101 SNPs were left for the analysis.

We ran ADMIXTURE in 10 replicates with different random seeds, with 5-fold cross-validation exploring number of clusters (K) ranging from 2 to 7. The multiple runs were then aligned using the "greedy "algorithm of CLUMPP (*43*) and visualized with the software Distruct (*44*). To provide a better visualization of the admixture proportions, a enlarged plot of our Melanesian sample is provided on the right side of Figure S1. Black lines separates distinct populations.

We find that at low values of K, the dominant ancestry component is similar among all Oceanian populations and is shared with Asian populations. At K= 5, populations in Melanesia, along with other populations in Oceania (Papuans from New Guinea, Australians and Bouganville), can be distinguished from Asian populations. For values of K > 5, populations in Oceania are charactherized for the presence of a dominant Oceanian-related ancestry component and a small proportion of ancestry shared with populations in East Asia and Siberia. This East Asian-related patch appears to be absent in Papuans from New Guinea and in the Papuan-speakers Baining in Island Melanesia. These results are highly consistent with the highest proportions (~ 74%) of "Papuan" ancestry we estimated in the Baining (Table S7) and with previous studies (*16*).

PSMC Analysis

We used the Pairwise Sequentially Markovian Coalescent (PSMC) to infer long-term effective population sizes in our Melanesian individuals (Fig. S2; *38*). The PSMC was applied to

27 unrelated Melanesians along with 11 previously published high coverage genomes (*3*,*5*). Alignments to the hg19/GRCh37 were downloaded from http://cdna.eva.mpg.de/denisova/. Diploid consensus sequences were generated for each individual using the 'pileup' command of the SAMtools software (Version: 1.2) (*45*). The following sites were marked as missing based on recommendations in (*38*):

- sites where read depth is higher than 60 or below 10. Such thresholds, that approximate 2 times and 1/3 of the average depth respectively, were chosen to account for the lower average depth in the published genomes (~30x) compared to Melanesians;
- sites where the root-mean-square mapping quality is below 10;
- sites within 5bp of a short insertion or deletion;
- sites where the estimated consensus quality is <30;
- sites where less than 18 out 35 overlapping 35-bp oligonucleotides from the human reference sequence , can be mapped elsewhere with zero or one mismatch.

The PSMC analysis was performed using default parameters. Results were scaled assuming a mutation rate of 1.25×10^{-8} per site per generation (*19*,*46*,*47*) and assuming 25 years per generation.

Quantifying ILS between Neandertal and Denisovan lineages

Neandertals and Denisovans are sufficiently closely related to each other that a lineage may be introgressed from Denisovan but may in fact be more similar to a lineage from Neandertal (or vice versa). We modeled the probability of this pattern theoretically by explicitly accounting for incomplete lineage sorting and the probability that a mutation occurs that makes a lineage closer to the species through which it was not introgressed. We begin by assuming a

simple model in which the time, in coalescent units, between the introgression from Denisova and the population split of Denisova and Neandertal is denoted t_D. The probability that the introgressed lineage and the Denisovan lineage fail to coalesce in this time period is e^{-tD}. Once in the common Neandertal/Denisovan ancestral population, the introgressed lineage will coalesce with Neandertal 1/3 of the time. Finally, for the introgressed lineage to be closer to Neandertal, we require there to be at least one mutation before coalescence, which occurs with probability $\theta/(1 + \theta)$, with $\theta = 4N_e\mu L$ where N_e is the effective size of the Neandertal-Denisova common ancestor, μ is the per base mutation rate, and L is the length of the haplotype. Thus, the probability that a lineage introgresses through Denisova but is closer to Neandertal is:

$$\mathbb{P}(\text{lineage closer to Neandertal}|\text{introgressed through Denisova}) = \frac{1}{3}e^{-t_D}\frac{\theta}{1+\theta}.$$

A similar argument can be made for the reverse case of a lineage introgressing from Neandertal and being closer to Densiova. We demonstrate how this probability varies with the parameters in Fig. S4. For example, assuming an archaic admixture event 55 thousand years ago, a Neandertal-Denisova split 400 thousand years ago, and an archaic effective size of 4500, this predicts approximately a 6% chance that a 50kb fragment is introgressed through Denisova but closer to Neandertal. If we instead assume that the introgressing Denisovan diverged from the reference Denisovan 350 thousand years ago, this probability increases to approximately 25%.

Statistical method to identify and classify archaic sequences

We extended our previously described framework to identify archaic sequences in the genomes of modern humans *(11)*. Specifically, we used a two-stage approach to first identify candidate introgressed sequences using the statistic *S** (*8,11*) and then refined this set of haplotypes by calculating a *p-value* to quantify whether a putatively introgressed haplotype

matched an archaic sequence more than expected by chance. Previously, we were only concerned with identifying sequences inherited from Neandertal ancestors *(11)*. However, the identification of archaic lineages in Island Melanesian individuals is more complicated as they are expected to segregate both Neandertal and Denisovan sequences (*5,9,10*). Thus, we extended our two-stage framework to make inferences from the bivariate distribution of archaic match *p-values*. The inference steps are described in detail below.

Step 1: Calculating *S**

We used our previously described S* framework for initially identifying putatively introgressed archaic sequence, which does not use any information about available archaic reference genomes *(11,48)*. On average, introgressed haplotypes are expected to have an older TMRCA compared to non-introgressed lineages and therefore exhibit high levels of divergence. Moreover, because admixture occurred relatively recently, the introgressed haplotype will tend to persist over sizeable genomic regions (~50 kb in the case of Neandertal introgression; *7,18*). Finally, because Neandertal admixture is expected to have occurred only in non-African populations, variants on the introgressed haplotype should not be found in African individuals (an assumption that can be relaxed to allow for the possibility of gene flow between African and non-African individuals). Thus, *S** is designed to detect divergent haplotypes whose variants are in strong linkage disequilibrium and are not found in a "reference" population. It is important to use a reference population that contains a minimal amount of archaic introgressed sequences. We chose 107 Yoruban genomes as a reference population, as levels of Neandertal variation in Yorubans are not statistically enriched (as measured by the D statistic), as opposed to Sandawe, Maasai, and African Americans *(49)*.

*S** was originally developed to work on small samples of unrelated individuals from a homogenous population [12 or 22 individuals in (18)], and to look for evidence of introgression without identifying specific introgressed haplotypes. In Vernot and Akey *(11)*, we extended *S** to identify specific introgressed haplotypes, but still to operate on sets of 20 non-African individuals at a time. As we are analyzing 35 individuals from various locations in Papua New Guinea, and ~500 individuals each from Europeans, South Asians and East Asians, we have further extended *S** to operate on a single non-African individual, and a large reference panel of African individuals (to remove ancestral variation). This simplifies the *S** for the *i*th individual in a sample to $S_i^* = \max_{J \subseteq V_i} S(J)$, where:

$$S(J) = \sum_{j \in J} \begin{cases} -10000, & d(j, j+1) > 0 \\ 5000 + bp(j, j+1), & d(j, j+1) = 0 \\ 0, & j = max(J) \end{cases}$$

Where V_i is the set of all variants in individual *i* in this region, and *J* is a subset of those variants. Variants that are also found in the reference population are not included in this analysis. In the calculation of *S()*, we treat *J* as a list of variants ordered by genomic position. Thus, variants *j* and *j+1* denote adjacent variants. The term *d(j,j+1)* represents the genotype distance between two variants, where genotypes are coded as 0, 1, and 2, and the distance between two variants is the sum of the difference between their genotype values in individual *i*. The term *bp(j,j+1)* is the distance in base pairs between two variants. In the calculation of *S(J)*: -10000 is a penalty for consecutive variants with 1-5 genotype differences; variants with no genotype differences (perfect LD) are scored 5000 + the distance between them, which gives a higher score to variants in perfect LD that extend over larger distances; the final line allows the last variant to be added. To calculate *S**, we find the set of variants *J* that maximizes *S(J)*, using an efficient dynamic

programming algorithm that allows computation of *S** in genome-wide datasets, as explained in *(11)*.

We then estimate a null distribution of *S** values by simulating sequence data using ms, under the East Asian model from Vernot et al. *(11)*. We simulate under a grid of recombination rates and population diversity (represented by number of segregating sites in the 50kb window for one non-African individual and 107 Yoruban individuals), and build a generalized linear model to the grid of *S** quantiles using the R package mgcv *(50)* as described in *(11)*. For each putative introgressed haplotype, we then use this model to estimate the *S** percentile based on the population diversity and recombination rate. We thus retain putative introgressed haplotypes with an *S** score in the 99th percentile of null simulations, obtaining an *S** callset.

Step 2: Calculating archaic match *p-values*

We now take the *S** callset for each population, which is statistically enriched for archaic sequences but has not been compared to any archaic genome, and calculate archaic match *p-values* against both Neandertal and Denisovan in a method similar to that described in *(11)*. Specifically, we first build a large database of Yoruban haplotypes by stepping every 10kb along the genome in 50kb windows, and for each haplotype in each individual, store a number of values including: a) the number of SNVs on the haplotype, b) the length of the haplotype, c) the number of SNVs shared with an archaic genome (either Neandertal or Denisovan), counting a 1 if the archaic genome is homozygous derived, and 0.5 if the archaic genome is heterozygous and one allele matches the derived allele on the modern human haplotype, d) the number of variable sites on the haplotype and in the archaic species (i.e., the total number of SNVs at which it would be possible to match the archaic), and e) the number of unmasked bases in the 50kb window. For

each *S** haplotype, we select all Yoruban haplotypes in the database with exactly the same number of SNVs, the same length (±1000bp), and the same number of unmasked bases in the region (±5000bp). We then generate an empirical distribution of the expected archaic match percentage (the number of matches to the given archaic / the number of variable sites) in a population without substantial Neandertal introgression given the characteristics of the putative introgressed haplotype, and use this distribution to obtain an empirical archaic match *p-value* (Fig. 2, Fig. S7).

Step 3: Calling and classifying archaic sequence

We developed a likelihood method, which operates on the bivariate distribution of Neandertal and Denisovan match *p-values*, and also leverages simulations with and without archaic admixture. Specifically, our method estimates the proportion of Neandertal, Denisovan and null sequence (i.e., not introgressed) in the set of *S** significant haplotypes, and converts archaic match *p-values* into posterior probabilities, allowing us to identify introgressed haplotypes at a desired FDR and probabilistically assign them the labels of "Neandertal", "Denisovan", or "Ambiguous" (archaic haplotypes that cannot be robustly distinguished as Neandertal or Denisovan).

Formally, let categories 0, 1, and 2 denote significant *S** haplotypes that are non-introgressed (false positives), introgressed from Neandertals, and introgressed from Denisovans, respectively. For the i^{th} significant *S** haplotype, let the joint density of the pair of archaic match *p-values* $p_i = (p_{iN}, p_{iD})$ be denoted as f_0, f_1, and f_2 for categories 1, 2, and 3 respectively. Note, p_{iN} and p_{iD} denote Neandertal and Denisovan match *p-values*, respectively. Furthermore, π_1 and π_2 denote the proportion of significant *S** haplotypes from categories 1 and 2, respectively. We

compute the likelihood of the data across all n significant S^* haplotypes $\{p_i, 1 \leq i \leq n\}$ given π_1, π_2, f_0, f_1, and f_2 as:

$$\prod_{i \in 1..n} (f_0(p_i)(1-\pi_1-\pi_2) + f_1(p_i)\pi_1 + f_2(p_i)\pi_2),$$

where the *p-value* densities f_0, f_1, and f_2 are obtained by simulations under a particular demographic model. We use the R package kde2d to perform 2-dimensional kernel density estimation for each category from the simulated data.

We consider a grid of values for π_1 and π_2 from 0 to 0.60 with a step of 0.005 (subject to the constraint $\pi_1 + \pi_2 \leq 1$) from and select π_1 and π_2 to maximize the likelihood of the data given a demographic model (Table S6; Fig. S5, S8).

Given a demographic model and estimate of π_1 and π_2, we can now compute the posterior probability that each putative introgressed haplotype I is drawn from the category Z_i, $p(Z_i = x)$ for $x \in \{0,1,2\}$:

$$p(Z_i = x|p_i) = \frac{f_x(p_i)\pi_x}{f_0(p_i)(1-\pi_1-\pi_2) + f_1(p_i)\pi_1 + f_2(p_i)\pi_2}$$

We then categorize haplotypes in two steps. First, we classify all haplotypes as to whether they are null (category 0) or non-null (categories 1 and 2) by selecting a threshold t_0 at which haplotypes with $p(Z_i = 0 \mid p_i) < t_0$ have an FDR ≤ 5%. Given such a threshold, and the number, N_0, of haplotypes with $p(Z_i = 0 \mid p_i)$, the FDR can be calculated as:

$$FDR = \frac{1}{N - N_0} \sum_{i=0}^{N-N_0} p(Z_i = 0|p_i)$$

The threshold t_0 is selected separately for each population such that FDR = 5%.

In the second step, we categorize haplotypes as Neandertal, Denisovan, or Ambiguous. Specifically, we called haplotype i Denisovan if $\frac{P(Z_i=2|p_i)}{P(Z_i=1|p_i)} > 2$, Neandertal if $\frac{P(Z_i=1|p_i)}{P(Z_i=2|p_i)} > 2$, and

Ambiguous otherwise. Additionally, we noticed that a small subset of haplotypes with poor matching to Neandertal or Denisovan genomes were initially categorized as non-null because they fell outside the bounds of the simulated null *p-values*, i.e., (0.99, 0.99). To remove these, we set an arbitrary threshold and require any Neandertal called haplotype to have $logit(p_{iN}) < -3$ and any Denisovan haplotype to have $logit(p_{iN}) < -3$. This corresponds approximately to a 5% FDR threshold obtained by using the R qvalue package on the univariate archaic match *p-value* distributions.

We validated our entire pipeline by running it on four African populations, as discussed below. African populations are thought to contain little if any Neandertal admixture *(20,49)*, and as expected we identified either no or very low levels of Neandertal admixture in these populations (Table S6; Fig. S8). We also ran extensive simulations of introgression into a non-African population, and calculated the false positive and true positive rates at sliding *S** thresholds (Fig. S6). We estimate that at a threshold where 50% of the *S** callset is composed of archaic haplotypes (a 25x enrichment over the genome-wide 2%), we should recover ~60% of the archaic sequence in a population (Fig. S6). These results correspond strikingly to the amount of Neandertal sequence recovered in Eurasians, in that we estimate that each *S** callset is composed of ~50% Neandertal, and we recover on average 51-65 Mb of Neandertal sequence, corresponding to 1.1-1.3% of the queryable genome, and approximately 55-60% of the total estimated Neandertal sequence in these populations. Note, the overall FDR of the final call set is much lower (5%) as a consequence of subjecting significant *S** haplotypes to further refinement by calculating archaic match *p-values*.

Identifying Neandertal and Denisovan sequences

To better understand the bivariate archaic match *p-value* distributions of introgressed and non-introgressed sequence, we simulated sequence data under a variation of standard demographic models (*11*,*21*; details below). Specifically, we simulated an African population, a non-African population, an archaic species that was sequenced, and a related archaic species that introgressed into the non-African population. We were particularly concerned about the effects of population structure in the archaic population - that is, there was some amount of divergence between the sequenced archaic individuals and the introgressing archaic population. This divergence is thought to be more extreme in the case of Denisovans than Neandertals *(3)*. To account for this, we varied the divergence times of the sequenced archaic and the introgressing archaic population, between 150kya and 350kya. For each pair (e.g., 150kya for Neandertal and 200kya for Denisovan), we ran the full likelihood method on the Melanesian 99th percentile *S** callset, estimated the proportion of the *S** callset that is introgressed from Neandertal or Denisovan, or is non-introgressed, and obtained a likelihood for the *S** callset under the demographic model under consideration (Fig. S5). In this way, we can select both the demographic model and proportions with the highest maximum likelihood.

The divergence times with the highest likelihood are 150kya for Neandertals, and 350kya for Denisovans (Fig. S5). These are somewhat older than the previously estimated divergence times *(3)*, but as any such estimates operate on ascertained introgressed haplotypes, considerable uncertainty is expected. Additionally, models with similarly high likelihoods resulted in similar callsets. For example, for the top four alternative models, between 97.7% and 99.1% of all introgressed calls in the top callset were represented, and between 0% and 0.5% of Neandertal or Denisovan calls were categorized as the opposite archaic species (Fig. S5).

We next applied our likelihood method to the 99th percentile *S** callsets for four African populations (Luhya, Gambian, Mende, Esan), along with Europeans, South Asians and East Asians, using the demographic parameters selected above. For each population, we estimated the proportion of the *S** callset from Neandertal and Denisovan (Table S6; Fig. S8). As expected, we estimated no Denisovan admixture in Europeans or in any of the four African populations. We did find small levels of Denisovan admixture in East and South Asian *S** callsets (1.0% and 0.75% respectively), but it was not a large enough proportion of the *S** callset to make a substantial number of confident calls, and this would be an interesting area of future work. We also identified substantial proportions of Neandertal admixture in European, South Asian and East Asian populations, and between 0.5% and 3% of the *S** callsets for the four African populations (Table S6). Note, this is not equivalent to 0.5%-3% Neandertal ancestry in those populations as it is conditional on significant *S** (and not all) haplotypes.

Bioinformatics analyses

For all analyses using genomic regions, either BEDOPS *(51)* or the R package GenomicRanges *(52)* were used. Where applicable, derived and ancestral state were inferred with respect to chimpanzee state in the Ensembl v64 EPO 6 primate alignment (53). Variant annotations were obtained using the SeattleSeq pipeline. Coding and transcribed regions were obtained from UCSC Genome Browser *(54)* using the RefSeq database and refFlat table *(55)*.

The spatial and temporal expression patterns of genes involved in prenatal and postnatal human brain development have been measured by the "BrainSpan: Atlas of the Developing Human Brain" consortium (http://developinghumanbrain.org). To determine whether the genes in regions depleted of archaic ancestry have specific patterns of expression during brain

development we carried out an enrichment analysis across developmental stages and brain regions using the ABAEnrichment R package (https://www.bioconductor.org/packages/release/bioc/html/ABAEnrichment.html) requiring a FWER < 0.05. We binned the developmental stages into five categories: prenatal, infant (0-2yrs), child (3-11yrs), adolescent (12-19yrs), and adult (>19yrs). For each of these, we tested each brain region as well as testing whether the genes change significantly in their expression over development. We identified an enrichment for genes in regions depleted of archaic ancestry in particular regions of the brain in three age categories (Table S10).

We used the hypergeometric test implemented in the FUNC package *(56)* to determine whether any particular Gene Ontology functional category *(57)* is over-represented among the genes in regions depleted of archaic lineages compared to genes in the rest of the queryable genome (i.e., regions that were not masked out as described above and therefore could be tested for archaic sequences. For GO enrichments we calculated the family-wise error rate (FWER) from 1000 permutations and report categories that are significant at a FWER<0.05 (Table S11). We also used WebGestalt *(58)* to test for significant enrichment of disease related genes using the default parameters.

Finally, to determine whether the genes in regions with low archaic ancestry are enriched for tissue-specific expression patterns we analyzed the Illumina BodyMap 2.0 data which provides expression data for 16 tissues *(59)*. We define as "tissue-specific" those genes that are significantly more highly expressed (DESeq *p-value*<0.05; *60*). We find no evidence for any tissue-specific expression of the genes in regions of significantly reduced archaic ancestry when examining the adult tissues represented in BodyMap.

Categorizing homozygous archaic haplotypes

We identified genomic intervals in individual genomes with archaic introgressed sequence inherited from both parents. These sequences may be homozygous (e.g. Neandertal/Neandertal) or heterozygous (e.g. Neandertal/Denisovan) for archaic-introgressed sequence, and may also contain ambiguous sequence that was identified as archaic but could not be confidently assigned to either archaic hominin species. We first used bedtools (v2.23.0; *61*) to merge over overlapping sliding windows called as introgressed from a particular archaic hominin species on a given human haplotype (Fig. S9). We then used the 'bedtools intersect' command to look for all possible combinations of overlapping archaic/archaic haplotypes within each human genome. We overlapped these archaic/archaic homozygous and heterozygous introgressed sequence intervals with the coordinates of 20,345 protein-coding genes in GENCODE (Release 19; *62*). We annotated the overlaps according to the set of archaic/archaic intervals in the region (Fig. S9). If multiple archaic/archaic combinations containing Neandertal or Denisovan sequence overlapped the same gene, the overlap was annotated as "multiple". If the set of archaic/archaic intervals were attributable to a single hominin species (or ambiguous), then the overlap was annotated as homozygous for introgression from that hominin species. Most genes overlapped a single archaic/archaic interval and simply inherited the annotation of that interval. Melanesian individuals had between 94 and 266 genes overlapped by archaic/archaic intervals (median = 130; Fig. S9), strongly correlated with the overall number of introgressed base pairs per individual genome ($r = 0.74$; *p-value* $= 1.13 \times 10^{-5}$) and exceeding amounts found in genomes of individuals from other populations. Summed across all Island Melanesia samples, genic regions homozygous for Neandertal-introgressed sequence exceeded those homozygous for Denisovan-introgressed sequence, potentially reflecting differences in the frequency spectra of introgressed

segments from each archaic hominin species (Fig. S9). We also identified a substantial number of regions in individual Island Melanesian genomes (4.5% of all gene-overlapping archaic/archaic intervals) where one haplotype can be attributed to Neandertal introgression and the other haplotype to Denisovan introgression.

