Supplementary Materials for

Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes

PingHsun Hsieh, Mitchell R. Vollger, Vy Dang, David Porubsky, Carl Baker, Stuart Cantsilieris, Kendra Hoekzema, Alexandra P. Lewis, Katherine M. Munson, Melanie Sorensen, Zev N. Kronenberg, Shwetha Murali, Bradley J. Nelson, Giorgia Chiatante, Flavia Angela Maria Maggiolini, Hélène Blanché, Jason G. Underwood, Francesca Antonacci, Jean-François Deleuze, Evan E. Eichler*

*Corresponding author. Email: eee@gs.washington.edu

Published 18 October 2019, *Science* **366**, eaax2083 (2019)
DOI: 10.1126/science.aax2083

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at science.sciencemag.org/content/366/6463/eaax2083/suppl/DC1)

Tables S1 to S21 (.xlsx)

Materials and Methods

Genotyping single-nucleotide variants (SNVs) and indels for SGDP samples

Paired-end Illumina data for 266 fully public samples were downloaded via the Simons Genome Diversity Project (SGDP) (*27*). These data were aligned to human reference genome GRCh37 (hs37d5) using BWA-MEM (v0.7.12) as described in Mallick et al. (*27*). We applied the HaplotypeCaller function in Genome Analysis Toolkit (GATK, version: 3.5-0-g36282e4) to call variants in each SGDP sample separately using the following command. "java -Xmx25g -XX:+UseSerialGC -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R hg19.fasta -variant_index_type LINEAR --variant_index_parameter 128000 -nct 5 -D dbsnp_138.b37.vcf --emitRefConfidence GVCF -rf BadCigar --min_base_quality_score 20 -I SGDP_sampleID.bam -o SGDP_sampleID.gvcf". We trained the VQSR model in GATK in order to recalibrate variant quality scores using "-an QD -an DP -an FS -an SOR -an ReadPosRankSum -an MQRankSum" and used ts_filter_level of 99.9.

To ensure genotype quality, we excluded variants that were 1) in low complexity regions (RepeatMasker, UCSC Genome Browser), 2) segmental duplications (SDs; WGAC, GRCh37), 3) in telomeric or centromeric regions (UCSC Genome Browser), 4) known indels with 10 bp flanking both ends, and 5) of quality score (QUAL) < 20. In addition, we removed variants that do not have valid human–chimpanzee alignment (*53*). Together these filters account for 378,779,062 autosomal bases in the genome. We identified a total of 34,532,567 autosomal SNVs in our SGDP samples after filtering. We noted that one of the East Asian samples (Daur_HGDP01217_M) has a missing genotype rate of 6.3% after filtering. Because we required that variants must be fully called in all of the downstream analyses and inferences, this sample was removed to maximize the number of sites remaining in our analyses.

We downloaded genotypes and BAM files for three published archaic hominin genomes—a Denisovan (*25*) and a Neanderthal from the Altai Mountains in Siberia (*26*) as well as a European Neanderthal from Croatia (*24*) via http://cdna.eva.mpg.de/Neanderthal/. Genotypes of this archaic panel of three genomes and the SGDP samples were combined using BCFtools (v1.5). An in-house Python script was supplied to ensure sites that are variable in one set, but missing in the other due to monomorphic in reference alleles, were properly merged. After merging, a total of 23,103,829 fully called autosomal SNVs remain in our data for downstream analyses.

Analysis of hominin-specific CNVs and shared CNVs between archaic and non-African samples

We performed an exploratory analysis to identify hominin-specific CNVs and those shared specifically between archaic and modern Eurasians. We applied the digital comparative genomic hybridization (dCGH) (*9*) CNV discovery method to a discovery panel of 20 publically available genomes, including the three archaic hominin and 17 diverse SGDP genomes, which were selected for their lowest variances

in sequencing coverage (*9*). Note that the FDR estimates for the dCGH calls for low-coverage genomes from the 1000 Genomes Pilot study (~4X and most reads with lengths <50 bp) are as low as 10% for a 1 kbp variant, and for regions >10 kbp it reduces to <2% (*54*). Given that the three archaic genomes are high-coverage (>30X) and have longer read length (median read length >88 bp), we expect a lower FDR for archaic CNVs. We identified 5,135 CNVs in this discovery panel and determined each individual copy number by rounding to the nearest integer. To infer the ancestral copy number and the distribution of copy number in contemporary humans, we genotyped these CNVs in 72 nonhuman primate (chimpanzee, gorilla, and orangutan) and 249 SGDP genomes. We determined hominin-specific CNVs using a parsimony approach: a CNV is hominin-specific if all nonhuman primate samples are fixed in diploid copy number 2 (CN2), but it is variable in copy number in at least one of the hominin samples (archaic and modern humans). In addition, a CNV is specifically shared between archaic and non-African samples if it is hominin-specific and at least one copy number (CN) genotype is only found in archaic and non-African Eurasian samples. The significances of the numbers of hominin-specific and archaic-and-non-African-specifically-shared CNVs were tested using 100,000 permutation simulations.

Structural variant calling and genotyping

To generate a maximally sensitive set of copy number variants (CNVs) in the SGDP samples, we carried out CNV calling for each sample using WHAMG (*55*), LUMPY (v0.2.13) (*56*), DELLY2 (v0.7.2) (*57*), digital comparative genomic hybridization (dCGH) (*9*), and Genome STRiP (v2.00.1611) (*58*). In short, while dCGH computationally infers copy numbers based on read-depth information across repeat-masked genomes, the others identify CNVs using read-mapping information, such as discordant reads, soft-clipped reads, and/or unmapped reads, etc. To identify CNVs in the three archaic genomes, we were able to apply dCGH (*9*), but not the other approaches, to each of these genomes. This is primarily for two reasons: <0.5% reads of these archaic genomes are paired-end reads and there are no unmapped reads in these downloaded BAM files. Read-depth profiles for the three archaic genomes were generated by realigning the BWA-MEM aligned reads to the reference genome using mrsFAST-ultra (*59*). GC-corrected read-depth coverage across the genome was done through a regression procedure previously described in Sudmant et al. (*9*).

Deletions, duplications, inversions and CNVs identified by the five CNV callers were merged with 'mergeSVcallers' (https://github.com/zeeev/mergeSVcallers; commit: 746c6d2). This method merges CNVs by type, requiring that the start and end of the overlapping CNV begin and finish within 1,000 bp of one another with a reciprocal overlap of 60%. One iteration of merging was done to avoid collapsing unique alleles into the same call. Unless mentioned otherwise, we genotyped CNVs using the sequence read-depth genotyper (*54*) and integrated the call set for the entire panel of samples. CNVs <50

bp or >10 Mbp were excluded because of poor genotyping quality. We further constructed a conserved CNV call set of 19,211 variants by only including CNVs if they (i) are identified by at least two different CNV callers and/or dCGH because of its low false discovery rate and unique ability to infer aggregate paralogous copy number in repetitive regions (*9*), (ii) have missing genotype rates <0.1, (iii) are polymorphic in copy number across samples, and (iv) have >500 unmasked bases in sequence.

Population structure

Population structure within the SGDP samples was examined using both ADMIXTURE (v1.23) (*51*). For these analyses, we excluded variants with minor allele frequency < 0.01, thinned the data (--thin 0.2), and pruned linked variants (--indep-pairwise 50 10 0.1) using PLINK (v1.9), resulting in 264,848 SNVs. ADMIXTURE analyses were applied using the number of ancestral populations (K) between 2 and 12 and using fivefold cross-validation and 20 bootstrapping replicates for each K. Based on the results of ADMIXTURE (K = 5 – 8) and geographic locations, we noted that seven African samples—Somali_Ayodo81S_F, MasaiMKK_NA21490_M, MasaiMKK_NA21581_M, Mozabite_HGDP01253_M, Mozabite_HGDP01274_F, Saharawi_SAH31_M, and Saharawi_SAH41_M—were estimated to have 25–77% Eurasian ancestries and were excluded from downstream analyses. We grouped the SGDP samples into eight focal populations: sub-Saharan Africans (AFR, n=33), Native Americans (AMR, n=20), East Asians (EA, n=47), Europeans (EUR, n=51), Melanesians (MEL, n=16), Middle Easterners (ME, n=22), South Asians (SA, n=38), and Siberians (SIB, n=22) for downstream population genetic inferences. We also did not further analyze the two Australian and six other Oceanian samples due to limited sample sizes.

Demographic inferences

To infer the demographic history of Melanesians, we used ∂a∂i (*50*) to build and fit demographic models for the population trio of AFR-EA-MEL. To ensure genotype quality for proper demographic inferences, we further excluded data in UCSC Genome Browser Self Chain database (if sequence identity >90%) as well as any known/called structural variants (Database of Genomic Variants, as of September 2016). To avoid possible biases in our demographic inferences due to natural selection, we also excluded coding sequences with 1000 bp flanking on both ends (RefSeq genes database from UCSC Genome Browser, downloaded September 2016). This results in an unfolded, non-genic joint allele frequency spectrum (AFS) of 3,632,680 SNVs from 409,234,894 autosomal bases, polarized using human–chimpanzee alignment (*53*).

A variety of models for a population trio were considered. We added an additional parameter, P_{flip}, in each model to account for the fraction of sites in the data, whose ancestral states are misidentified.

We estimated demographic parameters using derivative-based BFGS and fmin algorithms implemented in SciPy to optimize the composite likelihood. We used the Godambe information matrix to estimate the confidence intervals for model parameters and to adjust the statistics of likelihood ratio tests for model selection (*60*). The Godambe information matrix of each AFS was calculated through 100 bootstrap replicates generated from nonoverlapping 1 Mbp regions across the entire genotype data. All parameter point estimates reported in physical units were converted using a mutation rate of 1.5×10^{-8} per base per generation (*61*) and a generation time of 29 years.

Coalescent simulations

We used MaCS (*62*) to carry out whole-genome coalescent simulations. To explicitly account for local mutation rate heterogeneity in the genome, we followed the framework published in (*63*). Briefly, it is a three-step procedure. First, we estimated the population genetic mutation parameter $\hat{\theta}_j$ using ∂a∂i for each locus of 25,000 bases under the best-fit demographic model. We then simulate genomes using MaCS with a mutation parameter $\hat{\theta}_{max}$, the largest θ estimated among all of the windows. Finally, for each locus we adjusted its mutation rate by dropping $1 - (\frac{\hat{\theta}_j}{\hat{\theta}_{max}})$ of the simulated variants. To simulate recombination variation across the genome, we incorporated HapMap recombination map in our simulations (*64*).

For all of the models we simulated, in addition to the three populations in each of the population trios, we also included the chimpanzee (n=1), the Siberian (Altai Neanderthal, NDL Altai, n=1) and European (Vindija Neanderthal, NDL Vindija, n=1) Neanderthal, and the Denisovan (DNS, n=1) branches (**Figure S8**). Confidence intervals and point estimates for these relevant demographic parameters were drawn from previous studies (**Table S8**). For each simulation, to account for uncertainties in parameter estimates we randomly sampled values from the confidence interval of each parameter, assuming that they had a multivariate normal distribution. Whenever conversions between genetic and physical units for parameters are required, we used a mutation rate and a generation time randomly drawn from $[1\times10^{-8}, 2\times10^{-8}]$ per site, per generation (*61*) and [25, 30] years (*65*).

Tests for natural selection, archaic introgression, and population-stratified CNVs

To identify population-stratified CNVs, we used three statistics comparing between two groups: a focal population and the rest of the SGDP samples. First, we computed $V_{ST}(focal, allOthers) = (1 - \frac{V_S}{V_T})$, where V_S and V_T are the weighted mean and total variances of the two groups, respectively (*66*). For each CNV, V_{ST} falls between 0 and 1 and the larger it is, the more difference the two groups are in copy number. Second, we compared the distributions of copy numbers in the two groups using the Mann-Whitney U (*MWU*) test (two-sided test). Finally, we calculated the statistic $D_{median} := |$ M(integer CN,

166 focal) – M(integer CN, allOthers) |, where M(integer CN, group) is the median copy number in integer
167 form for a group. We determined a CNV is population-stratified in a focal population if (i) $V_{ST} > 0.1$,
168 (ii) Bonferroni p of the *MWU* test < 0.05, and (iii) $D_{median} > 0.5$.
169 We further look for evidence for selection and introgression using multiple population genetic
170 statistics and SNVs from sequences flanking population-stratified CNVs. To search for evidence of
171 positive selection in a focal population, we computed population branch statistic (*PBS*) (*2*) and extended
172 haplotype homozygosity (EHH) (*67*) for each focal population. We applied BEAGLE v4.1 (*68*) to phase
173 haplotypes in both real and simulated data to account for possible biases due to phasing errors. To detect
174 signatures of archaic introgression, we used the f_D (*30*), which is designed specifically to find loci with
175 excess ancestry sharing with an archaic population due to admixture. Following the definition of Martin et
176 al. (*30*), we defined the population relationships among three SGDP populations and an outgroup
177 (chimpanzee) to be (((P1, P2), P3), O) := (((AFR, focal), ARC), Chimpanzee), where ARC ∈ {DNS
178 (n=1), NDL (n=2)}. We calculated the estimator for f_D using the derived allele frequency *f* (with respect
179 to the ancestor of chimpanzees and humans) at site *i* as the following:
180 $$\widehat{f_D} = \frac{S(((P1,P2),P3),Outgroup)}{S(((P1,X),X),Outgroup)}$$
181
182 Population-stratified CNVs
183 We are interested in identifying variants in a human population that significantly differ in copy number
184 because selection might have driven the observed differentiation. Because the SGDP samples are from
185 more than 100 diverse populations, we carried out ADMIXTURE to help delineate population
186 relationships. While the ADMIXTURE analysis suggests that our samples are mostly represented by sub-
187 Saharan African, Native American, East Asian, Sahul Oceanian, and Western Eurasian ancestries (**Figure**
188 **S4**), we noticed that three additional populations, including South Asians, Siberians, and Middle
189 Easterners, can be further separated from the others. In addition, two Australian and six other Oceanian
190 samples were excluded from further analyses due to limited sample size. Thus, for the rest of analyses, we
191 grouped the SGDP samples into eight focal populations: sub-Saharan Africans (AFR, n=33), Native
192 Americans (AMR, n=20), East Asians (EA, n=47), Europeans (EUR, n=51), Melanesians (MEL, n=16),
193 Middle Easterners (ME, n=22), South Asians (SA, n=38), and Siberians (SIB, n=22). Note that seven
194 African samples were excluded due to high Eurasian ancestries (>22%) (**Methods**), which can confound
195 downstream population genetic inferences.
196 To identify population-stratified CNVs, we used three statistics comparing each focal population
197 with the remaining global population samples: (i) V_{ST}, a measurement of copy number variation;
198 (ii) *MWU* test, comparing the distributions of copy numbers; and (iii) D_{median}, quantifying the average

difference in copy number (**Methods**). These three statistics quantify the differences in copy number between a focal population and all the other SGDP samples. In all cases, we found less than 10% of the CNVs with V_{ST} > 0.1, Bonferroni *p*-value of the *MWU* test < 0.05, or D_{median} > 0.5 (**Figure S5**). Note that while we found little to no correlation between the sample sizes of the focal populations and the numbers of stratified CNVs identified by V_{ST} (Pearson's correlation = -0.16, p = 0.699) and *MWU* (Pearson's correlation = -0.19, p = 0.638), those identified by D_{median} do negatively correlate with sample sizes (Pearson's correlation = -0.73, p = 0.039). To conservatively determine if a CNV is stratified in a focal population, we used the following criteria: (i) V_{ST} > 0.1, (ii) Bonferroni *p*-value of the *MWU* test < 0.05, and (iii) D_{median} > 0.5 (**Figure S6**). While the numbers of stratified CNVs vary among focal populations (**Figure S6; Table S5**), we found that in all cases, under the null expectation it is highly unlikely to observe the number of population-stratified CNVs (*p*-values < 0.0105, 10,000 non-parametric permutation simulations; **Table S5**). Our analysis suggests that these candidates are unlikely to be false positives due to sampling errors. Intriguingly, Melanesians carry the largest number of highly stratified CNVs (n = 162) among the eight focal populations despite having the smallest sample size (**Table S5**), most likely due to increasing statistical power in a more homogeneous group than other focal populations.

, where $S(((P1,P2),P3),Outgroup) = \sum_i C_{(((\mathrm{A,B}),\mathrm{B}),\mathrm{A})}(i) - C_{(((\mathrm{B,A}),\mathrm{B}),\mathrm{A})}(i)$, and

$$C_{(((\mathrm{A,B}),\mathrm{B}),\mathrm{A})}(i) = \left(1 - f_{i,P1}\right) \times f_{i,P2} \times f_{i,P3}$$

$$C_{(((\mathrm{B,A}),\mathrm{B}),\mathrm{A})}(i) = f_{i,P1} \times \left(1 - f_{i,P2}\right) \times f_{i,P3}$$

In theory, the function *S* reaches its maximum when the population P2 is completely replaced by the P3 lineage or vice versa, and thus the X in the denominator is dynamically determined for each site as which of P2 and P3 has the highest derived allele frequency. In addition, we also applied *S** (*49*), which utilizes linkage information, to detect archaic introgression. To increase statistical power and identify candidate regions for selection and/or archaic introgression, all test statistics were calculated and summarized using predefined windows of 100 SNVs, with a sliding size of 50 SNVs. We assessed the statistical significance of our inferences using coalescent simulations and calculated all test statistics for windows in simulated data that are homologous to those in the real data. The *p*-value of each window was defined as the fraction of simulations with test statistic values greater than or equal to the observed value in the real data. A test for a window is significant if its *p*-value < 0.05.

We used BEAGLE v4.1 phased SNVs (and the bi-allelic CNV if desired) from the putative unique (copy number [CN] = 2) sequences flanking a candidate CNV in order the study the haplotype pattern among the samples. To summarize the pattern of haplotypes, we used SNVs with *PBS* > 0.5 to further classify each haplotype into a specific haplogroup, where the pairwise mutation distance is at most

5. To simplify our inferences and ease the complexity of display, in most of cases we focused on the first four major haplogroups and pull the rest of haplotypes into the "others" group.

Population genetic inferences

We used the method of Thomson et al. (*69*) to estimate the time to most recent common ancestor (TMRCA) for each candidate archaic introgressed locus, assuming that a divergence of 6 million years between human and chimpanzee and a generation time of 29 years (*65*).

To examine if the sharing of this duplication polymorphism between Melanesian and Denisovan is a result of incomplete lineage sorting (ILS) or recent gene flow, we reconstructed haplotypes for the duplication polymorphic site in the Melanesian and Denisovan short-read genomes. Reads were re-mapped to the assembled Melanesian contig, along with GRCh37 and KV880768.1, the contig of the ancestral locus of DUP_{16p12}. To ensure enough Denisovan sequence coverage on the assembled Melanesian contig, we focused on sequences with at least five Denisovan reads with MAPQ >30. For simplicity, we focused on ~10 kbp sequences at the unique portion of the duplication locus (yellow arrow within the red-dashed box in **Figure 3**), where eight SNVs were called using FreeBayes (v1.0.2). Haplotypes were inferred by applying BEAGLE (v4.1) to these SNVs along with a bi-allelic CNV. Phylogeny and divergence for these 10 kbp sequences, along with homologous sequences from GRCh37 and published nonhuman great ape assemblies (*53*) were inferred using Thomson's TMRCA estimator and BEAST (v2.5.0).

To test ILS, we calculated the probability of observing a sequence of L_{obv} bases shared between modern and archaic humans using a model of sequence decay over time (*6*). In short, under a model of ILS and neutral evolution, the expected length L_{exp} of a shared sequence between two populations separated by t generations is $1/(r \times t)$, where r is the recombination rate per base per generation of the locus. The length distribution of the shared sequence evolved in the two lineages is a sum of two exponential distributions, which follows a Gamma distribution with shape parameter 2 and rate parameter $1/L_{exp}$. Thus, the probability of sharing a sequence of L_{obv} due to ILS is

$$1- CDF\{Gamma\left(L_{obv}, shape = 2, rate = \frac{1}{L_{exp}}\right)\}$$

To test if the observation of the high frequency $DEL_{MEL\text{-}NDL}$ deletion-linked variant in Melanesians is likely a result of positive selection, we used coalescent simulations under the best-fit demographic models and conditional on variants with a similar age to the $DEL_{MEL\text{-}NDL}$ allele. Specifically, we computed the expected distribution of *PBS* values for those with derived allele frequency being within 30% of the frequency of the $DEL_{MEL\text{-}NDL}$ deletion allele observed among the sampled Melanesian chromosomes (i.e., 0.306–0.568). We determined the significance of the selection signal by computing the rank of the

observed *PBS* of $DEL_{MEL\text{-}NDL}$ in the distributions generated using simulations from the Melanesians, Africans, and East Asians.

Phylogenetic analyses

To infer the phylogenetic relationships for loci of interest in primates, we performed both maximum likelihood (RAxML, v.8.2.10) and Bayesian phylogenetic-based (BEAST v2.5.0) (*71*) analyses. For RAxML, we used the command "-m GTRGAMMA -f a -x 13345 -N autoMRE -p 14801". To run BEAST, we used 1) HKY, with GAMMA Category Count = 5, for the Site Model and 2) random local clock for Clock Model to explicitly test mutation rate on individual branch in the tree. For tree priors, we tested both Calibrated Yule Model (if desired) and Coalescent Bayesian Skyline model for individual phylogenetic analyses. While we kept most of the parameters of the priors as default, for the prior distributions of clock rate, we used Gamma(0.001, 1000) and human–chimpanzee or human–rhesus macaque divergence as the calibration using a log-normal(M=1.8, S=0.12) or log-normal(M=3.35, S=0.085) distribution, respectively. For each locus, we performed five independent runs to infer the phylogeny using a chain length of 50,000,000 samples and recorded every 2,000 samples. We used the accompany program Tracer (v.1.7.1) to determine the quality of each run and, in general, we used the first 10% as burn-in. All phylogenetic trees were plot using Figtree (v1.4.3).

To test signals of selection among RNA transcripts, we used the codon substitution model, codeml, in the PAML package (v14.9) (*43*). To construct a frame-aware sequence alignment for PAML, we first translated predicted open reading frame (ORF) sequences into amino acid sequences, followed by performing amino acid alignment using MAFFT (v7.407) (*72*), which was then used to build the frame-aware sequence alignment using the Perl program pal2nal.pl (v14) (*73*). We began with the computation of pairwise *dN/dS* (ω) ratios (PAML with runmode=-2, CodonFreq=2) and estimated their 95% confidence intervals (C.I.) using 1,000 bootstraps constructed by sampling the codons of the input sequence alignment, as described in (*43, 74, 75*). To search for evidence for positive selection acting on sites along particular lineages or clades, we performed likelihood ratio tests using the following models: (i) the free-ratio model (model=1) vs. strictly neutral model (model=0, fix_omega=1, omega=1); (ii) branch-site test of positive selection (model=2, NSsites=2, fix_omega=0) against the null model (model=2, NSsites=2, fix_omega=1, omega=1); (iii) branch-site clade model C (model=3, NSsites=2) against the null model (model=0, NSsites=1) (*43, 76*). The input phylogenetic tree was inferred using BEAST (v2.5.0) as described above. The probability of a site being under positive selection was calculated using Bayes empirical Bayes (BEB) (*43*).

PCR-based validation for the Neanderthal–Melanesian-shared chromosome 8p21.3 deletion

We created a PCR-based assay to test for the presence of the chromosome 8p21.3 deletion at the approximate position chr8:22981814-22988247 based on read-depth estimates from the SGDP data. Primers were designed to flank this deletion and would create a PCR product of a few hundred base pairs if the deletion was present in the individual. The set of primers were located at chr8:22982187 (atctcgactcaccacaacgtc) and chr8:22988614 (catgttgaaatgagaaaagtttgg) and in individuals with the deletion, it creates a PCR product that is 501 bp. Sequencing this product and aligning to the human reference (GRCh37) gives the actual breakpoints of the deletion at chr8:22982302-22988251 (5,950 bp). A second set of primers were designed within the deletion region to test for presence (Forward: gttggcagtgtgaggttgtg, Reverse: cacccaccagaaggacaact), which amplifies a 300 bp fragment from chr8:22987306-22987605. With a combination of these two PCR assays, we can determine the copy number of individuals for this deletion: PCR product for the first assay will indicate at least one chromosome has the deletion and PCR product for the second assay will indicate at least one chromosome does not have the deletion. We applied this assay to 16 blood-derived DNA Melanesian samples and reported the results (**Figure S56**).

Fluorescence *in situ* hybridization (FISH) experiments

Metaphase spreads and interphase nuclei were obtained from lymphoblast cell lines from four human HapMap individuals. Three were Papuans from Bougainville Island (GM10541, GM1543 and GM10539) while one was the Caucasian (EUR) individual GM12878, used as control. All these cell lines were purchased from Coriell Cell Repository. FISH experiments were performed using human fosmid and bacterial artificial clones (BACs) (**Table S13**) directly labeled by nick-translation with Cy3-dUTP (Perkin-Elmer), Cy5-dUTP (Perkin-Elmer) and fluorescein-dUTP (Enzo) as described by Lichter et al. (*77*), with minor modifications. Briefly, 300 ng of labeled probe were used for the FISH experiments; hybridization was performed at 37^{o}C in 2xSSC, 50% (v/v) formamide, 10% (w/v) dextran sulphate and 3 mg sonicated salmon sperm DNA, in a volume of 10 mL. Posthybridization washing was at 60^{o}C in 0.1xSSC for three times. Nuclei were simultaneously DAPI stained. Digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments). DAPI, Cy3, Cy5 and fluorescein fluorescence signals, detected with specific filters, were recorded separately as grayscale images. Pseudo-coloring and merging of images were performed using Adobe Photoshop software.

DNA sample preparation and whole-genome sequencing using PacBio technology

Sample HGDP00550 was chosen from the HGDP-CEPH (Human Genome Diversity Project-Centre d'Étude du Polymorphisme Humain) panel for long-read sequencing, and frozen cells were provided by the CEPH. DNA was isolated as previously described (*53*) and genomic libraries were prepared for DNA sequencing. For PacBio sequencing, we prepared one DNA fragment library (40–50 kbp inserts) using Megaruptor (Diagenode) shearing at the 60 kbp setting. After SMRTbell preparation per the "Procedure & Checklist - Preparing >30 kb Libraries Using SMRTbell® Express Template Preparation Kit" (PacBio), the library was size-selected with the BluePippin™ system (Sage Science) at a minimum fragment length cutoff of 40 kbp. Single-molecule, real-time (SMRT) sequence data were generated using the PacBio Sequel instrument with Sequel Binding and Internal Ctrl Kit 2.1, Sequel Sequencing Kit 2.1 v2, MagBead cleanup, diffusion loading, and acquisition times of 10- or 20-hour movies. A total of 22 SMRT Cell 1M v2 and 3 SMRT Cell 1M v2 LR cells were processed yielding 73.8-fold (ROI/3.2 G) or 75.2-fold (raw/3.2G) whole-genome sequence data. The average subread length was 18.2 kbp with a median subread length of 12.9 kbp and N50 subread length of 34.8 kbp.

Identification of integration site for the Melanesian–Denisovan-specific 16p12.2 duplication

To identify the putative integration site, we constructed pseudo mate-pair reads using long-read data extending over duplication junctions at 16p12.2 (**Figure S40**). Specifically, we set to split individual long reads using an initial window size of 5 kbp and a step size of 1 kbp. Consider hypothetical PacBio read of length 12 kbp, we split this read such that we create a 5 kbp portion on the left and leaving the rest of the PacBio read (7 kbp) on the right. Then we move the cut site by 1 kbp to the right creating the left portion of the PacBio read of 6 kbp and the other 6 kbp portion on the right. We iterated this procedure until the right mate read equaled 5 kbp. The resulting pseudo mate-pairs were then mapped to the reference genome (hg38) in a paired-end fashion using BWA-MEM (version 0.7.15-r1140) with '-x pacbio' parameter. The putative integration site of this duplication was identified by discordant read pairs. Finally, we used Canu (v1.5) and the following command to assemble contigs with long reads showing mapping evidence to chromosome 16p11.2:

canu -pacbio-raw $1 genomeSize=50000 corOutCoverage=300 corMhapSensitivity=high corMinCoverage=1 gnuplotTested=true -p outFile useGrid=false -d outFile contigFilter="2 20000 1.0 .75 2"

BAC library construction, processing, and assembly

GM10539, a Melanesian cell line from Coriell, was grown to 10^8 cells and embedded in agarose plugs, then lysed. Plugs are partially digested with ECOR1, run on pulsed field gel and slices from 100-200 kbp are cut. DNA is electro-eluted, ligated, and transformed into *E. coli* cells. 350,000 clones are picked by Norgren picker into 96 well plates for a 10X BAC library and stamped onto Performa II Genetix filters.

Probes for regions of interest were designed, radioactively labeled and hybridized to the Performa filters, washed, exposed to Phosphor screens, and scanned on Typhoon scanner. Positives are called and corresponding clones picked from the BAC library.

DNA from positive BAC clones were extracted as described previously (*41*). We prepared barcoded libraries from clone DNA using Illumina-compatible Nextera DNA sample prep kits (Epicentre, catalog number GA09115) as described previously (*37*) and carried out paired-end sequencing (125 bp reads) on an Illumina HiSeq 2500. Reads were then mapped to the reference genome, GRCh37, to identify singly unique nucleotide k-mers (SUNKs) (*54*). Non-overlapping BACs were pooled and sheared as described previously (*41*). Libraries were processed using the PacBio SMRTbell Template Prep kit following the protocol "Procedure and Checklist—20 kb Template Preparation Using BluePippin Size-Selection System," with the addition of barcoded adaptors during ligation. Up to ten barcoded libraries were then pooled at equimolar amounts and size-selected as a pool on the Sage PippinHT with a start value of 10,000–12,000 and an end value of 50,000. The resulting library was then sequenced on one Sequel SMRT cell 1M by diffusion using Sequel v3.0 chemistry. We performed *de novo* assembly of pooled BAC inserts using Canu (v1.5). Reads were masked for vector sequence (pCC1BAC) and assembled with Canu, then subjected to consensus sequence calling with Arrow (https://github.com/PacificBiosciences/GenomicConsensus). We reviewed PacBio assemblies for misassembly by visualizing the read depth of PacBio reads in Parasight (http://baileylab.brown.edu/parasight/download.html), using coverage summaries generated during the resequencing protocol.

Assembly of the Melanesian 16p11.2 contig using Segmental Duplication Assembler (SDA)

Assembly of the Melanesian–Denisovan DUP_{16p12} duplication polymorphism at the 16p11.2 insertion site was accomplished using a pipeline based on the SDA method (https://github.com/mvollger/SDA) (*41*). Initially, long-read whole-genome data were mapped to a BAC haplotype (222 kbp) containing the core of the DUP_{16p12} duplication polymorphism. After this, paralogous sequence variants (PSVs) specific to the integration site were identified using SDA and reads containing these PSVs were phased and assembled resulting in a 252 kbp contig. Because the contig expanded upon the original BAC haplotype by 30 kbp, we repeated the experiment using the 252 kbp contig as the backbone. SDA was applied to the 252 kbp

contig and two of the resulting contigs shared 44/44 and 59/59 PSVs with the original backbone and extended the contig on the left and right side, respectively. This process was then iteratively reapplied to further extend the contig to an ultimate length of 787 kbp (**Table S14**), at which point it could be confidently overlapped with other locally assembled contigs, resulting in a ~1.8 Mbp contig.

PCR validation for the Melanesian–Denisovan-specific duplication at chromosome 16p12.2

In order to validate the presence or absence of the chromosome 16 Melanesian-Denisovan introgressed duplication variant, we developed a series of PCR and restriction enzyme assays. Across a 75 kbp region of the duplication (chr16:22,710,041-22,783,558), there were SNVs previously identified (*9*) that are present in the duplicated copy from Denisovan and are fixed alternatives from the human reference. Designing assays specific to these SNVs will allow us to identify which Melanesians have this duplication. We selected and tested 11 SNV sites across the 75 kbp duplicated region that show an alternate allele for the Papuans (matching the Denisovan allele) and are fixed for the reference allele in the control samples.

We designed PCR assays to amplify approximately 300 bp surrounding each of these sites and were able to successfully amplify and Sanger sequence those fragments in two Melanesians and three control samples (**Tables S19-S20**). We then selected three sites to create a restriction digests test, wherein we would select restriction enzymes that would cut our PCR product over the SNV site and nowhere else within the amplicon. Enzymes were found using NEB's product search tool and experiments followed the standard protocols listed by NEB. For site 13, we used AciI, which cuts the reference haplotype, creating fragments of 157 bp and 144 bp (and uncut = 301 bp). For site 22773497, we used MscI, which cuts the alternate (Papuan) haplotype, creating fragments 149 bp and 147 bp (uncut = 296 bp). For site 22768213, we used BsrDI, which also cuts the alternate (Papuan) haplotype, creating fragments 107 bp and 191 bp (uncut = 298 bp). The digested samples were visualized on a 2% agarose gel to visualize the size of DNA fragments present (**Figure S71**).

We also Sanger sequenced the precut PCR product from these three assays to determine if sequencing matched restriction digest results in a subset of 23 Papuan individuals; the results from the two assays matched. Sequencing these PCR products showed a variation in peak height at the SNV site based on the copy number of the duplication (CN3 showed half the peak height of the alternate allele compared to CN4; **Figure S71**). We tested a total of 242 additional Melanesian individuals for the presence of this Denisovan-introgressed duplication. These samples were extracted DNA obtained from the buffy coat from Melanesians across Papua New Guinea, mostly from New Britain, New Ireland, and Bougainville Islands (**Figure 2B; Table S11**), collected and housed by Dr. J. Friedlaender from the

Temple University as well as Drs. M. Brilliant and Dr. T. Carter at the Marshfield Clinic Research Institute.

Full-length non-chimeric (FLNC) transcripts for candidate regions using long-read cDNA sequencing

Total RNA was harvested from four human cell lines (GM10539, GM10541, GM10543, and GM12878) and one chimpanzee lymphoblast cell line and polyA RNA was purified by oligo-dT magnetic beads (Dynal: Thermo). Double-stranded cDNA with dual barcodes was prepared, amplified and subjected to hybridization capture in the manner detailed in (*52*). Hybridization probes were selected based on the genic candidate regions identified in Melanesian genomes. Probes were tiled along the exons of the following genes (Integrated DNA Technologies; IDT): *TNFRSF10A*, *TNFRSF10B*, *TNFRSF10C*, *TNFRSF10D*, *CHMP7*, *RHOBTB2*, and *NPIPB5*.

Following post-capture PCR, the amplified dsDNA was purified on magnetic beads (AMPure PB; PacBio) and then subjected to library preparation for long-read sequencing (SMRTbell Template Prep Kit 1.0; PacBio with barcoded SMRTbell adapters). SMRT sequencing was performed on the Sequel v2.1 chemistry (PacBio) with LR SMRT Cells with 2-hour pre-extension and 20-hour movies. Reads corresponding to each sample were extracted by their SMRTbell barcodes and circular consensus sequences were generated from the raw subreads using SMRT Link with minimum number of pass set to 1. The program *lima* in the *Iso-Seq3* pipeline (https://github.com/PacificBiosciences/IsoSeq3) was applied to remove the 5' and 3' dual barcodes and also obtain the unclustered FLNC reads. Parameters used were: lima --isoseq --dump-clips. We did not cluster the FLNC reads further because highly identical paralogous transcripts could undesirably cluster together in this step. Due to the variability of the yields of FLNC transcripts across samples and loci, we used data from various combinations of samples in subsequent analyses when applicable.

The resulting reads were mapped to the human reference (GRCh37) and/or other sequence contigs using minimap2 (v2.14-r883) with the option for long-read spliced alignment (-ax splice). Paralogs of the candidate genes were classified by identifying PSVs specific to the paralogs (*52*). In the case of identifying *NPIP* transcripts in the assembled Melanesian contig, we mapped all FLNC transcripts to two sets of reference sequences: (1) GRCh37 and the KV880768.1 contig (NCBI BioProject: PRJNA31257) and (2) the assembled Melanesian contig. We determined the best mapping location for each FLNC transcript by choosing the best mapping identity and focused on those with >99% identity in the alignment. In addition, we found that 14 fixed PSVs and two 13 bp indels can be used to identify *NPIPB16*.

Supplementary Text

Discovery of SNVs and CNVs

The analyses in this study were primarily based on genomic data of SGDP samples and three archaic hominin genomes—a Denisovan (*25*) and a Neanderthal (*26*) from the Altai Mountains in Siberia, and a Neanderthal from Croatia (*24*). Paired-end Illumina data for 266 fully public SGDP samples were generated and aligned to human reference genome GRCh37, as previously described in Mallick et al. (*27*). To call SNVs and indel variants in the SGDP samples, we used the GATK HaplotypeCaller (**Methods**). Diploid genotypes of the three archaic genomes were downloaded from Prufer et al. (*24*) and combined with the SGDP genotypes by position. After a series of quality control (QC) filtering procedures, 23,103,829 fully called autosomal SNVs remained in the data (**Methods**).

To maximize sensitivity for identifying CNVs, we applied a suite of five different CNV callers to each of the SGDP genomes (**Methods**). After initial filtering (**Methods)**, we discovered 368,256 autosomal CNVs (**Table S3**). Of these CNVs, 93.5% were identified by a single CNV caller (**Figure S2)**. To ensure the quality of CNV calls for downstream population genetic analyses, we focused on a conserved call set of 19,211 CNVs constructed by applying a variety of QC filters (**Methods**), including 4,705 bi-allelic deletions (length: median= 6424, s.d.=31278.74), 4,727 bi-allelic duplications (length: median=6117, s.d.=161628.3), and 10,234 multi-allelic CNVs (length: median=4257, s.d.=70565.54) (**Figure S3**). Overall validation rates of >83.07% (>68.18% and ~100% for duplications and deletions, respectively) for these 19,211 CNVs were determined using an orthogonal single-nucleotide polymorphism microarray-based CNV detection approach (**Table S4; Methods**).

Tests for positive selection and archaic introgression in Melanesians

A population-stratified CNV could be a result of a beneficial CNV targeted by natural selection and/or an initially introgressed variant from a diverged hominin species, subsequently raised to a high frequency through demographic processes. To test these hypotheses, we performed a window-based scan using F_{ST} (*78*), *PBS* (*2*), Tajima's *D* (*79*), nucleotide diversity (π), and EHH (*67*) to search for signatures of positive selection, and computing the f_D (*30*) and *S** (*49*) statistics to detect introgressed archaic hominin sequences (**Methods**). Note that we only calculated *S** for candidates of introgressed loci identified by f_D, if desired, as an orthogonal method to support the signals of archaic introgression.

To account for possible biases on the test statistics due to demographic processes, we performed large coalescent simulations based on 1,000 demographic models to construct the expected null distributions for the statistics (**Methods**). We used the site frequency spectrum-based demographic inference tool ∂a∂i (*50, 60*) (**Methods**) to infer the prehistory of Melanesians. To reduce the dimensions of parameter searching space, we chose to build models for the population trio AFR-EA-MEL. Among a

variety of models that we tested (**Table S6**), we found that the best-fit model with asymmetric migrations between the two populations (**Figure S7**, log-likelihood = -103386) significantly fit the data better (adjusted-D = 5.707, *p*-value of likelihood ratio test = 0.0168) than a model with symmetric migrations between EA and MEL (**Figure S7**, log-likelihood = -104590). Our best-fit model suggests that the ancestors of Africans and non-Africans diverged ~74 thousand years ago (kya; 95% C.I.: 73,321–75,199), followed by the divergence between Melanesians and East Asians at ~52 kya (95% C.I.: 51,450–52,591), a compatible to a recent estimate of split time between aboriginal Australians/Papuans and Eurasians (*19*). In addition, the complexity of our best-fit model, such as the moderately high gene flow from East Asians to Melanesians (95% C.I. for N_{AMH}*$m_{MEL\text{-}EA}$ = 1.109–1.128 > 1; **Table S7**), highlights the importance of controlling biases due to demographic processes in downstream selection and introgression inferences in Melanesians.

In our coalescent simulations, we added parameters for branches prior to all the modern human branches, including those of archaic humans as well as the one leading to chimpanzee (**Figure S8; Table S8**). Parameter values associated with these additional branches were uniformly drawn from the confidence intervals published in literature to account for the uncertainties of those parameter estimates (**Table S8**). Our simulations also account for genomic heterogeneity in mutation and recombination rates (**Methods**). In general, our whole-genome coalescent simulations recapitulate the genetic variation patterns of SNVs observed in real data (**Figure S9A**), while the genomic distributions of test statistics (e.g., F_{ST}) are remarkably different between simulated and real data (**Figure S9B-C**), demonstrating the power of our inference to identify candidates of non-neutrally evolved loci. Unless stated otherwise, all *p*-values for the tests of selection and introgression scans are based on these parametric coalescent simulations.

We determined selective and introgressed signals for each of the highly stratified CNVs if they are flanked by significant windows (*p*-value < 0.05). Within Melanesians, we identified signatures of positive selection at 37 distinct CNV loci, and signals of introgression at 24 and 28 CNV loci using Neanderthal and Denisovan genomes as archaic references, respectively; interestingly, 19 were found using either reference (**Tables 1, S9-S10**). Notably, we found that stratified CNVs are significantly associated with candidate loci of positive selection (*p*-value = 0.032, a permutation test of 10,000 non-parametric simulations), but not with archaic introgression signals (*p*-value = 0.228) (**Figure S10**). The strong association between stratified CNVs and selection candidates is consistent with the predicted large effect sizes of CNVs, highlighting their important roles in adaptive evolution.

Signals of the Melanesian–Denisovan-specific duplication on chromosome 16p12.2 consistent with positive selection and archaic introgression

Among the most differentiated Melanesian CNVs, the top two loci (Bonferroni's *p*-value of CNV stratification test < 2.5×10^{-32}) were previously reported by Sudmant et al. (2015) (*9*). First, at the locus of a 5 kbp duplication upstream of *METTL9* (chr16:21,596,722-21,601,720), we found significant signals for archaic introgression (*p*-value(f_D) ≤ 0.033; **Tables S9-S10**), but not for selection (*p*-value = 0.127). At the second region, a 73.5 kbp duplication variant, spanning both *MIR548AA2* and *MIR548D2*, was found only in the Melanesian and Denisovan samples (DUP_{16p12}, chr16:22,710,041-22,783,558; **Figure 2A)**. Metaphase and interphase FISH experiments of three Melanesian cell lines (**Methods**) confirm the presence of DUP_{16p12} (**Figures 2C, S38; Table S12**). We detected significantly elevated f_D scores using the Denisovan individual (*p*-value < 0.001; **Figures S34, S36**), but not the Neanderthals (*p*-value = 0.178; **Figure S36**), as the archaic reference. This is consistent with a result of interbreeding events between Melanesians and Denisovan-like archaic humans (*9*, *21*, *25*). Note that although there were putative selection signals at the DUP_{16p12} locus (**Figure S34**), we found elevated Tajima's *D* values across this locus (**Figure S37**), consistent with a pattern of an excess of heterozygosity, which is likely driven by the collapse of PSVs. To assess the prevalence of this duplication variant, we designed a sequencing assay to genotype an independent set of 242 diverse Melanesians from eight different population groups (**Methods**). We confirmed this variant is present at high frequency across these Melanesian populations (>79% samples; **Figure 2B; Table S11**). While DUP_{16p12} is present at high frequency in all Melanesian groups, introgression is nearly complete among lowland populations of West and East New Britain.

By performing additional FISH experiments, we were able to map the extra copies within a duplication block at chr16:28.3-30.4 Mbp, most likely at 16p11.2 between 29.03 and 29.66 Mbp (i.e., between the green and blue probes in **Figure S39**). In addition, we generated 75X coverage long-read sequence data targeting the 16p12.2 ancestral duplication locus from a Melanesian genome (**Methods**). To refine the putative integration site of the duplicate sequences, we constructed pseudo mate-pair reads by splitting long-read data extending over duplication junctions at 16p12.2, and then searched for read pairs linking 16p12.2 and 16p11.2 loci (**Methods; Figure S40**). Using this information, we further refined the integration locus to a 200 kbp interval (29.47 to 29.67 Mbp) mapping adjacent to an *NPIP* core duplicon and near *BOLA2* and *SMG1P2*. To sequence resolve the DUP_{16p12} copy number polymorphism, we generated a Melanesian large-insert BAC library (GM10539). We constructed two haplotypes of 222 kbp and 133 kbp, partially confirming the structure of DUP_{16p12} reported in (*9*) (**Figures S41-S42**) using five BAC contigs (NCBI BioProject: PRJNA522307). In order to fully assemble the entire locus *ab initio*, we used the haplotypes as the initial seeds to pull down long-read Melanesian whole-genome sequencing data and iteratively applied the SDA method (*41*) (**Figure S43**; **Table S14**).

We generated a ~1.8 Mbp sequence contig spanning more than 900 kbp of complex SDs (**Figure 3A**). To assess the quality of the *de novo* assembly of the contig, we aligned the sequences between the finished BAC haplotypes and the assembly, and confirmed the organization and sequence accuracy (99.86%) (**Figure S44**). We noted the observed sequence differences are likely due to the genome and the BAC library coming from two unrelated Melanesian samples. Notably, the sequence-resolved assembly shows that the actual length of DUP_{16p12} duplication polymorphism is ~383 kbp, which is larger than previous thought (*9*).

To reconstruct the evolutionary history of the DUP_{16p12} duplication polymorphism, we performed a series of phylogenetic analyses using BEAST (v2.5.0) (**Methods**) and a sample of loci across the duplication allele. Our inference results suggest that the variant originated from a series of complex structural changes involving duplication, deletion, and inversion events ~0.5–2.5 million years ago (Mya) within the Denisovan ancestral lineage, which subsequently inserted into chromosome 16p11.2 (chr16:29,640,235-29,640,459) between 0.2–0.5 Mya (**Figures 3B, S45-S46; Table S15**). To examine if the sharing of this CNV between Melanesian and Denisovan genomes is a result of ILS or recent gene flow, we remapped the short-read Illumina data to the assembled Melanesian contig, along with GRCh37 and KV880768.1, which is the contig of the ancestral locus of DUP_{16p12}. We inferred a phylogeny using ~10 kbp sequences at the duplication polymorphism site, where enough high-quality Denisovan reads were present (>5 reads with MAPQ > 30), along with homologous sequences from GRCh37 and published nonhuman great ape assemblies (**Methods, Figure S48**). Our inference results show that all Denisovan and Melanesian sequences that carry the DUP_{16p12} polymorphism forms a single clade and share TMRCA ~0.06–0.17 Mya. Importantly, the more recent ancestry of these duplication sequences than the Denisovan-modern human divergence (>400 kya) is consistent with the hypothesis of recent gene flow introducing this variant into populations ancestral to Melanesians.

Chromosome 16p11.2 is one of most complex regions in the human genome, where recurrent deletions and duplications, mediated by a complex set of SDs, have been known to associate with diseases (e.g., ~1% of cases of autism (*40*)) and implied their importance in human evolution (*15*). Interestingly, we observed both significantly elevated *PBS* (p-value < 0.012) and f_D (p-value < 0.021, archaic ref = Denisovan) scores at the unique, diploid sequences flanking the 16p11.2 complex region (**Figure S47**). Consistent with the analyses for the DUP_{16p12} locus above, the archaic introgression signals were completely diminished when Neanderthals were used as the archaic reference in the f_D calculation (p-value > 0.193), suggesting the scenario of adaptive introgression from Denisovan-like archaic humans into Melanesians. We also note that the Melanesian duplication polymorphism harbors extra copies of SD sequences that are absent from most human populations, including an additional member of the *NPIP* family (*42*). To explore the *NPIP* coding potential at this locus, we used FLNC transcripts from two

Melanesian (GM10539 and GM10541) fibroblast cell lines (**Methods**). We identified FLNC transcripts that maintain the same ORF and encode a novel member of the *NPIPB* family, *NPIPB16* (1,206 amino acids), mapping uniquely to the duplication polymorphism (**Figure 3C**). This Melanesian copy shows elevated pairwise *dN/dS* ratios when compared to other closely related *NPIPB* genes (RefSeq release 109) (**Figures S49-S50**). Using a phylogenetic branch site test (*43*), we identified 32 sites are likely positively selected, including a cluster of 28 sites locates at the last exon of *NPIPB16* (**Figure 3C**). Sequence analysis shows that this cluster is likely due to two indel events of a repeat motif (GAGCGTCTGCGGG)—an indel upstream to the cluster caused frameshifting, while the other one downstream to the cluster restored the original frame of the peptide sequence—resulting in a local novel amino acid sequence at the last exon that is unique and only found in *NPIPB16* (**Figure S51**), rather than a series of independent amino acid substitution events.

Archaic introgression of CNVs at chromosome 8p21.3 between Melanesians and Neanderthals

The most striking signals from our window-based selection (PBS/F_{ST}) scan and archaic introgression test (p-value < 0.005) in the Melanesian samples center at chr8:22,969,611-23,045,069 (**Figures 4A-B, S52**). This region encompasses two significantly stratified CNVs in Melanesians (**Figure 4B**), a 6 kbp deletion ($DEL_{MEL\text{-}NDL}$, chr8:22,982,302-22,988,251, Bonferroni's p-value < 8.9×10^{-11}) and a 31 kbp duplication of *TNFRSF10D* (DUP_{10D}, chr8:22,991,347-23,022,738; Bonferroni's p-value of *MWU* < 1.5×10^{-6}) at the centromeric side of the deletion locus. Interestingly, the copy number estimates of $DEL_{MEL\text{-}NDL}$ and DUP_{10D} are significantly and negatively correlated in Melanesian samples (Pearson's $\rho = -0.64$, $p < 0.05$), but not in other SGDP populations (**Figure S53**), showing a strong linkage between the two CNVs in Melanesians.

At the DUP_{10D} locus, we observed an excess of heterozygosity and a pattern of allelic imbalance at this locus only in individuals from Melanesian (15 out of 16 samples), African (7 out of 33 samples), and the three archaic genomes (**Figures 4B, S53-S54)**. Both patterns of excess of heterozygosity and allelic imbalance are consistent with the presence of PSVs due to the collapse of duplicate copies. With these lines of evidence, we determined that only individuals who show an excess of heterozygosity and allelic imbalance at this locus harbor the DUP_{10D} duplication allele. The deletion allele of $DEL_{MEL\text{-}NDL}$, on the other hand, was only observed in the Melanesian, the two Neanderthal, and the South Asian Punjabi genomes (allele counts = 14/32, 4/4, and 1/76, respectively). As an independent data set, we computationally genotyped 35 Papuans from Vernot et al. (*23*) and estimated a compatible frequency of 0.457 for the deletion allele. To further validate the $DEL_{MEL\text{-}NDL}$ variant, we designed a PCR assay and tested 16 randomly selected DNA samples from blood-derived materials, as well as three Melanesian

fibroblast derived cell lines (**Methods**), and confirmed the presence of the deletion allele in eight out of the 16 blood-derived DNA samples (**Figure S56**) and one of the three cell lines (**Table S12**).

To assess the temporal and spatial frequency distributions for both $DEL_{MEL\text{-}NDL}$ and DUP_{10D}, we used two additional large data sets from the Great Ape Project (GAP) (*28*) and the 1000 Genomes Project (1KG, Phase 3). Among the GAP genomes, while the absence of the $DEL_{MEL\text{-}NDL}$ indicates that this deletion variant is likely derived, we found evidence for the presence of the DUP_{10D} variant in all of the GAP genomes (**Figures S53, S55C**), suggesting that the duplication allele is ancestral to great apes. The DUP_{10D} variant segregates at low frequencies across the 1KG populations (<0.025) but is completely absent in the European populations (**Figure S57**). On the other hand, 64 out of the 2,504 samples from all five continents show a reduced level of sequence coverage at the $DEL_{MEL\text{-}NDL}$ locus (CN estimates < 1.5; **Figure S57; Table S17**). We performed two orthogonal approaches to examine the presence or absence of the deletion variant in other populations. First, we identified seven tag SNVs that are in nearly complete linkage with the $DEL_{MEL\text{-}NDL}$ variant in the SGDP samples ($r^2 > 0.9$ and $D' > 0.9$; **Table S18**) and used them as surrogates to understand the geographic allele frequency distributions of the deletion variant in the 1KG samples. In all seven cases, the deletion tag alleles were only found in South Asian populations, but at low frequencies (<0.07; **Figures 5A, S66**). In a low-coverage Neanderthal genome, Mezmaiskaya1 (*24*), the deletion tag alleles are fixed in all three of seven sites where it has sequence coverage, suggesting that Mezmaiskaya1 is homozygous for the deletion variant (**Table S18**). Similarly, our PCR experiment results also confirm that $DEL_{MEL\text{-}NDL}$ is geographically restricted to the South Asian populations at low frequencies (<0.068; **Table S17**).

The observation of $DEL_{MEL\text{-}NDL}$ in both Neanderthals and non-African populations as well as the significant archaic introgression signal (p-value of $f_D = 0.003$, p-value of $S^* = 0.043$; **Figure S67**) suggest that the sharing is likely a result of archaic introgression. We noted that the introgression signal around $DEL_{MEL\text{-}NDL}$ became insignificant when we used the Denisovan genome as the archaic reference in the f_D analysis (p-value=0.06; **Figure S68**). To further investigate this hypothesis, we analyzed genetic variation patterns of the unique sequence of 18,500 bp at the telomeric side of *TNFRSF10D* that spans the locus of $DEL_{MEL\text{-}NDL}$ (chr8:22,972,880-22,991,380). Phasing 56 SNVs in this region, along with the two bi-allelic CNVs for all SGDP samples, we found that the 15 $DEL_{MEL\text{-}NDL}$ linked haplotypes in SGDP samples are more closely related to the Neanderthal haplotypes than any other samples (**Figure S69**). Interestingly, all 14 Melanesian $DEL_{MEL\text{-}NDL}$ haplotypes are almost identical and equally distant to the four Neanderthal haplotypes.

Both the maximum likelihood estimated phylogenetic tree (log likelihood = -21578; **Figure S70A**) and haplotype network (**Figure 5B**) analyses show that all $DEL_{MEL\text{-}NDL}$ haplotypes form a monophyletic clade, suggesting a common ancestry of these haplotypes. We estimated that the time to

TMRCA for all modern and archaic sample haplotypes is 601 kya (95% C.I.: 430–853 kya) and is consistent with the divergence between modern humans and Neanderthal/Denisovan (*24*, *26*). We estimated that TMRCA of the Neanderthal and Melanesian $DEL_{MEL\text{-}NDL}$ haplotypes is 40 kya (95% C.I.: 0–122 kya) and that of all 19 $DEL_{MEL\text{-}NDL}$ haplotypes in SGDP is 120 kya (95% C.I.: 0–241 kya). The much younger TMRCA of these $DEL_{MEL\text{-}NDL}$ haplotypes than the Neanderthal-modern human divergence provides evidence for that the sharing of the $DEL_{MEL\text{-}NDL}$-linked haplotypes between the two species was a result of recent gene flow, as opposed to ILS. Moreover, under a model of ILS and reasonable demographic parameters we estimated that the probability of sharing a sequence of 18.5 kbp between modern humans and Neanderthals is highly unlikely (p-value ≤ 0.04; **Methods**). Finally, we hypothesized that the observed high frequency and homogeneity of the deletion haplotype in Melanesians is likely due to ongoing positive selection. Using the deletion variant as a surrogate for the haplotype, we performed a test that controls demography by comparing the observed *PBS* value of the deletion allele with a parametric *PBS* distribution, generated using SNVs from our coalescent simulations. To control the age of the variant, we required the derived allele frequency of simulated SNVs be within 30% of the frequency of the $DEL_{MEL\text{-}NDL}$ deletion allele among the simulated Melanesian chromosomes (i.e., 0.306–0.568; 7,850 SNVs for MEL). Compared with the parametric distribution of *PBS*, the observed *PBS* value of the $DEL_{MEL\text{-}NDL}$ deletion allele is significantly high (*PBS*=0.933, p-value=0.0082; **Figure 5C**). Together, our results are consistent with patterns expected under genetic introgression of $DEL_{MEL\text{-}NDL}$ haplotypes between the ancestors of non-Africans and Neanderthals and suggest that the unusually high frequency of this introgressed haplotype in Melanesians is likely a result of ongoing positive selection.

To understand the evolution of *TNFRSF10D* in primates, we generated high-quality sequences using BAC libraries of three nonhuman great ape lineages, including chimpanzee, gorilla, and orangutan, as well as an Old World rhesus macaque monkey (**Methods**). Sequence comparisons between human reference (GRCh37) and nonhuman primate BAC sequences revealed the same tandem organization of the duplication *TNFRSF10D1* of 30,394 bp and *TNFRSF10D2* of 33,022 bp in all nonhuman primate samples, which is not represented in the human reference (**Figures 4C, S60**). Because of the absence of duplication signals in most of modern human samples, these observations suggest that the haplotype of a single *TNFRSF10D* copy found in most humans is the derived form. We also tested the tandem organization of the duplications in Melanesians using interphase FISH experiments for three fibroblast cell lines. Indeed, our results showed that the tandem duplications are present in all two of the three cell lines (**Table S12**).

We inferred the breakpoints of the duplication using two complementary approaches. First, we mapped the reads of the SGDP samples to a high-quality chimpanzee assembly (*53*) and performed read-depth profiling for each sample across the DUP_{10D} region in chimpanzee. We identified a clear depletion

of read coverage in CN2 SGDP samples within the tandem duplication locus, of which about two-thirds of *TNFRSF10D1* (~18,400 bp, 000025F_1_22350596_quiver_pilon:20,763,000-20,781,405) and one-third of *TNFRSF10D2* (~11,900 bp, 000025F_1_22350596_quiver_pilon:20,751,031-20,762,948) sequences were seemingly deleted (**Figure S60**). As the second approach, we created a multiple sequence alignment using the homologous sequences of *TNFRSF10D* from the human reference and a chimpanzee BAC assembly (**Methods**). We found the most likely breakpoint at a region of 82 bp (chr8:23,003,123-23,003,255, GRCh37), partially overlapping the 5th intron and exon of *TNFRSF10D* in GRCh37, using a hidden Markov model (**Methods**). The top two best matches of the 82 bp sequences on the chimpanzee sequences were mapped to the two ends where the drop of read depth occurs (**Figure S59**).

We performed Bayesian phylogenetic analyses using the homologous sequences of the human reference (GRCh37) *TNFRSF10D1* and *TNFRSF10D2*, separately, from four primate lineages (**Figure S60; Methods**). We noticed an increase in high sequence identity between most of the rear portion of the DUP_{10D} sequences in orangutan (**Figure S60B**), a pattern consistent with interlocus gene conversion, which may obscure the true phylogenetic signals. Manually removing this and other low-quality alignment regions resulted in alignments of 3,934 and 4,215 bp for *TNFRSF10D1* and *TNFRSF10D2*, respectively. We performed two phylogenetic inferences, of which the human *TNFRSF10D1* and *TNFRSF10D2* sequences were used separately due to the lack of homology between the two sequences. The two phylogenies that we inferred are largely consistent to each other (**Figure S60**) and show at least two independent duplication events of *TNFRSF10D* in the evolution of primates: one in the lineage leading to the Old World rhesus macaque monkey and the other at 27.55 Mya (95% highest posterior density: 19.88–36.14), about the divergence between Old World monkeys and apes at 25–30 Mya. We noted a gene tree–species tree discordance among human, chimpanzee, and gorilla on the phylogeny using the human *TNFRSF10D2* data, likely due to ILS.

We determined gene expression and annotation for different copies of *TNFRSF10D* using full-length transcripts from Melanesian (GM10541, CN3), European (GM12878, CN2), and chimpanzee (PanTro, CN4) fibroblast cell line samples (*29*) (**Methods**). Complete transcripts, originated from full-length cDNA molecules, were generated following a framework based on long-read PacBio sequencing technology and mapped to the human reference assembly (ENST00000312584) (*52*) (**Methods**). Because the human reference *TNFRSF10D* is a product of gene fusion between *TNFRSF10D1* and *TNFRSF10D2*, to classify transcripts we leveraged the patterns of single-nucleotide mismatches in individual transcript alignments against the human reference copy. In the chimpanzee sample, our analyses reveal two types of transcripts corresponding to six and nine exons in *TNFRSF10D1* and *TNFRSF10D2*, respectively (**Figures S61-S62**). In addition, the Melanesian CN3 sample carries all three types of transcripts, while the European CN2 sample only possesses the fusion hybrid gene copy as expected (**Figure S61**).

Interestingly, while all *TNFRSF10D1* transcripts we examined carry a premature stop codon in exon 2, which truncates the protein after 59 amino acids, a 217 amino acid protein is likely translated in a different frame using a second start codon, upstream to the premature stop codon in exon 2 (**Figure S62**). In contrast, *TNFRSF10D2* transcripts maintain an ORF with all nine exons consistent with the annotation in the human reference assembly. Note that the common human fusion hybrid gene effectively deletes the *TNFRSF10D1* premature stop codon (**Figure S62**), thus restoring the ORF. To assess if positive selection has acted on protein-coding sequences across any of these *TNFRSF10D* lineages, we computed *dN/dS* ratios for the ORF sequences of the six common transcripts from the three cell line samples and two pseudo-transcripts extracted from rhesus macaque BAC sequences (**Methods**). Pairwise *dN/dS* comparisons among these protein-coding sequences indicate significant large *dN/dS* ratios for the *TNFRSF10D1* lineages, suggesting an excess of nonsynonymous changes observed within these lineages (**Figure S63**). We used a phylogenetic branch model of positive selection (*43*) and found that the inferred *dN/dS* ratios are significantly greater than 1 at the clade of *TNFRSF10D1* lineages ($p = 0.017$; **Figures S64-S65**), but not at any other clades, suggesting adaptive protein evolution acting on the clade of *TNFRSF10D1* lineages. In addition, using a branch-site test (*43*), we find evidence of positive selection for both the fusion gene and the *TNFRSF10D1* copy on the human lineage and more broadly for both *D1* and *D2* in other nonhuman apes ($p = 0.005$; **Figures 4D, S64**). Specifically, we identify a cluster of positively selected sites corresponding to the predicted transmembrane domain of the genes (**Figures 4D, S64**).

Supporting evidence for selection signals at known CNV loci in Melanesians

Here we provided a detailed discussion for four of the Melanesian candidate CNV loci that are in close proximity to known selected copy number variable regions in other populations. The deletion variant at the *APOBEC3* gene cluster (**Figure S15**) is commonly found among human populations (frequency of 22.5%), especially in Oceanian populations (92.9%) (*32*). Because the putative deletion boundaries that we inferred (chr22:39,388,950-39,483,917) overlap with SDs, which confound copy numbers estimated using read-depth-based whole-genome shotgun sequencing detection (WSSD), we utilized a SUNK genotyping method to infer paralog-specific copy number (*54*). Consistent with the previous studies, we estimated that 24.5% of the SGDP haplotypes (122 out of 498 haplotypes) carry this variant (**Figure S15**). While the deletion allele is highly variable in frequency among the SGDP samples and observed in nonhuman great apes, it is fixed only in the Melanesian samples (**Figure S16, left panel**). Here we provided evidence for significant signals of positive selection (parametric *p*-value of the *PBS* test < 0.017, blue dots in **Figure S17**) at the unique (CN = 2) sequences of the flanking regions around the deletion variant (chr22:39,340,000-39,450,000), although we did not observe any highly differentiate SNVs that

766 are potentially functional in Melanesians. To investigate the haplotype pattern around this candidate
767 locus, we phased 266 SNVs along with the deletion variants from the flanking sequences in Melanesians
768 and other populations (**Figure S17**). Haplotypes were further grouped using nine SNVs with *PBS* > 0.5
769 and classified into haplogroups according to the pairwise mutation distances and the deletion status
770 (**Methods**). The deletion-linked haplogroup is nearly absent in sub-Saharan Africans and found mostly in
771 low (4%, EUR) to intermediate (40%, AMR) frequencies in other non-Africans, suggesting a recent
772 origin of this haplotype (**Figures S17-S19**). Strikingly, we found that Melanesians carry a single deletion-
773 linked haplogroup with extended homozygosity (**Figures S17, S19**), in accordance with the hypothesis of
774 selective sweep.
775 The highly stratified multi-allelic CNV at the alpha-defensin (*DEFA*) gene family
776 (chr8:6,839,960-6,878,169; **Figure S20**) locates at the telomeric side of the chromosome 8p23.1 region,
777 one of the most structurally dynamic regions in the human genome, where recurrent rearrangements,
778 including microdeletions, interchromosomal transpositions, and inversions, have occurred over primate
779 evolution and have been associated with disease (*36, 80*). Our selection candidate of CNV overlaps with a
780 known SD pair, each encompassing the *DEFA1-T1* subfamily ($CNP_{DEFA1\text{-}T1}$), and thus $CNP_{DEFA1\text{-}T1}$ has a
781 diploid copy number four in the human reference genome (**Figure S20**). There is a great variability in
782 copy number from 3 to 18 copies for $CNP_{DEFA1\text{-}T1}$ across SGDP and nonhuman primate samples (**Figures**
783 **S20-S21**). Interestingly, we observed Melanesians are less variable in copy number, and 87.5% of the
784 samples carry more than eight copies of $CNP_{DEFA1\text{-}T1}$ (**Figures S20-S21**). Applying both metaphase and
785 interphase FISH experiments to three Melanesian cell lines (**Methods**), we determined all the three
786 samples have the direct orientation of 8p23.1 and the tandem organization of $CNP_{DEFA1\text{-}T1}$ duplications
787 (**Figure S22**). Our selection test using the *PBS* statistic provided evidence for positive selection at the
788 sequences flanking the SD region (*p*-value < 0.035, chr8:6,819,244-6,921,178; **Figure S23**). Haplotype
789 analysis using 855 SNVs from the selection candidate region reveals that most Melanesian samples carry
790 Haplogroup2 haplotypes (94%; **Figures S24-S25**), which are completely absent from sub-Saharan
791 African samples and observed at low (7%, SIB) to intermediate (23%, ME) frequencies in the rest of non-
792 African populations. The high frequency Haplogroup2 haplotypes in Melanesians also show a slow decay
793 of homozygosity, a pattern expected under positive selection acting at this locus (**Figures S23, S25**).
794 Another selective CNV candidate we identified in Melanesians is at the chromosome 17q21.31
795 locus (chr17:44,170,850-45,157,111; **Figure S26**), one of the most dynamic and complex regions in the
796 human genome. Previous studies reported both direct and inverted haplotypes and three large copy
797 number polymorphic duplications (CNP155, CNP205, and CNP210; **Figure S26**) at this locus and
798 showed associations of inverted form with the 17q21.31 microdeletion syndrome (*37, 38*). Analyses of
799 pairwise WSSD-based copy number estimates showed that all human populations are highly variable in

the copy number configuration for the three large SDs, except the Melanesians, of which 94% of the samples have diploid copy number two for all three variants (**Figure S26**). To further investigate the copy number configuration of these CNVs, we applied the SUNK-based genotyping (**Methods**) to all SGDP samples and nonhuman great apes. We found that the observed high copies of CNP210 variants are likely a result of duplication events for CNP210-dup1, not CNP210-dup2 (**Figures S26-S27**). Furthermore, our analysis suggests that 97% of Melanesian haplotypes carry a single copy for both CNP210-dup2 and CNP155/CNP205, but no CNP210-dup1 sequences (**Figures S26-S27**). The frequencies of this haplotype vary from 11.7% (EUR) to 36.3% (SIB) across SGDP populations as well as archaic and nonhuman great ape samples (**Figure S27**). Using FISH experiments, we confirmed that all three Melanesian cell lines carry the direct haplotype, with GM10541 and GM10543 carrying only one copy of CNP210 on both chromosomes (homozygous H1.1/ H1.1; nomenclature follows Steinberg et al. 2012 (*37*)) and GM10539 carrying one copy of the CNP210 duplication on one chromosome and two copies on the other chromosome (heterozygous H1.1/H1.2) (**Figure S28**). We detected signals of positive selection at the flanking sequences of the distal side of CNP210-dup2 region (*PBS* p-value < 0.02, chr17:44,784,657-44,854,722; **Figure S29)**. Haplotype analysis using 367 SNVs from this region showed that one of the common haplogroups, Haplogroup2, while is common in most non-Africans, particularly in ME (70%), EUR (75%), and SA (80%), is almost fixed in Melanesians (97%) (**Figure S30**). In addition, we also observed significantly negative Tajima's D (< –2.14, p-value = 0.007), low nucleotide diversity ($\pi < 4.3 \times 10^{-5}$, p-value = 0.042), and EHH across this region, patterns as expected under positive selection (**Figures S31-S32**). Together, these lines of evidence suggest selection acting on the high frequency Haplogroup2 in Melanesians.

One of the most significant signals for selection in Melanesians locates at chromosome 14q24 (p-value for the *PBS* test < 0.001), expanding ~510 kbp sequences (chr14:73,730,000-74,240,000, **Figure S11**). At this region, a CNV, $CNP_{ACOT1\text{-}2}$ (chr14:73,999,126-74,053,245), is highly differentiated in copy number between Melanesians and the rest of SGDP populations (Bonferroni's p-value for the CN stratification test = 0.027). Because $CNP_{ACOT1\text{-}2}$ spans over the SDs encompassing *ACOT1* and *ACOT2*, we applied the SUNK genotyping method to infer paralog-specific copy numbers underlying the *ACOT1* and *ACOT2* sequences. In short, SUNK genotyping leverages the presences of fixed sequence differences that uniquely found in each paralog and thus can infer a more accurate copy number for each paralog (Methods). The SUNK copy number heat maps for the SGDP samples revealed that the stratification signal at $CNP_{ACOT1\text{-}2}$ is primarily driven by the copy number variation at the *ACOT1* locus (chr14:73,999,126-74,018,293, Pearson's correlation $\rho = 0.938$, p-value < 2.2×10^{-16}, **Figure S12**), not the *ACOT2* locus (chr14:74,024,364-74,053,245, Pearson's correlation $\rho = -0.0005$, p-value = 0.9931). We estimated that the frequency of the presence of at least one copy of *ACOT1* (SUNK CN > 0) is the

834 highest in Melanesians (100%) compared with those of the rest of SGDP populations (27−75% samples).
835 In addition, the fraction of Melanesian samples carrying at least two copies of *ACOT1* is 87.5%, a much
836 higher rate than other SGDP populations (9–40%), which is consistent with the expectation of positive
837 selection acting on the *ACOT1* locus in Melanesians (**Figure S12**).
838 To further investigate the hypothesis of positive selection acting on *ACOT1* in Melanesians, we
839 phased the 1,460 SNVs spanning the 510 kbp region and constructed haplotypes for all SGDP samples.
840 We noted that among the four major haplogroups, Haplogroup2 is in high frequency among Melanesians
841 (72%), but in much lower presence in other populations (0–11%), while Haplogroup1 is the most
842 common in other populations, including the three archaic samples (**Figure S13**). Furthermore, under
843 positive selection, the haplotype that carries the beneficial variant is expected to show EHH (*67*). We
844 examined the pattern of haplotype homozygosity in Melanesians using two of the top *PBS* SNVs
845 (rs4903119, *PBS*=2.29; rs8015976, *PBS*=2.17, **Figure S11C**) as the core SNVs. At the two sites, 84% of
846 the Melanesian samples carry the T-G haplotype (T allele at rs4903119, ancestral, frequency=0.92; G
847 allele at 8015976, derived, frequency=0.84). **Figure S11C** shows that the T-G haplotype clearly retains
848 long and high levels of EHH, centered at the $CNP_{ACOT1\text{-}2}$ locus. We noted that although a similar EHH
849 pattern was observed in some of the non-African SGDP populations, most of individuals in those
850 populations in fact carry a different allele at the core (**Figure S14**). We noticed that there are also
851 significant signals of archaic introgression around the $CNP_{ACOT1\text{-}2}$ locus in Melanesians regardless the
852 archaic reference sequence (p-value of $f_D < 0.023$, **Figure S11A**). Because the haplotypes of the three
853 archaic genomes belong to Haplogroup1, the selection signal is unlikely confounded by the introgression
854 signal.
855

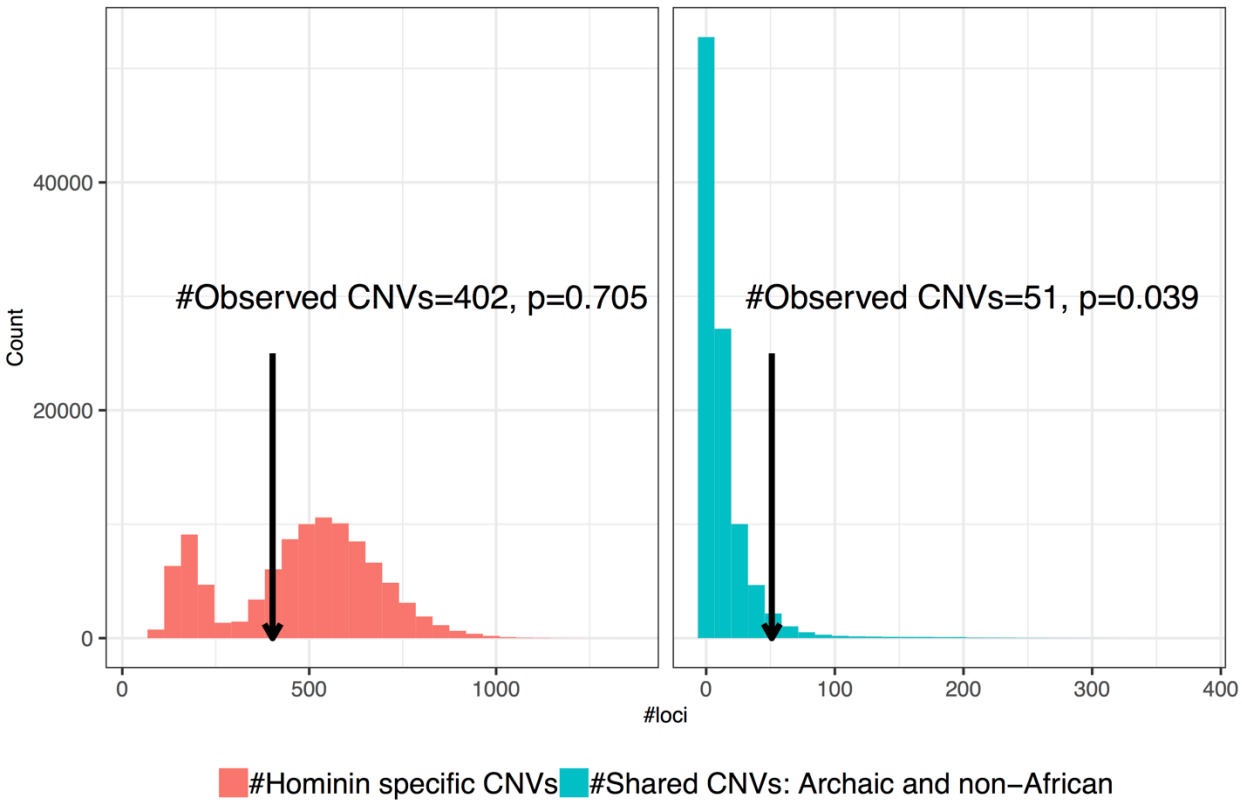


Figure S1. Significantly shared CNVs between archaic Eurasian hominins and non-Africans. Based on a database of 5,135 CNVs identified using a read-depth approach (*54*) and genotyped in the SGDP samples (n=224) (*27*), nonhuman great apes (n=72) (*28*), and archaic Eurasian hominins (n=3) (*24-26*). Lineage-specific or shared events are defined based on a comparison among species and/or populations as described in the Methods section. The *p*-values were based on 100,000 permutation simulations shuffling the labels of samples.

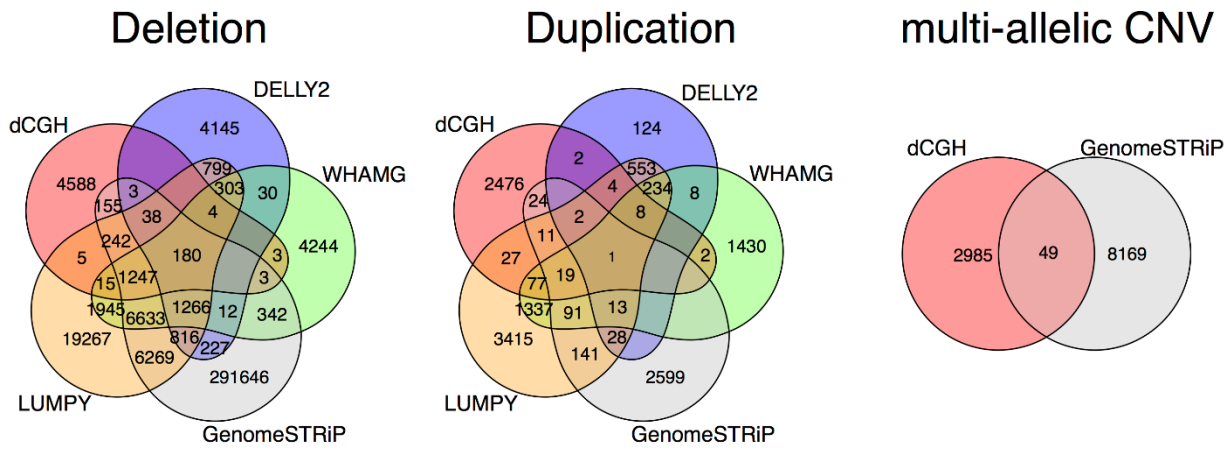


Figure S2. Venn diagrams of the 368,256 CNVs identified in SGDP samples using five callers. Note that only dCGH and Genome STRiP identify multi-allelic CNVs.