Inferences and analyses of archaic deserts

To better understand the heterogeneous distribution of Neandertal introgression, and the prevalence of significantly depleted regions of introgressed sequence, we performed extensive coalescent simulations. Specifically, the goal of these simulations were to test whether the large windows depleted of introgressed sequence that we observe in the empirical data can be explained from demographic models of human history without invoking selection. Additionally, we can then use these models to identify significant depletions of introgressed sequence. Finally, we identify regions where both Neandertal and Denisovan introgression are depleted, and show that depletions of Neandertal and from Denisovan overlap significantly more than expected by chance.

Coalescent simulations

We used the coalescent simulator MaCS, which simulates genealogies spatially across chromosomes as a Markovian process *(63)*. Five demographic models, modified from previously published and accepted models, were used for simulating the demographic history of European and East Asian populations (*64-67*; MaCS commands below). In each model, an introgression event was added at 50 kya from a separate Neandertal population into a larger Eurasia population, which subsequently split into Europeans and East Asians. The level of introgression was modified for each of the models so that modern sampled European and East

Asian populations would contain 2% introgressed sequence. The Neandertal population joined with the modern human population at 700 kya. For each model we simulated a sample of 503 European and 504 East Asian individuals (2,014 total haplotypes) over 1000 independent iterations of 15Mb of simulated sequence.

We then identified true introgressed haplotypes from reported coalescent trees, as in *(11)*. In empirical analyses, we are underpowered to identify short introgressed haplotypes and recover only a percentage of all introgressed bases. Therefore, we downsampled simulated introgressed haplotypes to match the haplotype length distribution and mean percentage introgression in our European and East Asian call sets. The mean percentage of introgressed bases in a 15Mb windows for the empirical data was 1.25%, while in simulated data it was ~2% for all models. The simulated call sets were down sampled by ~63% to match the percentage of introgressed bases recovered in the empirical data. In addition, simulated introgressed haplotypes <15kb were down sampled to represent 0.57% of the total distribution, haplotypes that fell between 15kb and 30kb in size were down sampled to represent 5.6% of the total distribution, and haplotypes that fell between 30kb and 45kb were sampled to represent 20% of the total distribution.

After down sampling, we identified "deserts" of introgression in the simulated sequences as regions where, for a given model, none of the individuals contained any introgressed sequence. We also calculated the percentage of introgressed bases in windows of varying size (1-15Mb) for the simulated data to compare with the empirical data for European and East Asian samples. When varying the window size, only one window per simulated chromosome was used so that each window represented an independent simulation of a given model (i.e. having simulated 1000 iterations of 15Mb window, if we wanted to look at percent introgression in only a 1Mb window, we only took the first 1Mb of the original 15Mb simulation).

Comparison to European and East Asian Neandertal callset

We then scanned our Neandertal introgression callset from 503 Europeans and 504 East Asians in windows from 1Mb to 15Mb, with a 100kb step, and similarly calculated the average amount of introgressed sequence in each window. We additionally require that at least 90% of each window is callable given the filters described above. For this dataset and for the above simulations, we counted the number of windows with average percentage introgression lower than $10^{-3.5}$, for each window size (Fig. 4A). Generally, we observe more depletions of Neandertal sequence in real data than in simulations. To estimate significance, we resample 5000 times from the simulated data, first correcting for the fact that we use sliding windows in real data by sampling a number of windows equal to: (total size of genomic regions considered) / (window size). For example, for 10Mb windows we consider 2418.8Mb of genomic sequence, and we resample 241 windows at a time from the simulated data. From these resamples, we can calculate an empirical *p-value* for the significance of the observed number of depleted windows (Fig. 4A). At window sizes 8Mb and larger, we see significantly more windows depleted of Neandertal sequence as compared to simulations.

We next identified regions 10Mb or larger and significantly depleted of Neandertal sequence in Europeans, East Asians, South Asians and Melanesians, at a threshold of $10^{-3.5}$; these regions total 85.3Mb of the genome (Table S8; Fig. S19). We then compared these regions to patterns of Denisovan introgression in Island Melanesians, by identifying similar large depletions of Denisovan sequence. A complicating factor in this comparison is the relatively small number of Island Melanesian individuals. This results in many large regions with no identified Denisovan introgression, the largest of which is 21.8Mb, totaling 253Mb of the genome. However, by strictly considering windows with no Denisovan introgression, we would be

unnecessarily splitting up large regions of depletion but with a small number of false positive calls. Thus, we considered a threshold of 0.0001 average Denisovan introgression (2.1% of all windows 10Mb or larger are significant at this threshold) in regions of 10Mb or larger - totaling 356.6Mb of the genome. The 85.3Mb of Neandertal depletion and 356.6Mb of Denisovan depletion overlap by 47.8Mb - over four regions of 10Mb or larger (Table S9). It is interesting to note that these overlaps are all larger than 10Mb, and not partial overlaps.

We next employed a "sliding genome" permutation algorithm to estimate the significance of the overlap between Neandertal and Denisovan depletions. We first collapsed the genome by merging all chromosomes and removing uncallable regions, shifting the genomic positions of each Neandertal and Denisovan depleted region appropriately. We advanced the positions of the Denisovan deserts by 1Mb steps, and for each step calculated the overlap between Neandertal and Denisovan depletions (Figure S20). Using this distribution, we find that the 47.8Mb of overlapping Neandertal and Denisovan depletions is significantly larger than random overlaps (empirical *p-value* = 0.0008).

Background selection and archaic depletions

It has previously been reported that genomic regions experiencing background selection, as measured by B-values (*68*), have lower levels of introgression (*12*). This could be due to increased purifying selection against Neandertal sequence, Neandertal sequence being removed due to increased drift in these regions, or biases against detecting Neandertal sequence. To determine the impact of background selection on depletions of archaic sequence, we compared B-values across introgressed and non-introgressed sequence, and in desert and non-desert sequence.

Although there is a small shift towards higher B-values in introgressed sequence (Fig. S22), we identify introgressed sequence across all B-values, suggesting that there is not a large bias against detecting sequence in regions experiencing. Nonetheless, to further explore whether stronger background selection influences the probability of observing archaic deserts, we considered a range of demographic models of varying N_e. Specifically, we performed our simulations under a mixture of demographic models, which include N_e during and after the time of introgression ranging from 1861-7700. [2100, 2758, 1861, 1861, and 7700 for (*66*), (*65*) with exons, (*65*) with low coverage + exons, (*64*), and (*67*), respectively). We find that similar to our pooled results (Fig. 4A), there is an excess of observed deserts compared to all models individually, indicating that these results are robust to large variations in N_e (Fig. S18).

The B-value distributions of archaic depletions are still of interest, as they may indicate higher levels of selection against archaic haplotypes. Indeed, two of the four depletions shared in all populations have lower mean B-values than genomic sequence, with the Chromosome 7 depletion having the lowest mean B-value (705.6), 11% lower than the genomic average (792.3) (Fig. S21). It is worth noting that this 11% reduction in B-values is much less than the four-fold range of simulated N_e discussed above.

Coalescent simulations for five standard models: MaCS commands

Tennessen model *(64)*:

```
macs 2025 15000000 -s ${RANDOM}${SGE_TASK_ID} -i 10 -r 3.0e-04 -t 0.00069 -T -I 4 10 1006 1008 1 0 -n 4 0.205 -n 1 58.00274 -n 2 70.041 -n 3 187.55 -eg 0.9e-10 1 482.46 -eg 1.0e-10 2 570.18 -eg 1.1e-10 3 720.23 -em 1.2e-10 1 2 0.731 -em 1.3e-10 2 1 0.731 -em 1.4e-10 3 1 0.2281 -em 1.5e-10 1 3 0.2281 -em 1.6e-10 2 3 0.9094 -em 1.7e-10 3 2 0.9094 -eg 0.007 1 0 -en 0.007001 1 1.98 -eg 0.007002 2 89.7668 -eg 0.007003 3 113.3896 -eG 0.031456 0 -en 0.031457 2 0.1412 -en 0.031458 3 0.07579 -eM 0.031459 0 -ej 0.03146 3 2 -en 0.0314601 2 0.2546 -em 0.0314602 2 1 4.386 -em 0.0314603 1 2 4.386 -eM 0.0697669 0 -ej 0.069767 2 1 -en 0.0697671 1 1.98 -en 0.2025 1 1 -ej 0.9575923 4 1 -em 0.06765 2 4 32 -em 0.06840 2 4 0
```

Gravel low coverage + exon model *(65)*:

macs 2025 15000000 -s ${RANDOM}${SGE_TASK_ID} -i 10 -r 3.0e-04 -t 0.00069 -T -I 4 10 1006 1008 1 0 -n 4 0.205 -n 1 2.12 -n 2 4.911 -n 3 6.703 -eg 1.0e-10 2 111.11 -eg 1.1e-10 3 140.35 -em 1.2e-10 1 2 0.731 -em 1.3e-10 2 1 0.731 -em 1.4e-10 3 1 0.228 -em 1.5e-10 1 3 0.228 -em 1.6e-10 2 3 0.9094 -em 1.7e-10 3 2 0.9094 -eG 0.031456 0 -en 0.031457 2 0.1412 -en 0.031458 3 0.07579 -eM 0.031459 0 -ej 0.03146 3 2 -en 0.0314601 2 0.2546 -em 0.0314602 2 1 4.386 -em 0.314603 1 2 4.386 -eM 0.0697669 0 -ej 0.069767 2 1 -en 0.0697671 1 1.98 -en 0.2025 1 1 -ej 0.9575923 4 1 -em 0.06765 2 4 32 -em 0.06840 2 4 0

Gravel low coverage model *(65)*:

macs 2025 15000000 -s ${RANDOM}${SGE_TASK_ID} -i 10 -r 3.0e-04 -t 0.00069 -T -I 4 10 1006 1008 1 0 -n 4 0.205 -n 1 2.12 -n 2 4.911 -n 3 6.703 -eg 1.0e-10 2 78.95 -eg 1.1e-10 3 90.64 -em 1.2e-10 1 2 0.491 -em 1.3e-10 2 1 0.491 -em 1.4e-10 3 1 0.1696 -em 1.5e-10 1 3 0.1696 -em 1.6e-10 2 3 1.725 -em 1.7e-10 3 2 1.725 -eG 0.03826 0 -en 0.03827 2 0.2216 -en 0.03828 3 0.1123 -eM 0.03829 0 -ej 0.03830 3 2 -en 0.03831 2 0.3773 -em 0.03832 2 1 5.848 -em 0.03833 1 2 5.848 -eM 0.1340 0 -ej 0.1341 2 1 -en 0.1342 1 2.105 -en 0.4322 1 1 -ej 0.9575923 4 1 -em 0.06765 2 4 32 -em 0.06840 2 4 0

Gutenkunst model *(66)*:

macs 2025 15000000 -s ${RANDOM}${SGE_TASK_ID} -i 10 -r 3.0e-04 -t 0.00069 -T -I 4 10 1006 1008 1 0 -n 4 0.205 -n 1 1.685 -n 2 4.4 -n 3 8.6 -eg 1.0e-10 2 116.8 -eg 1.1e-10 3 160.6 -em 1.2e-10 1 2 0.876 -em 1.3e-10 2 1 0.876 -em 1.4e-10 3 1 0.5548 -em 1.5e-10 1 3 0.5548 -em 1.6e-10 2 3 2.8032 -em 1.7e-10 3 2 .8032 -eG 0.0290 0 -en 0.02901 2 0.1370 -en 0.02902 3 0.06986 -eM 0.02903 0 -ej 0.02904 3 2 -en 0.0290401 2 0.2877 -em 0.0290402 2 1 7.3 -em 0.0290403 1 2 7.3 -eM 0.19149 0 -ej 0.1915 2 1 -en 0.191501 1 1.685 -en 0.3014 1 1 -ej 0.9575923 4 1 -em 0.06774 2 4 34 -em 0.06849 2 4 0

Schaffner model *(67)*:

macs 2025 15000000 -s ${RANDOM}${SGE_TASK_ID} -i 10 -r 3.0e-04 -t 0.00075 -T -I 4 10 1006 1008 1 0 -n 4 0.205 -n 1 8 -n 2 8 -n 3 8 -em 1.2e-10 1 2 1.6 -em 1.3e-10 2 1 1.6 -em 1.4e-10 3 1 0.4 -em 1.5e-10 1 3 0.4 -en 0.004 1 1.92 -en 0.007 2 0.616 -en 0.008 3 0.616 -en 0.03942 2 0.0574 -en 0.03998 2 0.616 -en 0.038 3 0.058 -en 0.03997 3 0.616 -eM 0.03999 0 -ej 0.040 3 2 -en 0.04001 2 0.616 -en 0.0686 2 0.032 -en 0.0696 1 0.0996 -ej 0.07 2 1 -en 0.07001 1 1.92 -en 0.34 1 1 -ej 0.56 4 1 -em 0.04002 2 4 29 -em 0.04077 2 4 0

Statistical method to identify distinct admixture pulses

For every pair of modern human populations represented in our data, we sought to ask whether they shared a common archaic admixture event or if distinct admixture histories for each

population were required to explain the data. As a way of mitigating power issues attributable to variation due to biological factors such as coalescence times and mutation rates, as well technical factors such as sequenceability, across the genome, we developed an approach in which we identified archaic sequence in a given individual in one population, and then asked for the probability that another individual in a different population shares that archaic sequence. We first motivate this approach theoretically and then explain how we applied it to the observed data.

Theory: single, shared admixture

Consider a model as shown in Fig. S10A. The time between when the two populations join together and the archaic admixture event is t_1, and the time between the admixture event and when the archaic population splits from modern humans is t_2. A lineage is introgressed at the admixture event with probability f. We wish to compute the probability that a locus is introgressed in an individual from population j given that it is introgressed in an individual from population i, which we denote as $P(j|i)$. By the definition of conditional probability, we have that

$$\mathbb{P}(j|i) = \frac{\mathbb{P}(j\&i)}{\mathbb{P}(i)},$$

where P(j&i) denotes the probability that the locus is introgressed in both the individual from population j and the individual from population i, and P(i) is the marginal probability that it is introgressed in the individual from population i. There are two mutually exclusive ways for both individuals to be introgressed:

1. The individuals coalesce in time t_1 with probability $1- e^{-t_1}$, and then the single lineage is introgressed with probability f. We are only capable of calling an introgressed fragment if it coalesces with the archaic lineage more recently than the split time of the archaic population and

modern humans, so we additionally require a coalescence with the archaic lineage during time t_2, an event which occurs with probability $1- e^{- t_2}$.

2. The individuals fail to coalesce in time t_1 with probability $e^{- t_1}$, and then both lineages are introgressed with probability f^2. We then require all three lineages to coalesce in time t_2, which occurs with probability $1- \frac{3}{2}e^{- t_2} + \frac{1}{2}e^{- 3t_2}$.

Putting these together, we have

$$\mathbb{P}(j\&i) = f\left(1 - e^{-t_1}\right)\left(1 - e^{-t_2}\right) + f^2 e^{-t_1}\left(1 - \frac{3}{2}e^{-t_2} + \frac{1}{2}e^{-3t_2}\right).$$

For just the individual from population i to be called introgressed simply requires it to be introgressed with probability f and coalesce with the archaic lineage with probability $1- e^{- t_2}$. Therefore:

$$\mathbb{P}(i) = f\left(1 - e^{-t_2}\right).$$

Thus, the conditional probability is:

$$\mathbb{P}(j|i) = 1 - e^{-t_1} - f\frac{1}{2}e^{-(t_1+2t_2)}\left(1 + e^{t_2} - 2e^{2t_2}\right).$$

For $f \ll 1$, we have

$$\mathbb{P}(j|i) \approx 1 - e^{-t_1},$$

i.e. that the conditional probability is mostly determined by whether the two lineages coalesce in time t_1. Note that because this situation is symmetric, it is the case that P(j|i) = P(i|j). Thus, by comparing reciprocal probabilities of Neandertal ancestry among individuals from different populations, we may be able to reject a single, ancestral admixture if we find that they are not equal.

Theory: shared ancestral admixture followed by population specific admixture

Consider a more complicated model, shown in Fig. S10B. Here, in addition to the model as before of an admixture event of intensity f_2 occurring t_1 time units more anciently than the divergence time of the two populations, we have another admixture event of intensity f_2 into population *i* occurring t_3 time units more recently than the shared admixture event. This situation is no longer symmetric, and we expect more admixture into population *i*. To compute the joint probability that both the individual from population *i* is introgressed and the individual from population *j* is introgressed, there are now three possibilities:

1. The lineage from population *i* is introgressed with probability f_1 and coalesces with the archaic lineage more recently than the shared inrogression with probability $1- e^{- t_3}$. Then, the lineage from population *j* is introgressed with probability f_2, and coalesces with the archaic lineage with probability $1- e^{- t_2}$. In total, this event has probability $p_1 = f_1(1 - e^{-t_3})f_2(1 - e^{t_3})$.

2. The lineage from population *i* is introgressed with probability f_1, and fails to coalesce with the archaic lineage more recently than the shared introgression, an event with probability $e^{- t_3}$. The lineage from population *j* is then introgressed with probability f_2 and all three lineages coalesce with probability $1- \frac{3}{2}e^{- t_2} + \frac{1}{2}e^{- 3t_2}$. This event has total probability $p_2 = f_1 e^{-t_3} f_2(1 - \frac{3}{2}e^{-t_2} - \frac{1}{2}e^{-3t_2})$.

3. The lineage from population *i* is not introgressed, which occurs with probability $1 - f_1$. Then, we can use the calculation for the single, shared ancestral admixture after making the substitution

$f \rightarrow f_2$. The total probability is then $p_3 = (1 - f_1)(f_2(1 - e^{-t_1})(1 - e^{t_2}) + f_2^2 e^{-t_1}(1 - \frac{3}{2}e^{-t_2} - \frac{1}{2}e^{-3t_2}))$.

Thus, the complete probability that both the sample from population *i* and the sample from population *j* are introgressed is:

$$P(i\&j) = p_1 + p_2 + p_3.$$

Because the model is no longer symmetrical, P(*i*) ≠ P(*j*), we compute each in turn. For the probability that a lineage from population *i* is called introgressed, we must consider two events:

1. The lineage introgresses with probability f_1 and then coalesces with the archaic lineage in time $t_3 + t_2$, which occurs with probability $1 - e^{-(t_3+t_2)}$.

2. The lineage fails to introgress with probability $1 - f_1$, introgresses in the ancestral population with probability f_2, and then coalesces with the archaic lineage with probability $1 - e^{-t_2}$.

Thus,

$$\mathbb{P}(i) = f_1\left(1 - e^{-(t_3+t_2)}\right) + (1 - f_1)f_2\left(1 - e^{-t_2}\right).$$

For the lineage from population *j* to be introgressed, it must introgress in the ancestral population with probability f_2 and then coalesce with the archaic lineage with probability $1 - e^{-t_2}$. Hence,

$$\mathbb{P}(j) = f_2\left(1 - e^{-t_2}\right).$$

We can now compute the reciprocal probabilities, although the expressions are unwieldy. Importantly, we can easily show that P(*i*|*j*) ≥ P(*j*|*i*), *i.e.* that the probability of an introgressed sequence being in the population with more admixture given that it is in the population with less admixture is always at least as big as the probability that an introgressed sequence is in the population with more admixture given that it is in the population with less admixture. To see

this, note that:

$$\begin{aligned}\frac{\mathbb{P}(i|j)}{\mathbb{P}(j|i)} &= \frac{\frac{\mathbb{P}(i\&j)}{\mathbb{P}(j)}}{\frac{\mathbb{P}(i\&j)}{\mathbb{P}(i)}} \\ &= \frac{\mathbb{P}(i)}{\mathbb{P}(j)} \\ &= 1 - f_1 + \frac{f_1}{f_2}\left(\frac{1 - e^{-(t_3+t_2)}}{1 - e^{-t_2}}\right).\end{aligned}$$

The term in parentheses is the relative probability that a lineage coalesces with an archaic lineage in time $t_3 + t_2$ compared to time t_2. Thus, it is always greater than 1. Because of this and the fact that $0 \leq f_2 \leq 1$, the term multiplying f_1 is always greater than 1. Hence, the ratio is always greater than 1 so long as f_1 is non-zero.

Calculating reciprocal match probabilities in real data

We initially identified introgressed haplotypes in sliding windows. However, each window does not represent an independent admixture. Therefore, we used bedops to merge adjacent haplotypes on a per chromosome basis. This resulted in a dataset in which each chromosome contained merged haplotypes. We then computed match probabilities by asking how many merged haplotypes were present in a given individual, and how many of those overlap by at least one base with a haplotype in a different individual. The match probability is simply the ratio of those two numbers. We computed the reciprocal match probabilities by doing this in both directions for every pair of individuals analyzed. For computational efficiency, we restricted our analysis to at 30 randomly sampled haplotypes per population (27 in the case of Island Melanesians as this is the number of unrelated individuals). In order to ensure that our results were robust to this sampling, we repeating the random subsampling 10 times for each of the continental population comparisons.