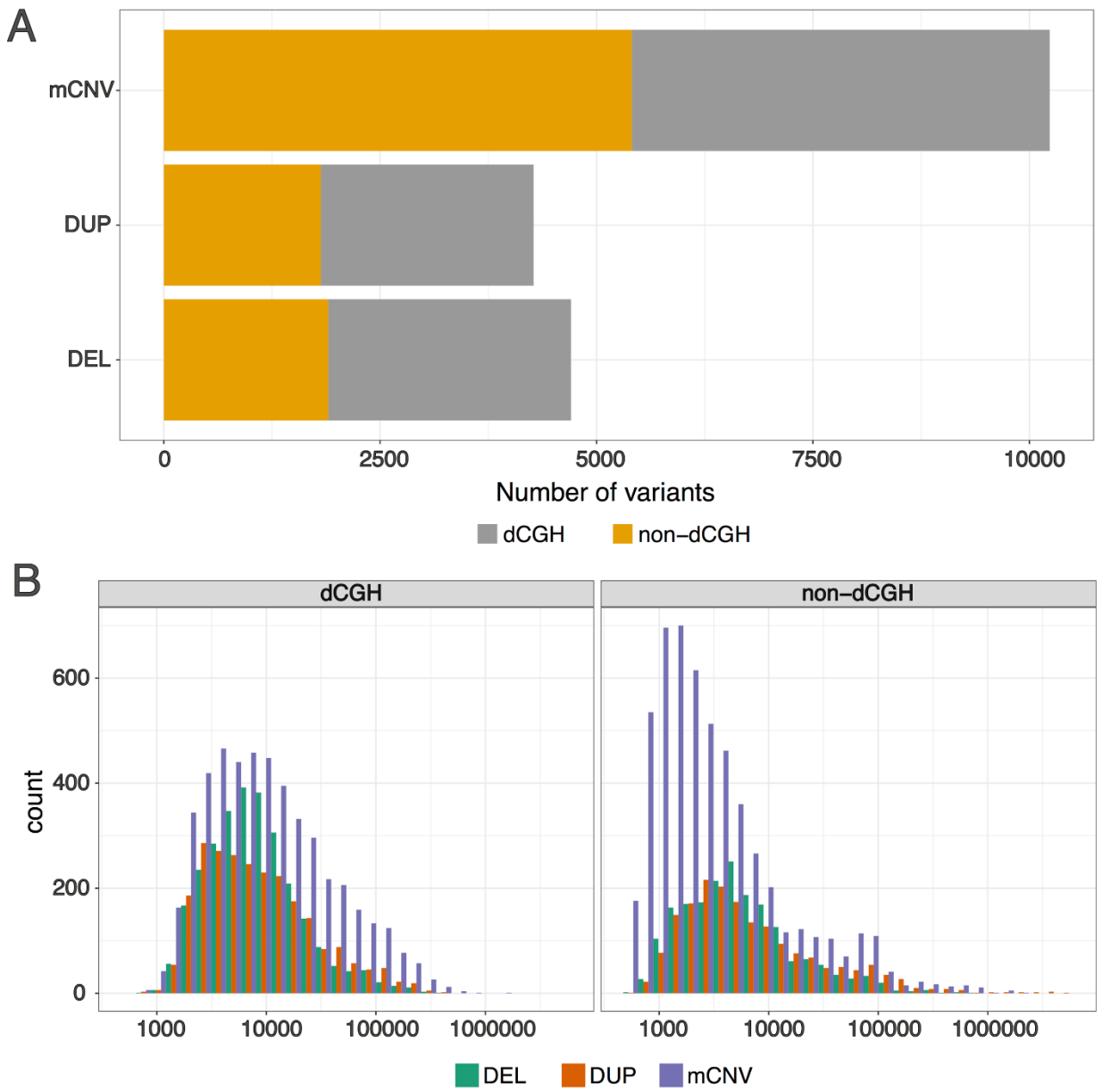


Figure S3. Distributions of 19,211 post-filtered CNVs in type and length. (A) The frequencies of dCGH and non-dCGH called variants, plotted in different CNV categories: mCNV: multi-allelic CNV, DUP: bi-allelic duplication, and DEL: bi-allelic deletion. (B) The length distributions of dCGH and non-dCGH called CNVs.

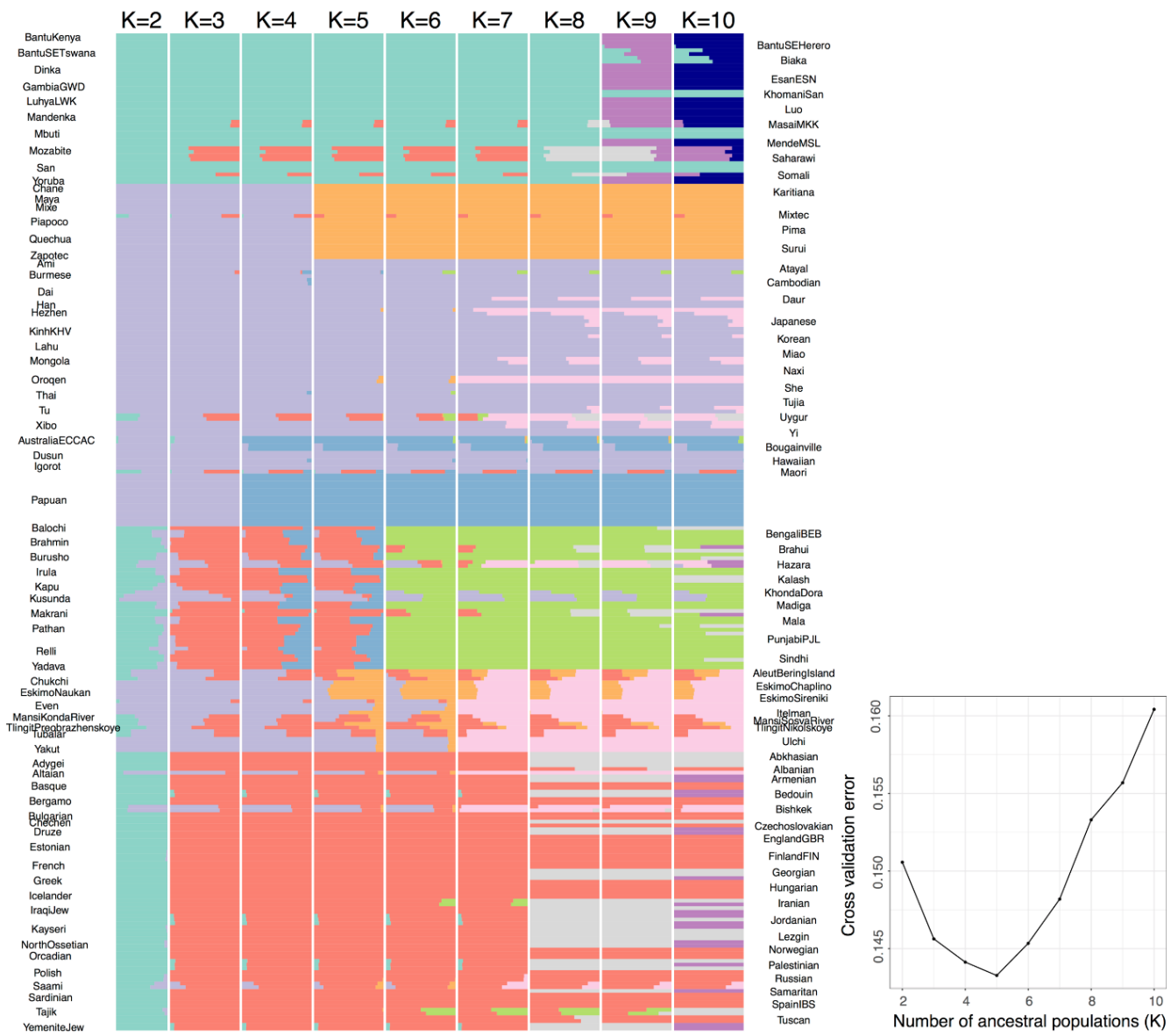


Figure S4. ADMIXTURE analysis. We applied ADMIXTURE to the SGDP samples using the number of putative ancestral populations, K, between 2 and 10. To ensure the convergence of the estimation, we performed 20 replicates for each K. Using the default fivefold cross-validation, we inferred the best K = 5 (CV error = 0.14327), corresponding to major populations: sub-Sahara Africans, Native Americans, East Asians, Sahul Oceanians, and West Eurasians.

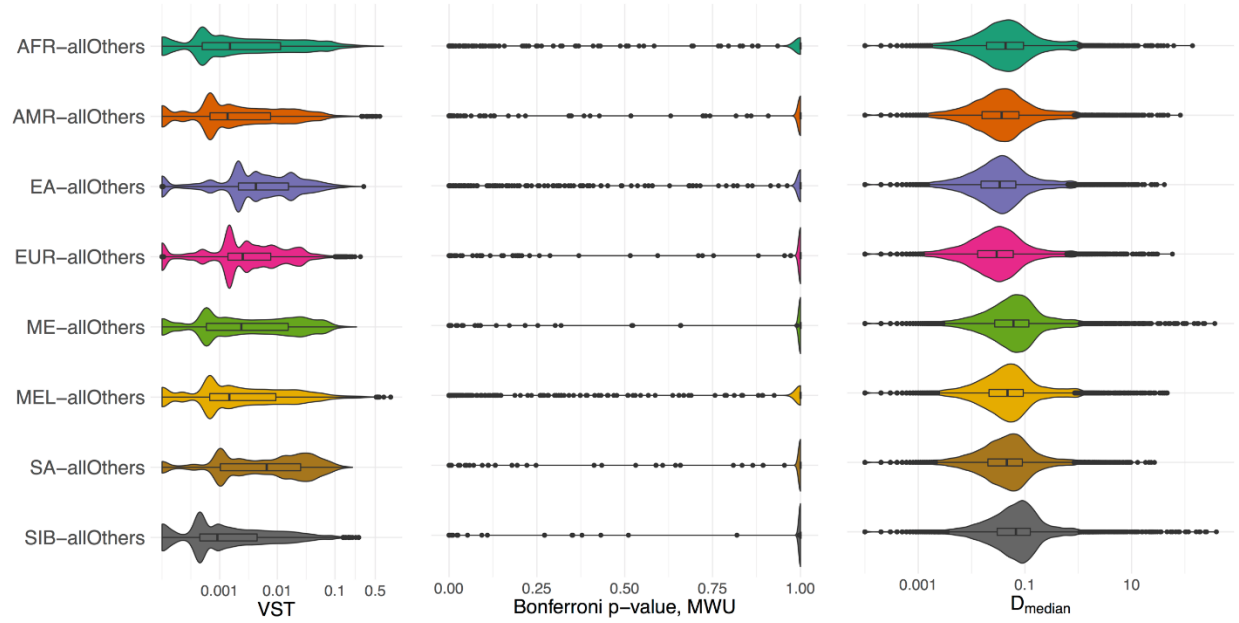


Figure S5. Distributions of statistics for identifying population-stratified CNVs. Statistics are calculated as described in Materials and Methods to compare focal populations X and all of the rest SGDP samples, where X = {AFR, AMR, EA, EUR, ME, MEL, SA, SIB}.

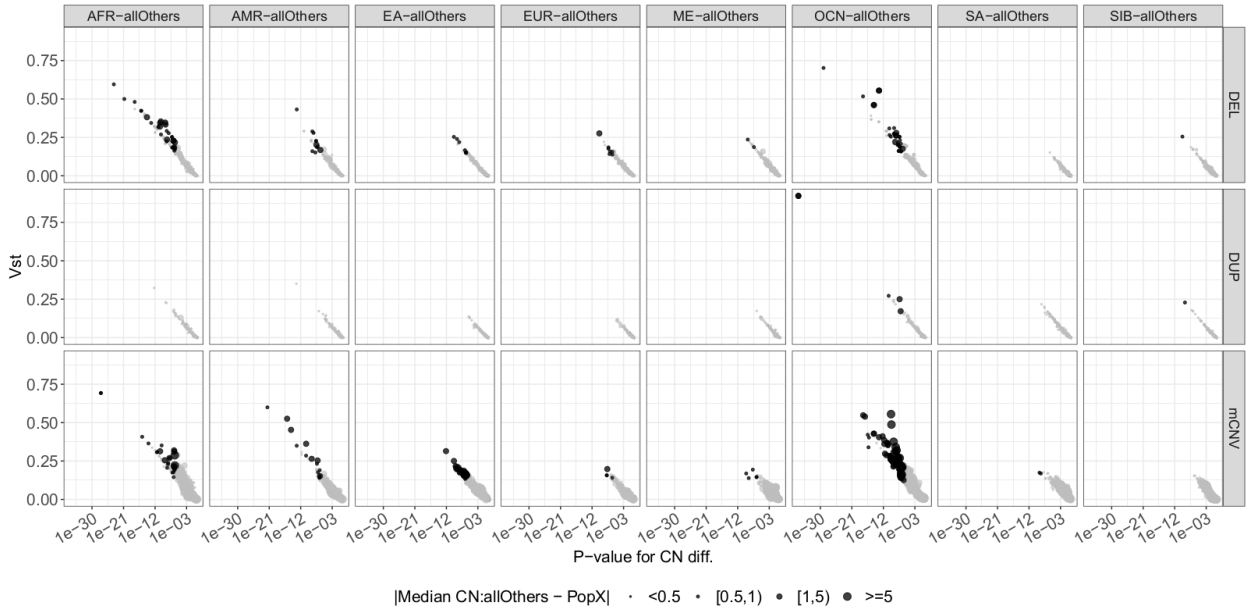


Figure S6. Joint distribution of test statistics for identifying population-stratified CNVs. Panels of each column show the three tests of copy number (CN) stratification for a focal population vs. the rest of the SGDP samples, while rows are the results for three different CN categories. Each dot is a CNV, whose size is determined by the D_{median} statistic. Significantly stratified CNVs, defined as (i) $V_{ST} > 0.1$, (ii) Bonferroni *p*-value of the CN differentiation (*MWU*) test < 0.05, and (iii) $D_{median} > 0.5$, are colored in black.

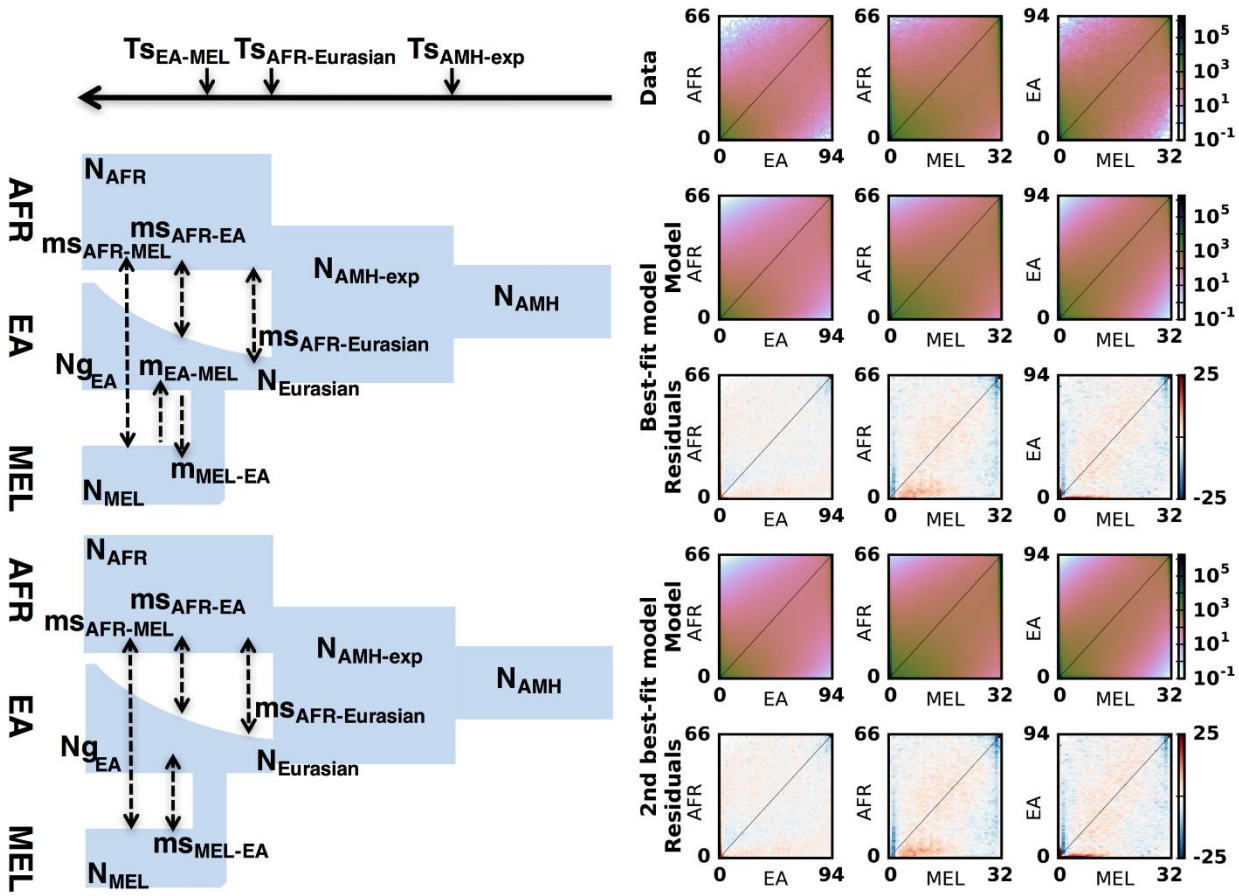


Figure S7. Demographic inferences for Melanesians (MEL), Africans (AFR), and East Asians (EA) using ∂a∂i (*50*). The left panel illustrates the best-fit demographic model with asymmetric gene flow between MEL and EA (top, 15 parameters, log-likelihood: -103386) and the 2nd best-fit model with symmetric gene flow between the populations (bottom, 14 parameters, log-likelihood: -104590). Corresponding maximum likelihood estimates for the parameters of the two models can be found in **Table S7**. The right panel shows observed and predicted frequency spectra for the two best-fit models. Row one is data, row two (the best-fit model) and four (the 2nd best-fit model) are models, and rows three and five are Anscombe residuals of model minus data for the best- and 2nd best-fit models, respectively. The range of Anscombe residuals is set to be [-25, 25] for better illustration.

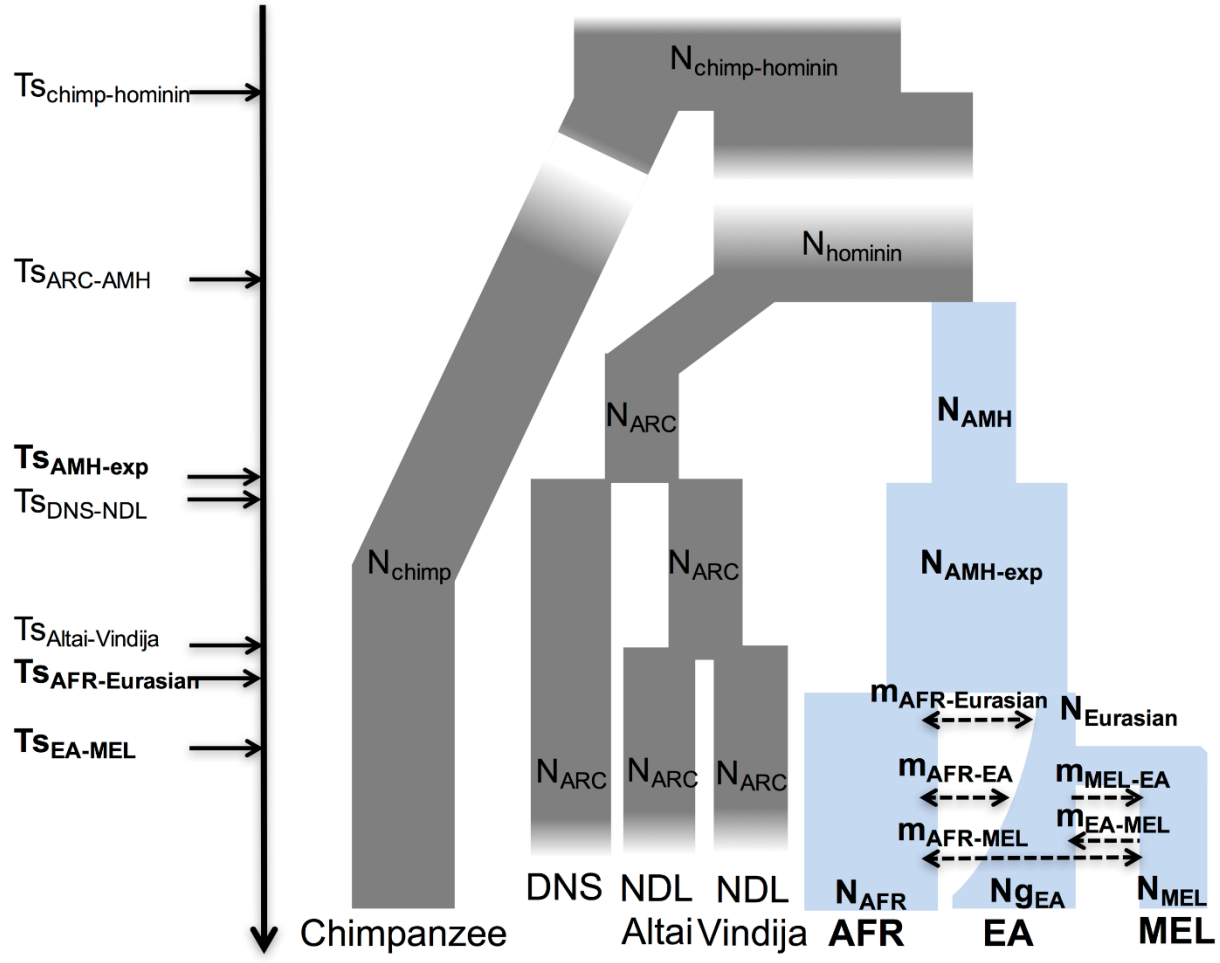


Figure S8. Demographic schematic for large-scale genomic coalescent simulations. Blue branches and bold parameters indicate the best demographic model inferred in this study for the African, East Asian, and Melanesian populations (**Figure S7; Table S7**). Parameter values of the gray branches were uniformly drawn from the 95% C.I. reported in previous studies (**Table S8**). The white gradients highlight the time scale for those branches are much larger than the others.

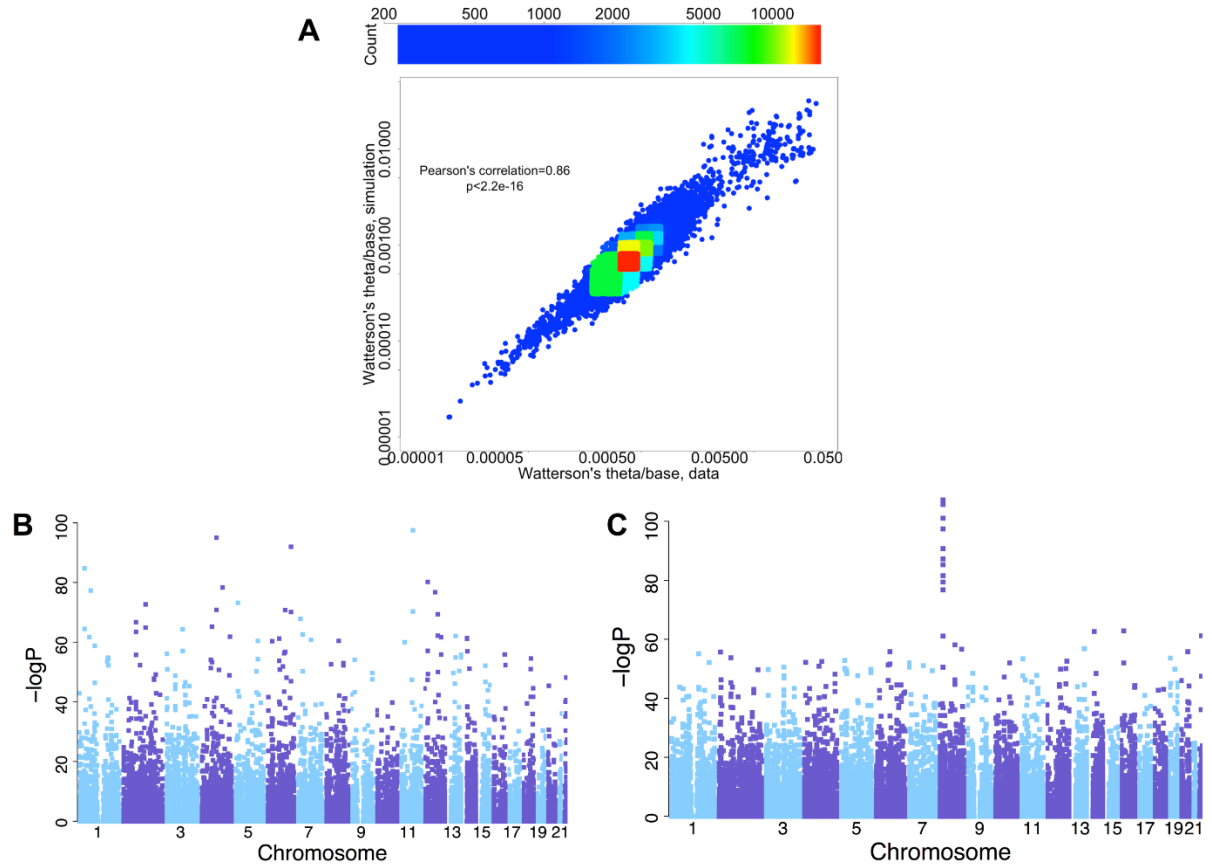


Figure S9. Whole-genome simulations accurately capture the local genetic diversity in the real data. (A) Correlation of per-base θ (Watterson's estimator) between windows in real and simulated whole-genome data**.** Window are defined as in our selection scans (100 SNVs per window). Pearson's correlation is 0.860. (B) Manhattan plot for the window-based F_{ST} test using a simulation based on one of the 1,000 models. (C) Manhattan plot for the window-based F_{ST} test using the real data (MEL vs. EA).

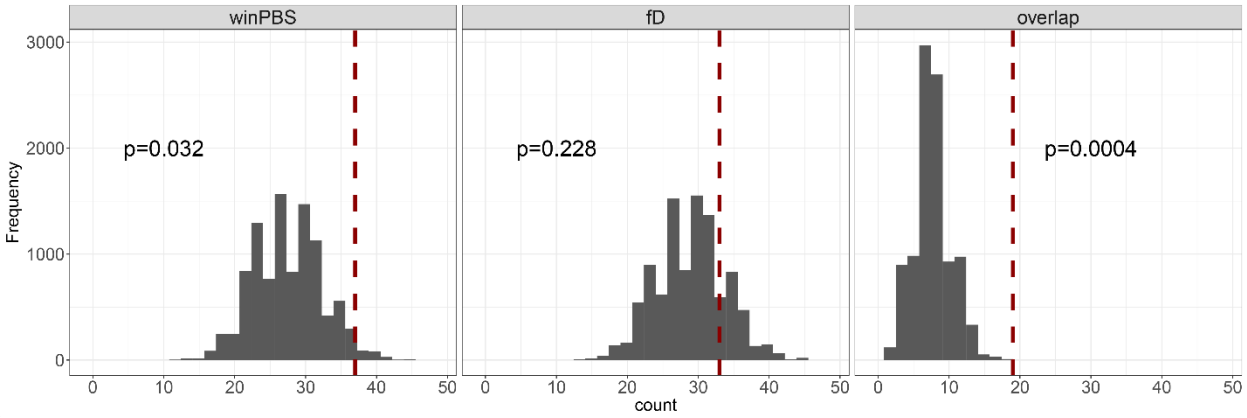


Figure S10. Melanesian-stratified CNVs are significantly associated with loci showing signals of positive selection, but not with those of archaic introgression. *P*-values were calculated using 10,000 simulations of the 162 stratified CNVs, which consists of 37 selective and 33 introgressed CNV candidate loci. Simulated CNV loci were generated by randomly shuffling these Melanesian-stratified CNVs across the unmasked sequences of the genome as defined in the real data (**Methods**). Red dashed lines are the observed numbers of candidate CNVs for selection and archaic introgression.

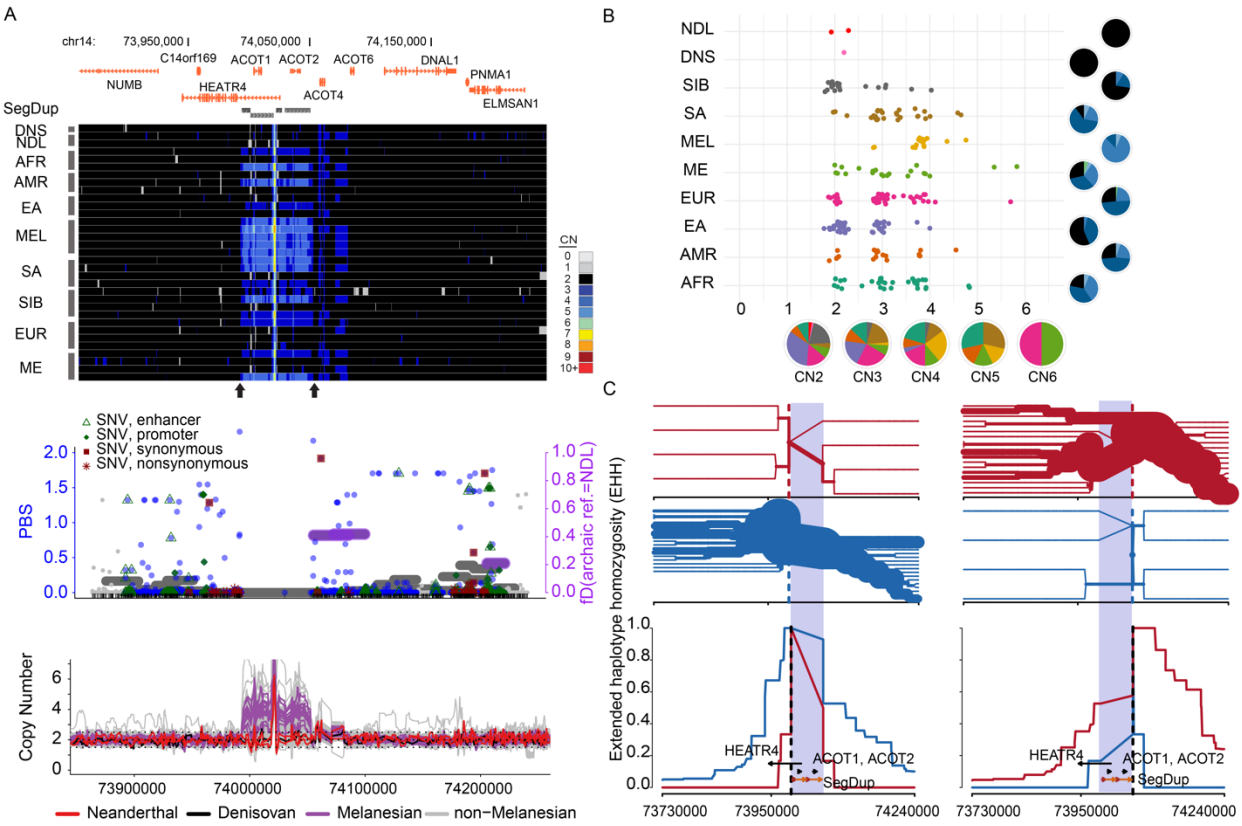


Figure 11. **Evidence for positive selection of the multi-allelic CNV locus in *ACOT* cluster ($CNP_{ACOT1-2}$, chr14:73,999,126-74,053,245) in Melanesians.** (**A**) As in Figure 2A, but for the $CNP_{ACOT1-2}$ locus. Functional annotation (RefSeq and ENCODE elements) denoted by green symbols. (**B**) Distribution of WSSD-based CN estimates of $CNP_{ACOT1-2}$ (between two black vertical arrows) for the eight SGDP populations and the three archaic samples. (**C**) The bifurcation diagrams (red: derived allele, blue: ancestral allele) and EHH (bottom) of Melanesians using 1,460 SNVs from chr14:73,730,000-74,240,000. Two dashed lines indicate two top PBS SNVs (left: rs4903119, PBS=2.29; right: rs8015976, PBS=2.17) on each side of the $CNP_{ACOT1-2}$. Ancestral states were polarized according to the human–chimpanzee alignment.

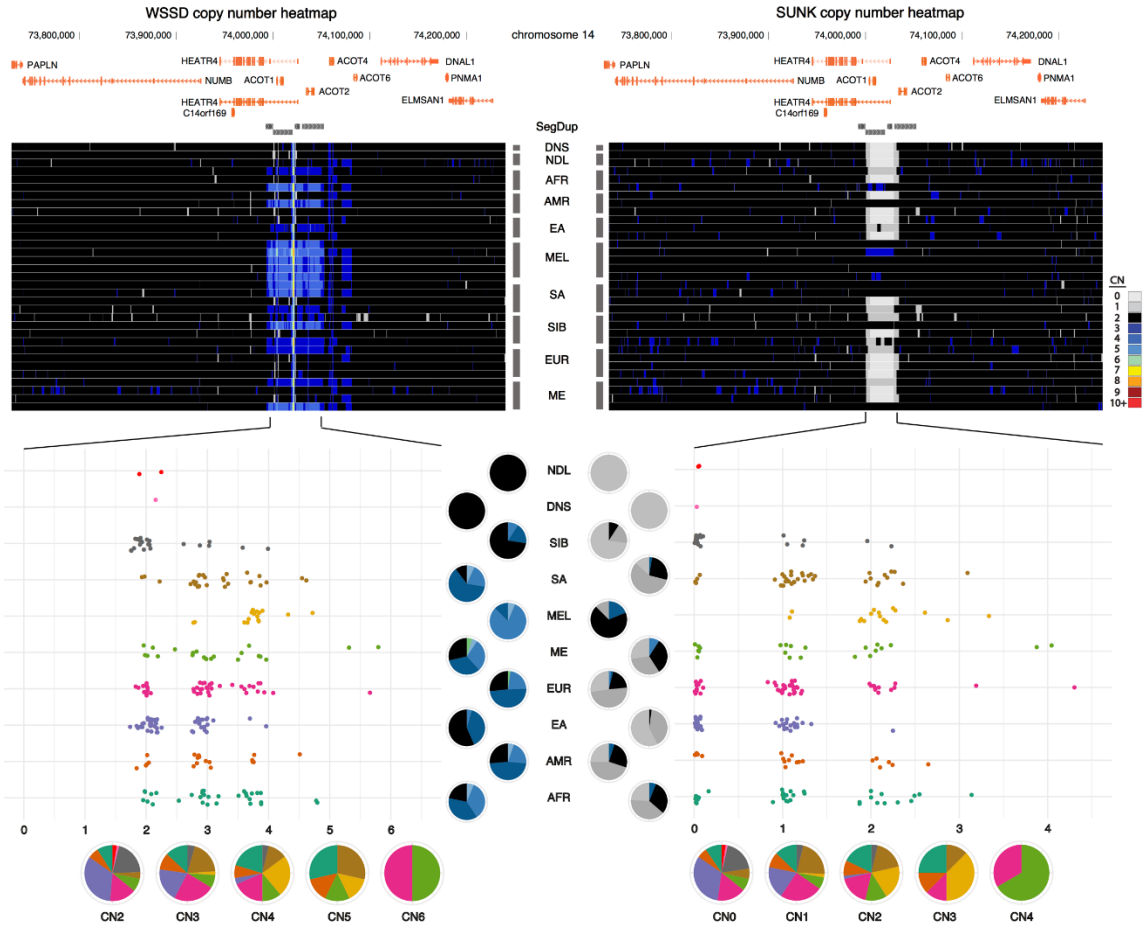


Figure S12. Copy number distributions for the multi-allelic copy number variant (chr14:73,999,126-74,053,245, $CNP_{ACOT1\text{-}2}$) at the *ACOT* gene family on chromosome 14q24. Left panels: copy numbers were estimated using read-depth WSSD for *ACOT1* and *ACOT2*; right panels: SUNK copy number estimates of the *ACOT1* locus. Top panels: copy number heat maps, where each row represents the copy numbers of a sample over the region. Bottom panels: copy number distributions for the variant among SGDP and three archaic samples. Pie charts on the x-axis indicate the population distributions in individual copy numbers (colors corresponding to those in the scatterplots), while each pie chart on the y-axis shows the frequency distribution of copy numbers for a given population (colors corresponding to those in the copy number heat maps).

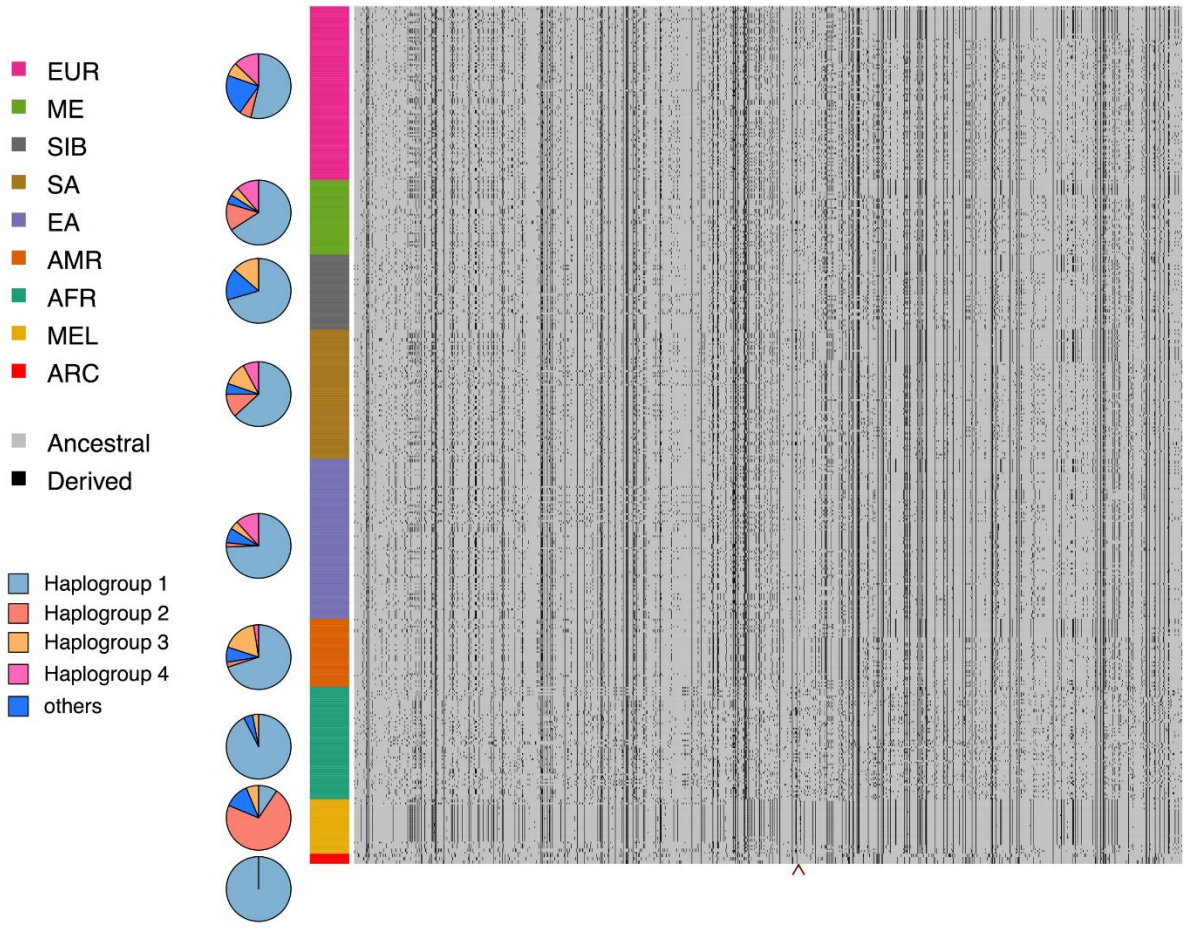


Figure S13. Haplotype pattern in the region showing signals of positive selection at the 510 kbp flanking sequences of the distal side of $CNP_{ACOT1\text{-}2}$ on chromosome 14q24 (chr14:73,730,000-74,240,000; 1,460 SNVs) among the SGDP and three archaic samples. The rows and columns are haplotypes and SNVs, respectively. Haplogroups were defined using all 1,460 SNVs. Haplogroups were formed by using 97 SNVs with *PBS* > 0.5 and grouping haplotypes with five mutations or less. To ease the complexity of the plot, we only used the first four major haplogroups and grouped the rest into the category "others" for display. Pie charts are the distribution of haplogroups in individual populations. The red arrow indicates the position of the first SNV after $CNP_{ACOT1\text{-}2}$.

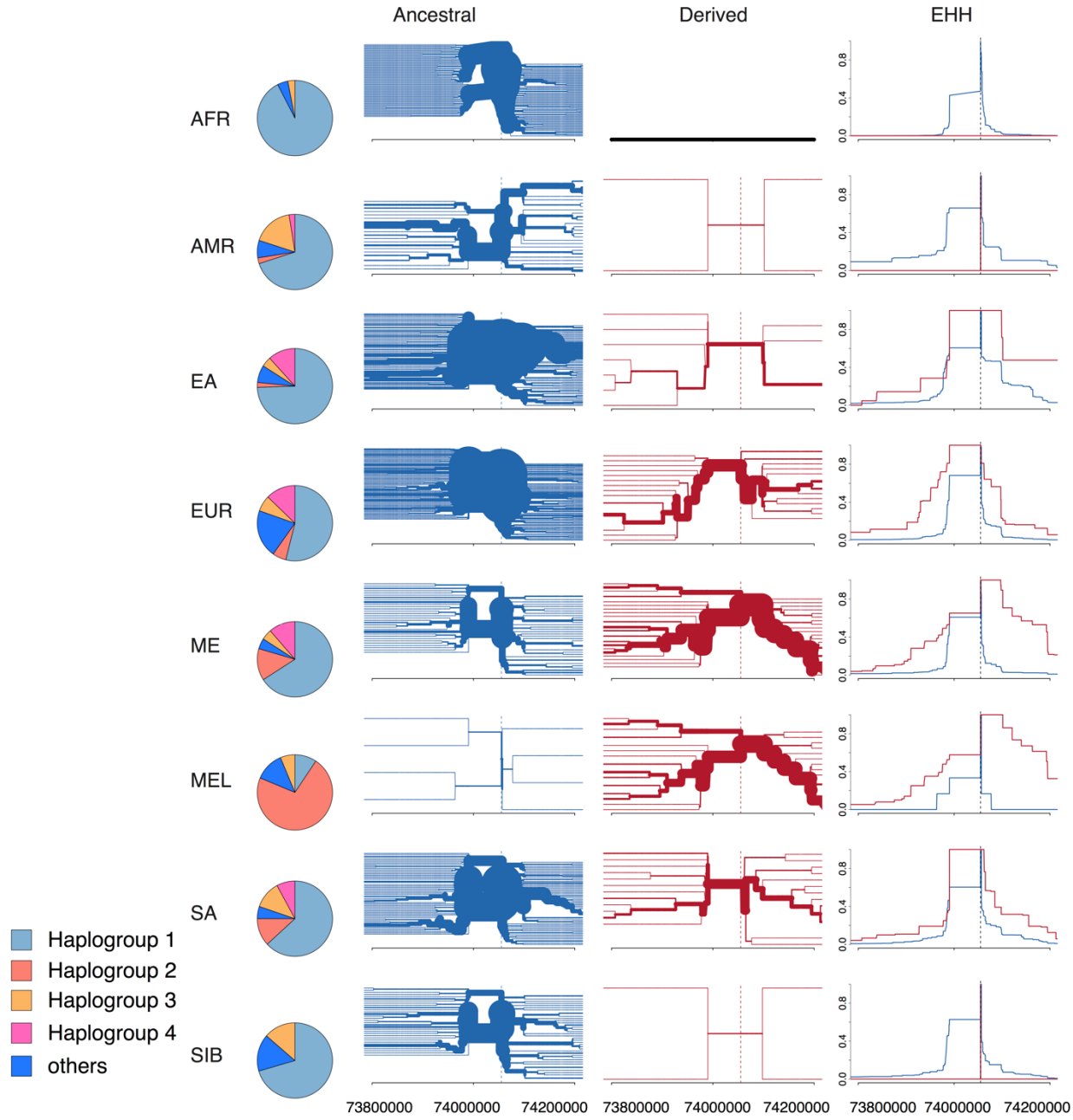


Figure S14. Distribution of haplogroups, bifurcation diagrams, and EHH in individual populations using 1,460 SNVs from the region showing signatures of positive selection on chromosome 14q24 (chr14:73,730,000-74,240,000). Dashed line indicates one of the SNVs with the largest *PBS* (rs8015976, *PBS*=2.17), whose ancestral state was polarized according to the human–chimpanzee alignment.

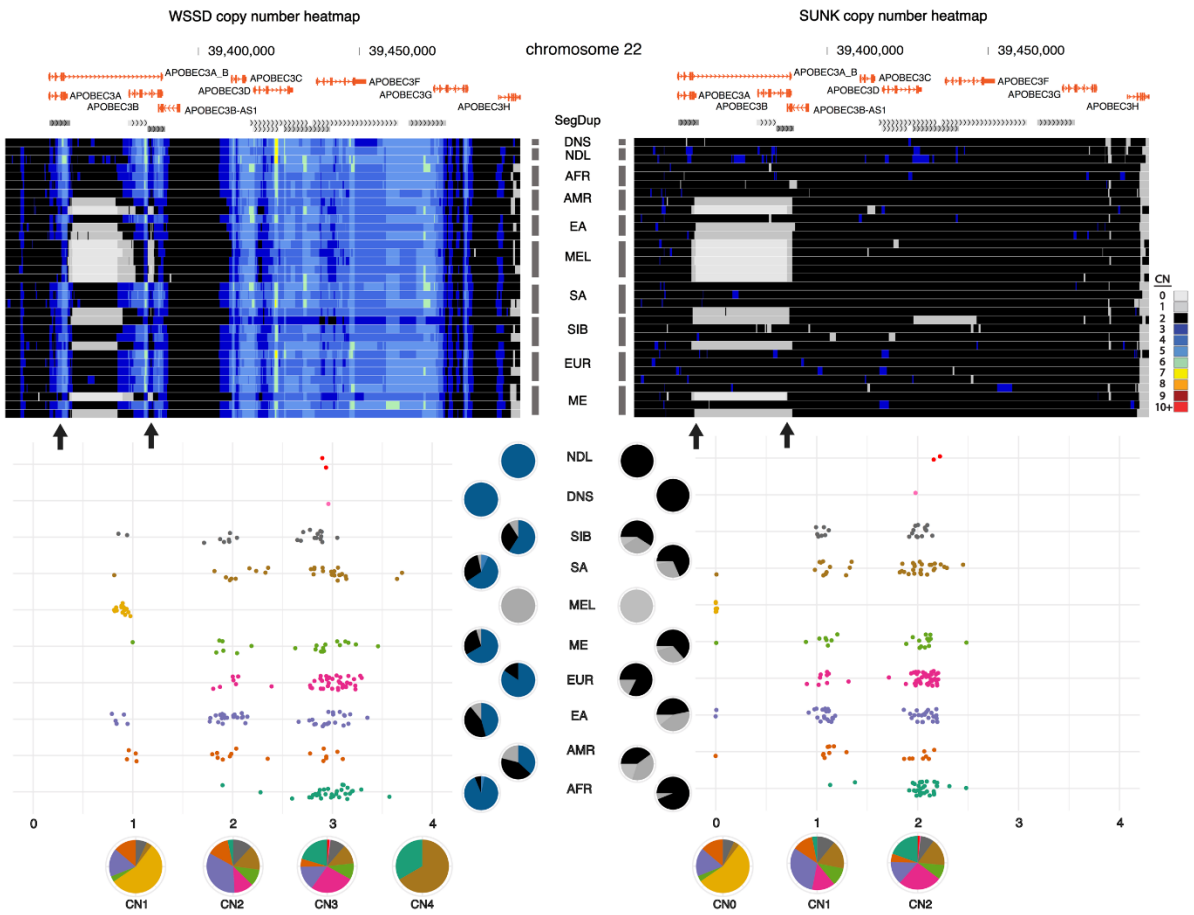


Figure S15. CN distributions for the common deletion variant (region between two black arrows, chr22:39,388,950-39,483,917) at the *APOBEC3* gene cluster on chromosome 22. Copy numbers were estimated using read-depth WSSD (left panels) and SUNK (right panels) genotyping methods. Top panels: CN heat maps, where each row represents the copy numbers of a sample over the region. Bottom panels: CN distributions for the variant among SGDP and three archaic samples. Pie charts on the x-axis indicate the population distributions in individual copy numbers (colors corresponding to those in the scatterplots), while each pie chart on the y-axis shows the frequency distribution of copy numbers for a given population (colors corresponding to those in the CN heat maps).

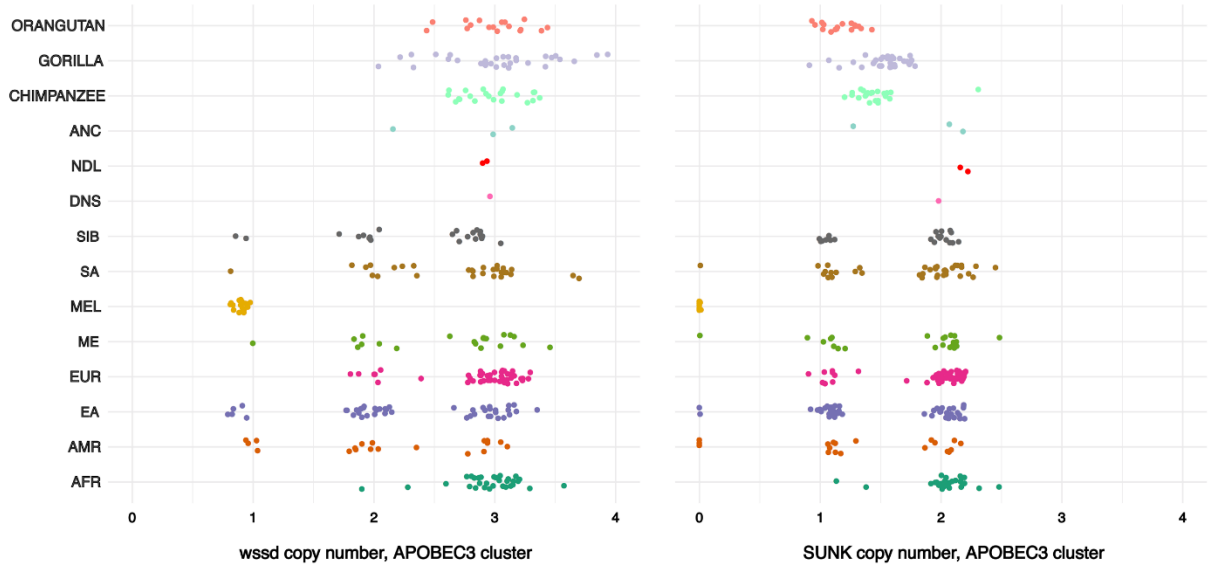


Figure S16. WSSD- and SUNK-based CN distributions for the deletion locus (chr22:39,388,950-39,483,917) at the *APOBEC3* gene cluster among extant modern humans (SGDP), ancient modern humans (Stuttgart, Loschbour, and Ust-Ishim), and nonhuman great apes (chimpanzees, gorilla, and orangutans). Each point is a CN estimate of a sample for the variant.

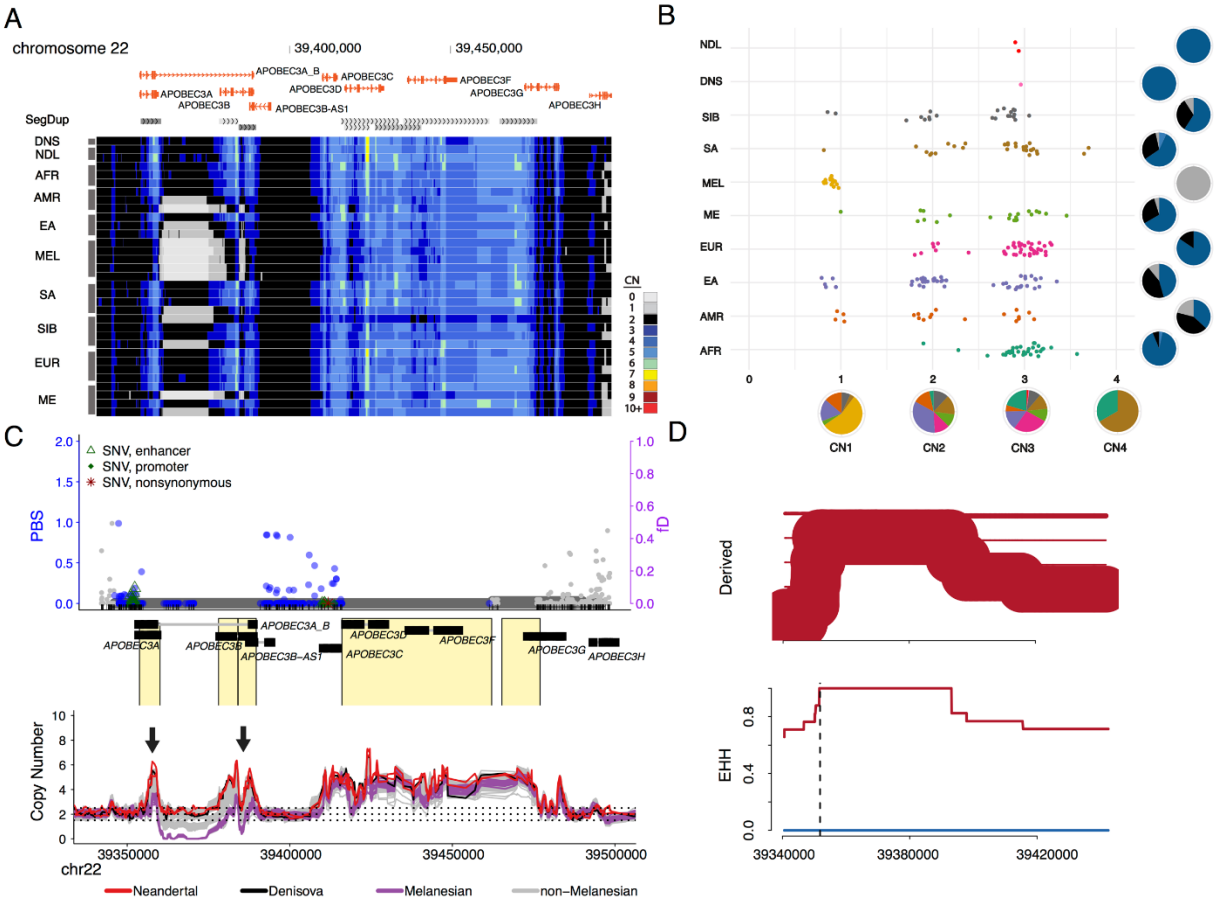


Figure S17. Evidence for positive selection of the deletion locus in the *APOBEC3* gene cluster (chr22:39,388,950-39,483,917) among Melanesians. (**A**) CN heat map around the candidate region and (**B**) the distribution of WSSD-based CN estimates for the eight SGDP populations and the three archaic samples. (**C**) Significant signals of positive selection at the flanking sequences of the deletion locus in Melanesians. Top panel: Distributions of *PBS* (left y-axis), functional annotation (RefSeq and ENCODE elements; **Methods**) for all SNVs (dots), and the f_D (horizontal bars, representing windows of 100 SNVs). Colored dots (blue) and horizontal bars (purple) indicate *p*-value < 0.05. Middle panel: SDs (light orange) and genes (gray lines: noncoding sequences, black boxes: exons). Bottom panel: CN line plot, where the trajectory of each line shows the CN variation across the region for a given sample. The two black arrows indicate the breakpoints of the deletion. These panels are aligned to the panel A above. (**D**) The bifurcation diagram (top) and EHH (bottom) using 266 SNVs from the flanking sequences of the deletion locus (dashed line) showing signals of positive selection (chr22:39,340,000-39,450,000) in Melanesians. Note that none of the Melanesian samples carry non-deletion versions of haplotypes (see **Figures S18-S19** for comparison among populations).

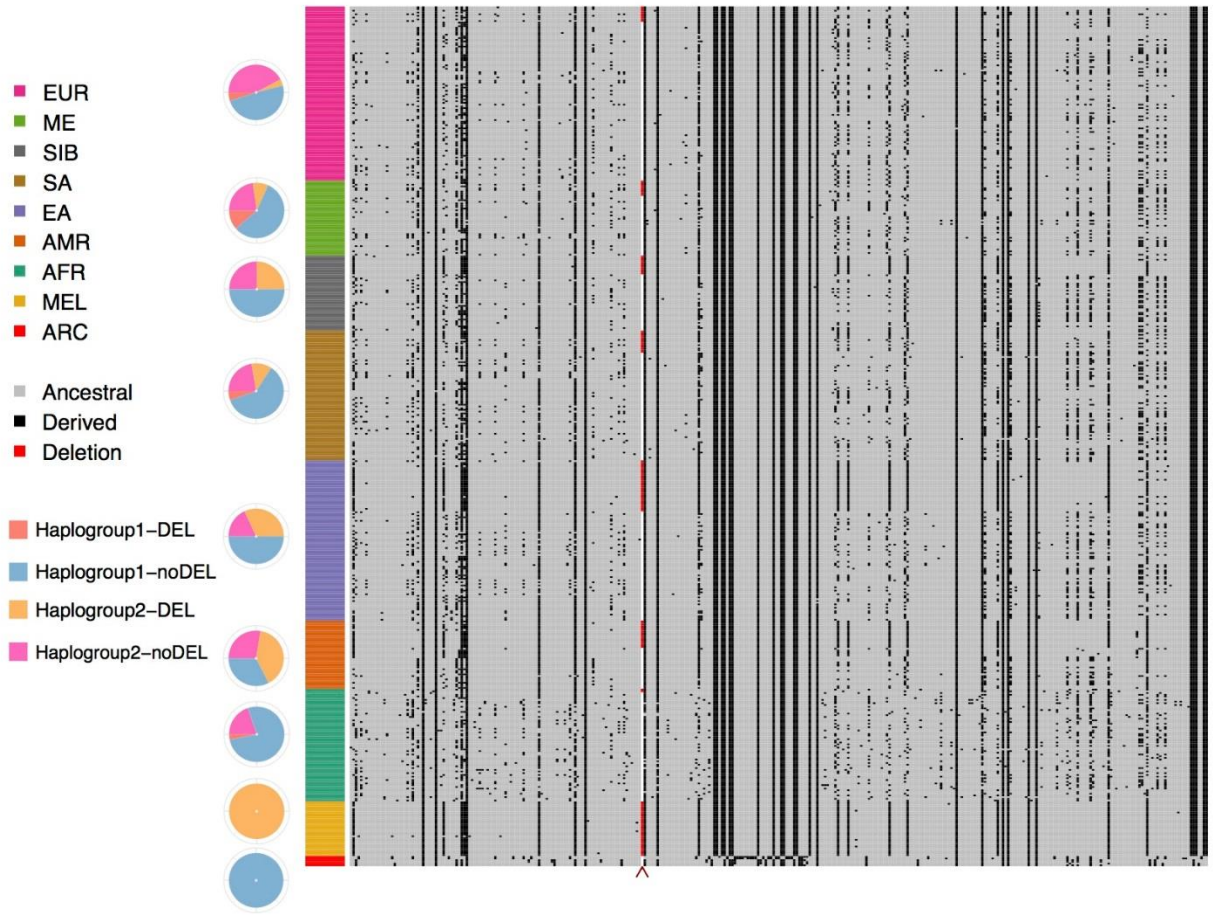


Figure S18. Haplotype pattern in the region shows signals of positive selection at the flanking sequences of the *APOBEC3* deletion (chr22:39,340,000-39,450,000; 266 SNVs) among the SGDP and three archaic samples. The rows and columns are haplotypes and SNVs, respectively. The red carrot indicates the position of the deletion. Haplogroups were defined using nine SNVs with *PBS* > 0.5, including the deletion locus. Haplogroups were formed by first grouping haplotypes with five mutations or less, followed by the status of the deletion locus. Pie charts are the distribution of haplogroups in individual populations.

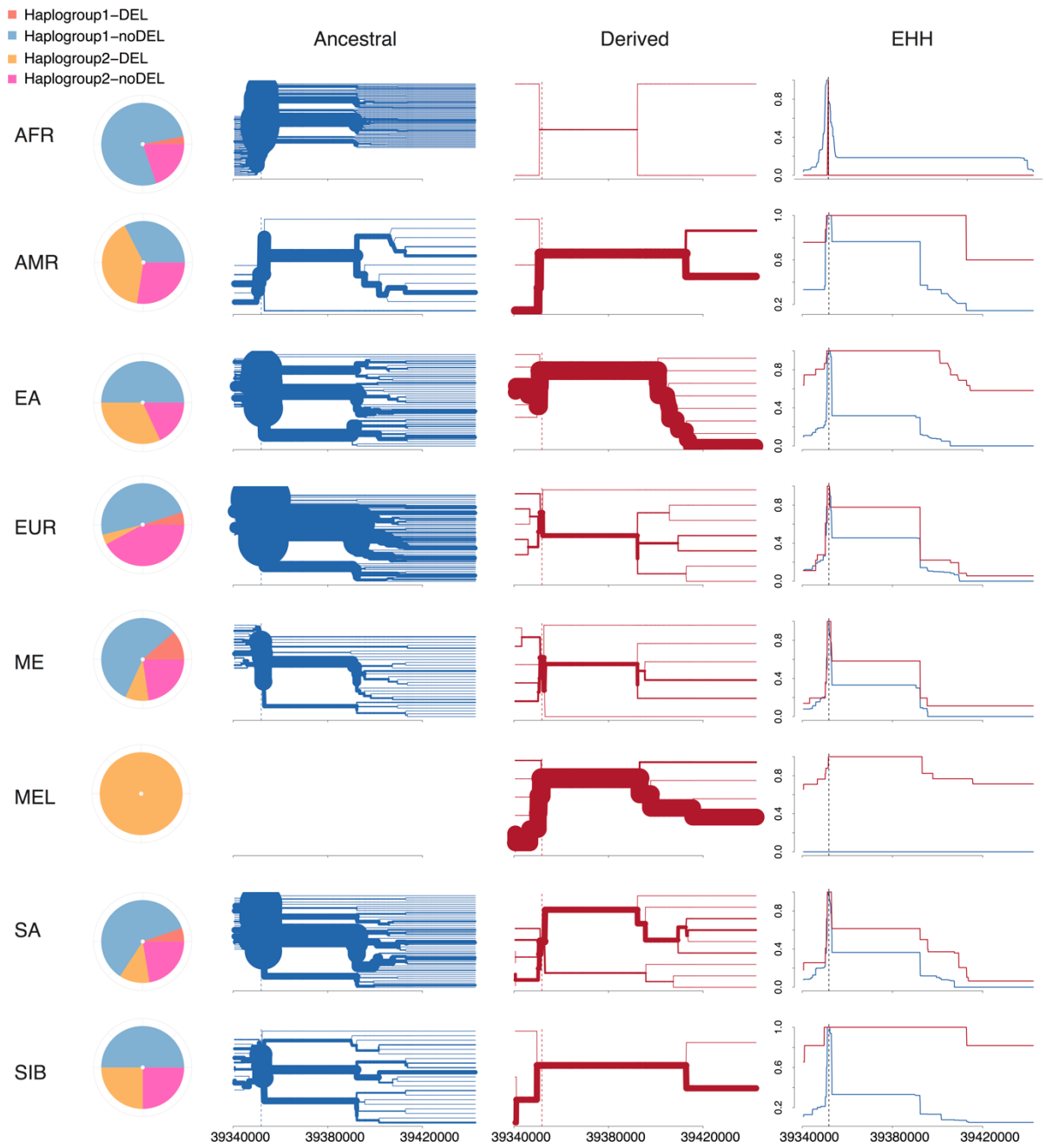


Figure S19. Distribution of haplogroups, bifurcation diagrams, and EHH in individual populations using 266 SNVs from the region showing signatures of positive selection (chr22:39,340,000-39,450,000) around the deletion at the *APOBEC3* gene cluster. Dashed line indicates the location of the *APOBEC3* deletion locus. Haplogroups were defined as described in **Figure S13**.

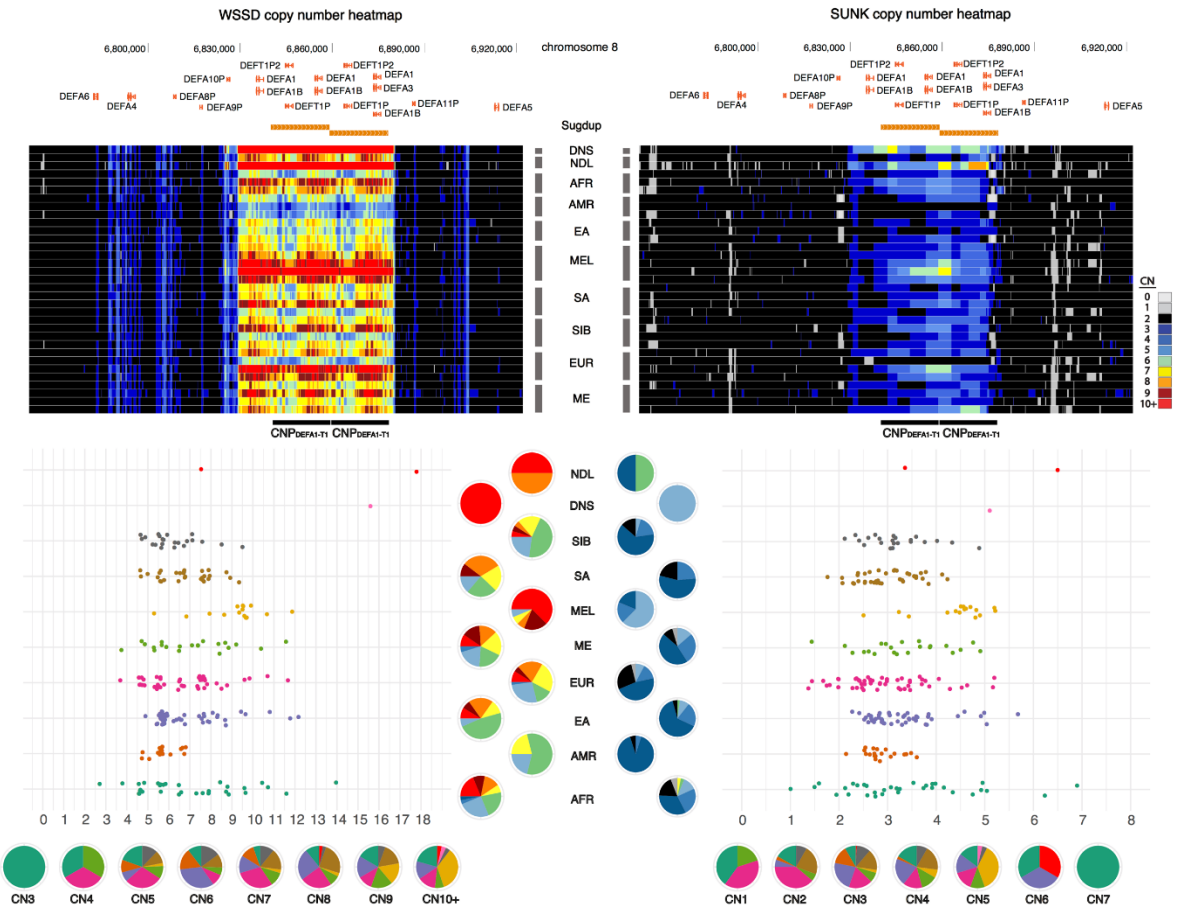


Figure S20. CN distributions for the multi-allelic CNV (chr8:6,839,960-6,878,169, $CNP_{DEFA1\text{-}T1}$) at the *DEFA* gene cluster on chromosome 8. Copy numbers were estimated using read-depth WSSD (left panels) and SUNK (right panels) genotyping methods. Top panels: CN heat maps, where each row represents the copy numbers of a sample over the region. Bottom panels: CN distributions for the variant among SGDP and three archaic samples. Pie charts on the x-axis indicate the population distributions in individual copy numbers (colors corresponding to those in the scatterplots), while each pie chart on the y-axis shows the frequency distribution of copy numbers for a given population (colors corresponding to those in the CN heat maps).

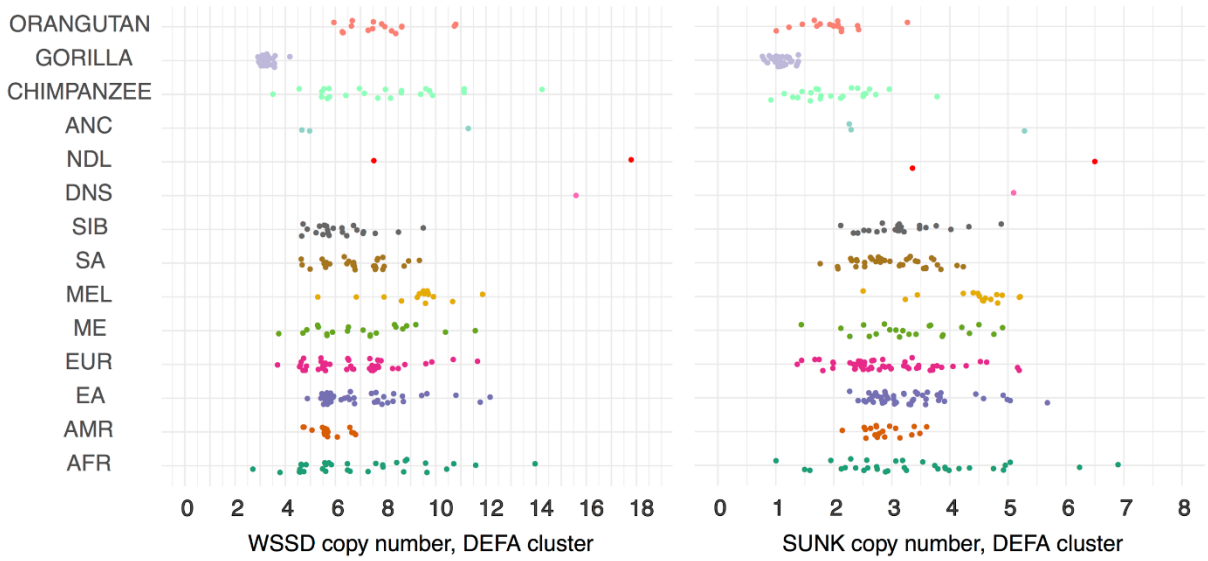


Figure S21. WSSD- and SUNK-based CN distributions for the multi-allelic CNV locus ($CNP_{DEFA1-T1}$; Figure S20) at the *DEFA* gene cluster among extant modern humans (SGDP), ancient modern humans (Stuttgart, Loschbour, and Ust-Ishim), and nonhuman great apes (chimpanzees, gorilla, and orangutans). Each point is a CN estimate of a sample for the variant.

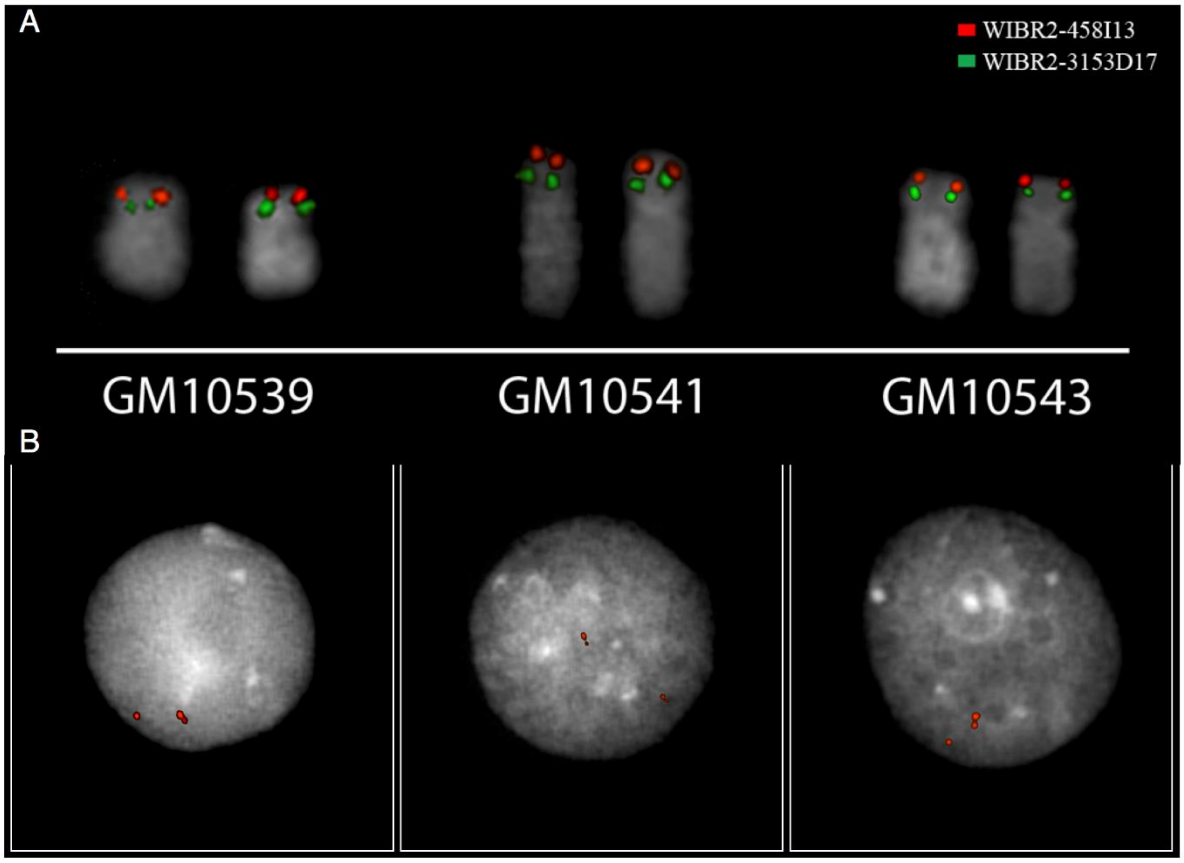


Figure S22. Metaphase (A) and interphase (B) FISH experiments for three Melanesian cell lines reveal the direct orientation of chromosome 8p23.1 and the tandem organization of the $CNP_{DEFA1\text{-}T1}$ duplications, respectively. (**A**) The order of the two fosmid clones, WIBR2-458I13 (red) and WIBR2-3153D17 (green), shows that all three individuals are in direct orientation for the 8p23.1 locus as represented in human reference genome (GRCh37). (**B**) The probe, WIBR2-2984I16 (chr8:6983382-7003217, red), queries the sequences at the junction of the two $CNP_{DEFA1\text{-}T1}$ segments.

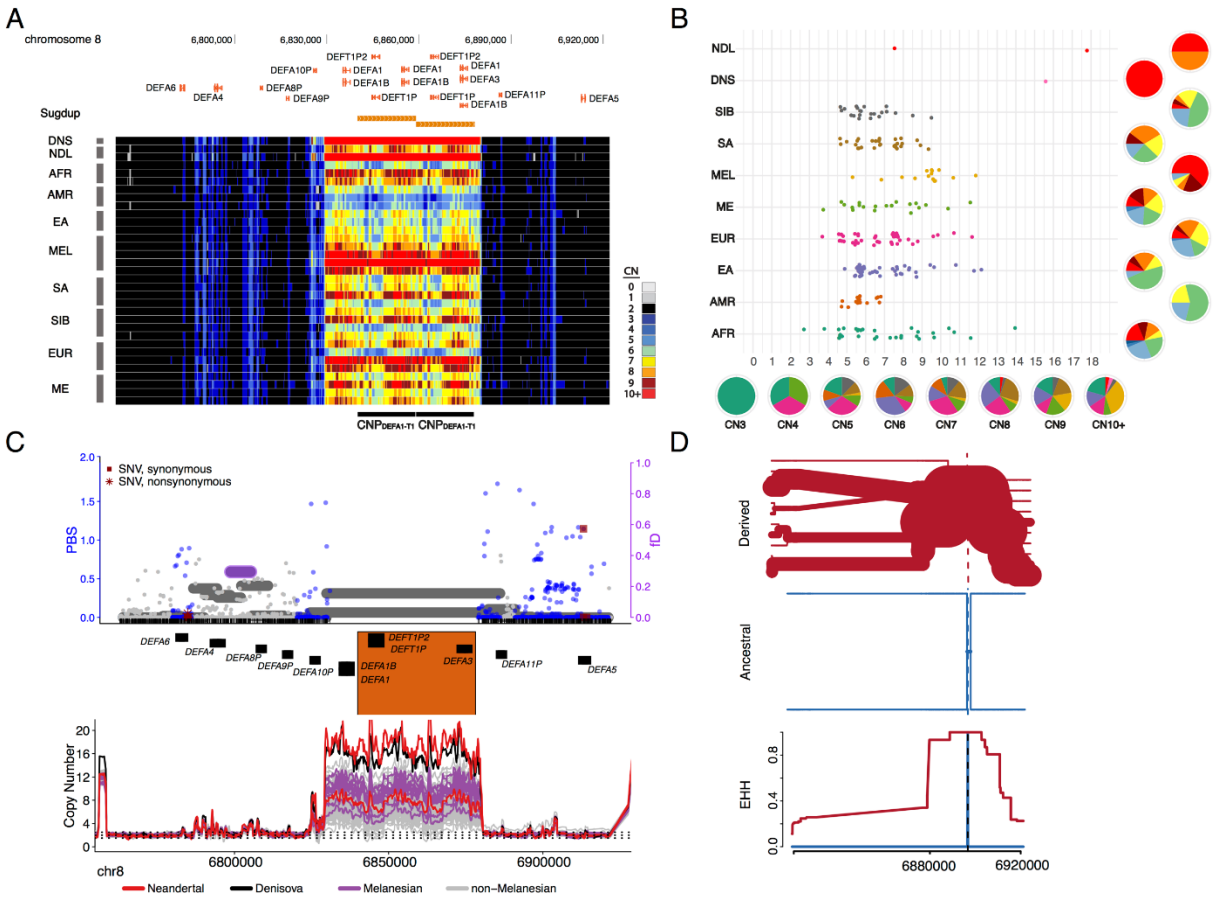


Figure S23. Evidence for positive selection of the multi-allelic CNV locus in the *DEFA* gene cluster ($CNP_{DEFA1\text{-}T1}$, chr8:6,839,960-6,878,169) in Melanesians. (**A**) CN heat map around the candidate region and (**B**) the distribution of WSSD-based CN estimates for the eight SGDP populations and the three archaic samples. (**C**) Significant signals of positive selection at the flanking sequences of the candidate locus in Melanesians. Top panel: Distributions of *PBS* (left y-axis), functional annotation (RefSeq and ENCODE elements; **Methods**) for all SNVs (dots), and the f_D (horizontal bars, representing windows of 100 SNVs). Colored dots (blue) and horizontal bars (purple) indicate *p*-value < 0.05. Middle panel: SDs (light orange) and genes (gray lines: noncoding sequences, black boxes: exons); Bottom panel: CN line plot, where the trajectory of each line shows the CN variation across the region for a given sample. These panels are aligned to panel A above. (**D**) The bifurcation diagram (top) and EHH (bottom) using 855 SNVs from the flanking sequences of the candidate CNV locus showing signals of positive selection (chr8:6,819,244-6,921,178) in Melanesians. Dashed line indicates one of the SNVs of the largest *PBS* values (chr8:6,896,688, *PBS*=1.5), whose ancestral state was polarized according to the human–chimpanzee alignment.