To determine if there was a statistically significant difference in the reciprocal match probabilities for a given population comparison, we note that under the null hypothesis of a single, shared admxiture event, the conditional match probabilities in each population should be equal. Hence, we used a binomial test of the null hypothesis that the distribution of $log(P(\text{Pop 1} \mid \text{Pop 2})/P(\text{Pop 2} \mid \text{Pop1}))$ is centered around zero. Specifically, we computed the number of comparisons for which P(Pop 1 | Pop 2) > P(Pop 2 | Pop 1) and asked if it was consistent with a binomial distribution with $p = 0.5$.

To ensure that our match probabilities in Island Melanesians were robust to our ability to correctly separate loci introgressed from Neandertal from loci introgressed from Denisovan, we used three different cutoffs. In the first, we used the haplotype calls reported in the majority of analyses, in which a sequence is called Neandertal if $\frac{P(Z_i=1|p_i)}{P(Z_i=2|p_i)} > 2$ (i.e., ratio of posterior probabilities of being Neandertal versus Denisovan is larger than two). This results in highly significant differences among continental populations (Fig. S11). However, as this subset of archaic sequences does not include any ambiguously labeled archaic haplotypes, it is potentially missing a substantial amount of Neandertal sequence in Island Melanesians. Thus, in a second approach, we took all significant archaic sequences called that had a higher Neandertal, compared to Denisovan, posterior probability (i.e., $PP_N - PP_D > 0$, where PP_N and PP_D denote posterior probabilities of an archaic sequence being Neandertal or Denisovan, respectively). Although this is a conservative analysis (as this set of archaic sequences likely contains a substantial amount of Denisovan sequence in Island Melanesians), we still observe highly significant results among continental populations (Figs. 3B) consistent with an additional pulse of Neandertal admixture in European, East Asians, and South Asians compared to Melanesians. Finally, as a positive control, we also calculated reciprocal match probabilities using all archaic

haplotypes in Island Melanesians, and as expected, found additional admixture events in Melanesians, corresponding to the Denisovan admixture that is not present in mainland Eurasians (Fig. S12). In addition to testing differences in admixture histories between continental populations (Fig. 3, Fig. S7), we also compared South Asian (Fig. S14), East Asian (Fig. S15), and European (Fig. S16) populations.

Finally, we evaluated non-parametric approaches to evaluating statistical significance of reciprocal match probabilities. The rationale of exploring non-parametric approaches is that individuals within a population are correlated due to their shared evolutionary history, and thus a binomial *p-value* might be anti-conservative. In order to assess the magnitude of potential bias and ensure interpretations from the binomial test were robust, we recalibrated the *p-values* using a permutation approach. Specifically, for each of the comparisons between continental populations, we permuted population labels 200 times and repeated our procedure of computing overlap on a random subset of 30 individuals from each population and computing a binomial *p-value*. The distributions of *p-values* obtained from permutations showed that the binomial test was modestly anti-conservative (Fig. S13). To compare the distribution of *p-values* generated from the permutations to the *p-values* observed in the real data, we used the mean log *p-value* from each of the 10 random subsamplings for each population comparison. In these analyses, we focused on the set of conservatively defined Neandertal sequences (i.e., those with $PP_N - PP_D > 0$).

Our inference of admixture histories is consistent with previous work suggesting that Melanesian populations diverged from Eurasians more anciently than the split between European and East Asian populations (*69-71*).

Detecting signatures of adaptive introgression

To identify putative substrates of adaptive introgression, we first identified high frequency archaic haplotypes in Melanesians. Because we identified introgressed sequence on a per individual basis, we needed to merge haplotype calls across individuals in order to determine population frequencies. To this end, we calculated r^2 using *vcftools (36)* to quantify linkage disequilibrium between all pairs of tag SNPs identified in Melanesians. To be included in this analysis, we required that tag SNPs match the correct archaic sequence, belong to a 50kb window with at least two other matching tag SNPs, and be within +/- 5% the median tag SNP frequency of its window. All SNPs at $r^2 > 0.1$ were initially clustered in to one large haplotype. Within this larger haplotype, we clustered SNPs again based on an r^2 cutoff of 0.3 in to smaller, distinct haplotypes, and the median derived allele frequency of all tag SNPs in each small haplotype was used to approximate the haplotype frequency. For subsequent analyses, we retained only the small haplotype with the maximum allele frequency within each larger haplotype. This allowed us to ensure that all high frequency haplotypes reported truly are independent of each other. In an effort to accurately establish the ends of high frequency haplotypes, regions reported in Table S12 were extended to include the coordinates of the left-most and right-most variants in LD ($r^2 > 0.8$) with tag SNPs.

We performed extensive coalescent simulations to determine how unusual these haplotypes are under neutral models of evolution. Simulations were performed using ms *(72)*. We used a base demographic model *(11)* and varied several parameters. The model consists of: a) splitting between modern humans and Neandertals/Densiovans at 700kya, b) Neandertal/Denisovan N_e of 1500 b) African ancestral population size of 7310, c) splitting between Africans and non-Africans at 95kya. d) A non-African N_e of 2k, 4k, 6k, or 8k, e) a

single 500 year introgression event from Neandertals/Denisovans in to modern humans. We simulated two scenarios for introgression: in the first, introgression occurs 85kya, followed by splitting of Melanesian populations from Europeans at 80kya. In the second, introgression occurred 55kya, and splitting of Melanesians and Europeans occurred at 40kya. We also varied across introgression rates of 7.5×10^{-4}, 1.5×10^{-3}, and 2.25×10^{-3}, f) European and Melanesian N_e were set to the original non-African N_e after splitting, and grew gradually to 10000 at 5115ya, g) Rapid population growth starting 5115 years ago, reaching 500,000 in Melanesians, 512,000 in Europeans, and 424,000 in Africans, h) migration rates were set as follows: 1.498975×10^{-4} between Africans and the ancestors of Europeans and Melanesians, 2.498291×10^{-5} between Africans and Europeans, 7.794668×10^{-6} between Africans and Melanesians, and 3.107874×10^{-5} between Europeans and Melanesians. These parameters were adapted from *(11)* to reflect uncertainty in the population history of Island Melanesians.

We ran 20000 simulations of a single locus across each combination of parameters for a total of 24 distinct models and 480,000 simulations. The 99^{th} percentile of simulated archaic haplotype frequencies across all demographic models considered corresponds to a frequency threshold of 0.556. Note, this percentile is determined conditional on the archaic haplotype segregating in present day individuals, as most introgressed loci in simulations are lost due to drift. Although this is a potentially conservative way of determining significance, it properly accounts for how archaic haplotypes are ascertained in the observed data.

To explore the relationship between background selection and haplotype frequency, we downloaded genome wide B values from (68) and used the UCSC liftover tool (54) to convert coordinates to hg19. Note, for convenience we multiplied B values by 1000. We calculated the median B value for all introgressed haplotypes in Melanesians, and used the lm() function in R to

create a linear model with haplotype frequency as a function of B value. There is a small but significant positive correlation (linear regression p-value = 2.77×10^{-5}; Adjusted R^2=0.0056). Thus, on average, regions experiencing stronger background selection have lower frequency archaic haplotypes. This suggests that the distribution of introgressed haplotype frequencies in the simulated data is skewed towards higher frequencies than we expect in real data, making our estimated statistical significance conservative.

Supplementary Tables

Table S1. Summary of sequenced Island Melanesian individuals.

Sample ID	Population	Neighborhood	Language classification	Island	Latitide	Longitude
UV043	Baining	Mali	Papuan*	West_New_Britain	-4.51619	151.99585
UV1003	Nakanai Loso	Loso	Austronesian*	West_New_Britain	-5.703448	150.839539
UV1042	Mamusi	Kisiluvi	Austronesian*	West_New_Britain	-5.872868	150.968628
UV1043	Mamusi	Kisiluvi	Austronesian*	West_New_Britain	-5.872868	150.968628
UV1134	Ata	Luge	Papuan*	West_New_Britain	-5.566783	151.034546
UV1196	Melamala	Ubili	Austronesian*	West_New_Britain	-5.019	151.331177
UV1224	Mamusi	Paleabu	Austronesian*	West_New_Britain	-5.872868	150.968628
UV1230	Ata	Uasilau	Papuan*	West_New_Britain	-5.566783	151.034546
UV1260	Mangseng	Mangseng	Austronesian*	West_New_Britain	-6.020385	150.584106
UV1263	Pasismanua	Poronga	Austronesian*	West_New_Britain	-6.274348	150.088348
UV1266	Pasismanua	Poronga	Austronesian*	West_New_Britain	-6.274348	150.088348
UV305	Baining	Kaket	Papuan*	West_New_Britain	-4.51619	151.99585
UV500	Lavongai	North Lavongai	Austronesian*	New_Hanover	-2.534268	150.26825
UV518	Mussau	Kaupgu	Austronesian*	Mussau	-1.58	149.73
UV573	Nailik	Nailik	Austronesian*	New_Ireland	-2.943041	151.303711
UV580	Nailik	Nailik	Austronesian*	New_Ireland	-2.943041	151.303711
UV886	Nakanai Bileki	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV897	Nakanai Bileki	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV910	Nakanai Bileki	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV919	Nakanai Bileki	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV923	Nakanai Bileki	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV925	Nakanai Bileki	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV927	Nakanai Bileki	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV929	Nakanai Bileki	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV931	Nakanai Bileki	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV940	Nakanai Bileki	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011

UV944	Nakanai Bileki	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV946	Nakanai	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV952	Nakanai	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV956	Nakanai	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV958	Nakanai	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV964	Nakanai	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV971	Nakanai	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV979	Nakanai	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011
UV986	Nakanai	Bileki	Austronesian*	West_New_Britain	-5.665185	150.661011

***Note:** unlike the Austronesian languages, linguists are uncertain that all the "Papuan" languages are related. They are best viewed as a residual category with a number of language isolates. Some linguists therefore prefer the term "non-Austronesian". However, it is quite likely that the "Papuan" languages from the Bismarck Archipelago in our study are related to each other (*31*).

Table S2. Coverage statistics for 35 Island Melanesia samples. Asterisks (*) denote samples removed in analyses of unrelated individuals.

Sample ID	Mean coverage (genome)	Mean coverage (autosomes)	Median coverage (genome)
UV1042*	33.520	34.730	34
UV952*	32.561	33.808	34
UV956*	33.372	34.535	34
UV1043	34.532	35.734	35
UV043	34.701	36.106	36
UV1266*	35.602	36.779	36
UV500	36.256	37.547	37
UV518	35.779	37.129	37
UV1224	37.181	38.525	38
UV305	37.687	39.145	38
UV946*	36.968	38.383	38
UV958	36.708	38.078	38
UV1230	38.423	39.853	39
UV1263	37.865	39.264	39
UV964	38.014	39.475	39
UV925	38.685	40.104	40
UV944	39.673	41.148	40
UV979*	39.50	40.929	40
UV573*	39.362	40.896	41
UV910	40.47	41.936	41
UV929	40.611	41.984	41
UV931	40.444	41.861	41
UV986	39.794	41.339	41
UV1003	41.454	42.906	42
UV897	41.302	42.744	42
UV927*	40.889	42.467	42
UV940	40.801	42.359	42
UV886	41.808	43.364	43
UV919	41.340	42.886	43
UV923	41.939	43.465	43
UV971	42.278	43.797	43
UV580	43.067	44.774	44
UV1134	44.888	46.433	46
UV1260	45.300	46.944	47
UV1196	46.579	48.309	49

Table S3. Summary of worldwide populations used for PCA, ADMIXTURE, and *f4* analyses.

Population ID	Region	Number of Individuals	Latitude	Longitude	Reference
Algerian	Africa	7	36.8	3	(16)
BantuKenya	Africa	6	-3	37	(16)
BantuSA	Africa	8	-29	29	(16)
Biaka	Africa	20	4	17	(16)
Datog	Africa	3	-3.3	35.7	(16)
Egyptian	Africa	18	31	31.2	(16)
Esan	Africa	8	6.5	6	(16)
Ethiopian_Jew	Africa	7	9	38.7	(16)
Gambian	Africa	6	13.4	16.7	(16)
Hadza	Africa	5	-3.6	35.1	(16)
Ju_hoan_North	Africa	5	-18.9	21.5	(16)
Khomani	Africa	11	-27.8	21.1	(16)
Kikuyu	Africa	4	-0.4	36.9	(16)
Libyan_Jew	Africa	9	32.9	13.2	(16)
Luhya	Africa	8	1.3	36.8	(16)
Luo	Africa	8	-0.1	34.3	(16)
Mandenka	Africa	17	12	-12	(16)
Masai	Africa	12	-1.5	35.2	(16)
Mbuti	Africa	10	1	29	(16)
Mende	Africa	8	8.5	-13.2	(16)
Moroccan_Jew	Africa	6	34	-6.8	(16)
Mozabite	Africa	21	32	3	(16)
Saharawi	Africa	6	27.3	-8.9	(16)
Somali	Africa	13	5.6	48.3	(16)
Tunisian	Africa	8	36.8	10.2	(16)
Tunisian_Jew	Africa	7	36.8	10.2	(16)
Yoruba	Africa	70	7.4	3.9	(16)
AA	America	12	39.7	-105	(16)
Bolivian	America	7	-16.5	-68.2	(16)
Karitiana	America	12	-10	-63	(16)
Mayan	America	18	19	-91	(16)
Mixe	America	10	17	-96.6	(16)
Mixtec	America	10	16.7	-97.2	(16)
Piapoco	America	4	3	-68	(16)
Pima	America	14	29	-108	(16)
Quechua	America	5	-13.5	-72	(16)
Surui	America	8	-11	-62	(16)
Zapotec	America	10	17	-96.5	(16)
Aleut	Central_Asia_Siberia	7	53.6	160.8	(16)

Altaian	Central_Asia_Siberia	7	51.9	86	(16)
Chukchi	Central_Asia_Siberia	23	69.5	168.8	(16)
Dolgan	Central_Asia_Siberia	3	73	115.4	(16)
Eskimo	Central_Asia_Siberia	22	64.5	172.9	(16)
Even	Central_Asia_Siberia	10	57.5	135.9	(16)
Itelmen	Central_Asia_Siberia	6	57.2	156.9	(16)
Kalmyk	Central_Asia_Siberia	10	46.2	45.3	(16)
Koryak	Central_Asia_Siberia	9	58.1	159	(16)
Kyrgyz	Central_Asia_Siberia	9	42.9	74.6	(16)
Mansi	Central_Asia_Siberia	8	62.5	63.3	(16)
Mongola	Central_Asia_Siberia	6	45	111	(16)
Nganasan	Central_Asia_Siberia	11	71.1	96.1	(16)
Selkup	Central_Asia_Siberia	10	65.5	82.3	(16)
Tajik_Pomiri	Central_Asia_Siberia	8	37.4	71.7	(16)
Tlingit	Central_Asia_Siberia	4	54.7	164.5	(16)
Tubalar	Central_Asia_Siberia	22	51.1	87	(16)
Turkmen	Central_Asia_Siberia	7	42.5	59.6	(16)
Tuvinian	Central_Asia_Siberia	10	50.3	95.2	(16)
Ulchi	Central_Asia_Siberia	25	52.2	140.4	(16)
Uzbek	Central_Asia_Siberia	10	41.3	69.2	(16)
Yakut	Central_Asia_Siberia	20	63	129.5	(16)
Yukagir	Central_Asia_Siberia	19	65.5	151	(16)
Ami	East_Asia	10	22.8	121.2	(16)
Atayal	East_Asia	9	24.6	121.3	(16)
Cambodian	East_Asia	8	12	105	(16)
Dai	East_Asia	10	21	100	(16)
Daur	East_Asia	9	48.5	124	(16)
Han	East_Asia	33	32.3	114	(16)
Han_NChina	East_Asia	10	32.3	114	(16)
Hezhen	East_Asia	8	47.5	133.5	(16)
Japanese	East_Asia	29	38	138	(16)
Kinh	East_Asia	8	21	105.9	(16)
Korean	East_Asia	6	37.6	127	(16)
Lahu	East_Asia	8	22	100	(16)
Miao	East_Asia	10	28	109	(16)
Naxi	East_Asia	9	26	100	(16)
Oroqen	East_Asia	9	50.4	126.5	(16)
She	East_Asia	10	27	119	(16)
Thai	East_Asia	10	13.8	100.5	(16)
Tu	East_Asia	10	36	101	(16)
Tujia	East_Asia	10	29	109	(16)
Uygur	East_Asia	10	44	81	(16)
Xibo	East_Asia	7	43.5	81.5	(16)

Yi	East_Asia	10	28	103	(16)
Ata	Oceania	2	-5.57	151.03	This study
Australian	Oceania	3	-13	143	(16)
Baining	Oceania	2	-4.52	152.00	This study
Bougainville	Oceania	10	-6	155	(16)
Lavongai	Oceania	1	-2.53	150.27	This study
Mamusi	Oceania	2	-5.87	150.97	This study
Mangseng	Oceania	1	-6.02	150.58	This study
Melamala	Oceania	1	-5.02	151.33	This study
Mussau	Oceania	1	-1.58	149.73	This study
Nailik	Oceania	1	-2.94	151.30	This study
Nakanai Bileki	Oceania	14	-5.67	150.66	This study
Nakanai Loso	Oceania	1	-5.70	150.84	This study
Papuan	Oceania	14	-4	143	(16)
Pasismanua	Oceania	1	-6.27	150.09	This study
Balochi	South_Asia	20	30.5	66.5	(16)
Bengali	South_Asia	7	23.7	90.4	(16)
Brahui	South_Asia	21	30.5	66.5	(16)
Burusho	South_Asia	23	36.5	74	(16)
Cochin_Jew	South_Asia	5	10	76.3	(16)
GujaratiA	South_Asia	5	23.2	72.7	(16)
GujaratiB	South_Asia	5	23.2	72.7	(16)
GujaratiC	South_Asia	5	23.2	72.7	(16)
GujaratiD	South_Asia	5	23.2	72.7	(16)
Hazara	South_Asia	14	33.5	70	(16)
Kalash	South_Asia	18	36	71.5	(16)
Kusunda	South_Asia	10	28.1	82.5	(16)
Makrani	South_Asia	20	26	64	(16)
Pathan	South_Asia	19	33.5	70.5	(16)
Punjabi	South_Asia	8	31.5	74.3	(16)
Sindhi	South_Asia	18	25.5	69	(16)
Abkhasian	West_Eurasia	9	43	41	(16)
Adygei	West_Eurasia	17	44	39	(16)
Albanian	West_Eurasia	6	41.3	19.8	(16)
Armenian	West_Eurasia	10	40.2	44.5	(16)
Ashkenazi_Jew	West_Eurasia	7	52.2	21	(16)
Balkar	West_Eurasia	10	43.5	43.6	(16)
Basque	West_Eurasia	29	43	0	(16)
BedouinA	West_Eurasia	25	31	35	(16)
BedouinB	West_Eurasia	19	31	35	(16)
Belarusian	West_Eurasia	10	53.9	28	(16)
Bergamo	West_Eurasia	12	46	10	(16)
Bulgarian	West_Eurasia	10	42.2	24.7	(16)

Chechen	West_Eurasia	9	43.3	45.7	(16)
Chuvash	West_Eurasia	10	56.1	47.3	(16)
Croatian	West_Eurasia	10	43.5	16.4	(16)
Cypriot	West_Eurasia	8	35.1	33.4	(16)
Czech	West_Eurasia	10	50.1	14.4	(16)
Druze	West_Eurasia	39	32	35	(16)
English	West_Eurasia	10	51.2	0.7	(16)
Estonian	West_Eurasia	10	58.5	24.9	(16)
Finnish	West_Eurasia	7	60.2	24.9	(16)
French	West_Eurasia	25	46	2	(16)
French_South	West_Eurasia	7	43.4	-0.6	(16)
Georgian	West_Eurasia	10	42.5	41.9	(16)
Georgian_Jew	West_Eurasia	7	41.7	44.8	(16)
Greek	West_Eurasia	20	40.6	22.9	(16)
Hungarian	West_Eurasia	20	47.5	19.1	(16)
Icelandic	West_Eurasia	12	64.1	-21.9	(16)
Iranian	West_Eurasia	8	35.6	51.5	(16)
Iranian_Jew	West_Eurasia	9	35.7	51.4	(16)
Iraqi_Jew	West_Eurasia	6	33.3	44.4	(16)
Jordanian	West_Eurasia	9	32.1	35.9	(16)
Kumyk	West_Eurasia	8	43.3	46.6	(16)
Lebanese	West_Eurasia	8	33.8	35.6	(16)
Lezgin	West_Eurasia	9	42.1	48.2	(16)
Lithuanian	West_Eurasia	10	54.9	23.9	(16)
Maltese	West_Eurasia	8	35.9	14.4	(16)
Mordovian	West_Eurasia	10	54.2	45.2	(16)
Nogai	West_Eurasia	9	44.4	41.9	(16)
North_Ossetian	West_Eurasia	10	43	44.7	(16)
Norwegian	West_Eurasia	11	60.4	5.4	(16)
Orcadian	West_Eurasia	13	59	-3	(16)
Palestinian	West_Eurasia	38	32	35	(16)
Russian	West_Eurasia	22	61	40	(16)
Sardinian	West_Eurasia	27	40	9	(16)
Saudi	West_Eurasia	8	18.5	42.5	(16)
Scottish	West_Eurasia	4	56	-3.9	(16)
Sicilian	West_Eurasia	11	37.1	15.3	(16)
Spanish	West_Eurasia	53	37.4	-6	(16)
Spanish_North	West_Eurasia	5	43.3	-4	(16)
Syrian	West_Eurasia	8	35.1	36.9	(16)
Turkish	West_Eurasia	56	39.6	28.5	(16)
Turkish_Jew	West_Eurasia	8	41	29	(16)
Tuscan	West_Eurasia	8	43	11	(16)
Ukrainian	West_Eurasia	9	50.3	31.6	(16)

Yemen	West_Eurasia	6	14	44.6	(16)
Yemenite_Jew	West_Eurasia	8	15.4	44.2	(16)
Chimp	Chimpanzee reference sequence	1			(73)
Denisovan	Denisova	1			(5)
Altai	Altai Neandertal	1			(3)

Table S4. Population estimates of the proportion of Denisovan admixture measured as a ratio of *f4* statistics. Populations marked with an asterisk (*) are from the Human Origins data set.

Population	*PD(X)*	Stardard Error	Z score
Mussau	0.018963	0.004057	4.674
Mangseng	0.021914	0.004254	5.151
Melamala	0.024321	0.004075	5.969
Nakanai Bileki	0.0244	0.003016	8.089
Ata	0.024976	0.003495	7.146
Nakanai Loso	0.026943	0.00465	5.795
Nailik	0.02758	0.004482	6.153
Pasismanua	0.029613	0.004421	6.698
Mamusi	0.031535	0.004185	7.535
Lavongai	0.032053	0.004496	7.129
Baining	0.034269	0.004567	7.503
Bougainville*	0.030587	0.003633	8.418
Papuan*	0.032161	0.003898	8.251
Australian*	0.033945	0.004264	7.96

Table S5. Summary of 1000 Genomes Project populations analyzed.

Population Code	Population Description	Super Population Code	Number of Individuals
ESN	Esan in Nigeria	AFR	99
GWD	Gambian in Western Divisions in the Gambia	AFR	113
LWK	Luhya in Webuye, Kenya	AFR	99
MSL	Mende in Sierra Leone	AFR	85
YRI	Yoruba in Ibadan, Nigeria	AFR	108
CDX	Chinese Dai in Xishuangbanna, China	EAS	93
CHB	Han Chinese in Bejing, China	EAS	103
CHS	Southern Han Chinese	EAS	105
JPT	Japanese in Tokyo, Japan	EAS	104
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS	99
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	EUR	99
FIN	Finnish in Finland	EUR	99
GBR	British in England and Scotland	EUR	91
IBS	Iberian Population in Spain	EUR	107
TSI	Toscani in Italia	EUR	107
BEB	Bengali from Bangladesh	SAS	86
GIH	Gujarati Indian from Houston, Texas	SAS	103
ITU	Indian Telugu from the UK	SAS	102
PJL	Punjabi from Lahore, Pakistan	SAS	96
STU	Sri Lankan Tamil from the UK	SAS	102

Table S6. Proportion of *S callset estimated to be Neandertal or Denisovan haplotypes for eight populations.**

Population	Neandertal Proportion	Denisovan Proportion
EAS	0.485	0.01
EUR	0.445	0.0
SAS	0.435	0.0075
MEL	0.23	0.29
GWD	0.03	0.0
LWK	0.02	0.0
ESN	0.005	0.0
MSL	0.005	0.0

Table S7. Proportion of Denisovan $P_D(X)$ and "Papuan" ancestry $P_P(X)$ per individual inferred from a ratio of *f4* statistics.

Sample	$P_D(X)$	Standard Error	Z-score	$P_P(X)$	Standard Error	Z-score
UV043	0.032609	0.004895	6.662	0.747434	0.010462	71.443
UV1003	0.026943	0.00465	5.795	0.662328	0.010998	60.221
UV1043	0.031057	0.004673	6.646	0.669834	0.010591	63.247
UV1134	0.028375	0.00443	6.406	0.658334	0.010435	63.087
UV1196	0.024321	0.004075	5.969	0.581165	0.011336	51.268
UV1224	0.032012	0.004707	6.801	0.680196	0.010503	64.764
UV1230	0.021578	0.003794	5.687	0.609222	0.011092	54.923
UV1260	0.021914	0.004254	5.151	0.598067	0.011211	53.347
UV1263	0.029613	0.004421	6.698	0.643947	0.011356	56.708
UV305	0.03593	0.005119	7.019	0.744321	0.010204	72.947
UV500	0.032053	0.004496	7.129	0.578321	0.010992	52.611
UV518	0.018963	0.004057	4.674	0.584274	0.012153	48.078
UV580	0.02758	0.004482	6.153	0.580215	0.011061	52.458
UV886	0.022135	0.004005	5.527	0.554669	0.011208	49.49
UV897	0.019584	0.004072	4.809	0.576134	0.010765	53.517
UV910	0.02473	0.00417	5.93	0.555351	0.011529	48.17
UV919	0.024033	0.003887	6.183	0.576982	0.011305	51.039
UV923	0.024205	0.004361	5.551	0.590337	0.011022	53.562
UV925	0.023404	0.004262	5.492	0.550026	0.011373	48.361
UV929	0.023101	0.004199	5.501	0.573346	0.01182	48.506
UV931	0.024211	0.00435	5.566	0.570603	0.0118	48.356
UV940	0.023685	0.00437	5.419	0.566538	0.012205	46.418
UV944	0.022937	0.004167	5.504	0.602444	0.011003	54.755
UV958	0.024396	0.004226	5.773	0.625663	0.01092	57.295
UV964	0.029789	0.004165	7.152	0.571054	0.011413	50.035
UV971	0.023652	0.00429	5.513	0.580852	0.011592	50.108
UV986	0.031738	0.004238	7.489	0.593095	0.011316	52.41

Table S8. Regions significantly depleted of Neandertal sequence in all populations.

Chromosome	Start	End
1	102200000	114900000
2	201100000	211500000
3	76500000	90500000
7	106300000	124700000
8	53900000	66000000
18	25000000	41800000

Table S9. Regions significantly depleted of both Neandertal and Denisovan sequence.

Chromosome	Start	End
1	104000000	114900000
3	76500000	90500000
7	113600000	124700000
8	54500000	65400000

Table S10. Enrichment analysis of genes in regions depleted of archaic ancestry in during brain development. The name and Allen Ontology ID for each brain region.

Age category	Enriched region 1	Enriched region 2	Enriched region 3
Infant (0-2 yrs)	VFC_ventrolateral prefrontal cortex [Allen:10185]	Cx_cerebral cortex [Allen:10159]	
Adolescent (12-19yr)	STR_striatum [Allen:10333]		
Adult (>19 yrs)	STR_striatum [Allen:10333]	NCx_neocortex (isocortex) [Allen:10160]	CN_cerebral nuclei [Allen:101331]

Table S11. GO enrichments of genes in regions depleted of archaic sequence

GO ID	GO Term	*p-value*	FWER	Genes
GO:0004556	alpha-amylase activity	1.60E-09	<0.001	*AMY1C,AMY1A,AMY2BA MY2A,AMY1B*
GO:0016160	amylase activity	9.20E-09	<0.001	*AMY1C,AMY1A,AMY2BA MY2A,AMY1B*
GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	9.30E-09	<0.001	*MGEA5,AMY1C,AMY1A,A MY2B,HYAL4, AMY2A,OVGP1,SPAM1CH IA,GBE1, CHI3L2,AMY1B*
GO:0004364	glutathione transferase activity	1.90E-08	<0.001	*GSTO2,GSTM4,GSTM2,GS TO1,GSTM1, GSTM3,GSTM5*
GO:0016798	hydrolase activity, acting on glycosyl bonds	1.80E-07	<0.001	*MGEA5,AMY1C,AMY1AA MY2B,HYAL4, AMY2A,OVGP1,SPAM1CH IA,GBE1, CHI3L2,AMY1B*
GO:0015271	outward rectifier potassium channel activity	6.60E-07	<0.001	*KCNIP2,KCNA3,KCND3,KC NA2,KCND2*
GO:0005250	A-type (transient outward) potassium channel activity	5.30E-06	<0.001	*KCNIP2,KCND3,KCND2*
GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups	7.20E-06	0.001	*GSTO2,GSTM4,GSTM2,GS TO1,GSTM1, GSTM3,GSTM5*
GO:0048593	camera-type eye morphogenesis	2.00E-05	0.035	*RP1,PAX2,TSPAN12, WNT2B,HIPK1,PITX3, WNT16,GNAT2,WNT2*
GO:0005109	frizzled binding	5.00E-05	0.012	*MAGI3,WNT2B,SDCBP,W NT16,WNT2*
GO:0008076	voltage-gated potassium channel complex	7.10E-05	0.022	*KCNC4,KCNIP2,KCNA3,KC ND3,KCNA10, KCNA2,KCND2*
GO:0034705	potassium channel complex	7.10E-05	0.022	*KCNC4,KCNIP2,KCNA3,KC ND3,KCNA10, KCNA2,KCND2*

Table S12. Summary of candidate adaptive introgression regions in Island Melanesians

Chr	Start	Stop	MEL	EUR	SAS	EAS	Genes within archaic haplotype	Genes within 100kb of archaic haplotype
chr1	49505169	49590000	0.59259	0.07555	0.06033	0.07639	*AGBL4*	
chr1	86683344	86941488	0.55556	0.00497	0.09611	0.1121	*CLCA1, CLCA2, ODF2L*	*CLCA4, COL24A1*
chr1	89620000	89700855	0.57407	0	0	0	*GBP4, GBP7*	*GBP1, GBP2 GBP5, LOC729930*
chr1	208812029	208922357	0.7963	0.00199	0.1135	0.19048		
chr1	210501503	210566962	0.68519	0.44334	0.34969	0.6627	*HHAT*	*SERTAD4, SERTAD4-AS1*
chr13	109004102	109039765	0.55556	0.02883	0.1135	0.46329		*TNFSF13B*
chr18	60045603	60261310	0.62963	0.27833	0.03885	0.00099	*DKFZp451A18, TNFRSF11A, ZCCHC2*	*KIAA1468*
chr18	71039387	71085853	0.59259	0.24453	0.10225	0.02877		*LOC100505817*
chr18	71148873	71221457	0.7037	0.24006	0.165135	0.315975		
chr2	3815476	3874363	0.59259	0.14215	0.09714	0.28472		*ALLC*
chr2	163007574	163336914	0.57407	0.01093	0.02045	0.10714	*FAP, GCA, GCG, IFIH1, KCNH7*	*DPP4, GCA,*
chr2	241978100	242412975	0.62963	0.00398	0.09816	0.1369	*AK055601, ANO7, DKFZp686L08115, FARP2, HDLBP, MTERFD2, PASK, PPP1R7, SEPT2, SNED1*	*AK055890, BC017214, BOK, BOK-AS1, FARP2, LOC200772, STK25*
chr20	46579442	46588066	0.57407	0.00398	0.03579	0.10813	*BX648826*	
chr22	20767102	20789294	0.64815	0.00199	0.04703	0.48611	*SCARF2*	*JX456220, KLHL22, MED15, USP41, ZNF74*
chr3	15990892	16080545	0.57407	0.01292	0.0409	0.01488	*GALNT15*	*ANKRD28, DQ582763, DQ592427, MIR563*
chr5	54756003	54861928	0.68519	0	0	0	*MIR5687, PPAP2A, RNF138P1*	*SKIV2L2, SLC38A9*
chr5	56278198	56315225	0.703705	0.10636	0.20552	0.26984		*MAP3K1, MIER3, SETD9*
chr5	150568767	150595676	0.57407	0	0	0	*CCDC69*	*ANXA6, GM2A, SLC36A2, SLC36A3*
chr6	138037107	138119791	0.77778	0.000995	0.02198	0.00794		*LOC100130476, TNFAIP3*
chr8	121101757	121132175	0.88889	0.22664	0.28323	0.24504		*COL14A1, DEPTOR*
chr9	114803285	114830551	0.77778	0.08648	0.19274	0.22817	*MI3134, SUSD1*	

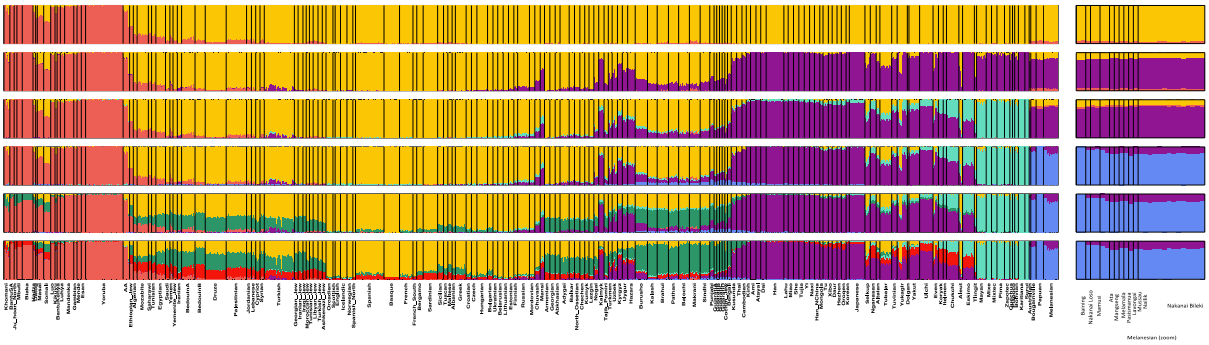

Figure S1. ADMIXTURE analysis of global populations. Ancestral clusters inferred using ADMIXTURE (K=2-7). For details about the geographic origin of the worldwide samples considered in the ADMIXTURE analysis, see map in Fig. 1A and legend in Fig. S3. Zoom of Island Melanesian samples is shown on the right.

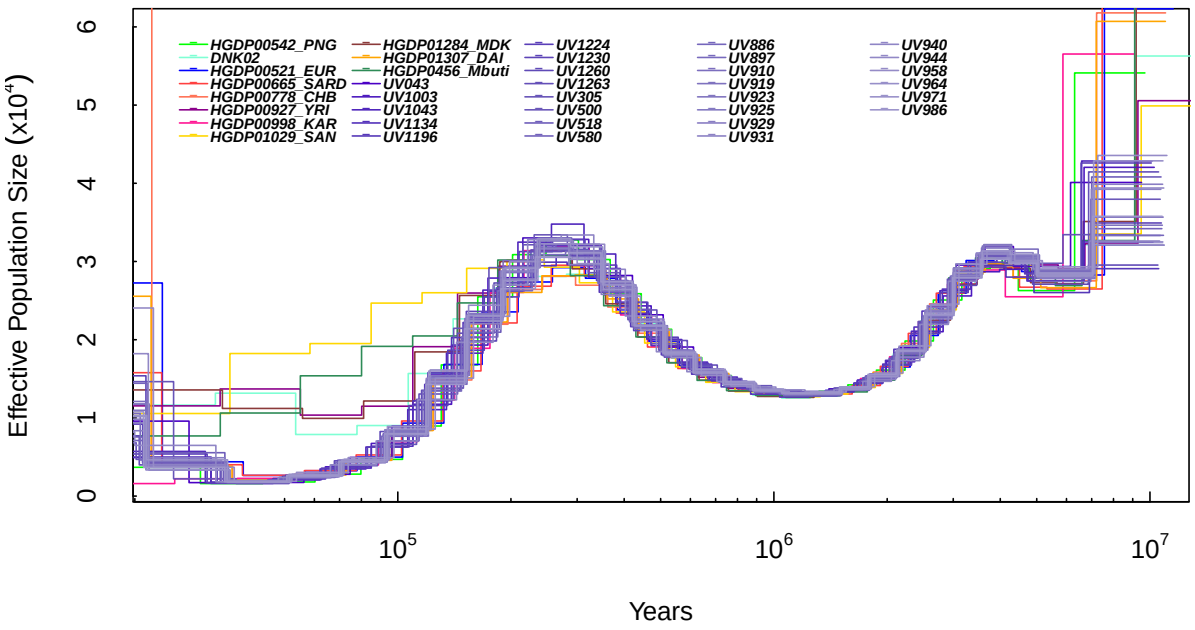

Figure S2. Inference of effective population sizes as a function of time. Inferred population size changes over time in 27 unrelated Melanesians and 11 published high coverage genomes from worldwide populations.

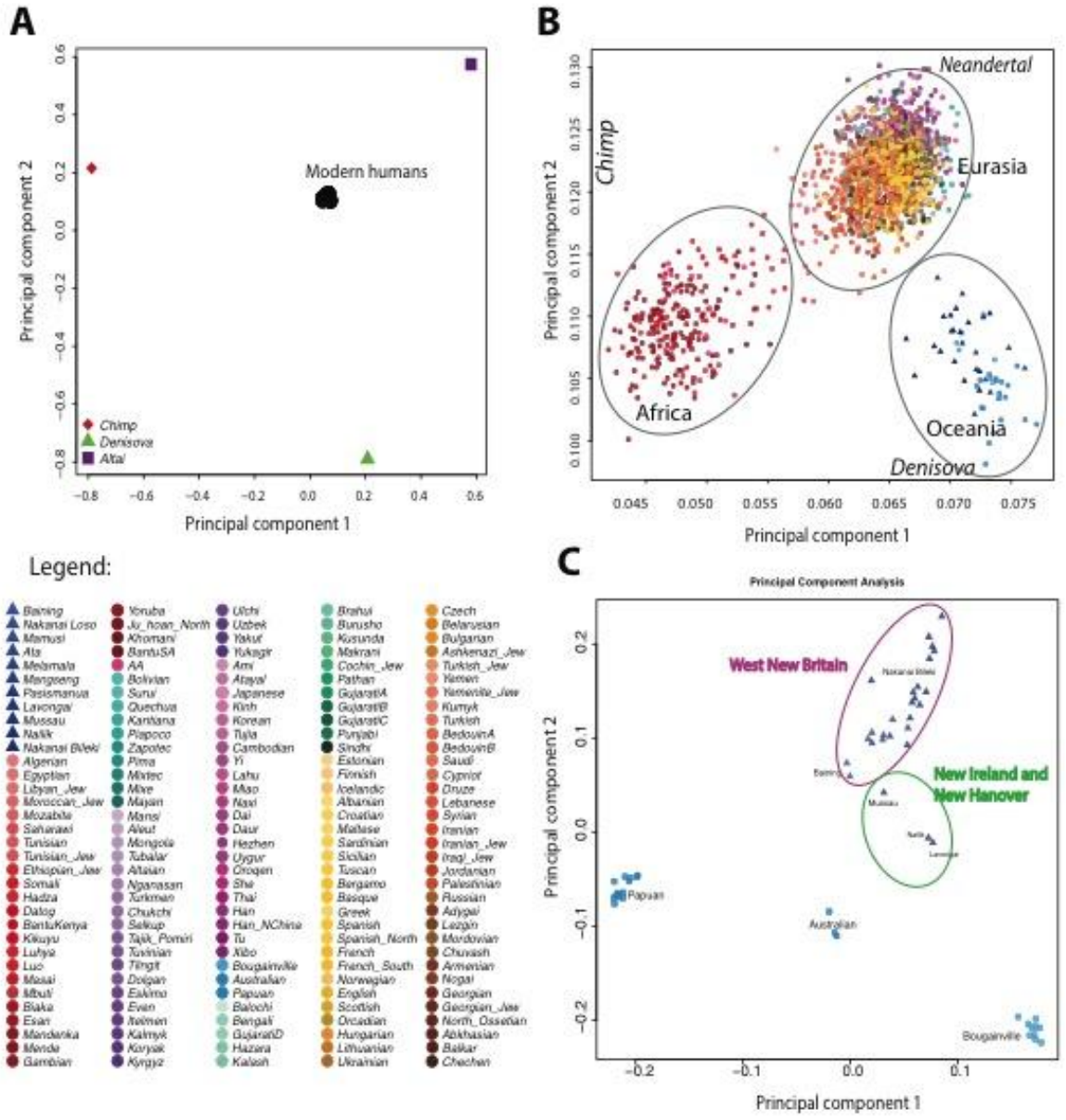

Figure S3. Principal Component Analysis to investigate genetic similarities of present-day humans and archaic species and population affinities in Oceania. (A) Axes of variation resulting from the PCA performed on the Altai Neandertal, the Denisovan and the chimpanzee, with all 1,964 modern humans projected into the resulting PCA space. Note, this is analogous to the figure shown in panel B and Fig. 1C, but is "unzoomed". (B) 1,964 present-day humans from 170 populations projected onto the variation of the Altai Neandertal, Denisovan, and Chimpanzee. (C) PCA of Oceanic individuals.