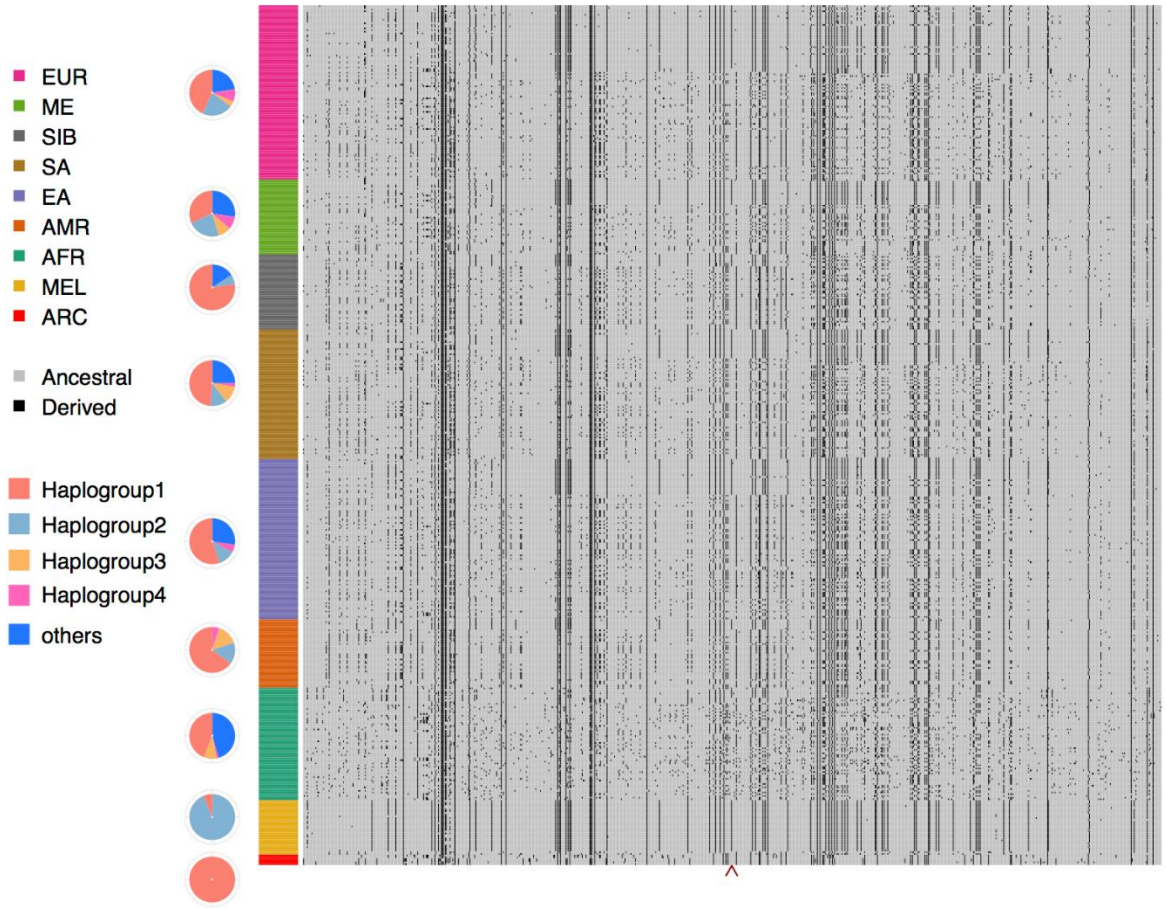


Figure S24. Haplotype pattern in the region showing signals of positive selection at the flanking sequences of the multi-allelic CNV within *DEFA* gene cluster (chr8:6,839,960-6,878,169; 855 SNVs) among the SGDP and three archaic samples. The rows and columns are haplotypes and SNVs, respectively. The red carrot indicates the position of the first SNV after the CNV locus. Haplogroups were defined using 36 SNVs with *PBS* > 0.5. Haplogroups were formed by grouping haplotypes with five mutations or less. To ease the complexity of the plot, we only used the first four major haplogroups (>23 haplotypes per group) and grouped the rest into the category "others" for display. Pie charts are the distribution of haplogroups in individual populations.

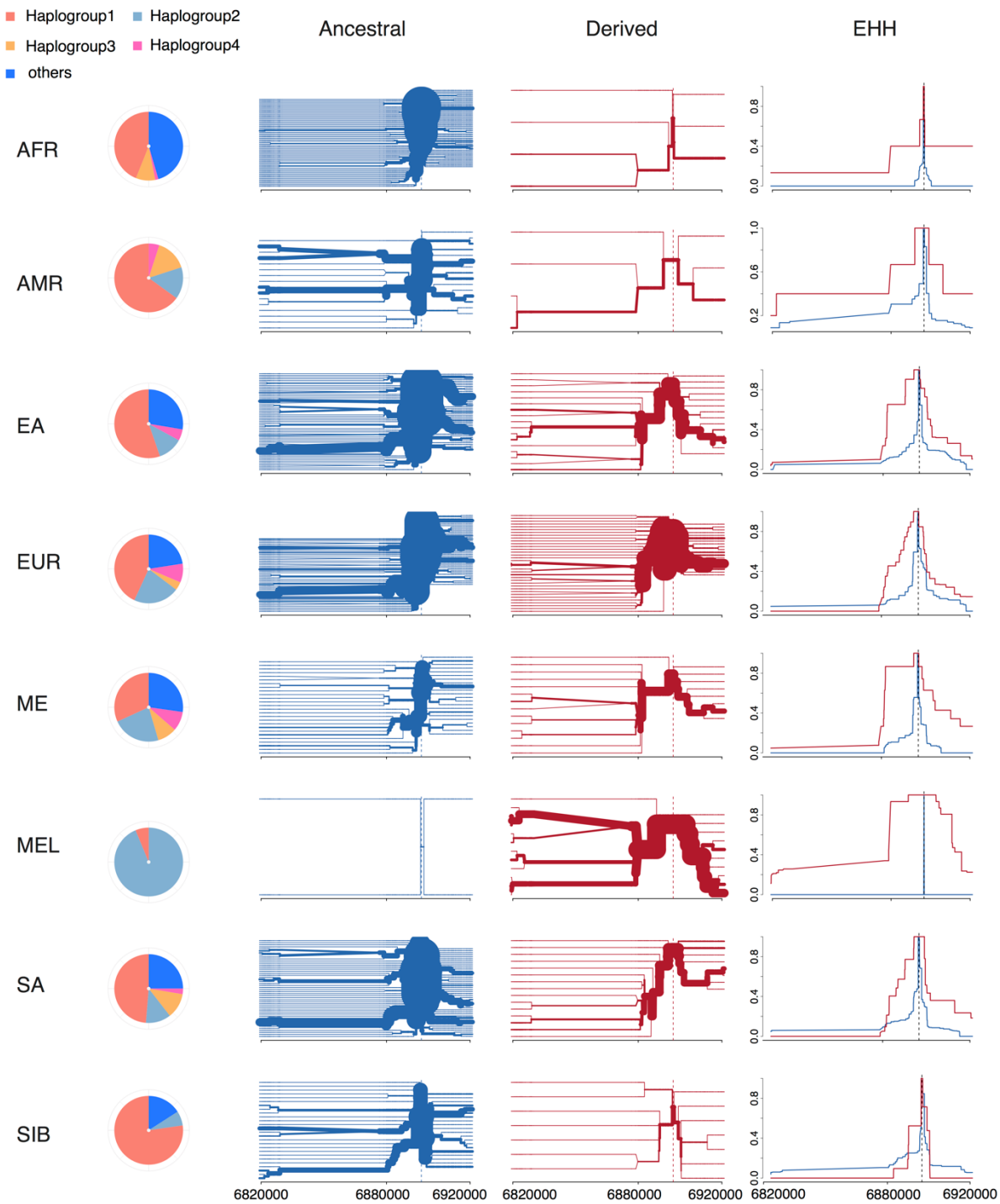


Figure S25. Distribution of haplogroups, bifurcation diagrams, and EHH in individual populations using 855 SNVs from the region showing signatures of positive selection (chr8:6,839,960-6,878,169) around the multi-allelic CNV locus at the *DEFA* gene cluster. Dashed line indicates one of the SNVs with the largest *PBS* (chr8:6,896,688, *PBS*=1.5), whose ancestral state was polarized according to the human–chimpanzee alignment.

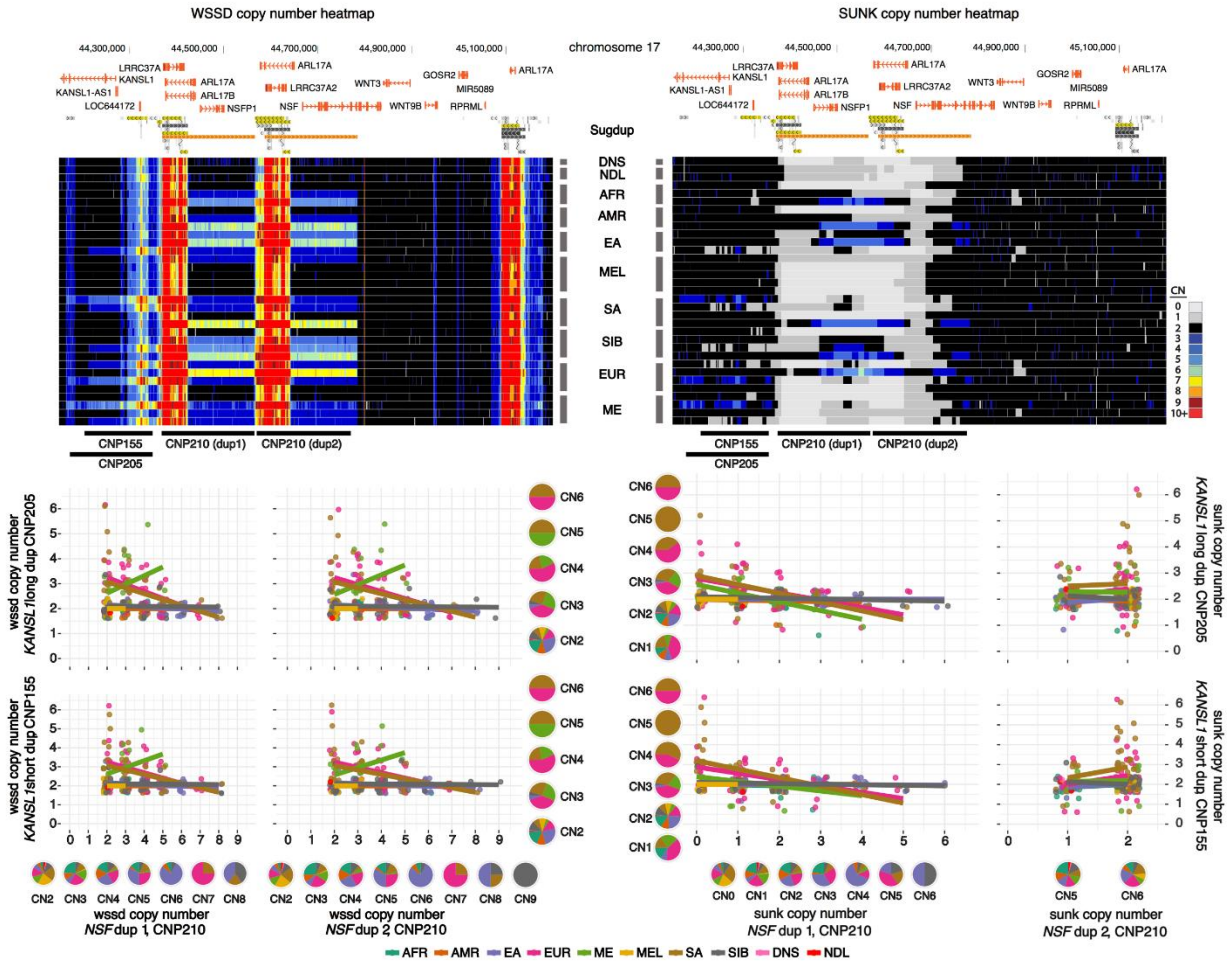


Figure S26. CN distributions for the three large CN polymorphic duplications (CNP155, CNP205, CNP210-dup1, and CNP210-dup2) at the chromosomal 17q21.31 locus. Copy numbers were estimated using read-depth WSSD (left panels) and SUNK (right panels) genotyping methods. Top panels: CN heat maps, where each row represents the copy numbers of a sample over the region. Bottom panels: Pairwise distributions of CN estimates for the three CN polymorphic loci among the SGDP and three archaic samples. Colors of dots represent individual populations, and regression lines for individual populations were shown. Pie charts on both axes indicate the population distributions in individual copy numbers (colors corresponding to those in the scatterplots).

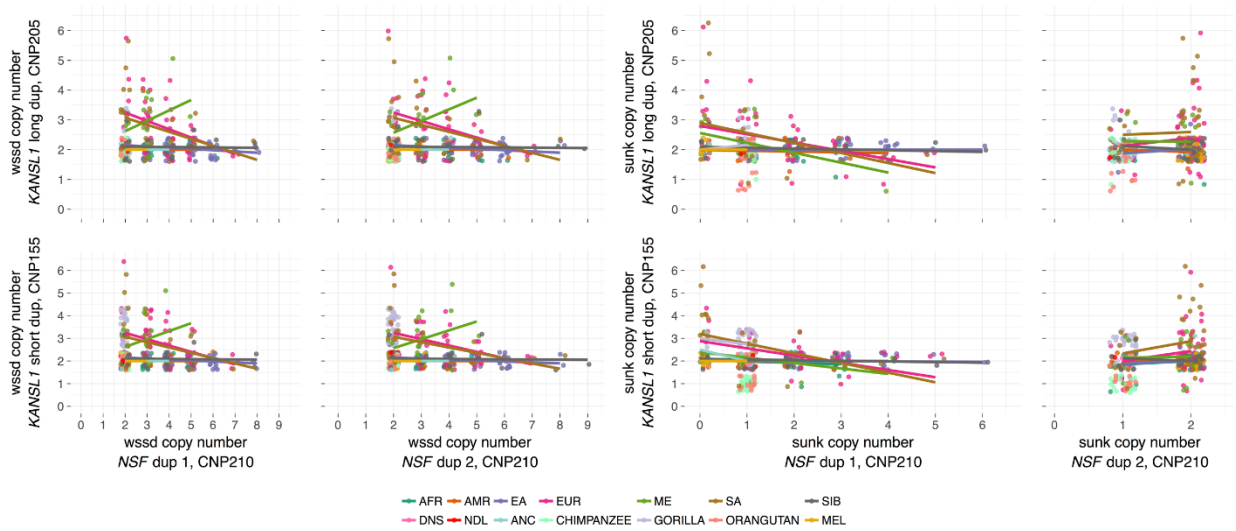


Figure S27. Pairwise distributions of WSSD- (left panels) and SUNK- (right panels) based CN estimates for CNP155, CNP205, CNP210-dup1, and CNP210-dup2 (Figure S21) among SGDP and nonhuman great ape samples. Colors of dots represent individual populations, and regression lines for individual populations were shown.

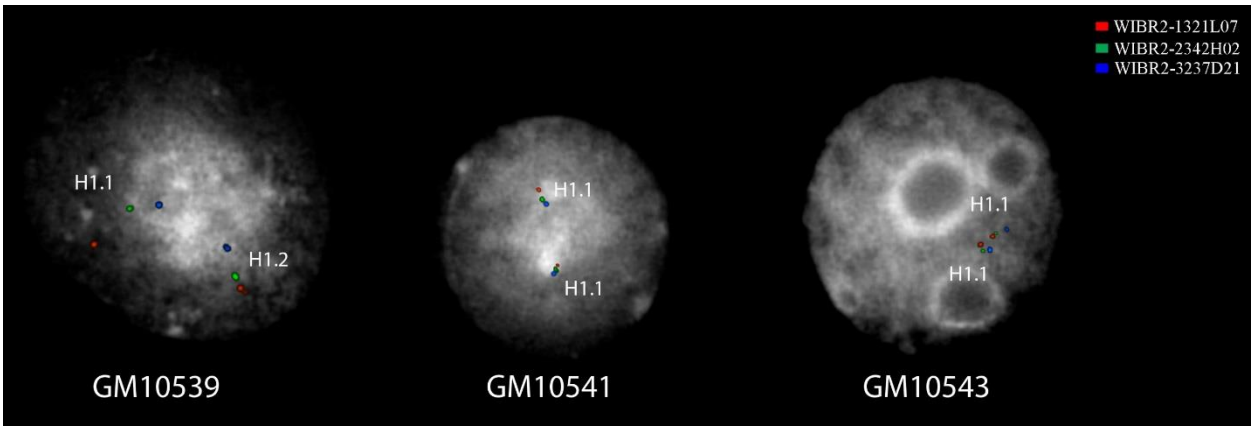


Figure S28. FISH experiments for the chromosome 17q21.31 locus using three Melanesian cell lines. All three individuals carry the direct haplotype, with GM10541 and GM10543 carrying only one copy of the CNP210 (WIBR2-1321L07, red) on both chromosomes (homozygous H1.1/ H1.1); GM10539 is heterozygous with one copy of the CNP210 on one chromosome and two copies on the other chromosome (heterozygous H1.1/H1.2).

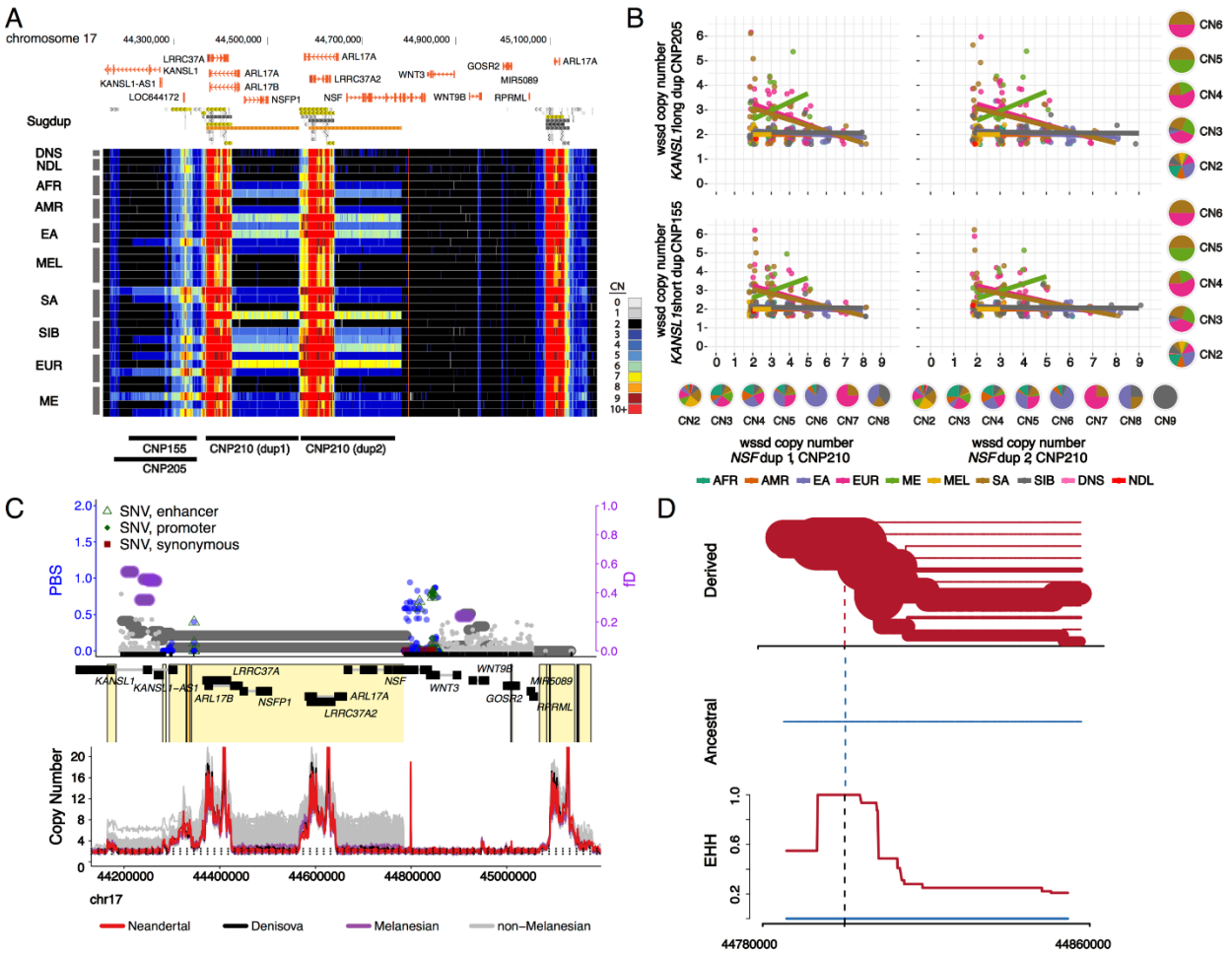


Figure S29. Evidence for positive selection at the chromosomal 17q21.31 locus in Melanesians. (**A**) CN heat map around the candidate region and (**B**) the distribution of WSSD-based CN estimates for the eight SGDP populations and the three archaic samples. (**C**) Significant signals of positive selection at the flanking sequences of the candidate locus in Melanesians. Top panel: Distributions of *PBS* (left y-axis), functional annotation (RefSeq and ENCODE elements; **Methods**) for all SNVs (dots), and the f_D (horizontal bars, representing windows of 100 SNVs). Colored dots (blue) and horizontal bars (purple) indicate *p*-value < 0.05. Middle panel: SDs (light orange) and genes (gray lines: noncoding sequences, black boxes: exons); Bottom panel: CN line plot, where the trajectory of each line shows the CN variation across the region for a given sample. These panels are aligned to panel A above. (**D**) The bifurcation diagram (top) and EHH (bottom) using 367 SNVs from the flanking sequences flanking the distal side of the CNP210-dup2 showing signals of positive selection (chr17:44,784,657-44,854,722) in Melanesians. Dashed line indicates one of the SNVs of the largest *PBS* values (chr8:44,800,046, *PBS*=0.94), whose ancestral state was polarized according to the human–chimpanzee alignment.

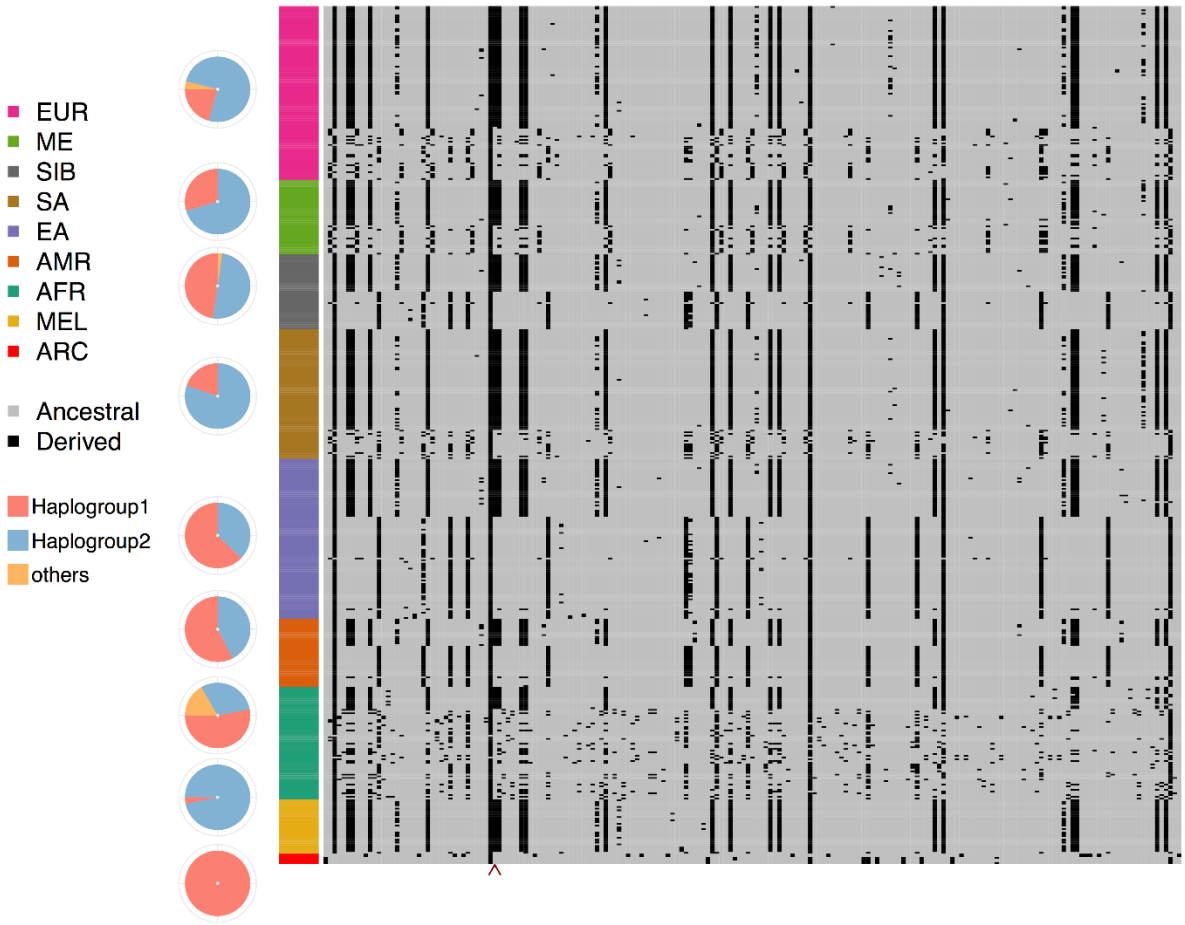


Figure S30. Haplotype pattern in the region showing signals of positive selection at the flanking sequences of the distal side of CNP210-dup2 on chromosome 17q21.31 (chr17:44,784,657-44,854,722; 367 SNVs) among the SGDP and three archaic samples. The rows and columns are haplotypes and SNVs, respectively. Haplogroups were defined using 24 SNVs with *PBS* > 0.5. Haplogroups were formed by grouping haplotypes with five mutations or less. To ease the complexity of the plot, we only used the first two major haplogroups and grouped the rest into the category "others" for display. Pie charts are the distribution of haplogroups in individual populations. The red carrot indicates the position of the SNV with largest *PBS* value at this region (chr8:44,800,046, *PBS*=0.94).

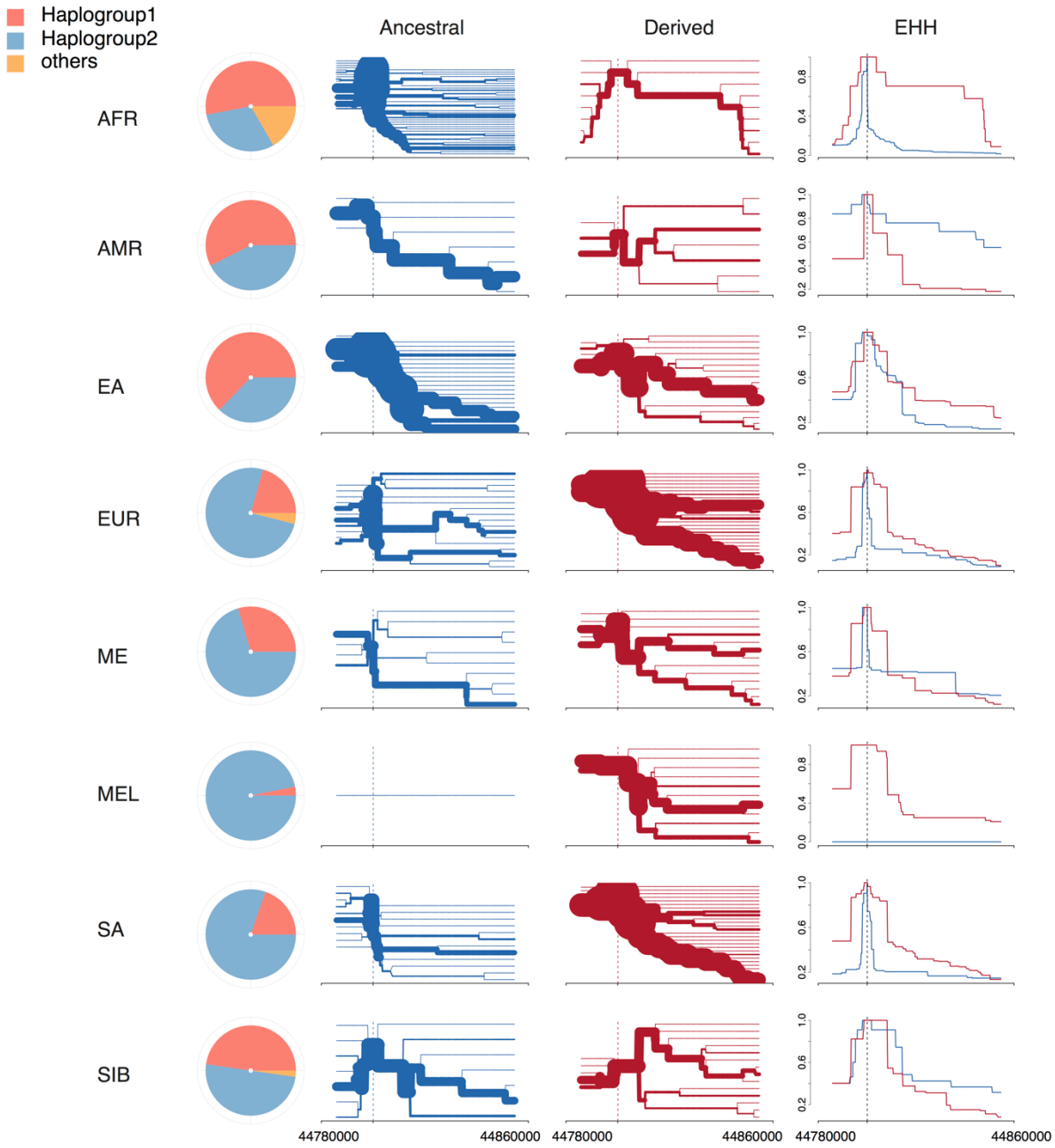


Figure S31. Distribution of haplogroups, bifurcation diagrams, and EHH in individual populations using 367 SNVs from the region showing signatures of positive selection on chromosome 17q21.31 (chr17:44,784,657-44,854,722; 367 SNVs). Dashed line indicates one of the SNVs with the largest *PBS* (chr8:44,800,046, *PBS*=0.94), whose ancestral state was polarized according to the human–chimpanzee alignment.

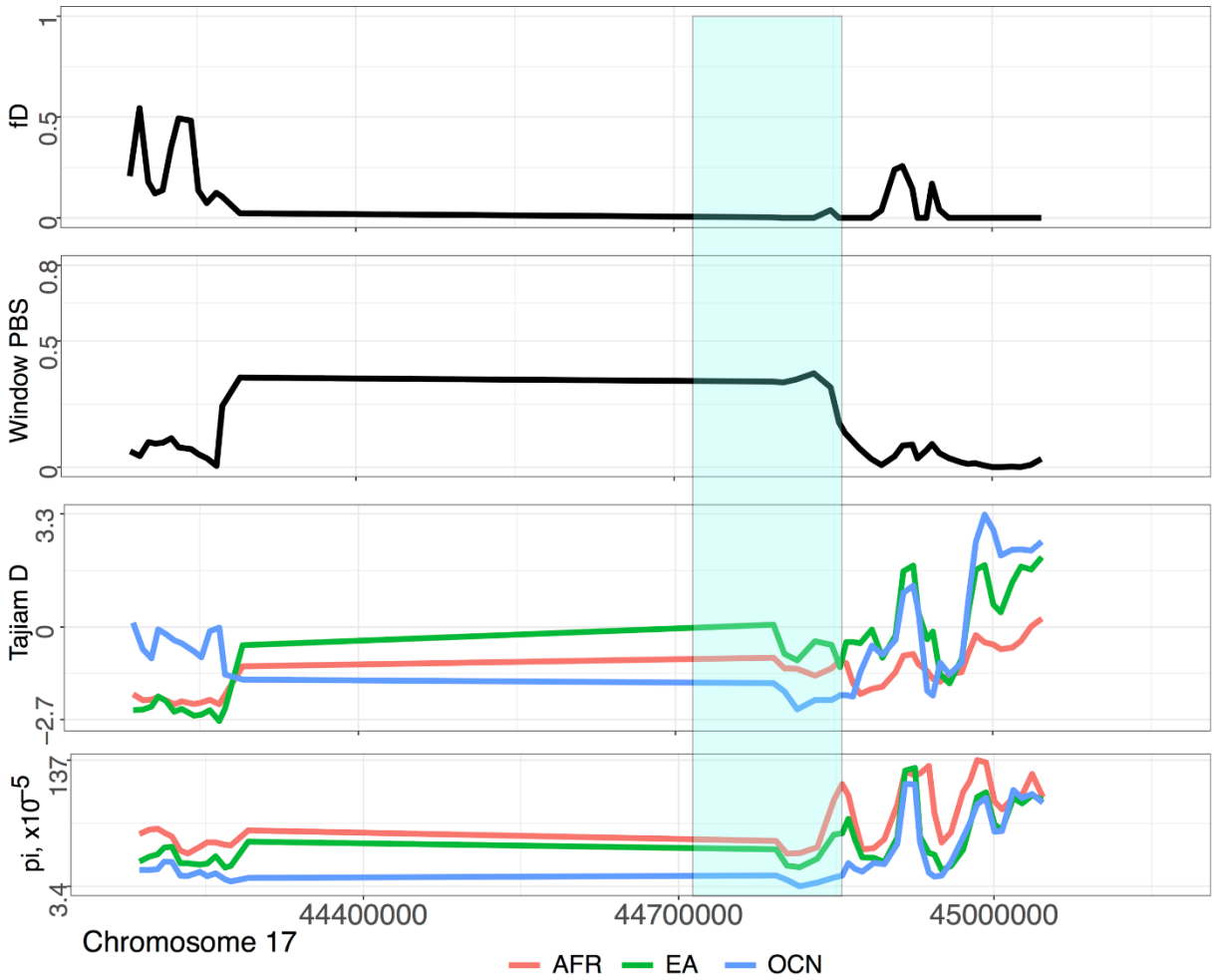
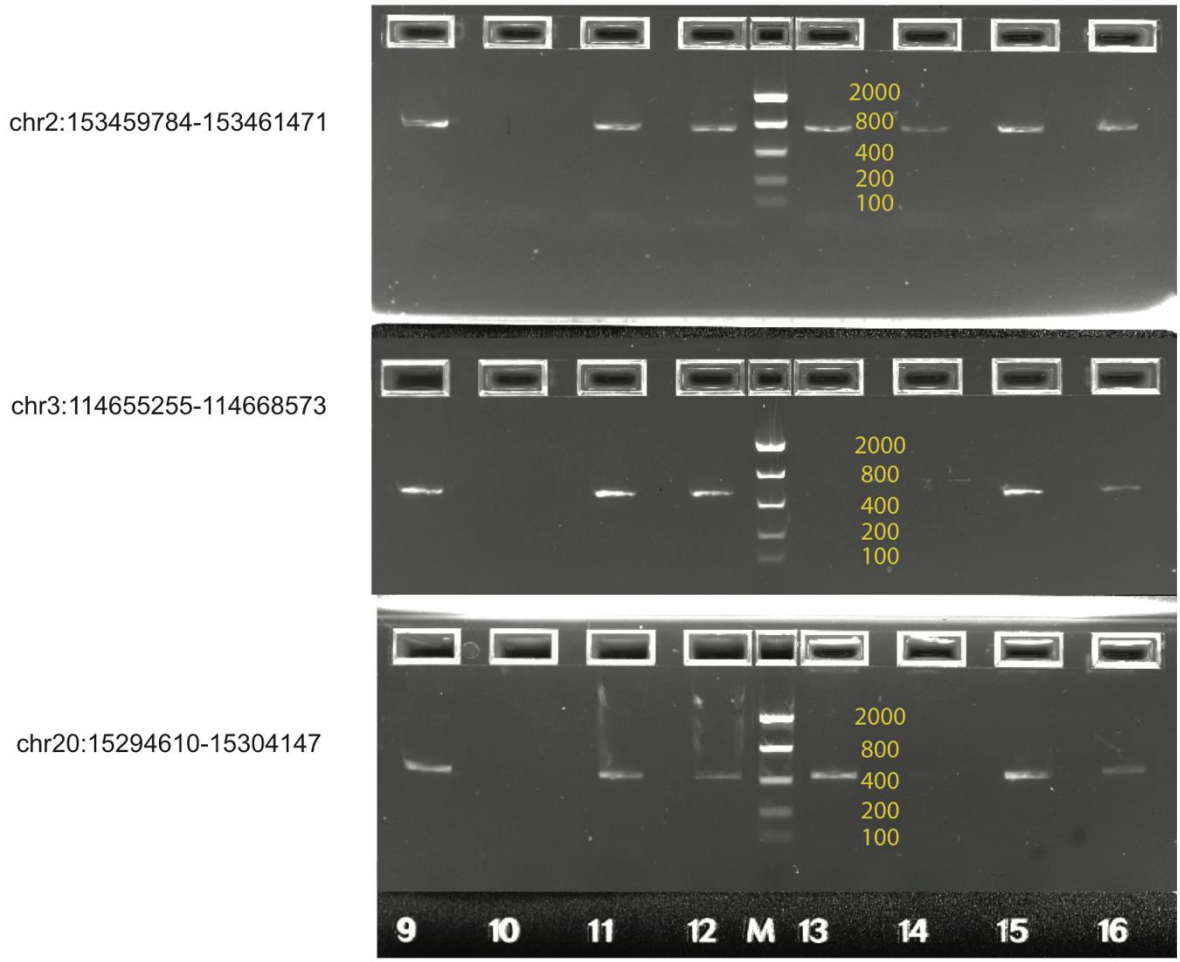


Figure S32. Test statistics for searching signatures of archaic introgression (f_D) and selection (*PBS*, Tajima's *D*, and *pi*) at the chromosome 17q21.31 locus in Melanesians. Note that all statistics were calculated using windows of 100 SNVs across the region. For comparison, statistics using AFR and EA samples were also plotted for both Tajima's *D* and *pi*. Light green shaded rectangle indicates the flanking region of the distal side of CNP210-dup2.



Gel #	Sample	DNA material	Location	chr2:153459784-153461471 deletion	chr3:114655255-114668573 deletion	chr20:15294610-15304147 deletion
9	GM10539	Cell line	Bougainville	Yes	Yes	Yes
10	UV0565	Blood	Nailik, North New Ireland AN	No	No	No
11	UV2196	Blood	Teop, North Bougainville	Yes	Yes	Yes
12	UV034	Blood	Baining (Marabu), East New Britain	Yes	Yes	Yes
13	UV0726	Blood	Kuot (Kabil), New Ireland PAP	Yes	No	Yes
14	UV1023	Blood	Mamusi (Lingite), West New Britain	Yes	No	No
15	UV2002	Blood	Tolai, East New Britain	Yes	Yes	Yes
16	UV001	Blood	Baining (Marabu), East New Britain	Yes	Yes	Yes

Figure S33. PCR validation of additional three adaptive CNV candidates for one cell line and seven blood-derived Melanesian DNA samples. For these CNVs, in the SGDP samples, one is the deletion variant (top panel) and two are multi-allelic CNVs (bottom two panels). Relevant selection statistics are reported in **Supplementary Table S9**. Note that most of Melanesian samples in the SGDP panel carry deletion alleles in all three loci, and thus our PCR experiments were designed so that PCR products are observed only if the deletion alleles are present. Ladder is at 2000 bp, 800 bp, 400 bp, 200 bp, and 100 bp. These PCR assays show 721, 519, and 438 bp products that is produced when the deletion alleles are present. Primers used in these assays are listed in **Supplementary Table S21**. The table at the bottom summarizes the results of our experiments. In all three cases, we confirm the presences of the deletion alleles in the cell line samples and several blood-derived DNA samples.