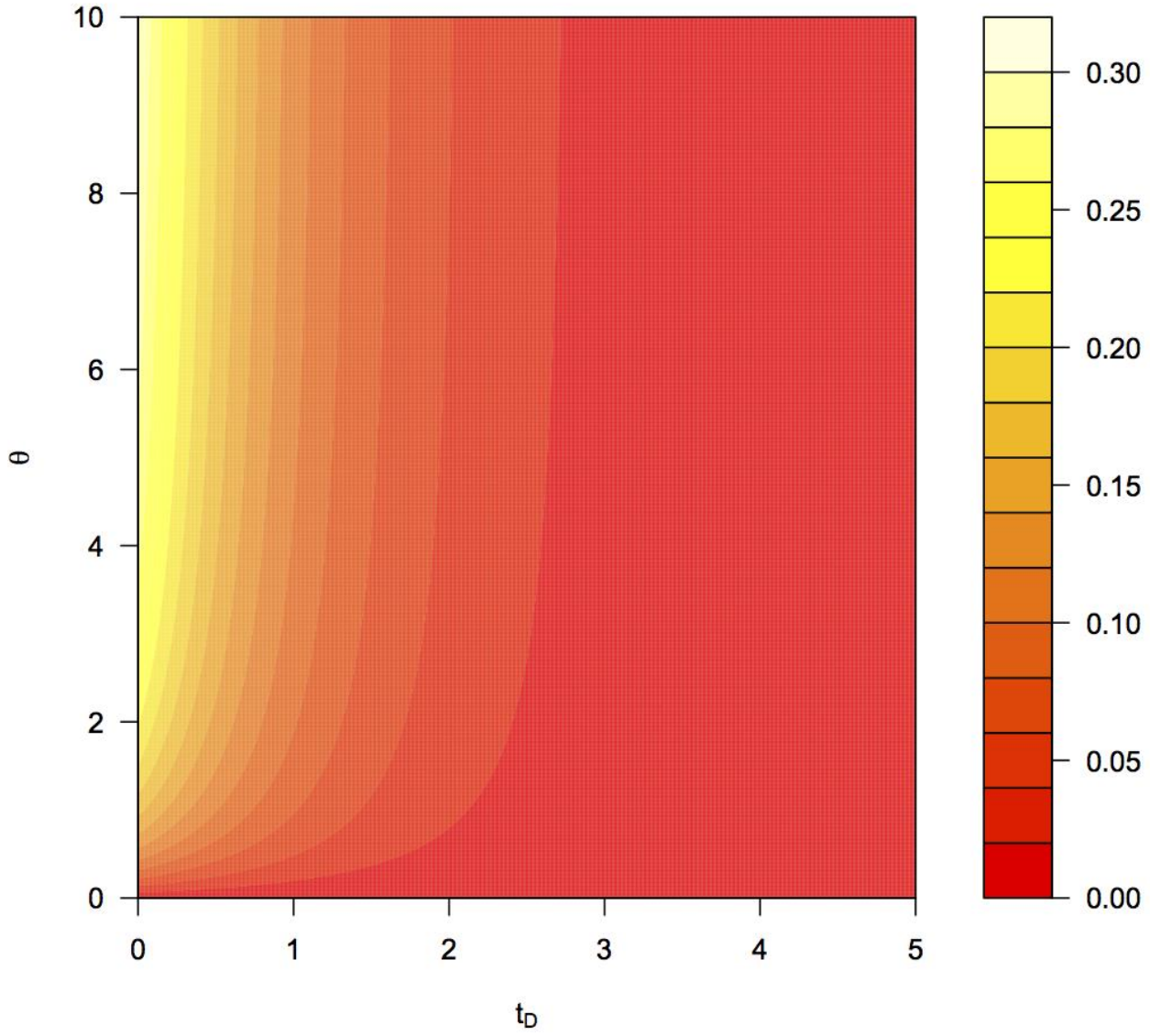

Figure S4. Expected proportion ILS between Neandertal and Denisovan introgressed haplotypes. Contour plot of the probability that a lineage is introgressed through Denisova but is closer to Neandertal. The x-axis represents the time between the introgression event and the Neandertal-Denisovan common ancestor, measured in units of $2N_e$ generations. The y-axis shows $\theta = 4N_e\mu L$. The color bar indicates proportion.

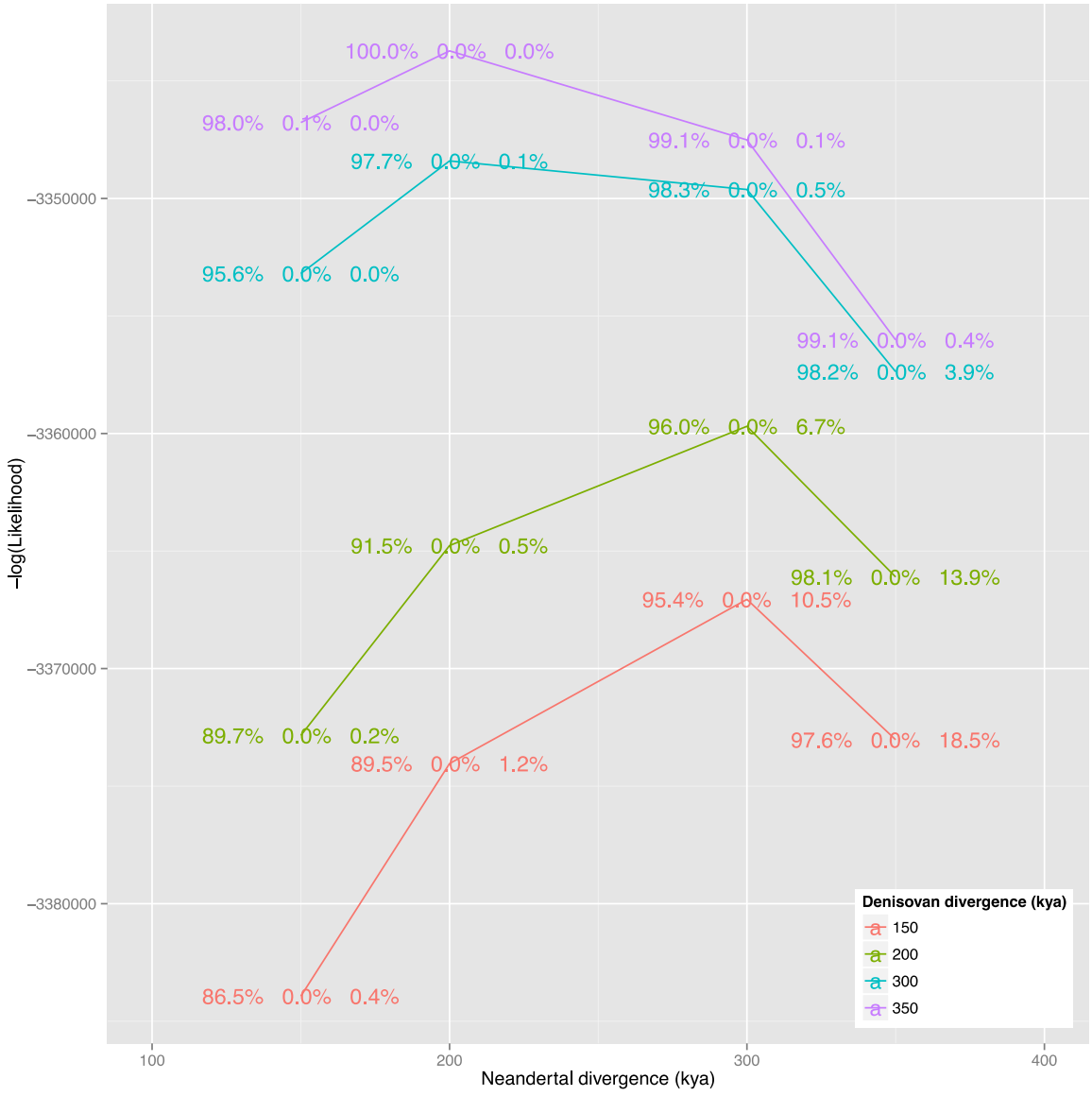

Figure S5. Likelihood and callset differences between demographic models. Likelihoods of the *S** callset given different Neandertal and Denisovan divergence times between the sequenced archaic individual and the archaic population that introgressed with modern humans. In the highest likelihood model, the Denisovan divergence occurred 350kya and the Neandertal divergence occurred 200kya. For each model, three values are given in comparison to the highest likelihood callset: 1) the percentage archaic sequence present in the callset, 2) the percentage of Neandertal calls labeled Denisovan in the callset, and 3) the percentage of Denisovan calls labeled Neandertal in the callset. The callsets for many alternative models are very similar to the highest likelihood callset.

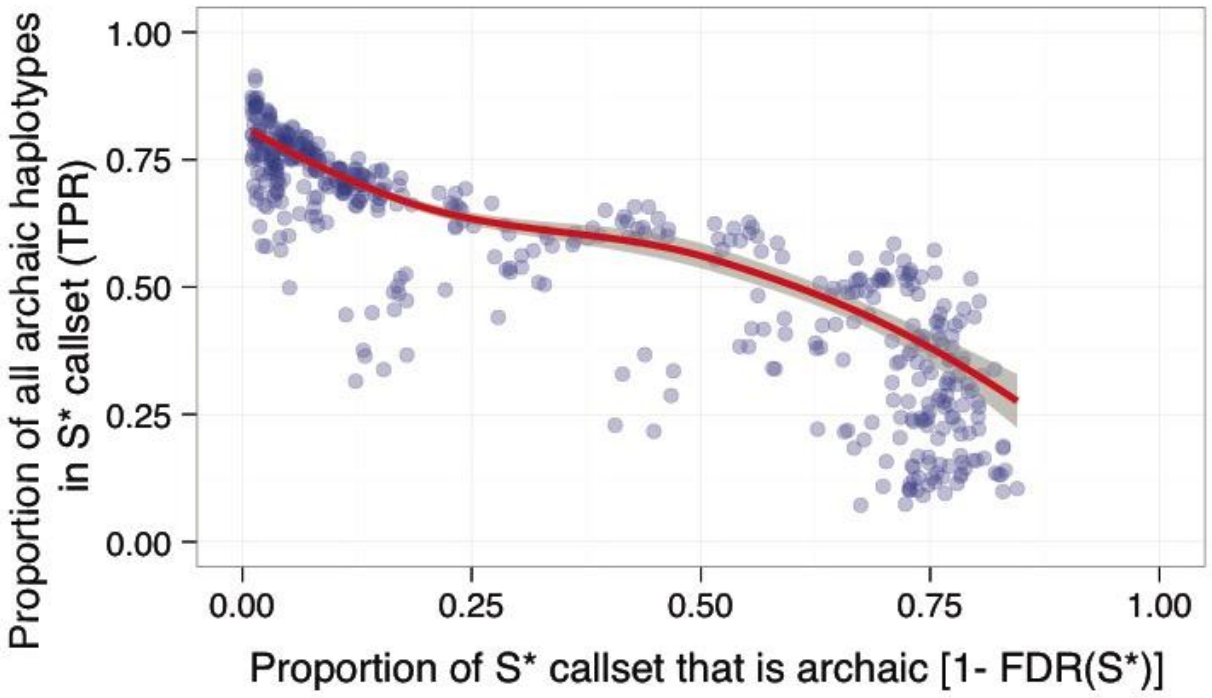

Figure S6. FDR vs TPR of *S in simulated sequence data.** Sensitivity and specificity of sliding S* thresholds on simulated sequence data with 2% introgression from an archaic species. At a 50% FDR, ~60% of introgressed haplotypes are identified. Note that this is in the absence of comparison to an archaic genome – this set is further refined via archaic match *p-values*. Each point represents simulations under varying levels of diversity. More challenging simulations (high FDR, low TPR) have low sequence diversity.

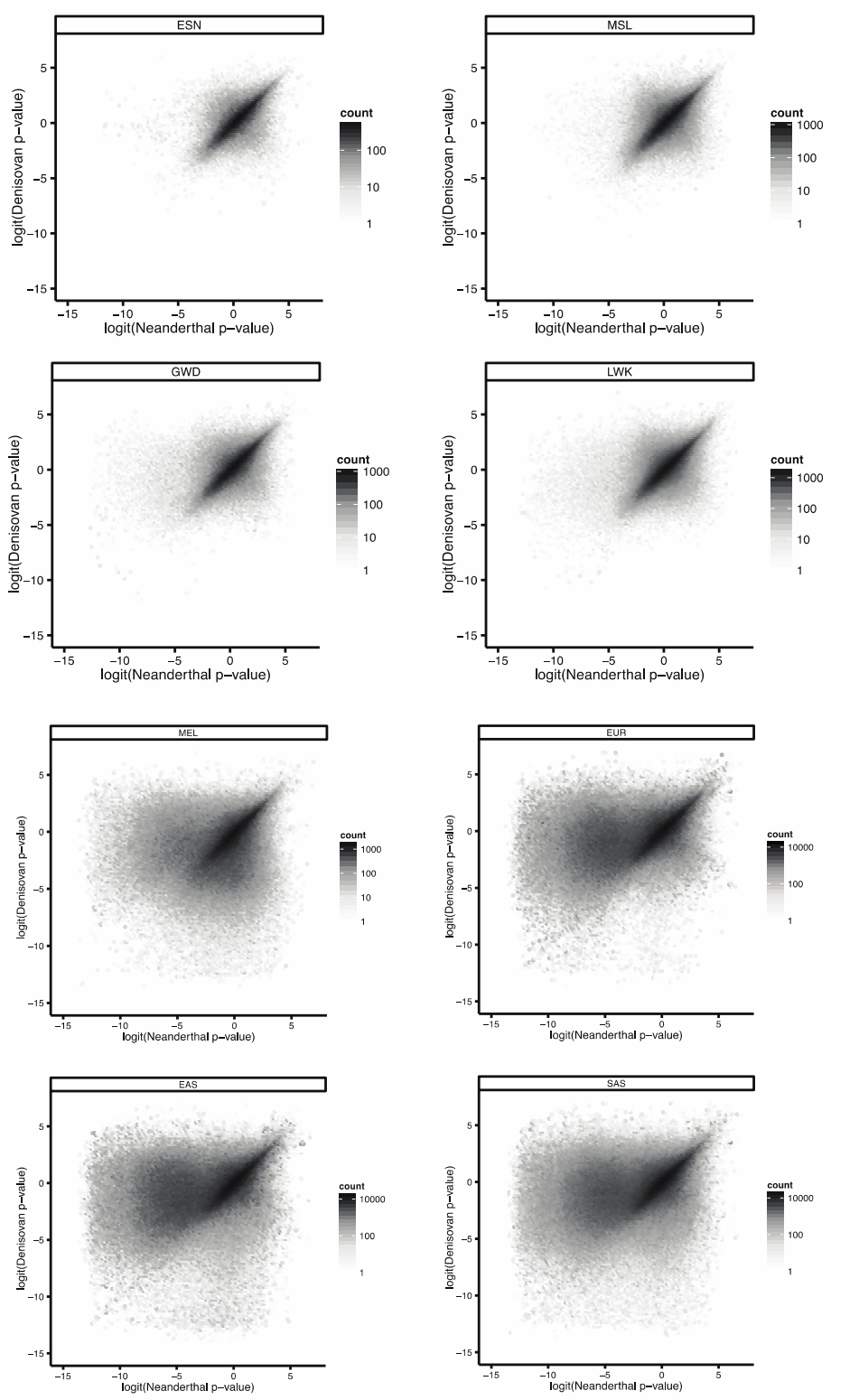

Figure S7. Bivariate archaic match *p-values* in all populations analyzed.

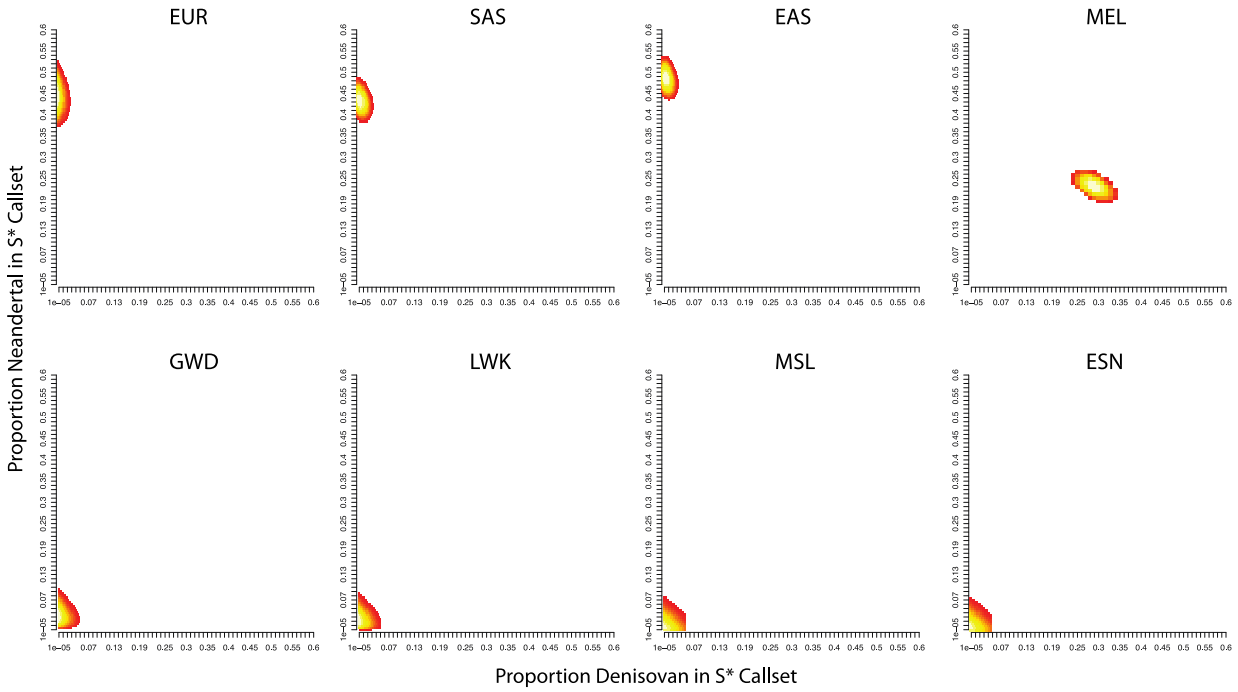

Figure S8. Maximum likelihood estimates of the proportion of Neandertal and Denisovan haplotypes in the set of *S haplotypes across populations.** Estimates of π_1 and π_2 (proportion Neandertal and Denisovan in significant S* callset) obtained through grid search using our likelihood framework. For clarity, only the top likelihood proportions are shown

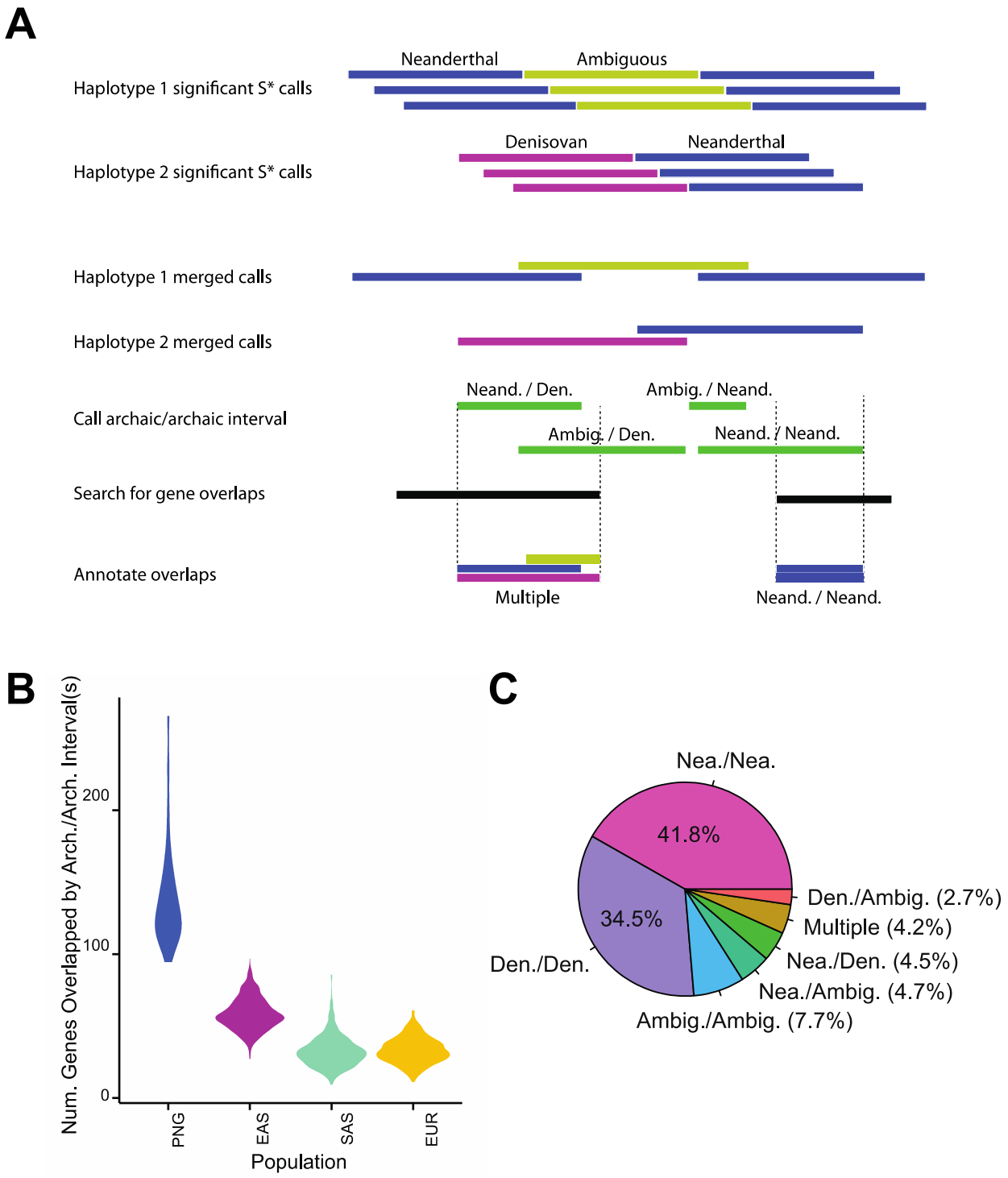

Figure S9. Distribution of homozygous archaic loci among individuals and populations. A) Schematic representation of approach to detecting and annotating archaic/archaic genomic intervals and their overlap with protein-coding genes. B) Violin plot of the per-individual counts of protein genes partially or fully overlapped by archaic-archaic intervals, stratified by population. C) Combining across 27 unrelated PNG individuals, proportions of different forms of archaic/archaic introgressed ancestry in gene-overlapping intervals.