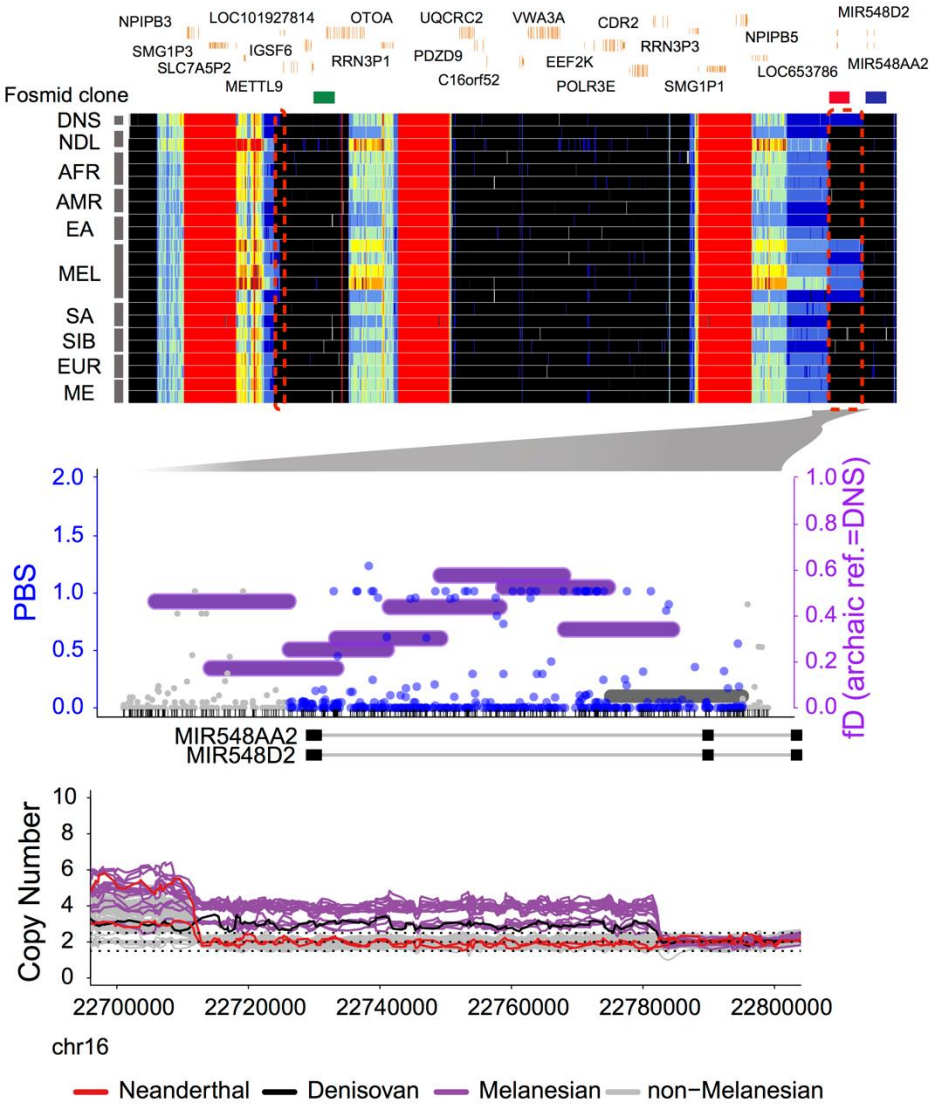


Figure S34. Melanesian–Denisovan-specific duplication at chromosome 16p12.2. Top: CN heat map for the chromosome 16p12.2 locus. Red dashed boxes indicate the 225 kbp Melanesian–Denisovan-specific duplication. Middle: The *PBS* (population branch statistic, left y-axis) for SNVs (dots) and f_D (horizontal lines, representing windows of 100 SNVs, computed using Denisovan (DNS) as the archaic reference, right axis) at DUP_{16p12}. Colored dots (blue) and/or horizontal lines (purple) indicate significant test statistics ($p < 0.05$). Bottom: CN line plot where the trajectory of each line shows CN variation for a given sample.

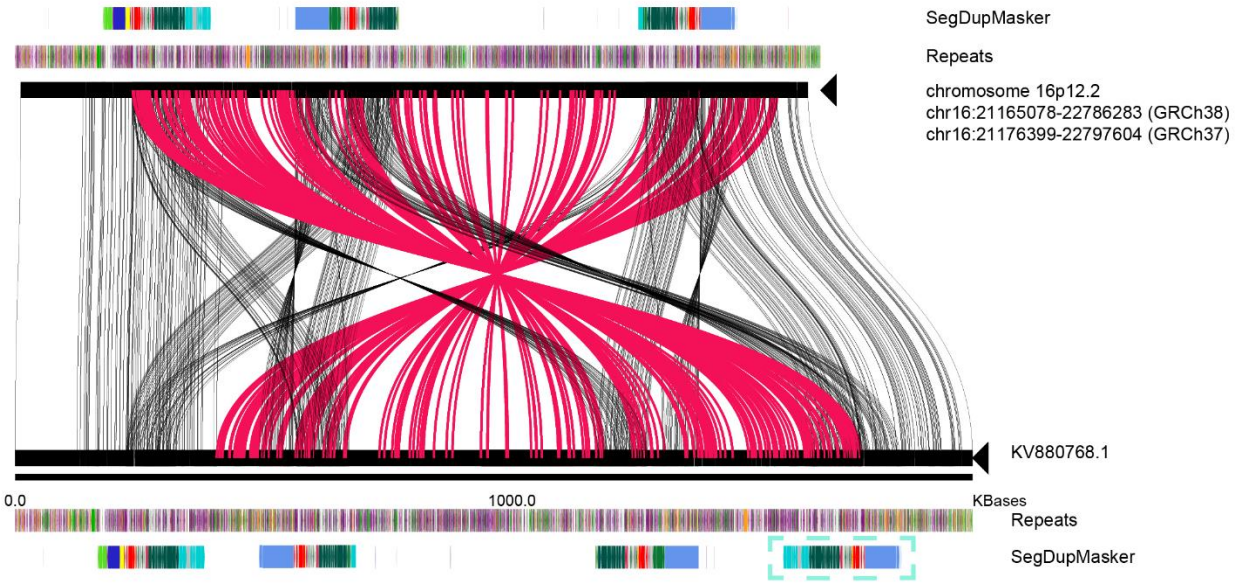


Figure S35. Miropeats analysis indicates genomic misassembly and misorientation at chromosome 16p12.2 in human reference (GRCh37 and GRCh38). The patch contig (KV880768.1) was downloaded from NCBI (BioProject: PRJNA31257). Colored boxes are annotated human SDs and lines connecting the sequences show regions of homology. Red lines and light green dashed box indicate misassembled regions for 16p12.2 in the human reference.

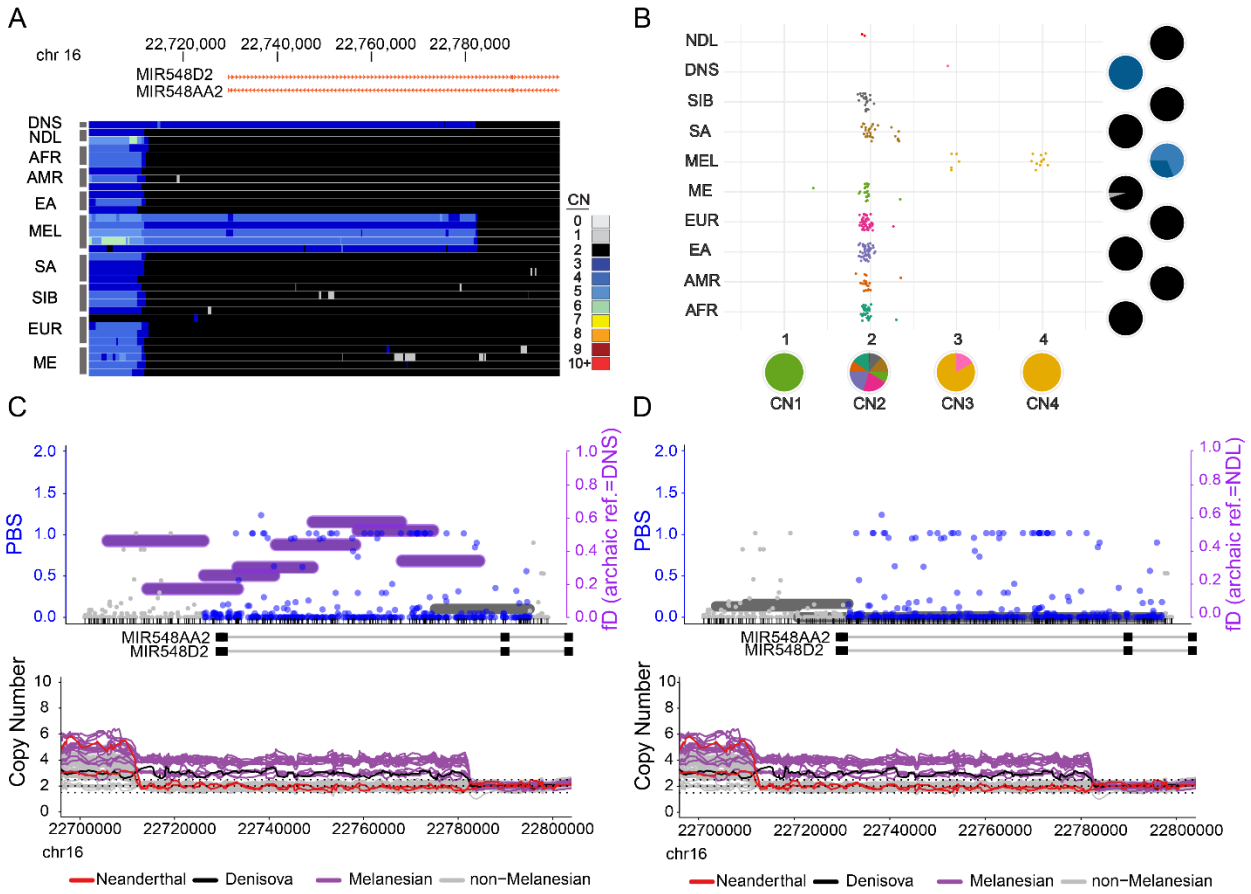


Figure S36. Evidence for Denisovan introgression at chromosome 16p12.2 duplication locus in Melanesians. (A) CN heat map around the candidate region (DUP_{16p12}, chr16:22,710,041-22,783,558, GRCh37) and (B) the distribution of WSSD-based CN estimates for the eight SGDP populations and the three archaic samples. (C and D) Significant signals of Denisovan introgression in the region shown in (A). Top panel: distributions of *PBS* (left y-axis), functional annotation (RefSeq and ENCODE elements; **Methods**) for all SNVs (dots), the f_D (horizontal bars, representing windows of 100 SNVs). Colored dots (blue) and horizontal bars (purple) indicate *p*-value < 0.05. Middle panel: genes (gray lines: noncoding sequences, black boxes: exons); bottom panel: the CN line plot, where the trajectory of each line shows the CN variation for a given sample across the region. These panels are aligned to subplot A. Note that introgression signals disappear when Neanderthals were used as the archaic reference in the f_D computation.

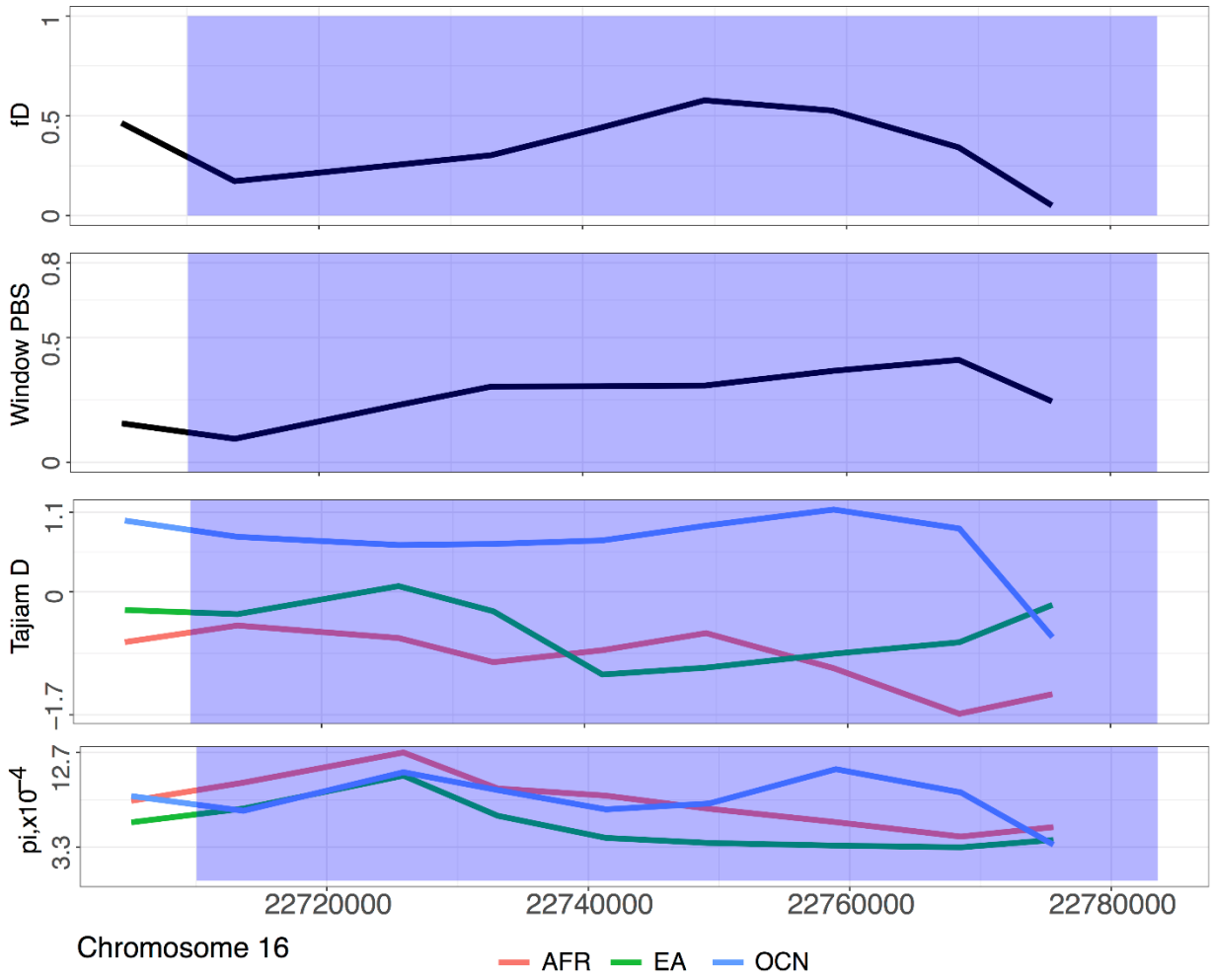


Figure S37. Elevated Tajima's *D* values at the Melanesian–Denisovan-specific duplication at chromosome 16p12.2 (GRCh37). Note that all statistics were calculated using windows of 100 SNVs across the region. For comparisons, statistics using AFR and EA samples were also plotted for both Tajima's *D* and *pi*. Purple shaded area indicates the region of interest.

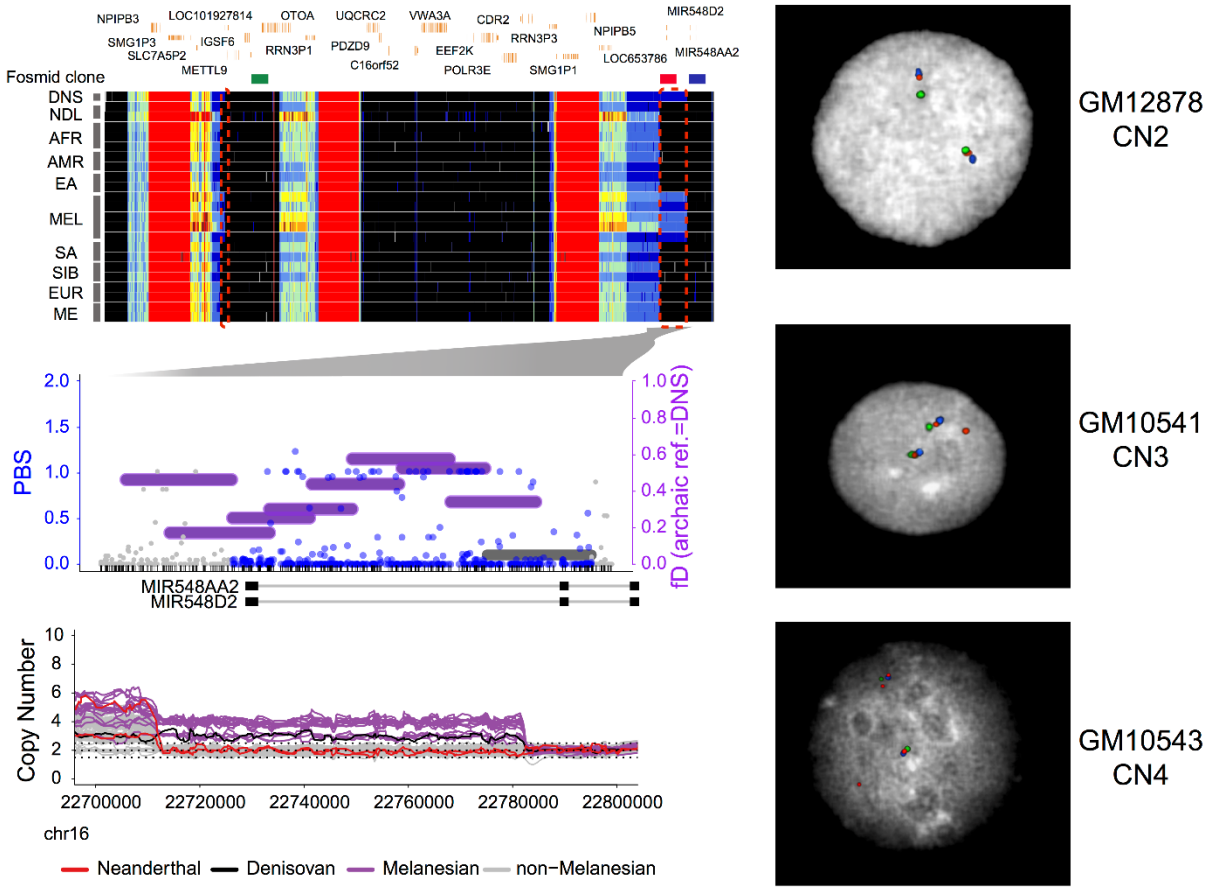


Figure S38. Testing for the presence of the chromosome 16p12.2 duplication in Melanesian cell lines. The browser track shows the location of the selected fosmid probes (**Table S13**) with colors corresponding to those in the FISH images on the right. See **Figure S34** for the detailed legends for the left panel. Right panel: FISH images show the presence or absence of an extra copy of the red fosmid clone (174222_ABC10_2_1_000044550500_M3; **Table S13**), which represents the DUP_{16p12} sequence, in three cell lines (GM12878, GM10541, and GM10543). Details of the experiments can be found at the Materials and Methods section.

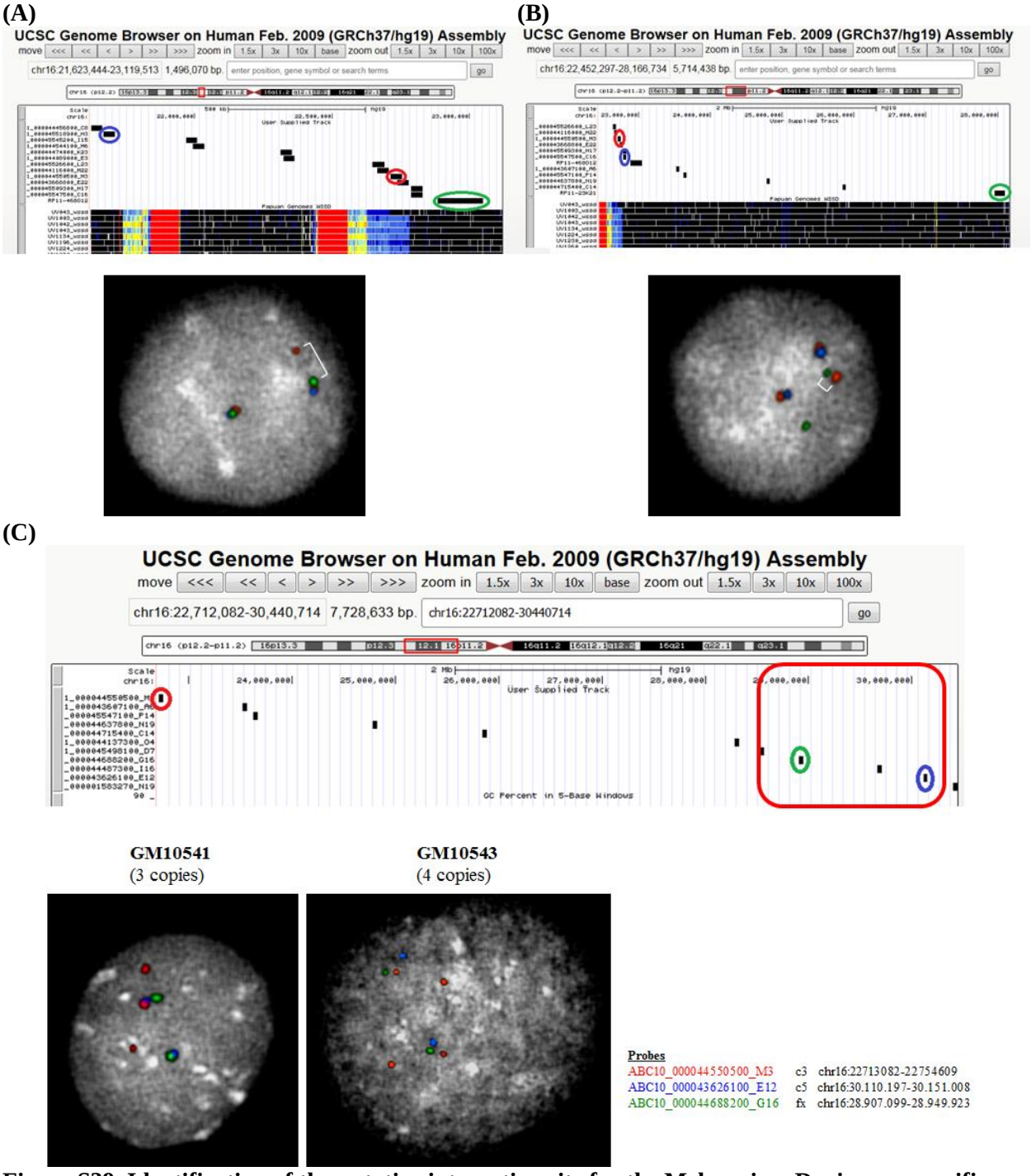


Figure S39. Identification of the putative integration site for the Melanesian–Denisovan-specific duplication using a series of FISH probes tiled along the short arm of chromosome 16. (**A**) FISH test to place the duplicated copy using GM10541 (CN3). Using a distal fosmid clone (ABC10-45518900-H3, blue), we see the duplicated copy land in a proximal location compared to the green BAC (RP11-468O12). (**B**) FISH test with an adjacent fosmid (ABC10-45547500-C16, blue) and more proximal BAC (RP11-23K21, green) shows the extra red copy mapping proximal to the BAC. (**C**) FISH test with two fosmids within the 16p11.2 region (ABC10-44688200-G16, green and ABC10-43626100-E12, blue) shows the duplicated red fosmid mapping within these two fosmid probes, indicating the location of the duplication.

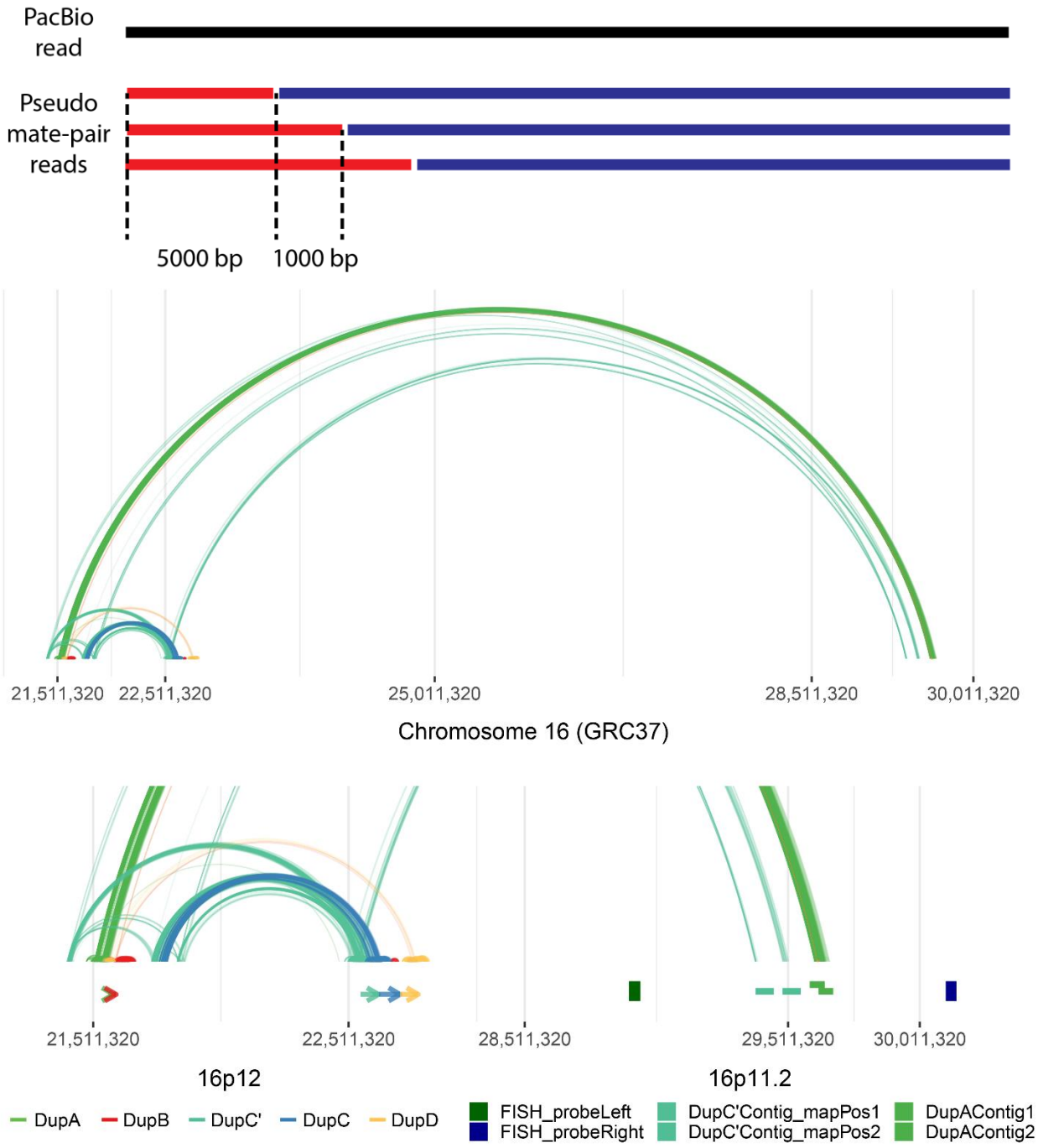


Figure S40. Link-plot between chromosome 16p12.2 and 16p11.2 using pseudo mate-pair reads constructed from long-read data. 75X long-read PacBio data of the chromosome 16p12.2 locus were generated for a Melanesian sample (HGDP00550), who is homozygous for the DUP_{16p12} duplication variant. Top: We selected reads extending over duplication segment junctions, constructed mate-pair reads by slicing individual reads using a window of 5 kbp with a step size of 1 kbp, and mapped these pseudo mate-pair reads using BWA-MEM in default settings. Bottom: Each arch line represents the link between a mate-pair and its color indicate to which the duplication segment in 16p12.2, its mate, mapped. Contigs at the putative integration site on chromosome 16p11.2 were assembled using Canu and reads showing links between 16p12.2 and 16p11.2. Note that the contig DupC'Contig can be mapped to two locations (DupC'Contig_mapPos1 and DupC'Contig_mapPos2) within complex SDs in 16p11.2.

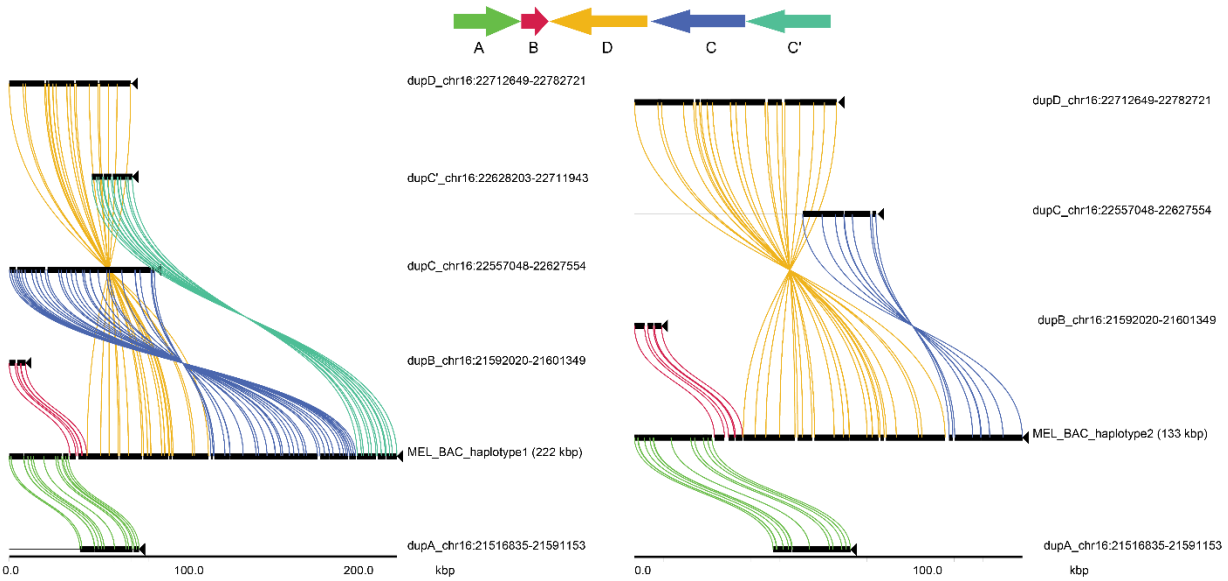


Figure S41. Miropeats analysis of two sequence-resolved BAC haplotypes for DUP_{16p12} against the five source sequences at the ancestral locus at 16p12.2 (GRCh37) as defined in (*9*). Our analysis confirms the predicted duplication structure (arrows, top panel) proposed in (*9*). BAC sequences for DUP_{16p12} from a Melanesian large-insert BAC library (GM10539) were generated using PacBio long-read sequencing technology.

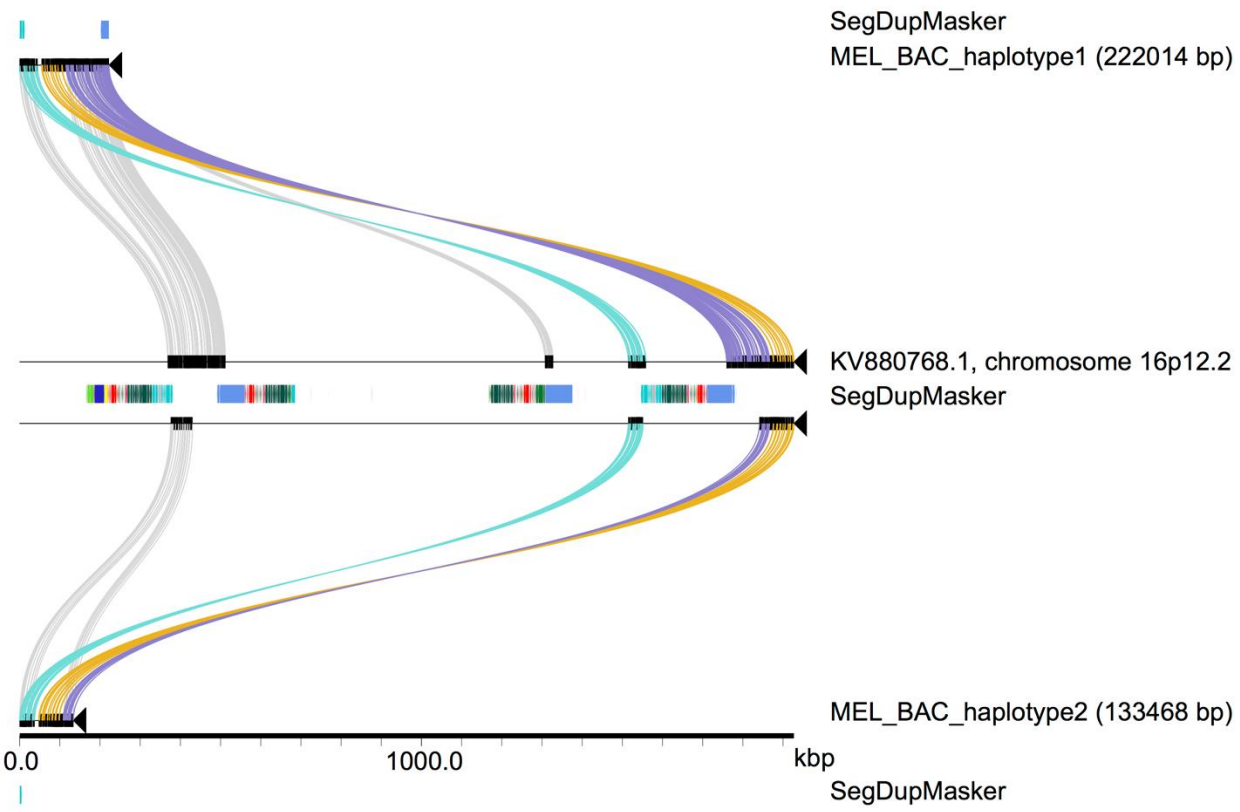


Figure S42. Miropeats analysis of two sequence-resolved DUP_{16p12} BAC haplotypes against the 16p12.2 patch contig (KV880768.1) (*39*). Colored boxes are annotated human SDs and lines connecting the sequences show regions of homology. Colored lines indicate regions corresponding to the DUP_{16p12} duplication sequence.

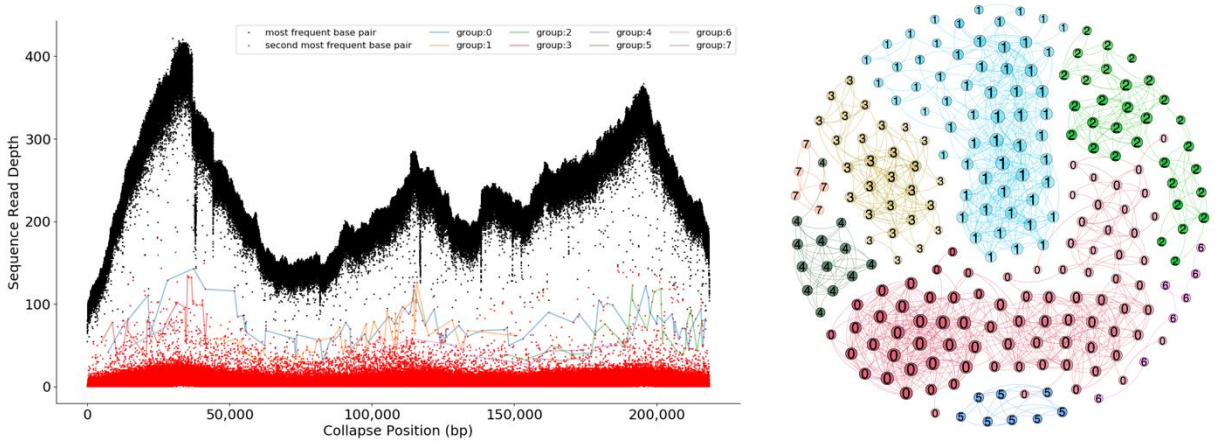


Figure S43. Segmental Duplication Assembler (SDA) constructing the Melanesian DUP_{16p12} contig *ab initio* using the 222 kbp BAC contig as its initial seed. In short, SDA explores and clusters PacBio reads that shared the same PSVs and performs local *de novo* sequence assembly for each cluster of reads. The left panel shows the read-depth profiling of the first run of the SDA iterations using the 222 kbp BAC haplotype as a seed to extract reads corresponding to the DUP_{16p12} duplication variant in the PacBio long-read genome. A contig of 252 kbp was built using reads from group:0 (blue line in the left panel; the red cluster on the right panel).

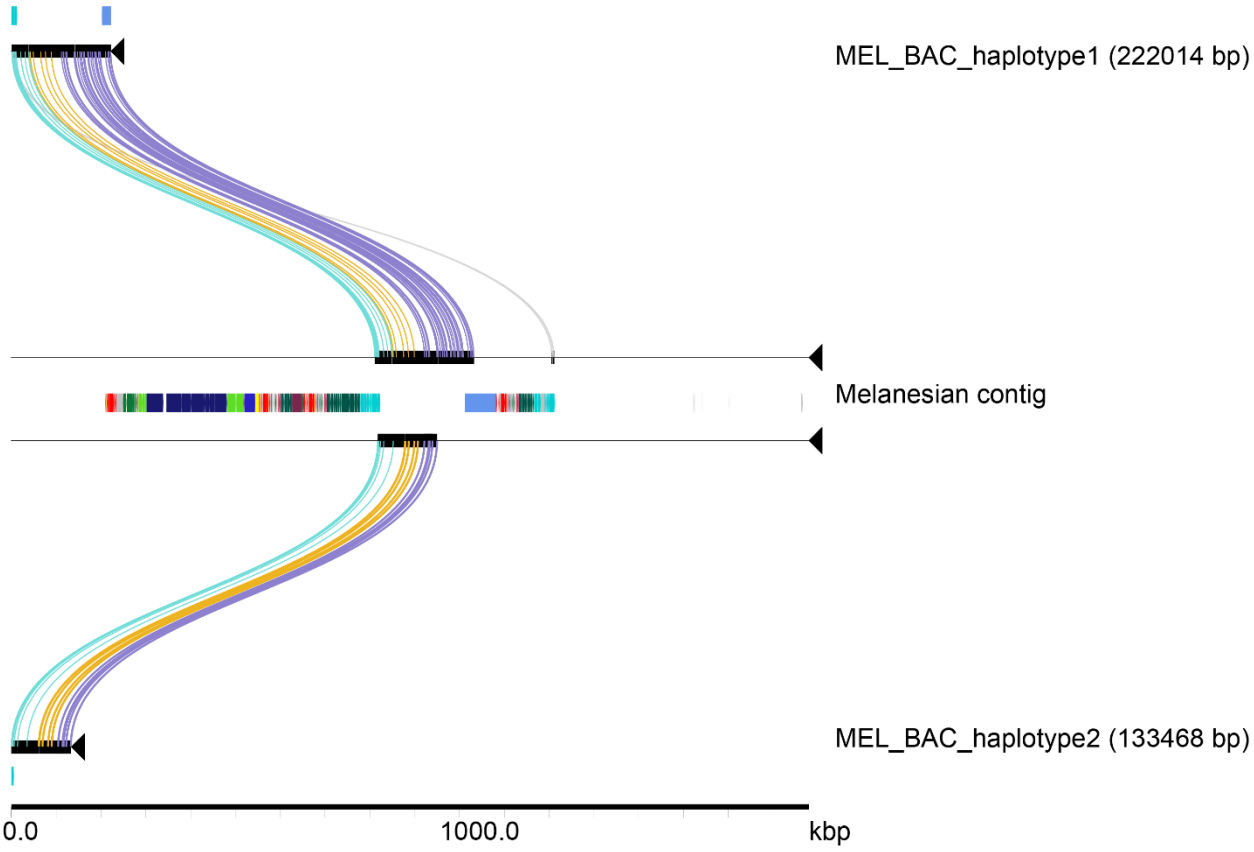


Figure S44. Miropeats analysis of two sequence-resolved DUP_{16p12} BAC haplotypes against the long-read assembled Melanesian contig by SDA (*41*). Colored boxes are annotated human SDs and lines connecting the sequences show regions of homology. Colored lines indicate regions corresponding to the DUP_{16p12} duplication sequence. The sequence accuracy of the Melanesian SDA contig is 99.86% using the 222 kbp BAC haplotype as the baseline. Note that the BAC (GM10539) and the genome (HGDP00550) were sequenced from two unrelated Melanesian individuals.

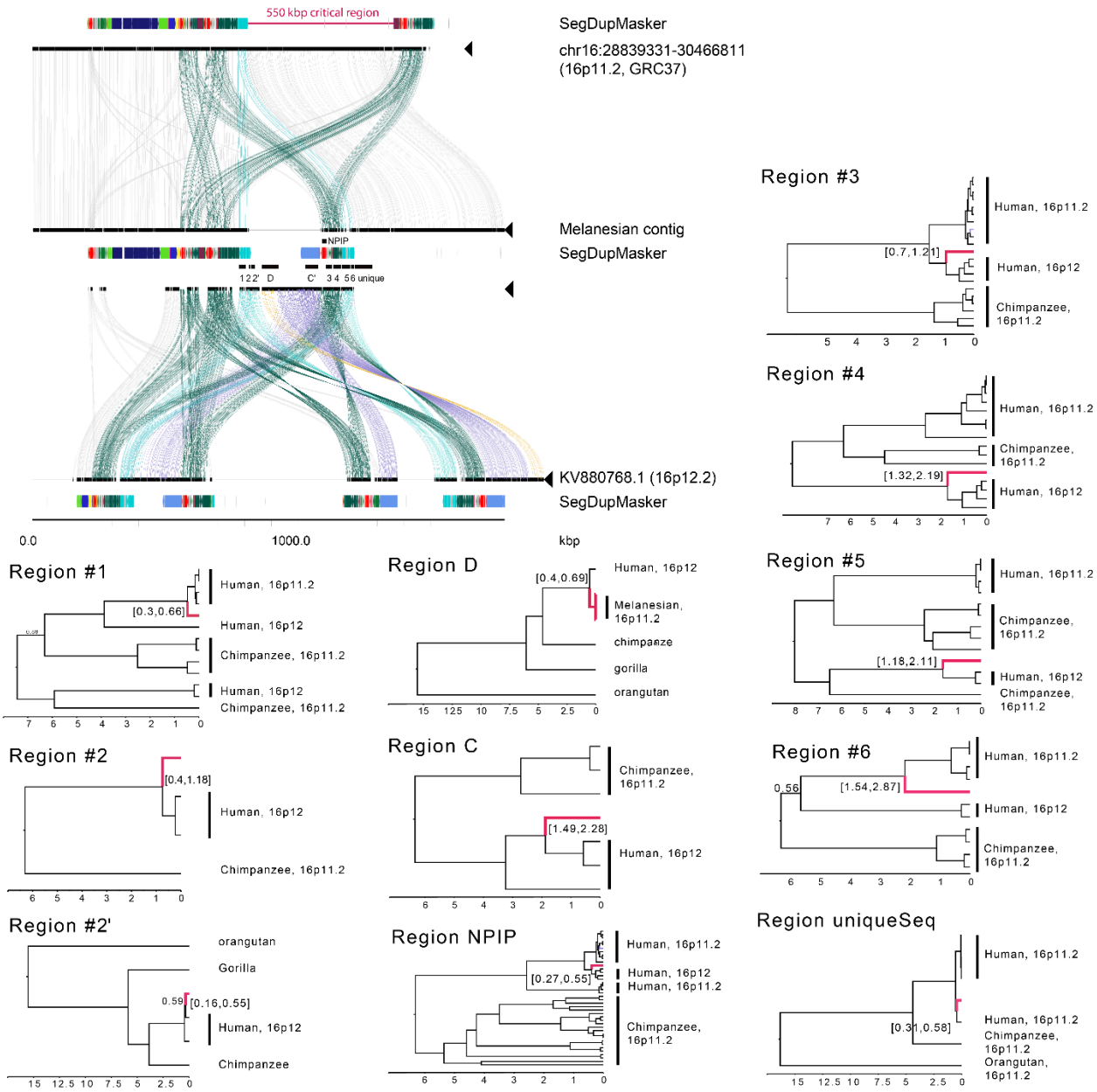


Figure S45. Phylogenetic reconstruction of the evolutionary history for DUP_{16p12}. Phylogenetic inferences were based on 11 regions sampled along the inserted DUP_{16p12} sequence at 16p11.2 (black boxes) along with homologous sequences from human (n=4), chimpanzee (n=2), and orangutan (n=2) BAC sequences for the 16p11.2 locus as well as from the 16p12.2 ancestral locus (KV880768.1). Phylogeny was inferred using BEAST (v.2.5.0) and five independent runs of 10 million iterations of Markov Chain Monte Carlo (**Methods**). Red branches indicate the lineages corresponding to the Melanesian DUP_{16p12} variant. The numbers within brackets show the 95% highest posterior density interval for the divergence (in Mya) between the Melanesian DUP_{16p12} variant and its closest related sequences (**Table S15**). Overall posterior branch supports are >99% unless otherwise labeled. Note that the homologous sequence from the 222 kbp DUP_{16p12} BAC haplotype 1 were also included in the analysis of Region D.

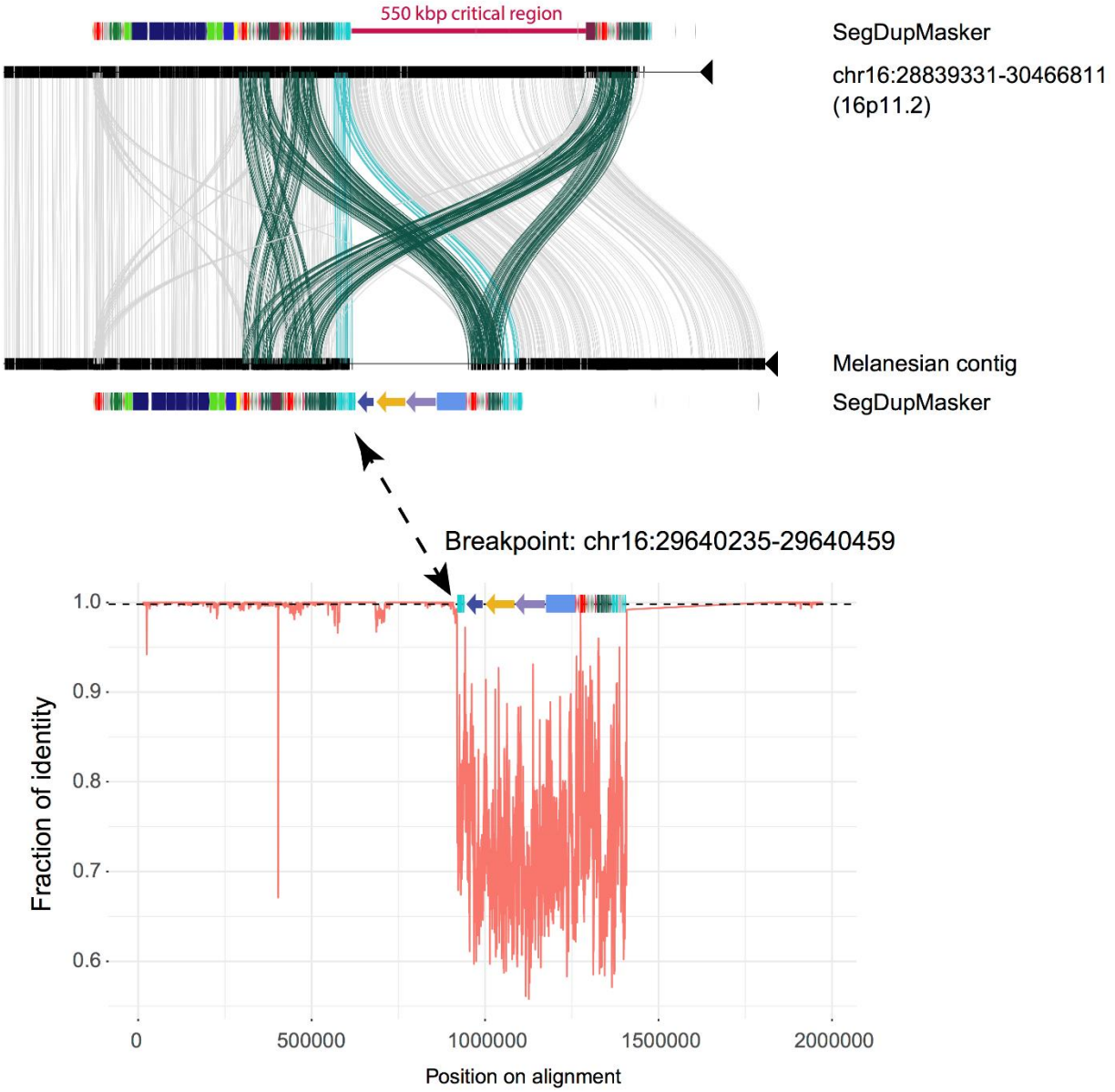


Figure S46. A 1.8 Mbp sequence-resolved Melanesian DUP_{16p12} contig. Miropeats analysis (top) shows a ~383 kbp insertion in the Melanesian contig with respect to the human reference GRCh37. Pairwise sequence alignment analysis using a window of 1000 bp and sliding by 100 bp (bottom) show the location of the insertion. We inferred the breakpoint of the insertion at chr16:29640235-29640459.

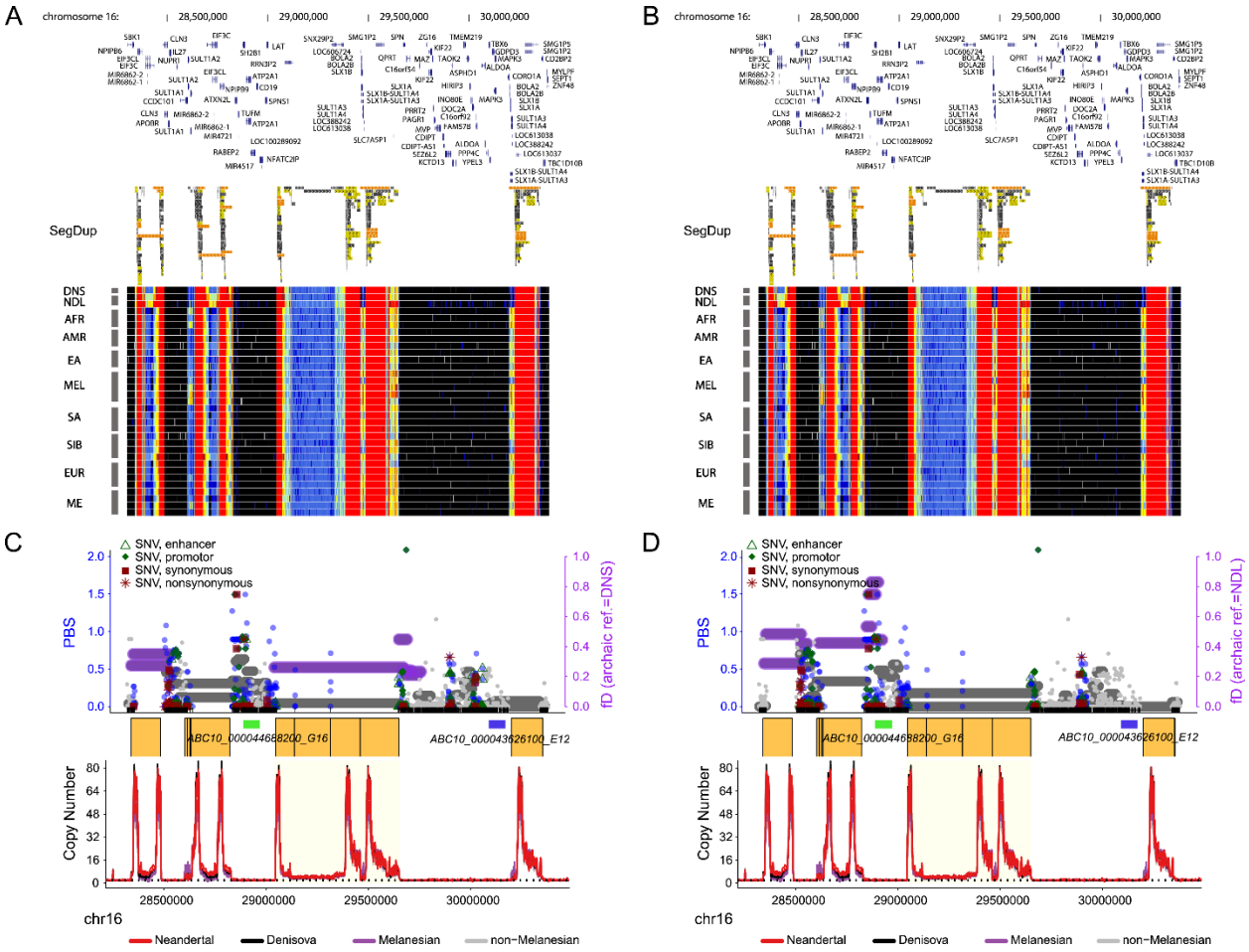


Figure S47. Adaptive introgression signals at chromosome 16p11.2, the putative location of the Melanesian-specific duplication DUP_{16p12}. (A and B) CN heat map around the chromosome 16p11.2 region. (C and D) Distributions of significant signals of both selection and archaic introgression, where top panels: distributions of *PBS* (left y-axis), functional annotation (RefSeq and ENCODE elements; **Methods**) for all SNVs (dots), the f_D (horizontal bars, representing windows of 100 SNVs). Note that the archaic references used in the f_D test in C and D are Denisovan (DNS) and Neanderthal (NDL), respectively. Colored dots (blue) and horizontal bars (purple) indicate *p*-value < 0.05. Middle panel: SDs (orange) and two fosmid probes (green and blue) used to locate the putative location of the DUP_{16p12} (**Figure S34**); bottom panel: the CN line plot, where the trajectory of each line shows the CN variation across the region for a given sample. These panels are aligned to subplots A and B above. Note that introgression signals disappear when Neanderthals were used as the archaic reference in the f_D computation.

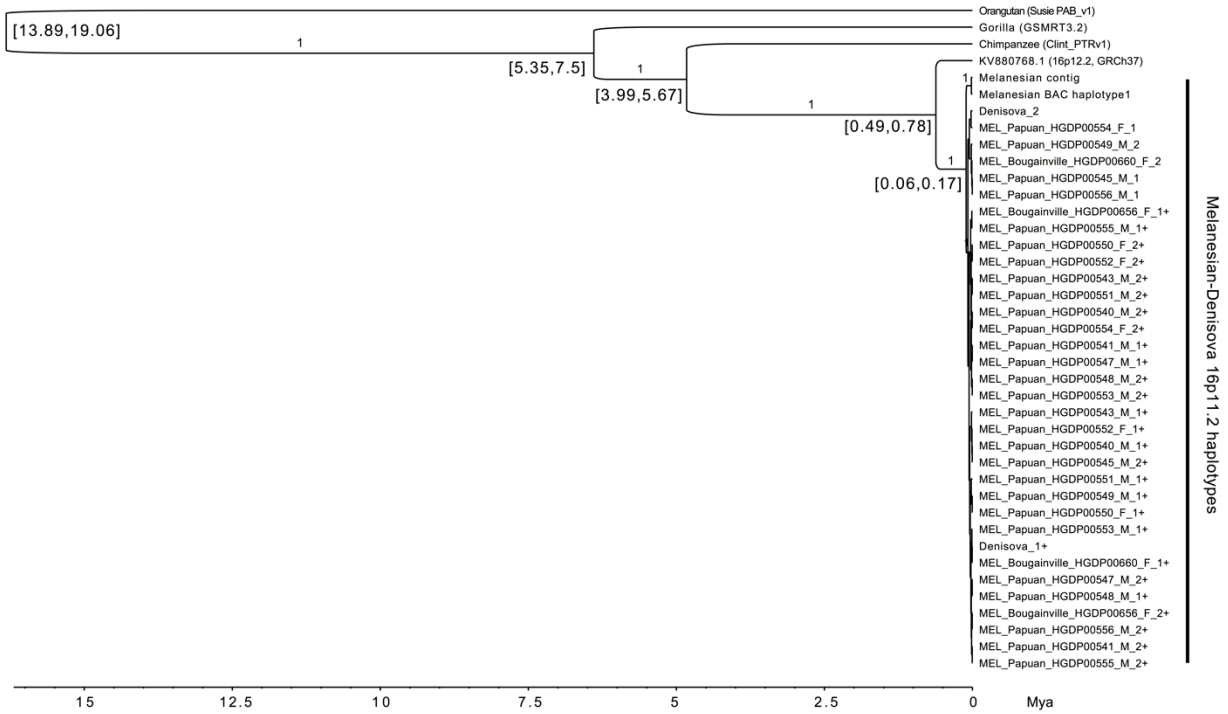


Figure S48. Evidence for supporting Denisovan introgression of the DUP_{16p12} duplication polymorphism in Melanesians. The Melanesian–Denisovan 16p11.2 haplotypes were based on variants from ~9 kbp, where >5 Denisovan reads were confidently aligned at positions between 865,000 and 927,000 on the assembled Melanesian contig. Eight SNVs were called using FreeBayes (v1.0.2) and Illumina short-read data that were confidently mapped to the assembled Melanesian 16p11.2 contig. BEAGLE (v4.1) was used to phase the eight SNVs along with a bi-allelic duplication variant for the Melanesian–Denisovan DUP_{16p12} duplication polymorphism. Sequences for the nonhuman great apes were based on the mapping of sequences with Denisovan coverage to published assemblies (*53*). BEAST (v2.5.0) was used to infer the phylogeny, the 95% high-probability density for the divergence (brackets), and the posterior supports (the numbers above the branches). A plus (+) sign at a given sample ID indicates the presence of the DUP_{16p12} duplication allele.

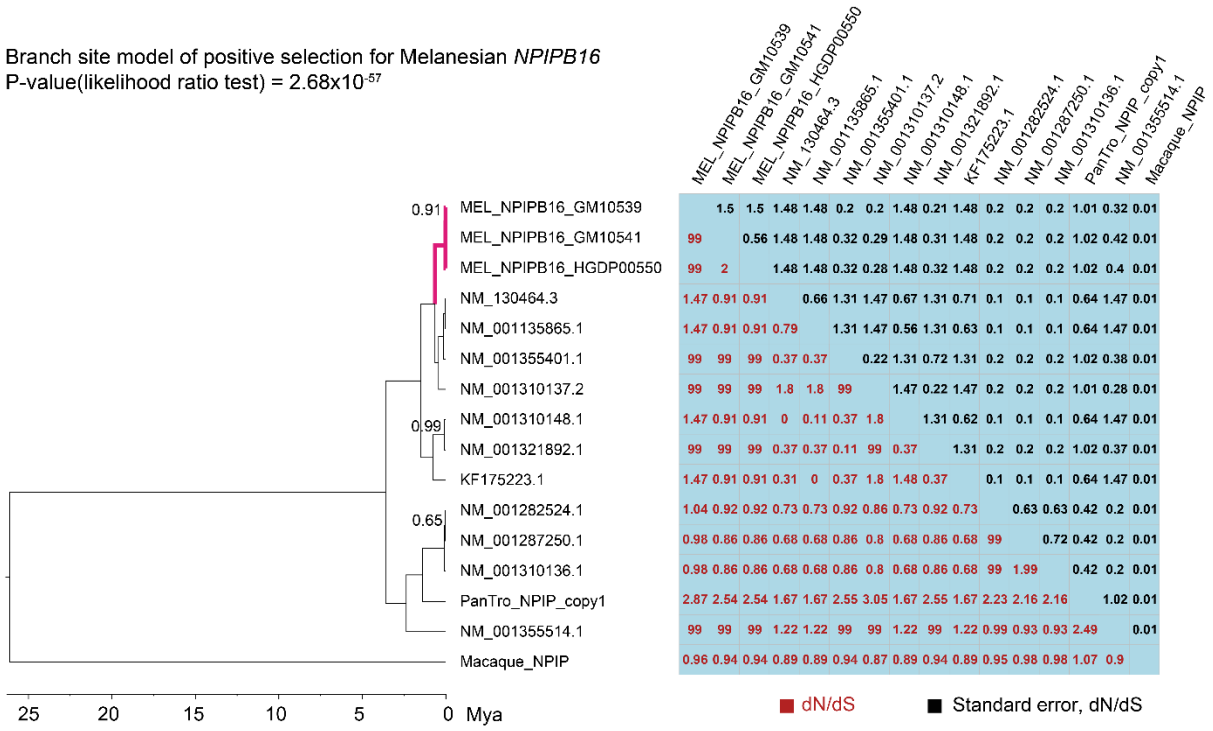


Figure S49. Phylogenetic analyses reveals positively selected amino acid substitutions at the Melanesian *NPIPB16* lineage. Two Melanesian *NPIPB16* FLNC transcripts (GM10539 and GM10541) were generated using PacBio IsoSeq technology. A third *NPIPB16* transcript was identified based on mapping in the assembled Melanesian contig (HGDP00550). Human *NPIPB* and *rhesus macaque* (XM_015125675.1) gene sequences were downloaded from RefSeq (release 109), while the chimpanzee (PanTro) *NPIP* copy was based on the chimpanzee BAC contigs from (*15*). Branch site tests of positive selection and pairwise *dN/dS* ratios were computed using PAML (v14.9). Standard errors of pairwise *dN/dS* were calculated using 1,000 bootstrap samples. Significance test of the branch site model was based on a chi-squared likelihood ratio test (d.f. = 1) against the null model of neutral evolution. Phylogeny of these *NPIPB* sequences was inferred using BEAST (v2.5.0). Note that the posterior probabilities for branch supports are all equal to 1.0, except for those shown.

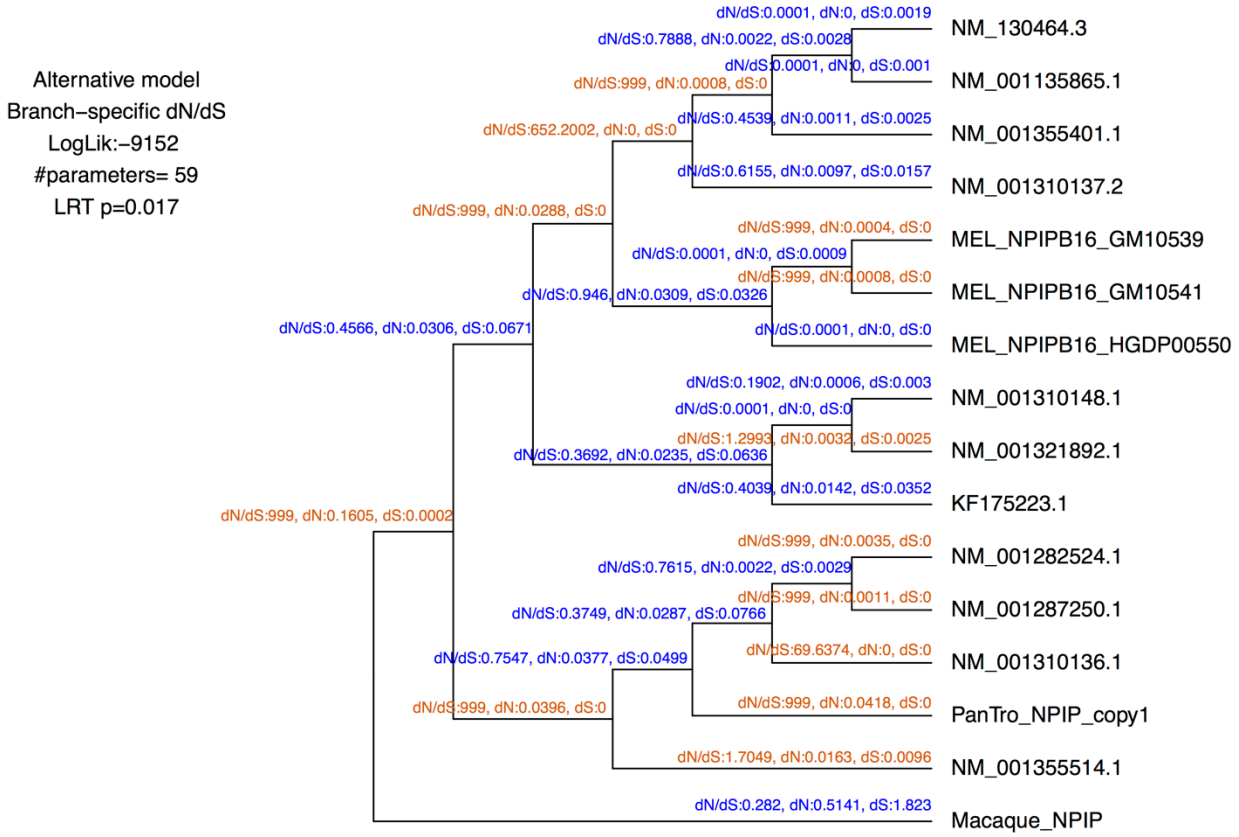
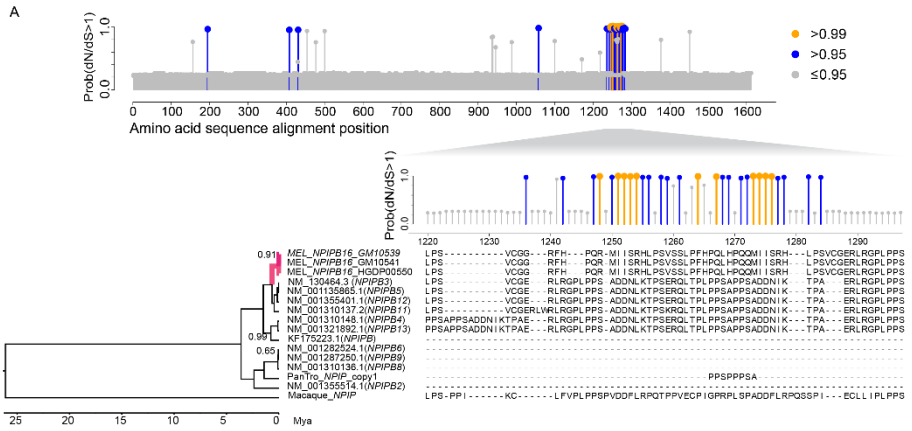
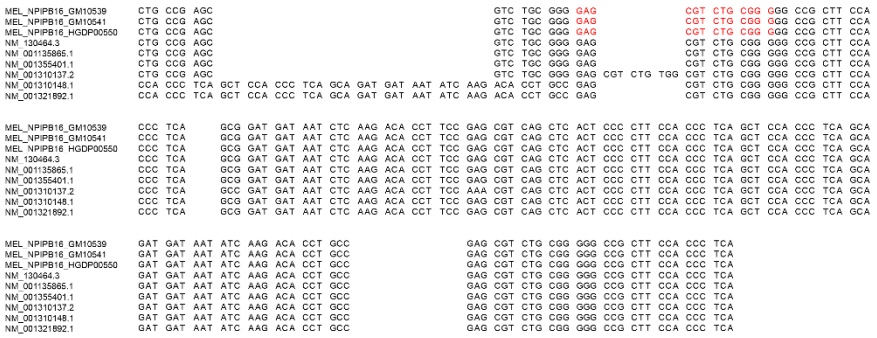


Figure S50. Cladogram of the *NPIPB* lineages showing *dN* and *dS* values and *dN/dS* ratios for individual branches. The same *NPIPB* sequences as in **Figure S49** were used to estimate the branch-specific *dN/dS*, *dN*, and *dS* values, shown above each branch, using PAML (v14.9). Orange and blue colors indicate if the test of *dN/dS* ratio ≥1 is significant ($p < 0.05$) or not ($p \geq 0.05$), respectively. A significance test of the free *dN/dS* ratios model was based on a chi-squared likelihood ratio test (d.f. = 1) against the null model of neutral evolution (*dN/dS* = 1). The phylogeny of these *NPIPB* sequences (**Figure S49**, left panel) was inferred using BEAST (v2.5.0). Note that PAML reports *dN/dS* = 99 or 999 for a branch when no synonymous mutation was inferred (*dS* = 0) along the lineage.



B. Hypothesized codon sequences among NPIPB16 and its close relatives before two indel events



In *NPIPB16*, the upstream 13-bp indel of GAGCGTCTGCGGG (red) results in frame-shift (purple), while the same indel allele downstream (blue) restores the original frame (black).

C. Observed codon sequences among NPIPB16 and its close relatives after the two indel events

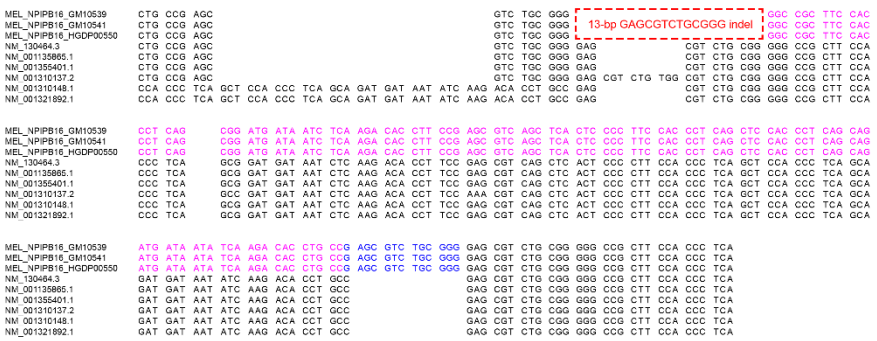


Figure S51. A unique peptide sequence structure in *NPIPB16* compared with its close relatives likely resulted from two indel events involving the same 13 bp repeat motif. (**A**) *dN/dS* analysis reveals a cluster of amino acid substitutions at position 1236-1284 (alignment space). Codon sequences corresponding to the cluster are shown in panels B and C. (**B**) The hypothesized codon sequences of *NPIPB16* prior to the two indel events. The highlighted (red) 13 bp repeat motif, GAGCGTCTGCGGG, appears in all sequences presented here. (**C**) The upstream indel (red) causes the *NPIPB16* codon sequences be out-of-frame, while the other indel (blue) downstream restores the original frame (black). Note that for the sake of simplicity, the resulting codon sequences were not realigned.

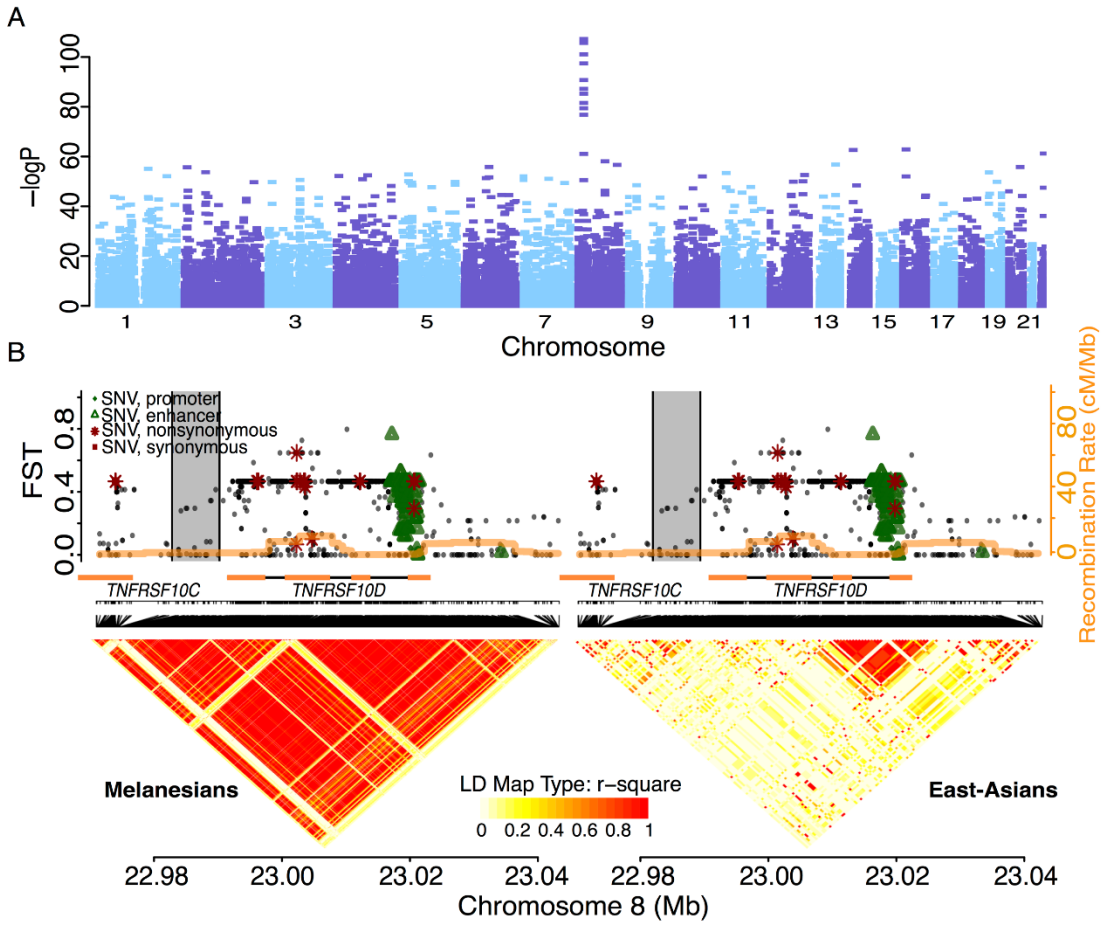


Figure S52. The strongest signal of selection in Melanesians intersecting two highly stratified deletion and duplication variants at 8p21.3. (**A**) The Manhattan plot of Bonferroni *p*-values of the window-based F_{ST} test (**Methods**). (**B**) Distribution of F_{ST}, functional annotation (RefSeq and ENCODE elements; **Methods**), and recombination rate (HapMap) for all variants (dots). Genes are shown under the plot with black lines (noncoding sequences) and orange boxes (exons). Bottom panels show the patterns of linkage disequilibrium, measured in r-square, across the locus.

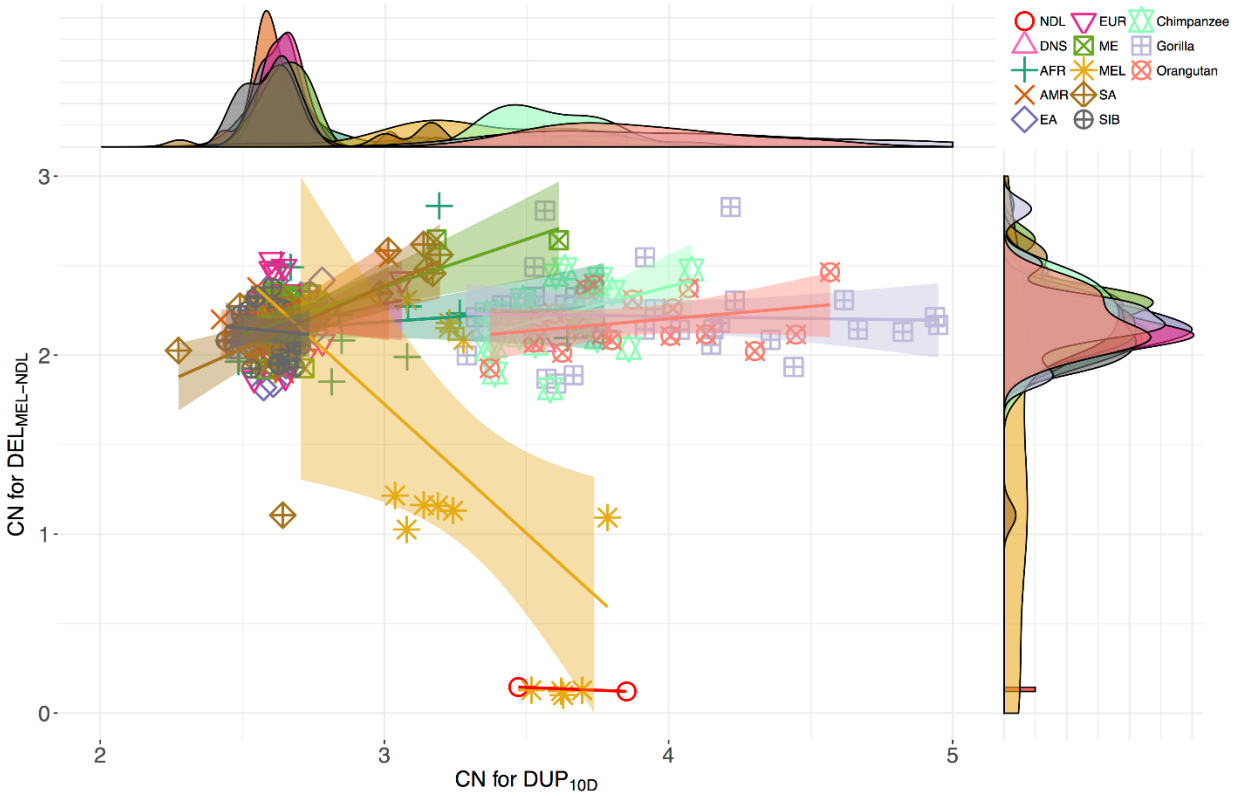


Figure S53. The joint distribution of CN estimates for $DEL_{MEL\text{-}NDL}$ and DUP_{10D} across great ape species. Linear regression lines and 95% C.I. were drawn for individual populations.

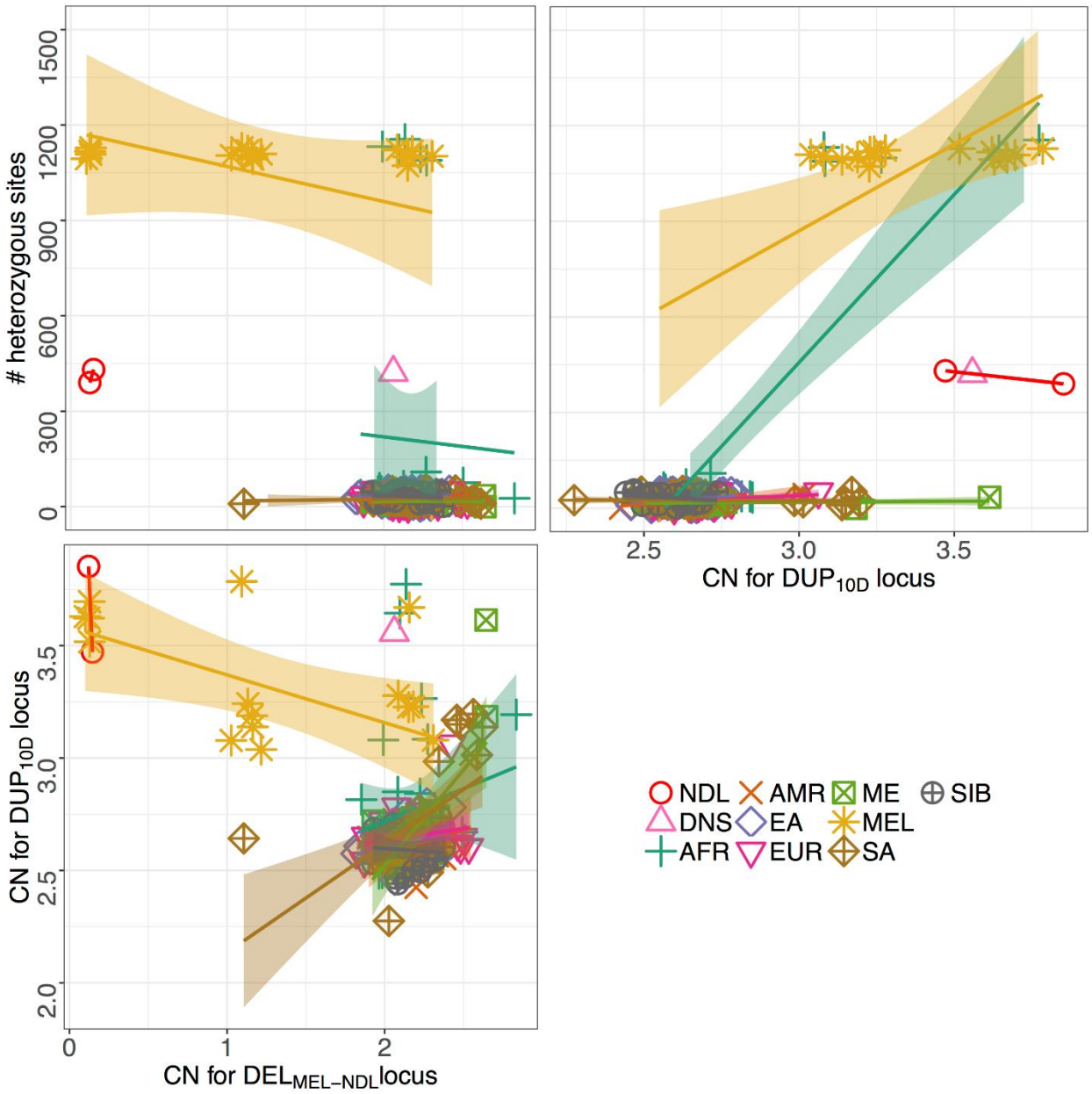


Figure S54. Pairwise joint distributions among CN estimates of $DEL_{MEL\text{-}NDL}$, DUP_{10D}, and the number of heterozygous sites of the DUP_{10D} locus for all SGDP and the three archaic samples. Each symbol is the data from an individual. Linear regression lines and their 95% C.I. region were drawn for individual populations.