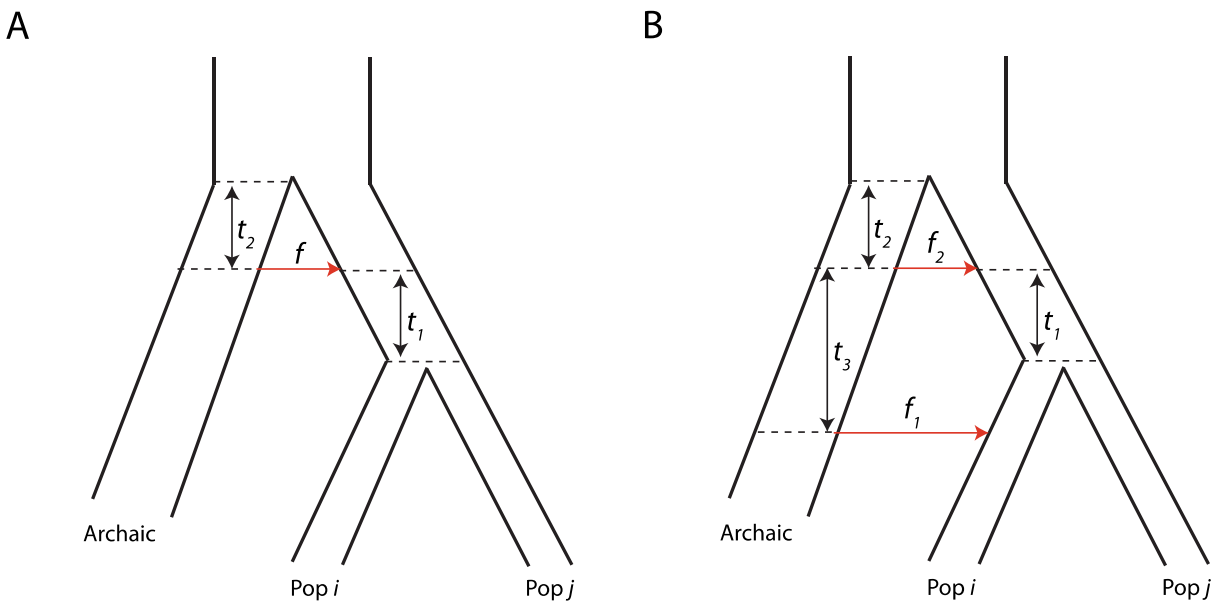

Figure S10. Demographic schematics for one and two pulse introgression events. A) Model of a single, shared ancestral admixture. Time t_1 elapses between the admixture event and the divergence of population i and population j. Admixture occurs with probability f per lineage. Between the ancestral admixture and the divergence of the archaic group from modern humans, time t_2 elapses. B) Model of two admixtures. Time t_1 elapses between the admixture event and the divergence of population i and population j. Admixture occurs with probability f_2 per lineage. Between the ancestral admixture and the divergence of the archaic group from modern humans, time t_2 elapses. In addition, an admixture of intensity f_1 occurs only into population i and time t_2 elapses between this admixture and the shared ancestral admixture.

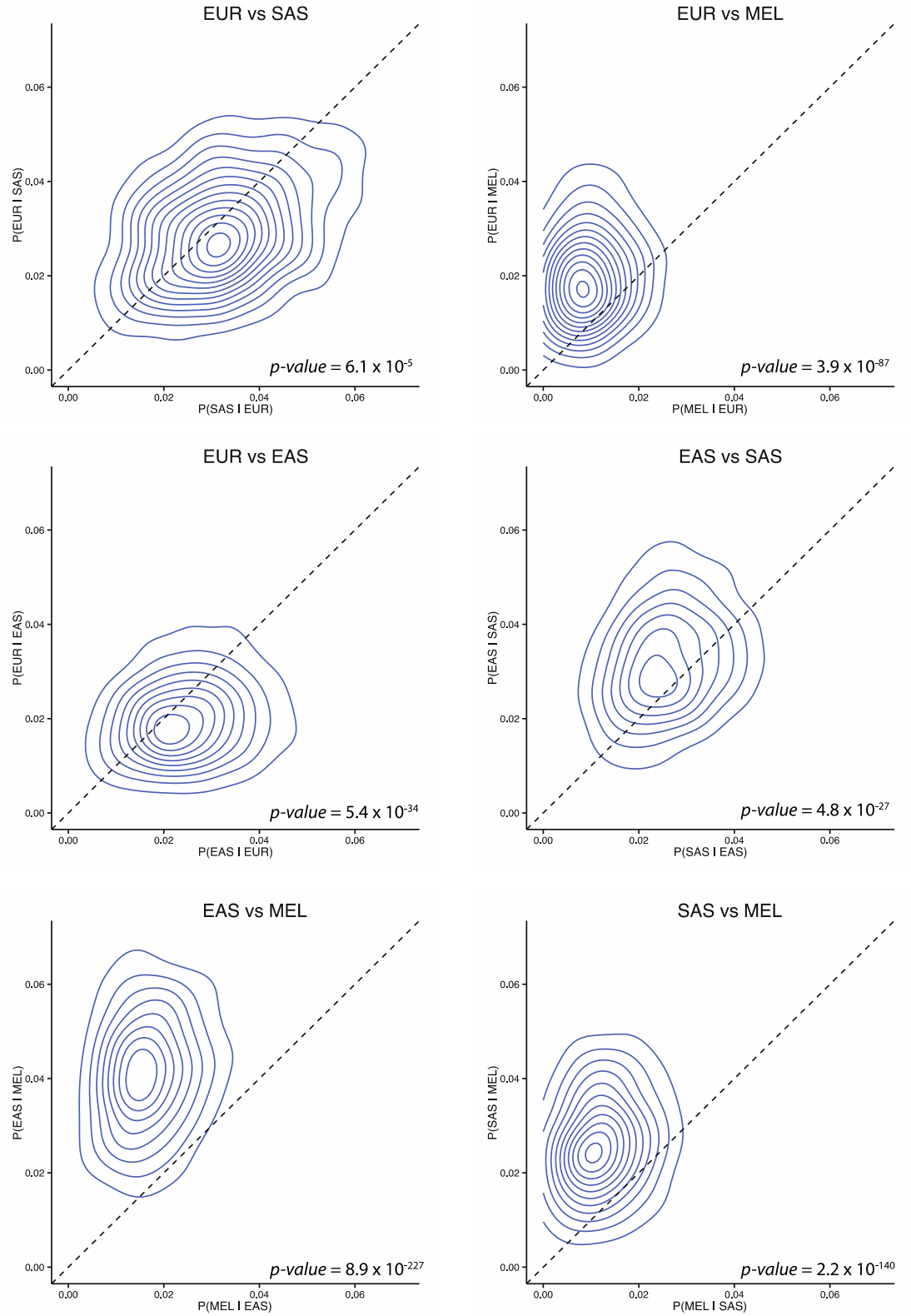

Figure S11. Reciprocal probability of sharing Neandertal sequences using only archaic sequences confidently called as Neandertal.

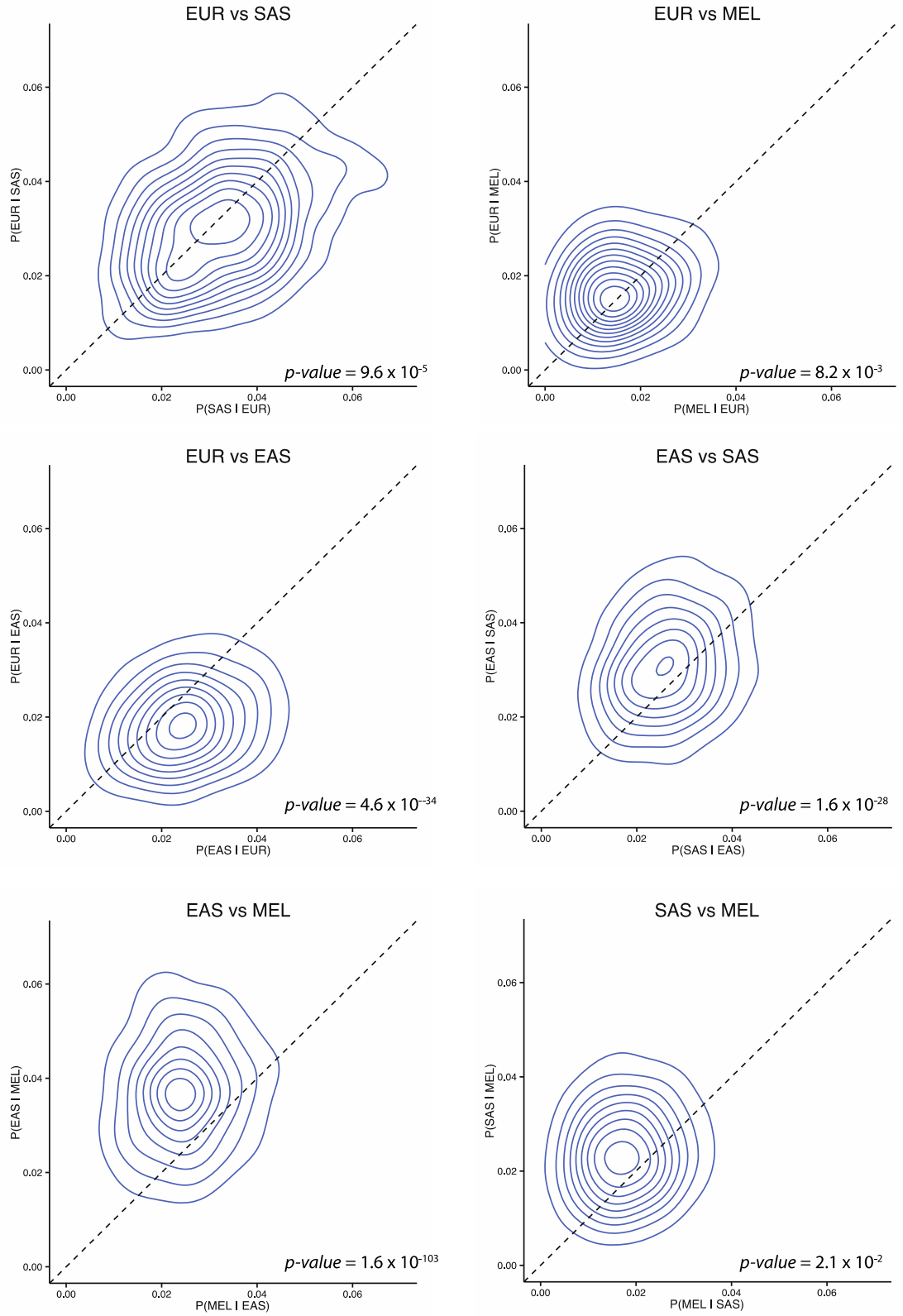

Fig. S12. Reciprocal probability of sharing Neandertal sequences using all archaic sequences (i.e., all Neandertal + Ambiguous sequence).

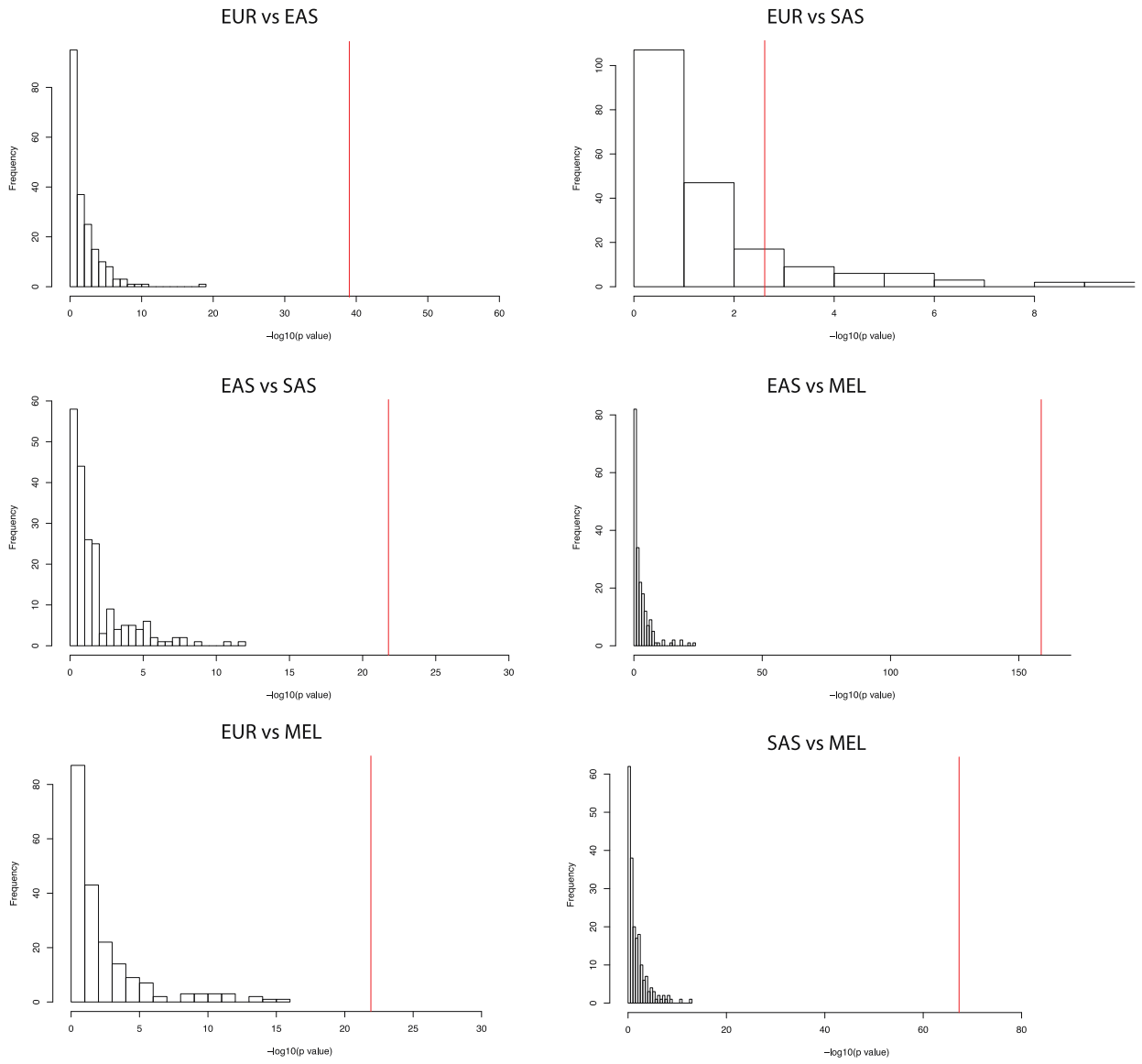

Figure S13. Evaluating significance of reciprocal match probabilities through permutations. The distribution of $-\log_{10}$ *p-values* calculated by the binomial test in 200 permutations is shown for each population comparison. Red lines indicate mean $-\log_{10}$ *p-values* from 10 random subsets of the data. In all comparisons, the permutation results result in the same interpretation as when evaluating reciprocal match probabilities through the parametric approach described in the supplementary material.

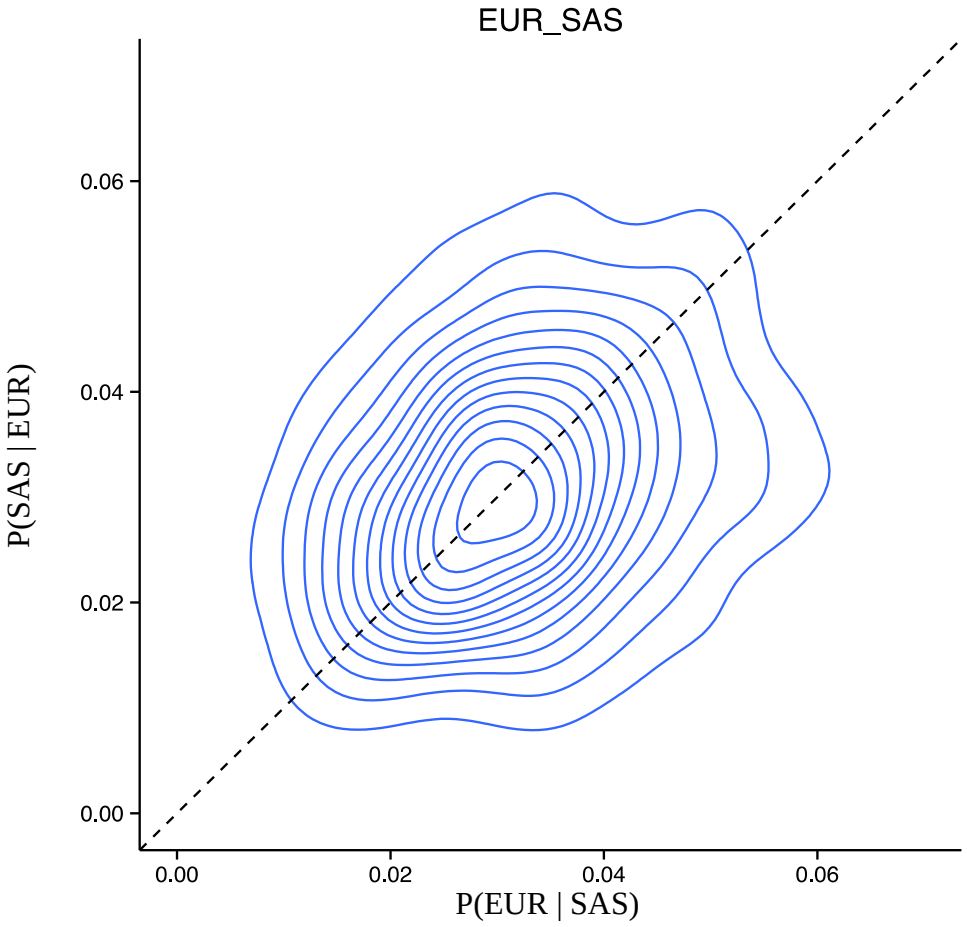

Figure S14. Reciprocal probability of sharing Neandertal sequences between European (EUR) and South Asian (SAS) individuals. No significant difference in reciprocal match probabilities was found (*p-value* = 0.8675).

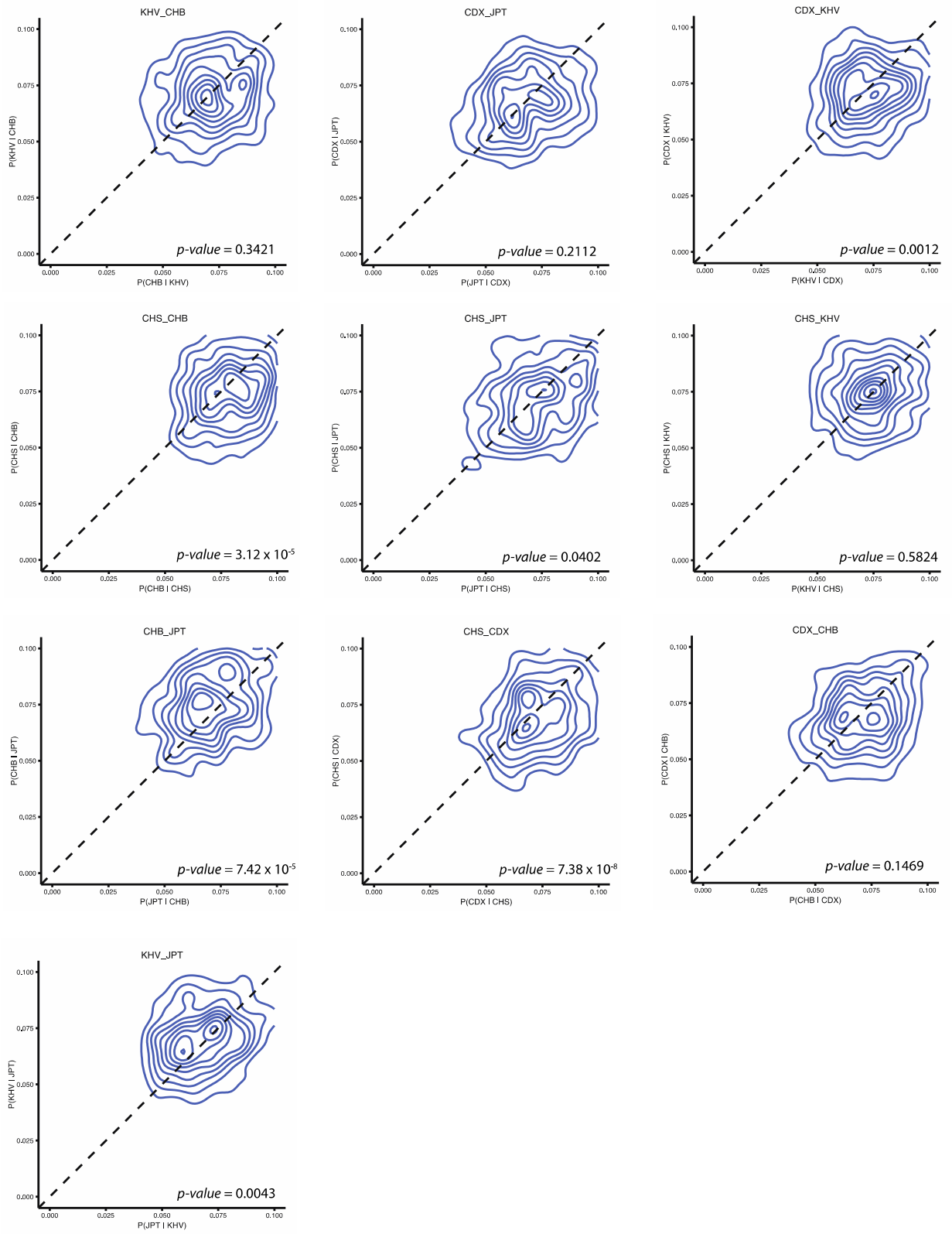

Figure S15. Reciprocal probability of sharing Neandertal sequences between East Asian populations. See Table S3 for population abbreviations.

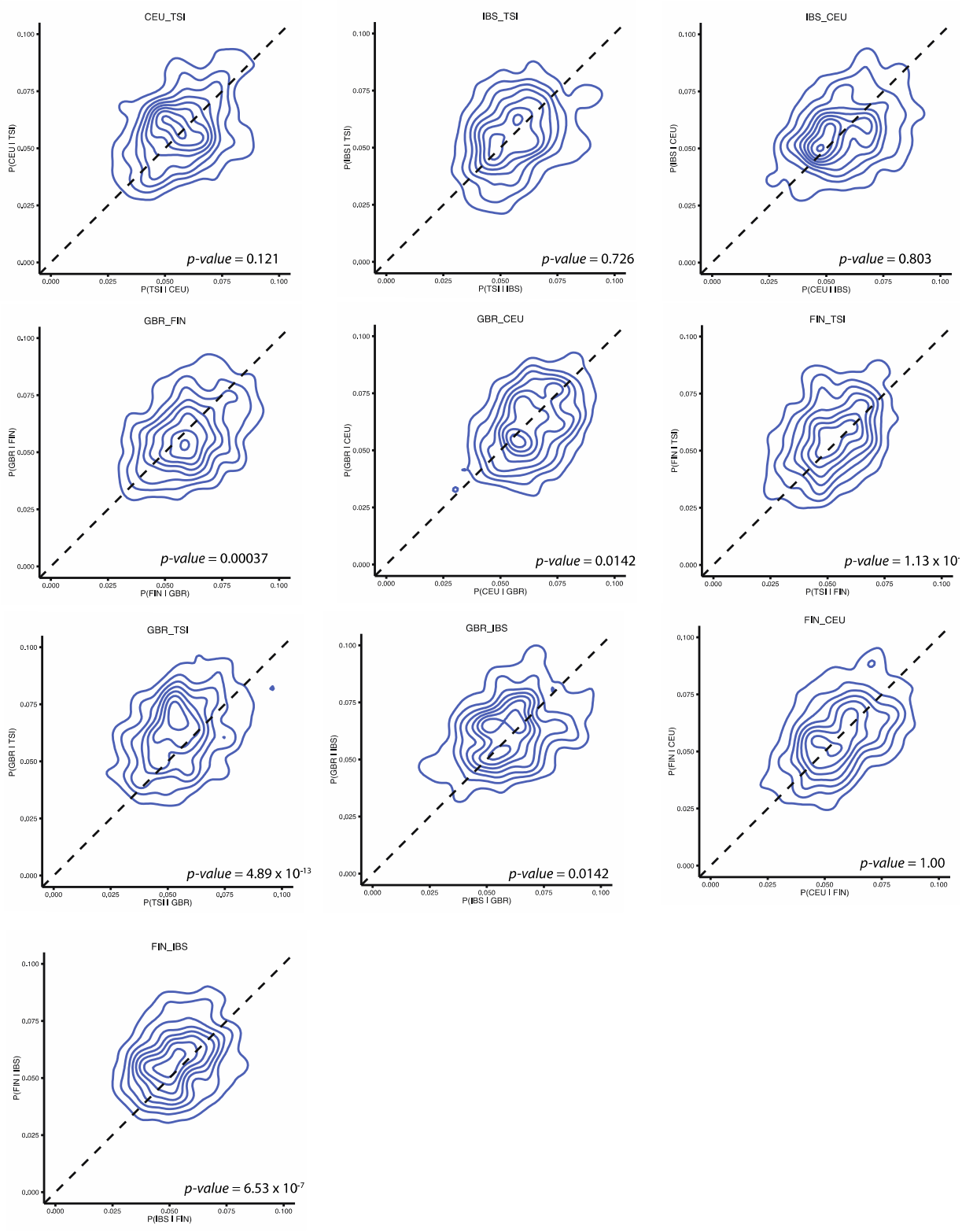

Figure S16. Reciprocal probability of sharing Neandertal sequences between European populations. See Table S3 for population abbreviations.

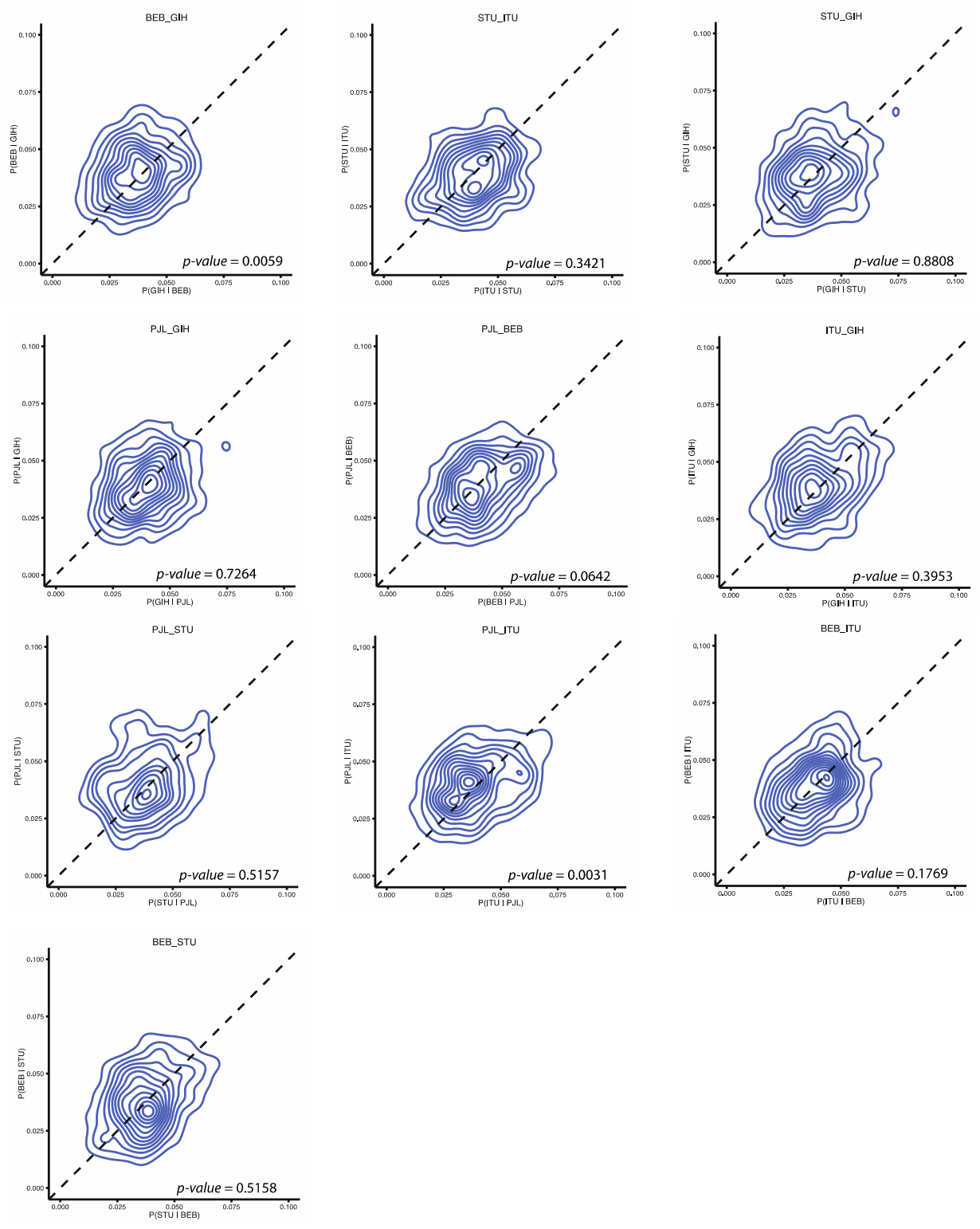

Figure S17. Reciprocal probability of sharing Neandertal sequences between South Asian populations. See Table S3 for population abbreviations.

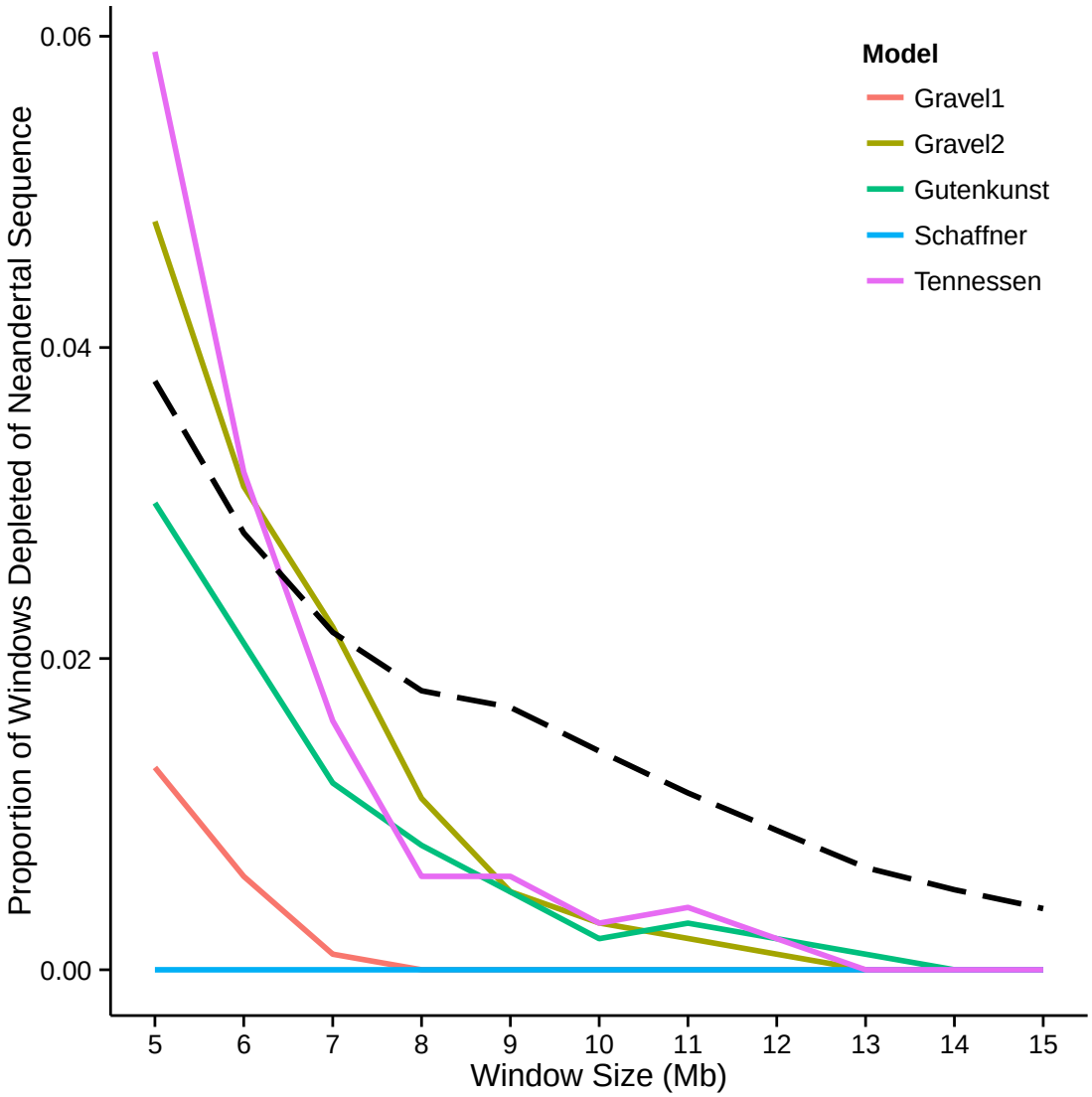

Figure S18. Proportion of windows significantly depleted of Neandertal introgression in Europeans and East Asians (dashed line) versus what is expected in five neutral demographic models. Due to uncertainty about modern human demography, we simulated sequence data under five demographic models for Africans, Europeans, and East Asians. The proportion of windows significantly depleted for Neandertal ancestry in Europeans and East Asians is shown for each model independently, and for observed data (dashed line). Gravel1 (*65*; exons); Gravel2 (*65;* low coverage + exons); Gutenkunst (*66*); Schaffner (*67*); Tennessen (*64*).

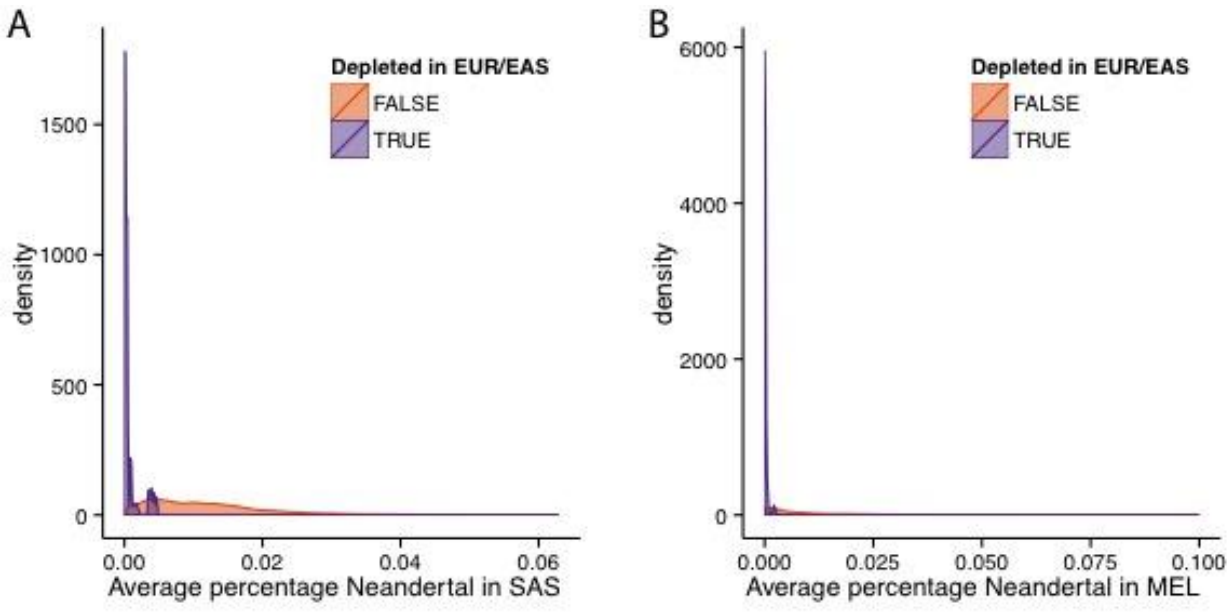

Figure S19. Shared depletions of Neandertal sequence between populations. Percentage Neandertal introgression in South Asians and Island Melanesians, for windows that are significantly depleted of Neandertal introgression in Europeans and East Asians (purple), and for regions that are not significantly depleted in Europeans and East Asians (orange).

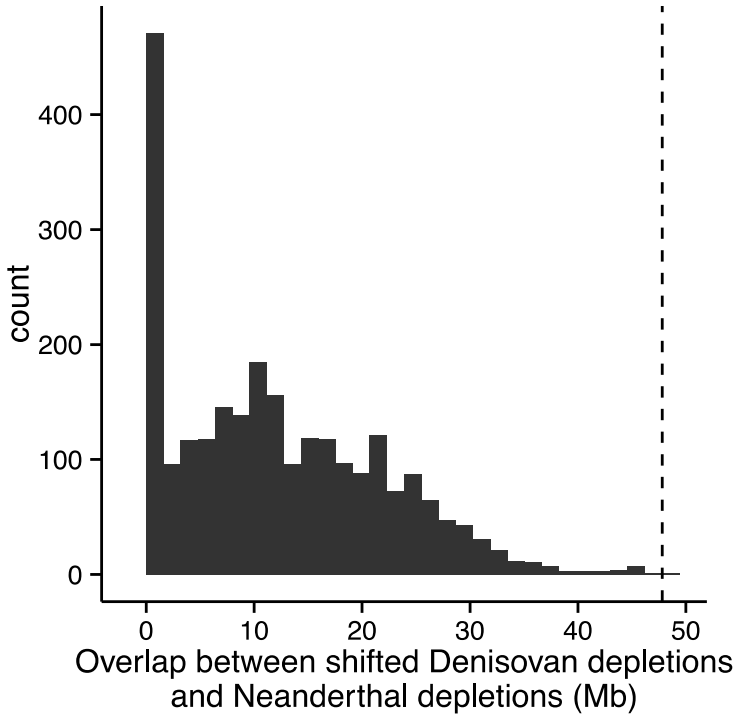

Figure S20. Neandertal and Denisovan overlap in randomized archaic deserts. Distribution of the amount of overlap between Neandertal and Denisovan deserts when randomized by shifting the location of Denisovan deserts by 1Mb along the length of the genome. This distribution is significantly below the observed overlap between these deserts, shown as a dotted line.

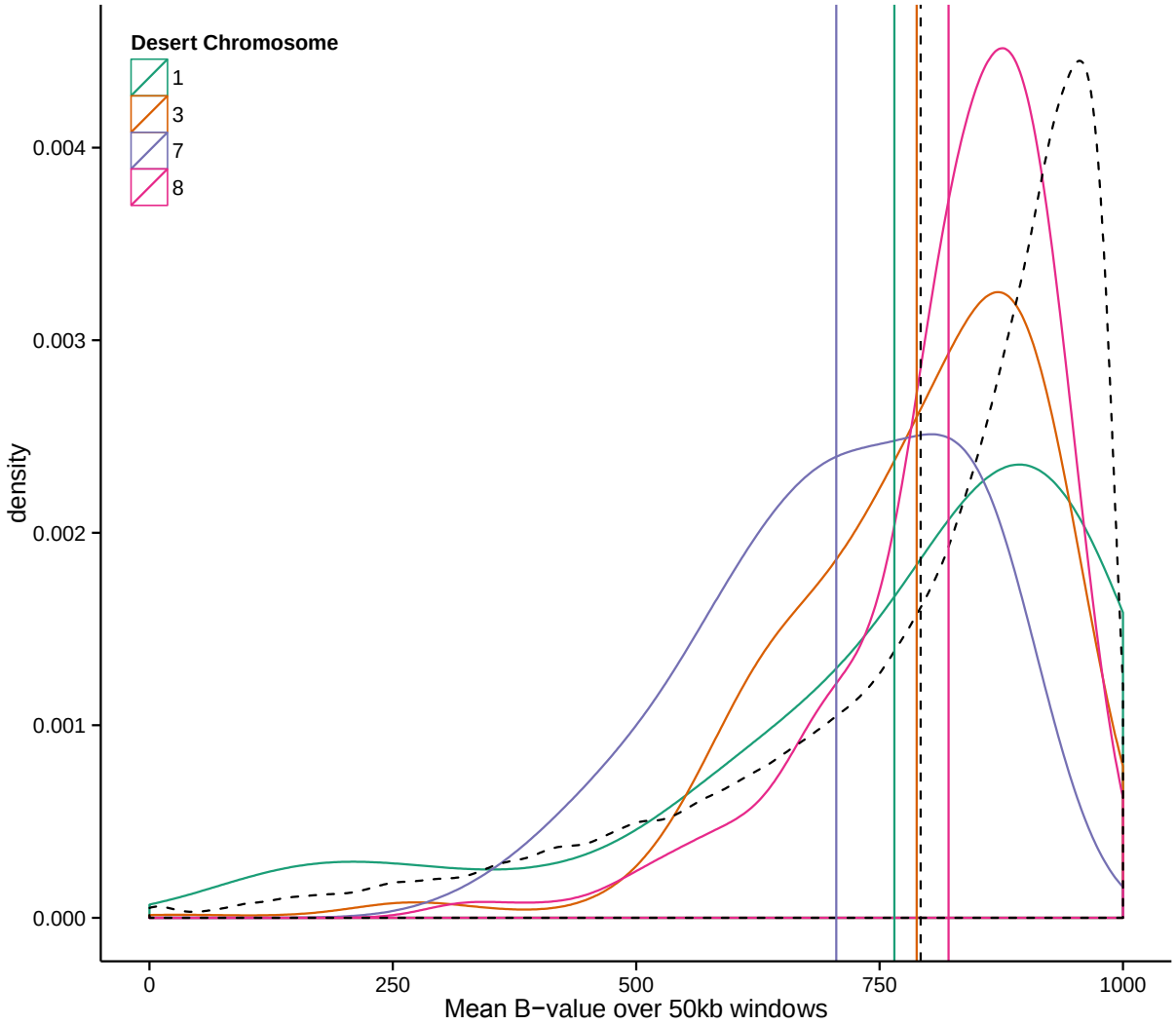

Figure S21. Levels of background selection in four large archaic depletions shared among all analyzed populations. Distributions of B-values for shared archaic depletions (solid curves), and genomic sequence (black dashed curve). Mean B-value for each desert and for genomic sequence is shown with vertical lines. The chromosome 7 desert on average has an 11% reduction in B-values from genomic sequence, representing an 11% reduction in diversity due to background selection.

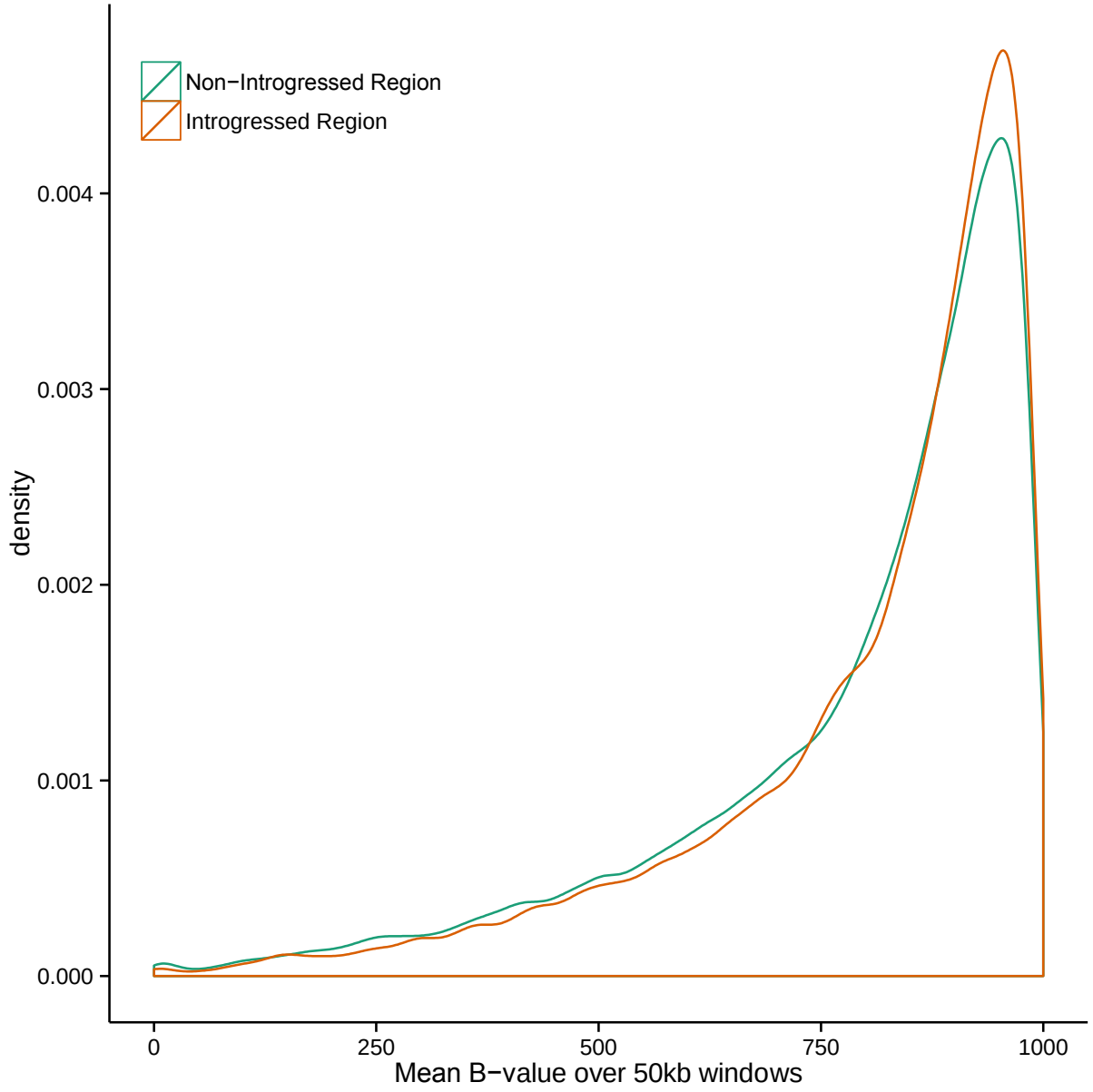

Figure S22. Background selection in introgressed and non-introgressed sequence. Distributions of B-values for introgressed (green) and non-introgressed (orange) regions.

References

1. S. Vattathil, J. M. Akey, Small amounts of archaic admixture provide big insights into human history. *Cell* **163**, 281–284 (2015). Medline doi:10.1016/j.cell.2015.09.042

2. R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Fritz, N. F. Hansen, E. Y. Durand, A. S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, S. Pääbo, A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010). Medline doi:10.1126/science.1188021

3. K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014). Medline doi:10.1038/nature12886

4. D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. Johnson, T. Maricic, J. M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. E. Eichler, M. Stoneking, M. Richards, S. Talamo, M. V. Shunkov, A. P. Derevianko, J. J. Hublin, J. Kelso, M. Slatkin, S. Pääbo, Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010). Medline doi:10.1038/nature09710

5. M. Meyer, M. Kircher, M. T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012). Medline

6. M. A. Yang, A. S. Malaspinas, E. Y. Durand, M. Slatkin, Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. *Mol. Biol. Evol.* **29**, 2987–2995 (2012). Medline doi:10.1093/molbev/mss117

7. S. Sankararaman, N. Patterson, H. Li, S. Pääbo, D. Reich, The date of interbreeding between Neandertals and modern humans. *PLOS Genet.* **8**, e1002947 (2012). Medline doi:10.1371/journal.pgen.1002947

8. J. D. Wall, M. A. Yang, F. Jay, S. K. Kim, E. Y. Durand, L. S. Stevison, C. Gignoux, A. Woerner, M. F. Hammer, M. Slatkin, Higher levels of Neanderthal ancestry in East

Asians than in Europeans. *Genetics* **194**, 199–209 (2013). Medline doi:10.1534/genetics.112.148213

9. P. Skoglund, M. Jakobsson, Archaic human ancestry in East Asia. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18301–18306 (2011). Medline doi:10.1073/pnas.1108181108

10. P. Qin, M. Stoneking, Denisovan ancestry in East Eurasian and Native American populations. *Mol. Biol. Evol.* **32**, 2665–2674 (2015). Medline doi:10.1093/molbev/msv141

11. B. Vernot, J. M. Akey, Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017–1021 (2014). Medline doi:10.1126/science.1245938

12. S. Sankararaman, S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, S. Pääbo, N. Patterson, D. Reich, The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014). Medline doi:10.1038/nature12961

13. J. S. Friedlaender, F. R. Friedlaender, F. A. Reed, K. K. Kidd, J. R. Kidd, G. K. Chambers, R. A. Lea, J. H. Loo, G. Koki, J. A. Hodgson, D. A. Merriwether, J. L. Weber, The genetic structure of Pacific Islanders. *PLOS Genet.* **4**, e19 (2008). Medline doi:10.1371/journal.pgen.0040019

14. Supplementary materials are available on *Science* Online.

15. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012). Medline doi:10.1534/genetics.112.145037

16. I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, B. Berger, C. Economou, R. Bollongino, Q. Fu, K. I. Bos, S. Nordenfelt, H. Li, C. de Filippo, K. Prüfer, S. Sawyer, C. Posth, W. Haak, F. Hallgren, E. Fornander, N. Rohland, D. Delsate, M. Francken, J. M. Guinet, J. Wahl, G. Ayodo, H. A. Babiker, G. Bailliet, E. Balanovska, O. Balanovsky, R. Barrantes, G. Bedoya, H. Ben-Ami, J. Bene, F. Berrada, C. M. Bravi, F. Brisighelli, G. B. Busby, F. Cali, M. Churnosov, D. E. Cole, D. Corach, L. Damba, G. van Driem, S. Dryomov, J. M. Dugoujon, S. A. Fedorova, I. Gallego Romero, M. Gubina, M. Hammer, B. M. Henn, T. Hervig, U. Hodoglugil, A. R. Jha, S. Karachanak-Yankova, R. Khusainova, E. Khusnutdinova, R. Kittles, T. Kivisild, W. Klitz, V. Kučinskas, A. Kushniarevich, L. Laredj, S. Litvinov, T. Loukidis, R. W. Mahley, B. Melegh, E. Metspalu, J. Molina, J. Mountain, K. Näkkäläjärvi, D. Nesheva, T. Nyambo, L. Osipova, J. Parik, F. Platonov, O. Posukh, V. Romano, F. Rothhammer, I. Rudan, R. Ruizbakiev, H. Sahakyan, A. Sajantila, A. Salas, E. B. Starikovskaya, A. Tarekegn, D. Toncheva, S. Turdikulova, I. Uktveryte, O. Utevska, R. Vasquez, M. Villena, M. Voevoda, C. A. Winkler, L. Yepiskoposyan, P. Zalloua, T. Zemunik, A. Cooper, C. Capelli, M. G. Thomas, A. Ruiz-Linares, S. A. Tishkoff, L. Singh, K. Thangaraj, R. Villems, D. Comas, R. Sukernik, M. Metspalu, M. Meyer, E. E. Eichler, J. Burger, M. Slatkin, S. Pääbo, J. Kelso, D. Reich, J. Krause, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014). Medline doi:10.1038/nature13673

17. D. Reich, K. Thangaraj, N. Patterson, A. L. Price, L. Singh, Reconstructing Indian population history. *Nature* **461**, 489–494 (2009). Medline doi:10.1038/nature08365

18. V. Plagnol, J. D. Wall, Possible ancestral structure in human populations. *PLOS Genet.* **2**, e105 (2006). Medline doi:10.1371/journal.pgen.0020105

19. A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis; 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). Medline

20. M. Sikora, M. L. Carpenter, A. Moreno-Estrada, B. M. Henn, P. A. Underhill, F. Sánchez-Quinto, I. Zara, M. Pitzalis, C. Sidore, F. Busonero, A. Maschio, A. Angius, C. Jones, J. Mendoza-Revilla, G. Nekhrizov, D. Dimitrova, N. Theodossiev, T. T. Harkins, A. Keller, F. Maixner, A. Zink, G. Abecasis, S. Sanna, F. Cucca, C. D. Bustamante, Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. *PLOS Genet.* **10**, e1004353 (2014). Medline doi:10.1371/journal.pgen.1004353

21. B. Vernot, J. M. Akey, Complex history of admixture between modern humans and Neandertals. *Am. J. Hum. Genet.* **96**, 448–453 (2015). Medline doi:10.1016/j.ajhg.2015.01.006

22. 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). Medline doi:10.1016/j.ajhg.2014.12.029

23. T. Maricic, V. Günther, O. Georgiev, S. Gehre, M. Curlin, C. Schreiweis, R. Naumann, H. A. Burbano, M. Meyer, C. Lalueza-Fox, M. de la Rasilla, A. Rosas, S. Gajovic, J. Kelso, W. Enard, W. Schaffner, S. Pääbo, A recent evolutionary change affects a regulatory element in the human *FOXP2* gene. *Mol. Biol. Evol.* **30**, 844–852 (2013). Medline doi:10.1093/molbev/mss271

24. S. Wickler, M. Spriggs, Pleistocene human occupation of the Solomon Islands, Melanesia. *Antiquity* **62**, 703–706 (1988). doi:10.1017/S0003598X00075104

25. G. R. Summerhayes, Island Melanesian pasts: A view from archaeology, in *Genes, Language, and Culture History in the Southwest Pacific*, J. S. Friedlaender, Ed. (Oxford Univ. Press, New York, NY, 2007; http://catalogue.nla.gov.au/Record/3992849), pp. 10–35.

26. M. G. Leavesley, J. Chappell, Buang Merabak: Additional early radiocarbon evidence of the colonisation of the Bismarck Archipelago, Papua New Guinea. *Antiquity* **78**, 301 available at http://www.antiquity.ac.uk/projgall/leavesley/ (2004).

27. R. Blust, The prehistory of the Austronesian-speaking peoples: A view from language. *J. World Prehist.* **9**, 453–510 (1995). doi:10.1007/BF02221119

28. R. D. Gray, A. J. Drummond, S. J. Greenhill, Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483 (2009). Medline doi:10.1126/science.1166858

29. M. Ross, A. Pawley, M. Osmond, *The Lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society: 2 The physical environment* (ANU Press, 2007; http://www.jstor.org/stable/j.ctt24hfkc).

30. M. Ross, Pronouns as a preliminary diagnostic for grouping Papuan languages, in *Papuan Pasts: Cultural, Linguistic and Biological Histories of Papuan-Speaking Peoples*, A. Pawley, R. Attenborough, J. Golson, R. Hide, Eds. (Pacific Linguistics, Canberra, 2005), pp. 15–65.

31. M. Dunn, A. Terrill, G. Reesink, R. A. Foley, S. C. Levinson, Structural phylogenetics and the reconstruction of ancient language history. *Science* **309**, 2072–2075 (2005). Medline doi:10.1126/science.1114615

32. S. A. Wurm, *Papuan Languages and the New Guinea Linguistic Scene* (Dept. of Linguistics, School of Pacific Studies, Australian National University, Canberra, 1975; http://nla.gov.au/nla.cat-vn2122276).

33. M. P. Cox, *The Encyclopedia of Global Human Migration* (Blackwell Publishing Ltd, 2013; http://onlinelibrary.wiley.com/doi/10.1002/9781444351071.wbeghm837/abstract).

34. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010). Medline doi:10.1101/gr.107524.110

35. P. Cingolani, A. Platts, L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012). Medline doi:10.4161/fly.19695

36. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin; 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011). Medline doi:10.1093/bioinformatics/btr330

37. J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, E. E. Eichler, Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002). Medline doi:10.1126/science.1072047

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

39. A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, W. M. Chen, Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010). Medline doi:10.1093/bioinformatics/btq559

40. S. R. Browning, B. L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007). Medline doi:10.1086/521987

41. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007). Medline doi:10.1086/519795

42. X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, B. S. Weir, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012). Medline doi:10.1093/bioinformatics/bts606

43. M. Jakobsson, N. A. Rosenberg, CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 (2007). Medline doi:10.1093/bioinformatics/btm233

44. N. A. Rosenberg, DISTRUCT: A program for the graphical display of population structure. *Mol. Ecol. Notes* **4**, 137–138 (2004). doi:10.1046/j.1471-8286.2003.00566.x

45. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). Medline doi:10.1093/bioinformatics/btp352

46. A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, W. S. Wong, G. Sigurdsson, G. B. Walters, S. Steinberg, H. Helgason, G. Thorleifsson, D. F. Gudbjartsson, A. Helgason, O. T. Magnusson, U. Thorsteinsdottir, K. Stefansson, Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012). Medline doi:10.1038/nature11396

47. C. D. Campbell, J. X. Chong, M. Malig, A. Ko, B. L. Dumont, L. Han, L. Vives, B. J. O'Roak, P. H. Sudmant, J. Shendure, M. Abney, C. Ober, E. E. Eichler, Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44**, 1277–1281 (2012). Medline doi:10.1038/ng.2418

48. J. Lachance, B. Vernot, C. C. Elbers, B. Ferwerda, A. Froment, J. M. Bodo, G. Lema, W. Fu, T. B. Nyambo, T. R. Rebbeck, K. Zhang, J. M. Akey, S. A. Tishkoff, Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150**, 457–469 (2012). Medline doi:10.1016/j.cell.2012.07.009

49. S. Wang, J. Lachance, S. A. Tishkoff, J. Hey, J. Xing, Apparent variation in Neanderthal admixture among African populations is consistent with gene flow from non-African populations. *Genome Biol. Evol.* **5**, 2075–2081 (2013). Medline doi:10.1093/gbe/evt160

50. S. N. Wood, Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Series B Stat. Methodol.* **73**, 3–36 (2011). doi:10.1111/j.1467-9868.2010.00749.x

51. S. Neph, M. S. Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman, A. K. Johnson, E. Rynes, M. T. Maurano, J. Vierstra, S. Thomas, R. Sandstrom, R. Humbert, J. A. Stamatoyannopoulos, BEDOPS: High-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012). Medline doi:10.1093/bioinformatics/bts277

52. M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, V. J. Carey, Software for computing and annotating genomic ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013). Medline doi:10.1371/journal.pcbi.1003118

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

54. L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, B. J. Raney, A. Pohl, V. S. Malladi, C. H. Li, B. T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R. A. Harte, M. Haeussler, L.

Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, T. R. Dreszer, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, W. J. Kent, The UCSC Genome Browser Database: Extensions and updates 2013. *Nucleic Acids Res.* **41** (D1), D64–D69 (2013). Medline doi:10.1093/nar/gks1048

55. K. D. Pruitt, T. Tatusova, D. R. Maglott, NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005). Medline doi:10.1093/nar/gki025

56. K. Prüfer, B. Muetzel, H. H. Do, G. Weiss, P. Khaitovich, E. Rahm, S. Pääbo, M. Lachmann, W. Enard, FUNC: A package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* **8**, 41 (2007). Medline doi:10.1186/1471-2105-8-41

57. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock; The Gene Ontology Consortium, Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000). Medline doi:10.1038/75556

58. B. Zhang, S. Kirov, J. Snoddy, WebGestalt: An integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **33** (Web Server), W741–W748 (2005). Medline doi:10.1093/nar/gki475

59. T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, R. Guigó, The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012). Medline doi:10.1101/gr.132159.111

60. S. Anders, W. Huber, Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010). Medline doi:10.1186/gb-2010-11-10-r106

61. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). Medline doi:10.1093/bioinformatics/btq033

62. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, T. J. Hubbard, GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012). Medline doi:10.1101/gr.135350.111

63. G. K. Chen, P. Marjoram, J. D. Wall, Fast and flexible simulation of DNA sequence data. *Genome Res.* **19**, 136–142 (2009). Medline doi:10.1101/gr.083634.108

64. J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G.

Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, Broad GO, Seattle GO, NHLBI Exome Sequencing Project, Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012). Medline doi:10.1126/science.1219240

65. S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, C. D. Bustamante; 1000 Genomes Project, Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11983–11988 (2011). Medline

66. 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). Medline doi:10.1371/journal.pgen.1000695

67. S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, D. Altshuler, Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576–1583 (2005). Medline doi:10.1101/gr.3709305

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

69. H. Reyes-Centeno, S. Ghirotto, F. Détroit, D. Grimaud-Hervé, G. Barbujani, K. Harvati, Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7248–7253 (2014). Medline doi:10.1073/pnas.1323666111

70. I. Pugach, F. Delfin, E. Gunnarsdóttir, M. Kayser, M. Stoneking, Genome-wide data substantiate Holocene gene flow from India to Australia. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 1803–1808 (2013). Medline doi:10.1073/pnas.1211927110

71. D. Reich, N. Patterson, M. Kircher, F. Delfin, M. R. Nandineni, I. Pugach, A. M. Ko, Y. C. Ko, T. A. Jinam, M. E. Phipps, N. Saitou, A. Wollstein, M. Kayser, S. Pääbo, M. Stoneking, Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* **89**, 516–528 (2011). Medline doi:10.1016/j.ajhg.2011.09.005

72. R. R. Hudson, Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002). Medline doi:10.1093/bioinformatics/18.2.337

73. The Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005). Medline doi:10.1038/nature04072