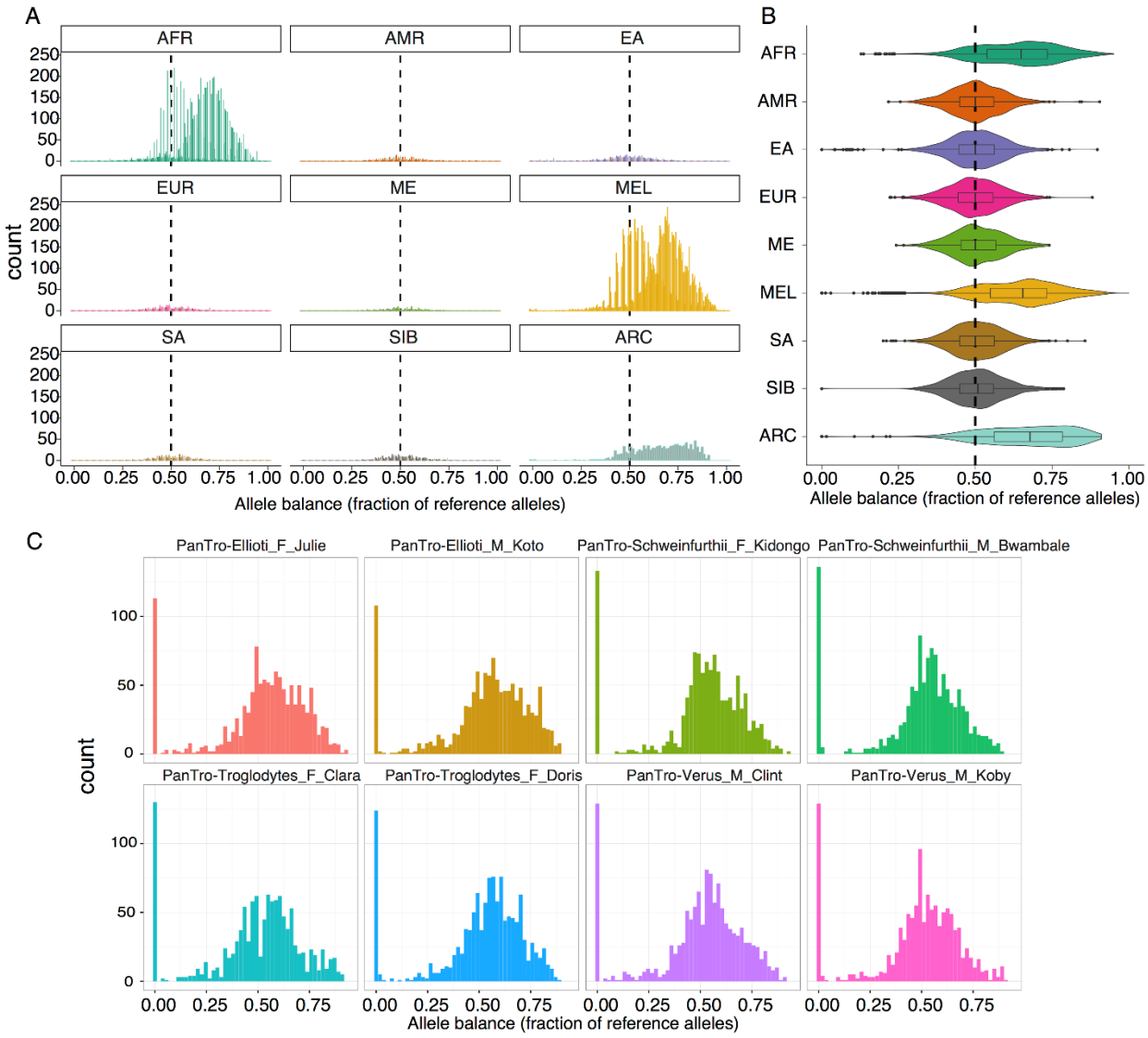


Figure S55. Observations of allele imbalance in individuals carrying the DUP_{10D} duplication variant: AFR (7/33), MEL (15/16), the archaic hominin (ARC, 3/3), and chimpanzee (8/8) samples. Distributions of allele balance for heterozygous sites at the DUP_{10D} locus were shown for individual SGDP populations (**A**) and chimpanzee genomes (**C**). For the eight chimpanzee genomes (Great Ape Project, 2013), reads were mapped to the human reference (GRCh37) using BWA-MEM (v0.7.12) and SNVs at the DUP_{10D} locus were called using FreeBayes (v1.0.2). (**B**) Violin plots of allele balance across the eight SGDP super-populations, along with the ARC samples, show a clear pattern of allele imbalance in populations, where the DUP_{10D} variants are present.

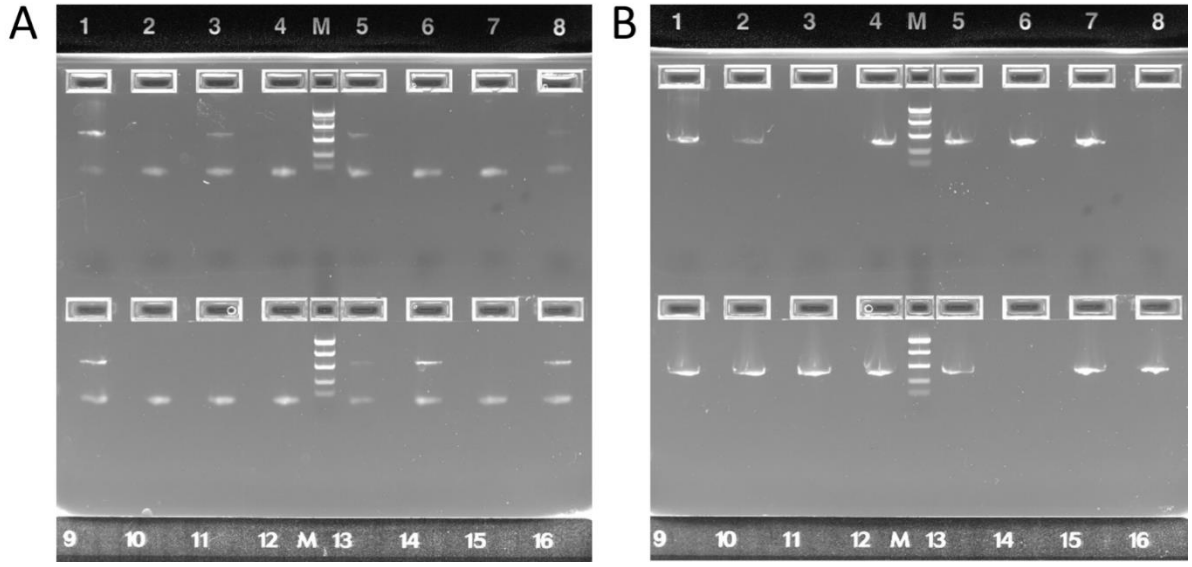


Figure S56. PCR validation of the $DEL_{MEL\text{-}NDL}$ variant using 16 randomly selected blood-derived Melanesian DNA samples. Ladder is at 2000 bp, 800 bp, 400 bp, 200 bp, and 100 bp. (**A**) Gel of PCR product from the first PCR assay showing the 501 bp product that is produced when the deletion is present (~50 bp band is primer dimer). Samples 1, 3, 5, 8, 9, 13, 14, and 16 all have at least one chromosome where the deletion is present. (**B**) Gel of PCR product from the second assay that amplifies a 300 bp fragment within the deletion region. Samples 3, 8, and 14 do not have product in this region and therefore are homozygous for the deletion (CN0). Samples 1, 5, 9, 13, and 16 have at least one copy of this region but also have the deletion band seen in A, so therefore are heterozygous for the deletion (CN1). Samples 2, 4, 6, 7, 10, 11, 12, and 15 do not have the deletion (CN2) because they did not have the deletion band (in **A**) and do have PCR product for the fragment within the deleted region.

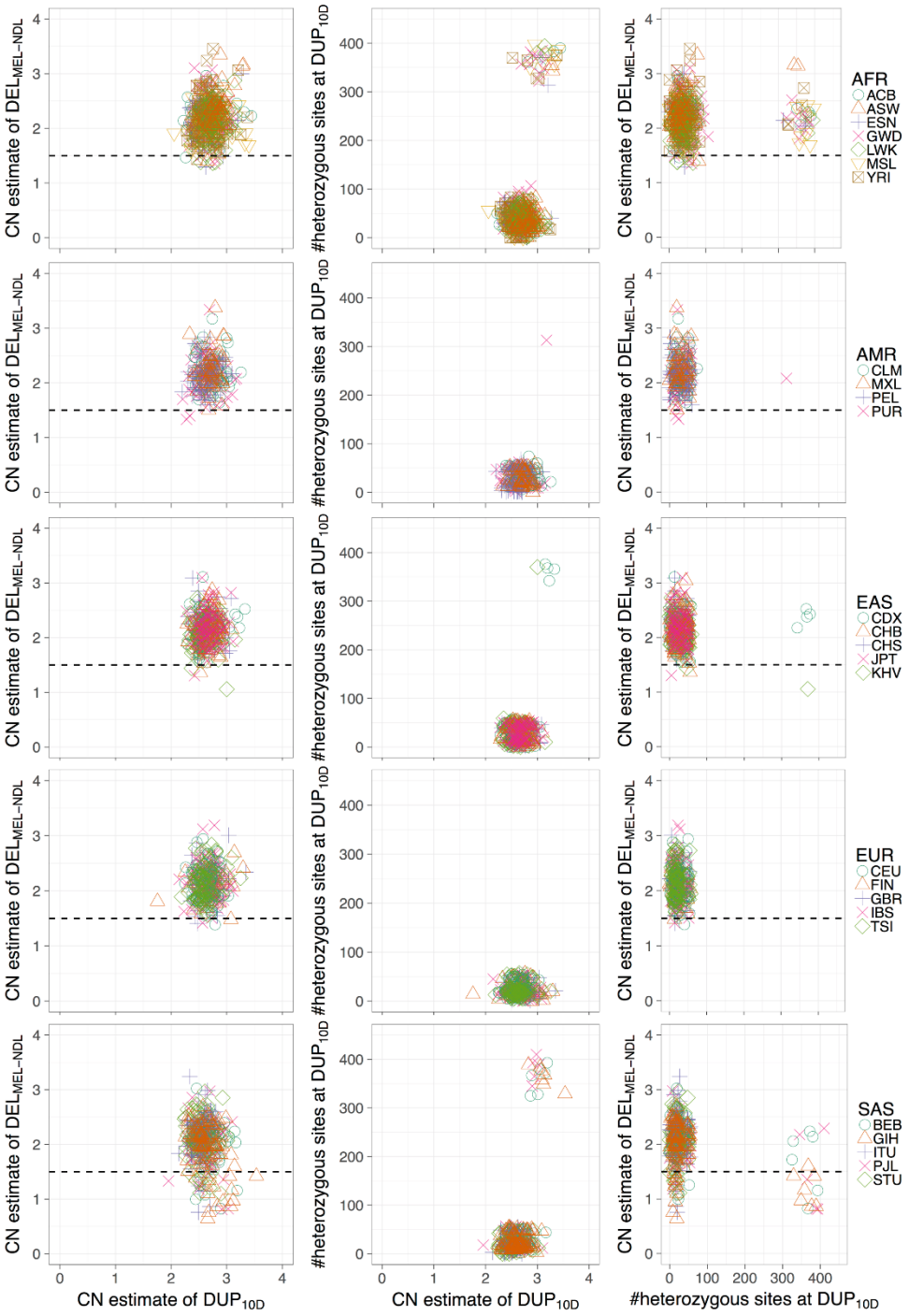


Figure S57. Patterns of heterozygosity, along with the CN estimates for both DUP$_{10D}$ and DEL$_{MEL-NDL}$, in all 1KG populations. Each symbol represents the estimated quantity for an individual.

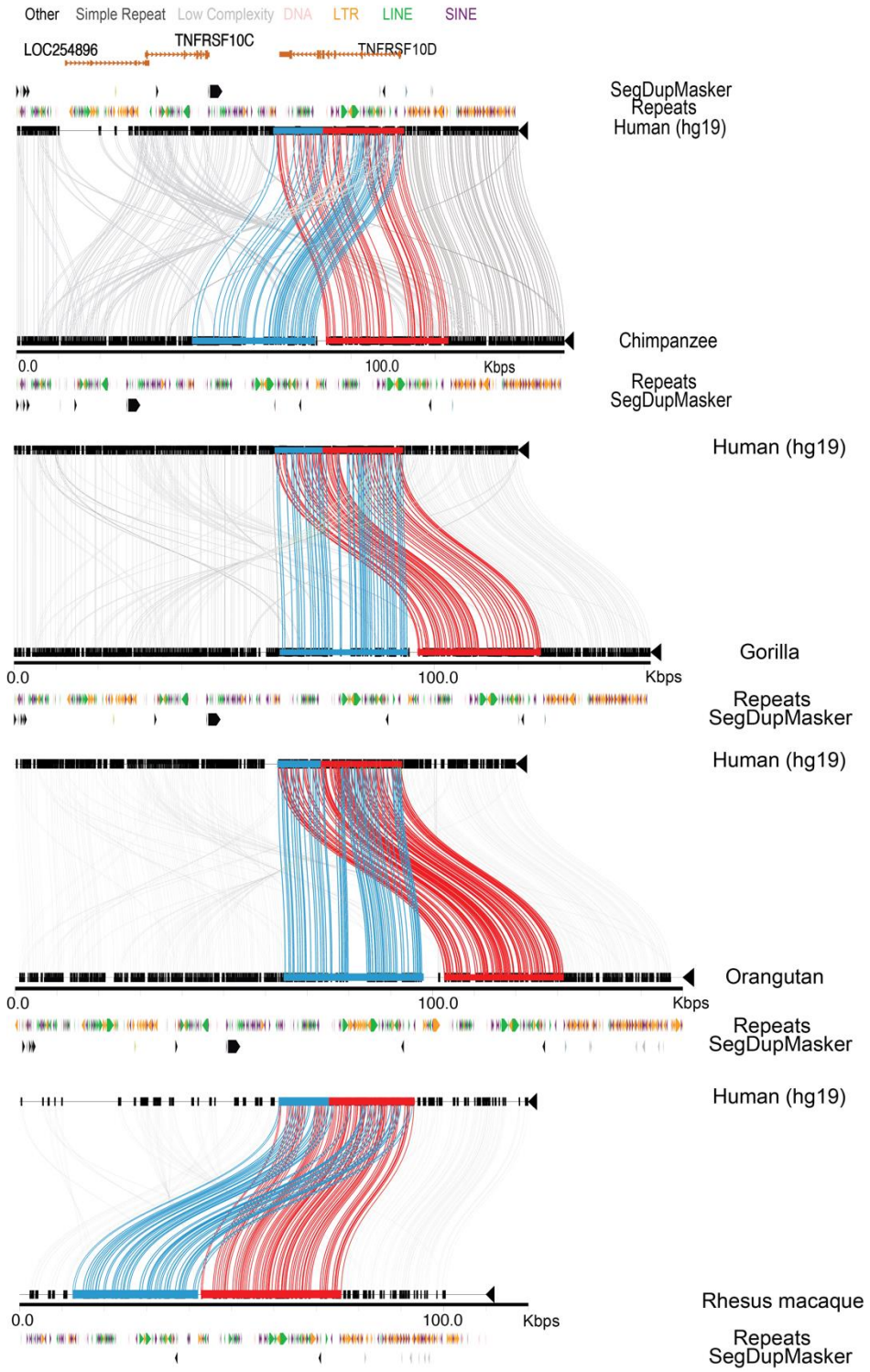


Figure S58. Miropeats reveals the tandem organization of *TNFRSF10D* duplications in nonhuman primates. High-quality homologous sequences to human *TNFRSF10D* (GRCh37) for four nonhuman primates were generated using BAC libraries and sequenced using PacBio technology. Blue and red traces aligned sequences between human reference and nonhuman primate *TNFRSF10D1* and *TNFRSF10D2*, respectively. Shown below the bottom track in each nonhuman primate are the repeat elements annotated using RepeatMasker (v3.3.0) as well as SDs.

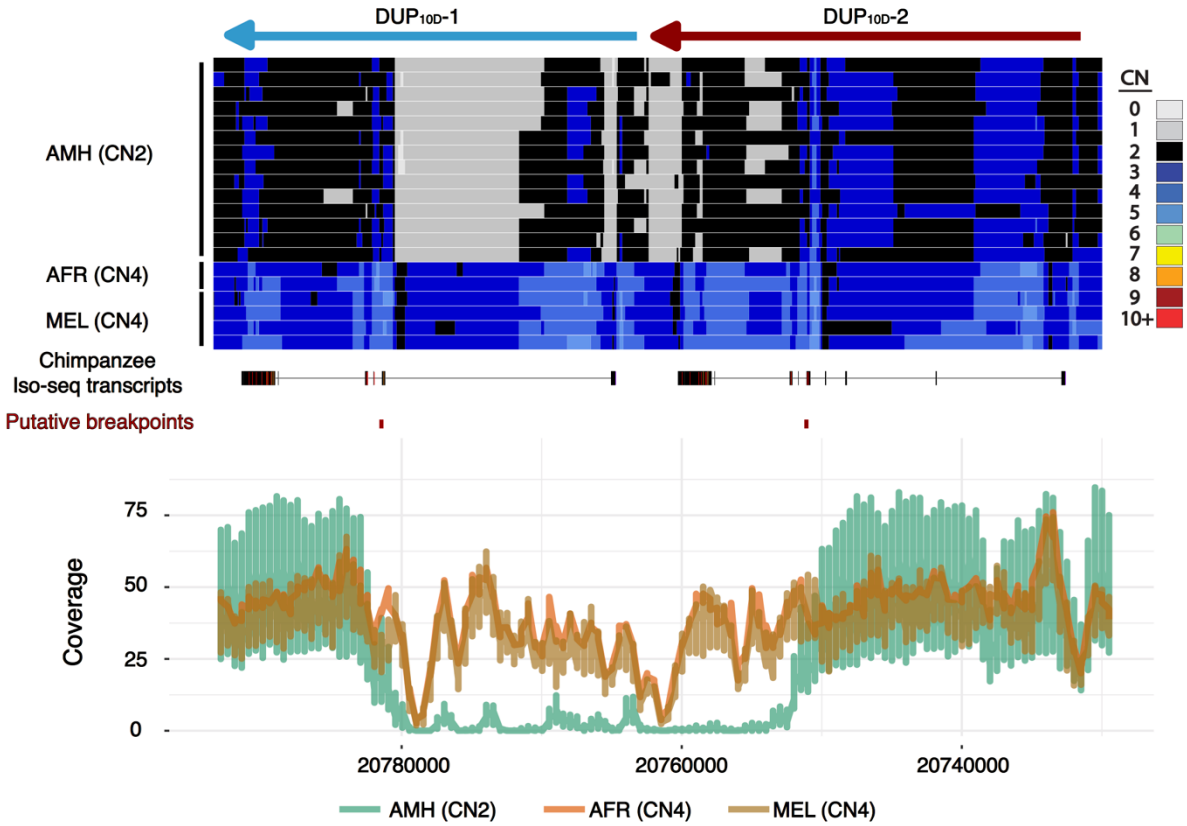


Figure S59. Organization of duplications at the *TNFRSF10D* locus in chimpanzee. *TNFRSF10D1* and *TNFRSF10D2* are in the same orientation as the human reference (GRCh37). Top panel: CN heat maps for a subset of CN2 modern human samples, followed by AFR and MEL duplication carriers. Gene models of *TNFRSF10D1* and *TNFRSF10D2* were inferred using PacBio IsoSeq technology (*53*). The breakpoints of DUP_{10D} in the chimpanzee reference sequence were inferred based on read-depth profile (bottom panel) and a hidden Markov Model utilizing a sequence alignment among human reference and the two chimpanzee *TNFRSF10D1* and *TNFRSF10D2* sequences (**Methods**). In the analysis of read-depth profile, reads of all SGDP samples were mapped to a recently published high-quality chimpanzee assembly (Kronenberg et al., 2018). Read-depth profile of a sample was estimated using 1000 bases with a step size of 500 bases sliding across the region. Each color bar represents the depth range of a given window.

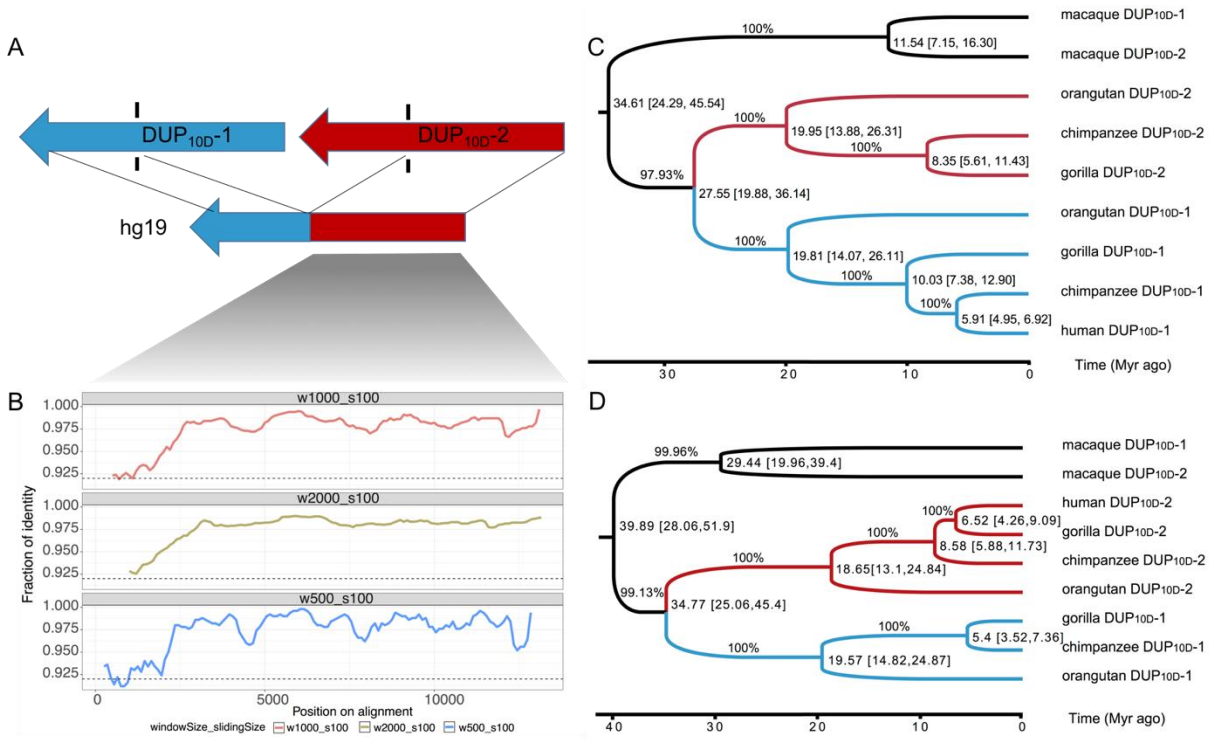


Figure S60. Phylogenetic analyses of the DUP_{10D} duplication sequences among primate species. Long-read BAC sequences of the DUP_{10D} locus for three great ape lineages (chimpanzee, gorilla, and orangutan) and one Old World rhesus macaque monkey were generated using BAC libraries. (**A**) Schematic of *TNFRSF10D* gene fusion in modern humans with respect to the putative ancestral tandem duplication form in other primates. (**B**) Compelling evidence for interlocus gene conversion between the rear (gray shaded) portions of the *TNFRSF10D1* and *TNFRSF10D2* sequences in orangutan indicated by high sequence identity. Fractions of sequence identity were calculated using three window sizes (1000, 2000, and 500 bp) and a sliding of 100 bp across the sequence alignment. Sharp increases in sequence identity, compared with the mean identity of 0.921 between the two sequences in orangutan, especially beyond the position 2,500, are consistent with the hypothesis of interlocus gene conversion. (**C and D**) Evolutionary history of DUP_{10D} sequences in primates inferred by Bayesian phylogenetic trees (BEAST v.2.5.0; **Methods**). Two trees were built separately using homologous sequences to the human reference (GRCh37) *TNFRSF10D1* and *TNFRSF10D2* sequences (blue and red portions in A). Numbers at the nodes are the divergence estimates with 95% high posterior density intervals, while the percentages on branches indicate the posterior probabilities supporting the branches. Note that a gene tree–species tree discordance within the *TNFRSF10D2* Homininae phylogeny likely arose as a result of incomplete lineage sorting.

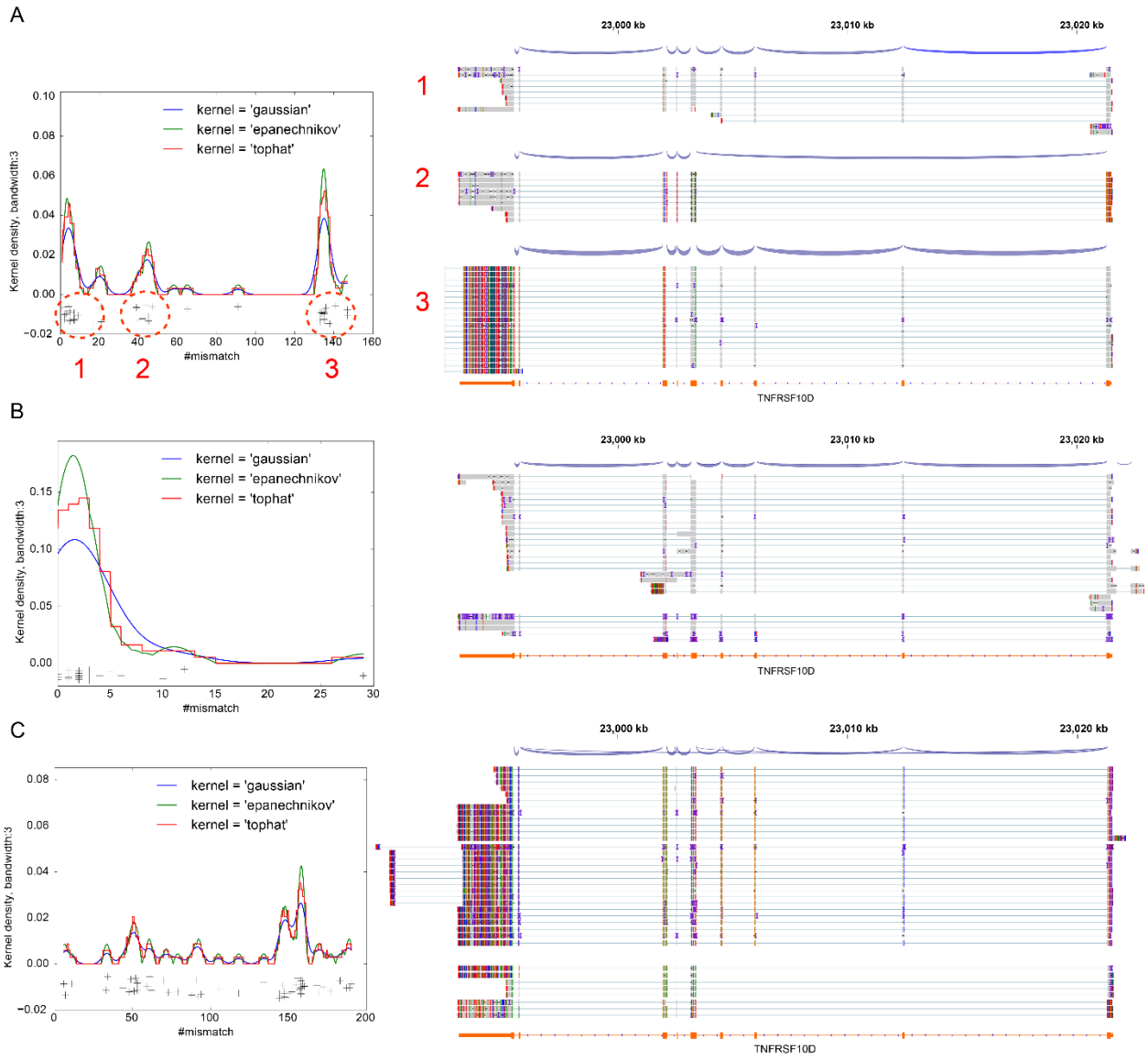


Figure S61. FLNC transcripts of *TNFRSF10D* locus for Melanesian (GM10541, CN3), European (GM12878, CN2), and chimpanzee (PanTro, CN4) fibroblast cell line samples. Full-length transcripts were generated using the PacBio long-read sequencing technology. The FLNC transcripts were mapped to human reference (GRCh37) using minimap2 (v2.1), and the number of mismatches for each transcript were computed against the human reference sequence. Left panels show the kernel density distributions of mismatches for the transcripts, while right panels are subsets of FLNC transcript reads mapped to the *TNFRSF10D* locus in GRCh37. Vertical color bars on each transcript indicate mismatches. (**A**) All three types of *TNFRSF10D* transcripts, including *TNFRSF10D1* (6 exons, cluster 2), *TNFRSF10D2* (9 exons, cluster 3), and the fusion gene *TNFRSF10D* (9 exons, cluster 1), are clearly present in the CN3 Melanesian sample. (**B**) In the CN2 European sample GM12878, all transcripts present 9 exons and low numbers of mismatches, suggesting that they are all the version of fusion gene *TNFRSF10D*. (**C**) Both 6-exon and 9-exon transcripts are present in the chimpanzee sample with high numbers of mismatches for the majority of the reads. Note that we did not observe clear clusters of these transcripts in the chimpanzee sample, likely due to the divergence between human and chimpanzee at this locus.

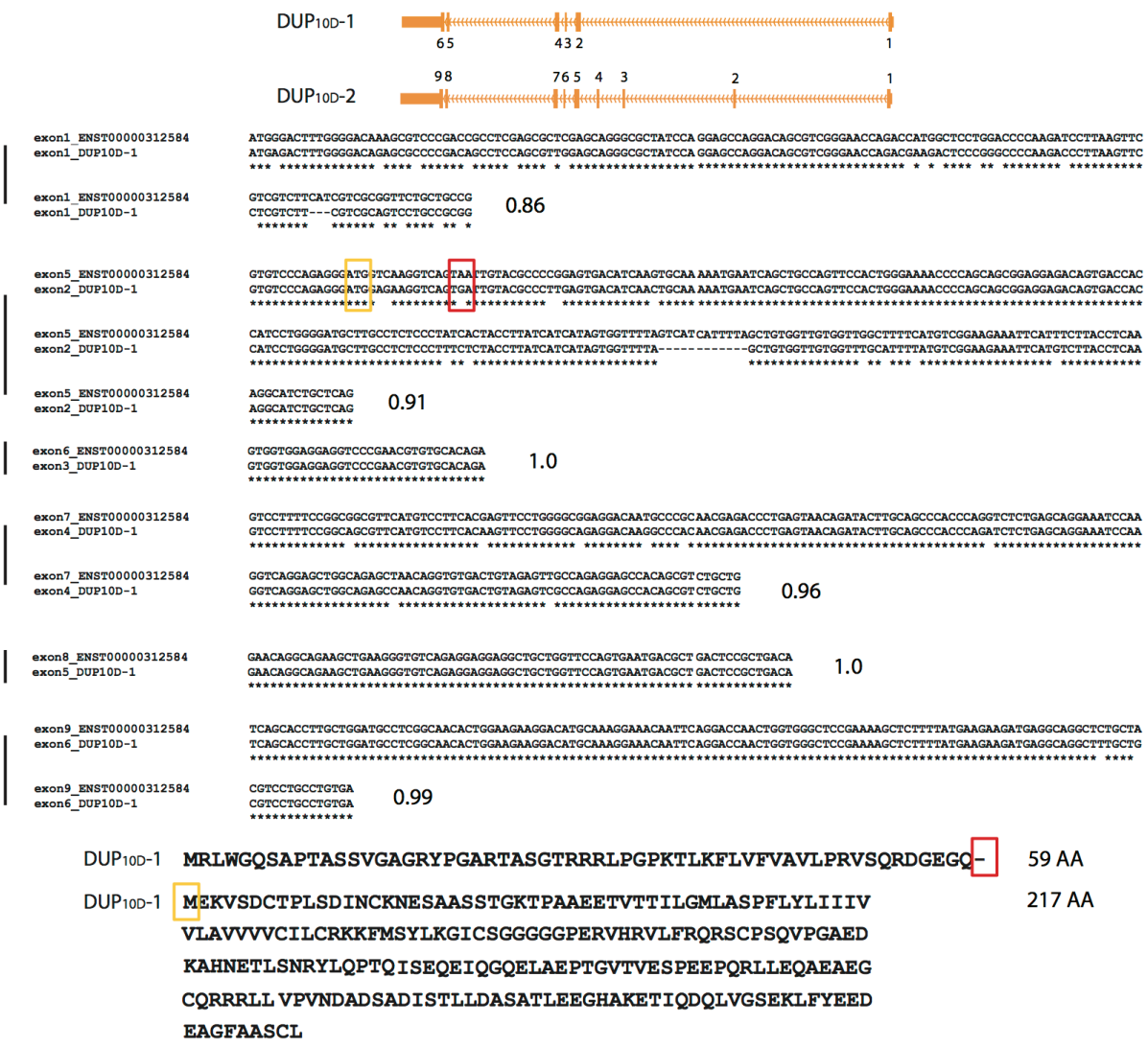


Figure S62. Premature stop codon in the chimpanzee *TNFRSF10D1* gene copy. Gene models (top) of the two copies of *TNFRSF10D* (*TNFRSF10D1* and *TNFRSF10D2*) in chimpanzee were inferred using PacBio Iso-Seq data from a chimpanzee sample and the resulting sequences were mapped to the corresponding chimpanzee assembly (Kronenberg et al., 2018). Middle panel shows the exon sequence alignments between human reference *TNFRSF10D* and chimpanzee *TNFRSF10D1*. The numbers after each alignment are the proportion of sequence identity (stars). *TNFRSF10D1* was truncated after 59 amino acids due to a stop codon in its second exon (highlighted in the red box) but is likely translated to a protein with 217 amino acids using a second start codon (highlighted in the orange box) in a different frame, upstream of the stop codon in exon 2.

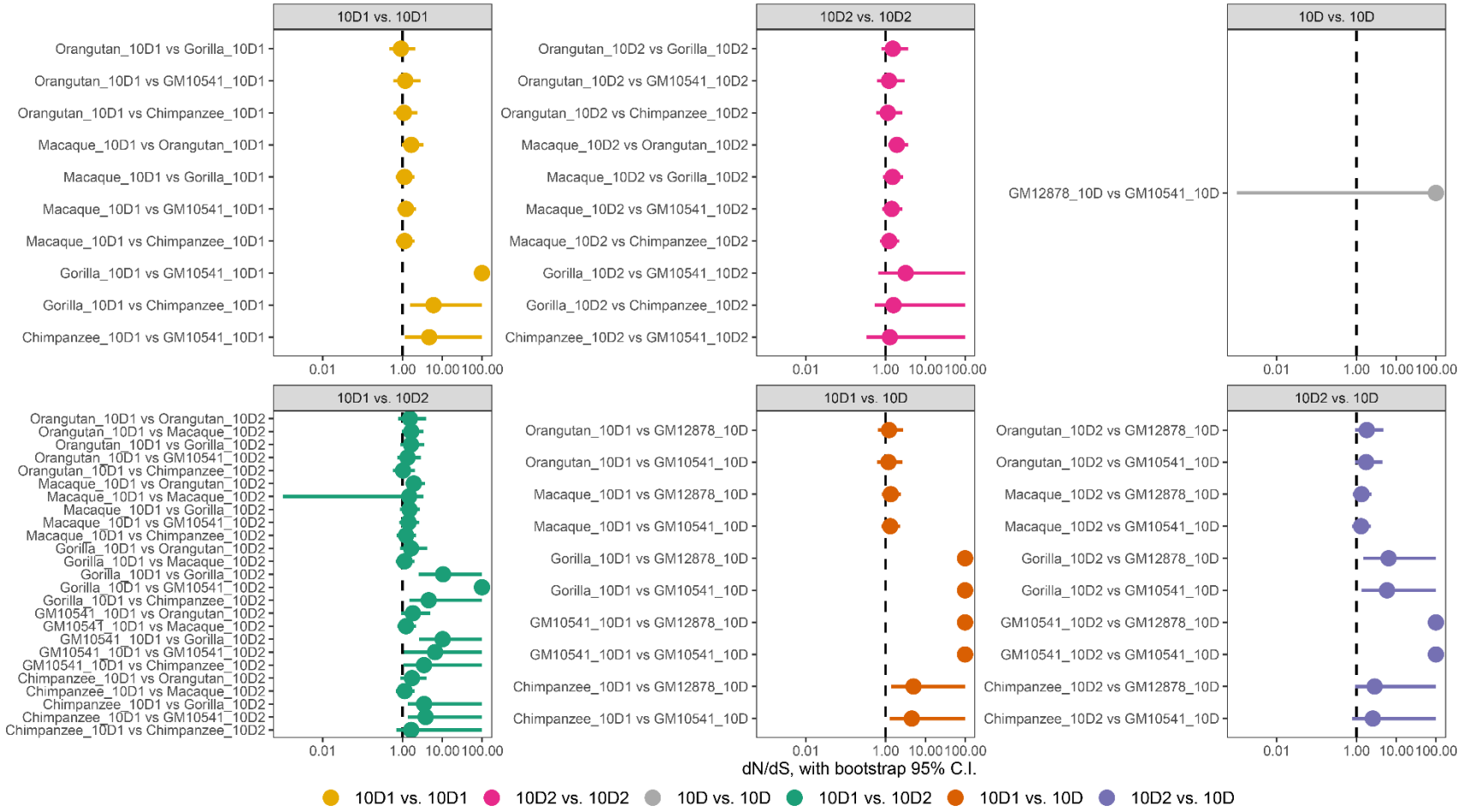
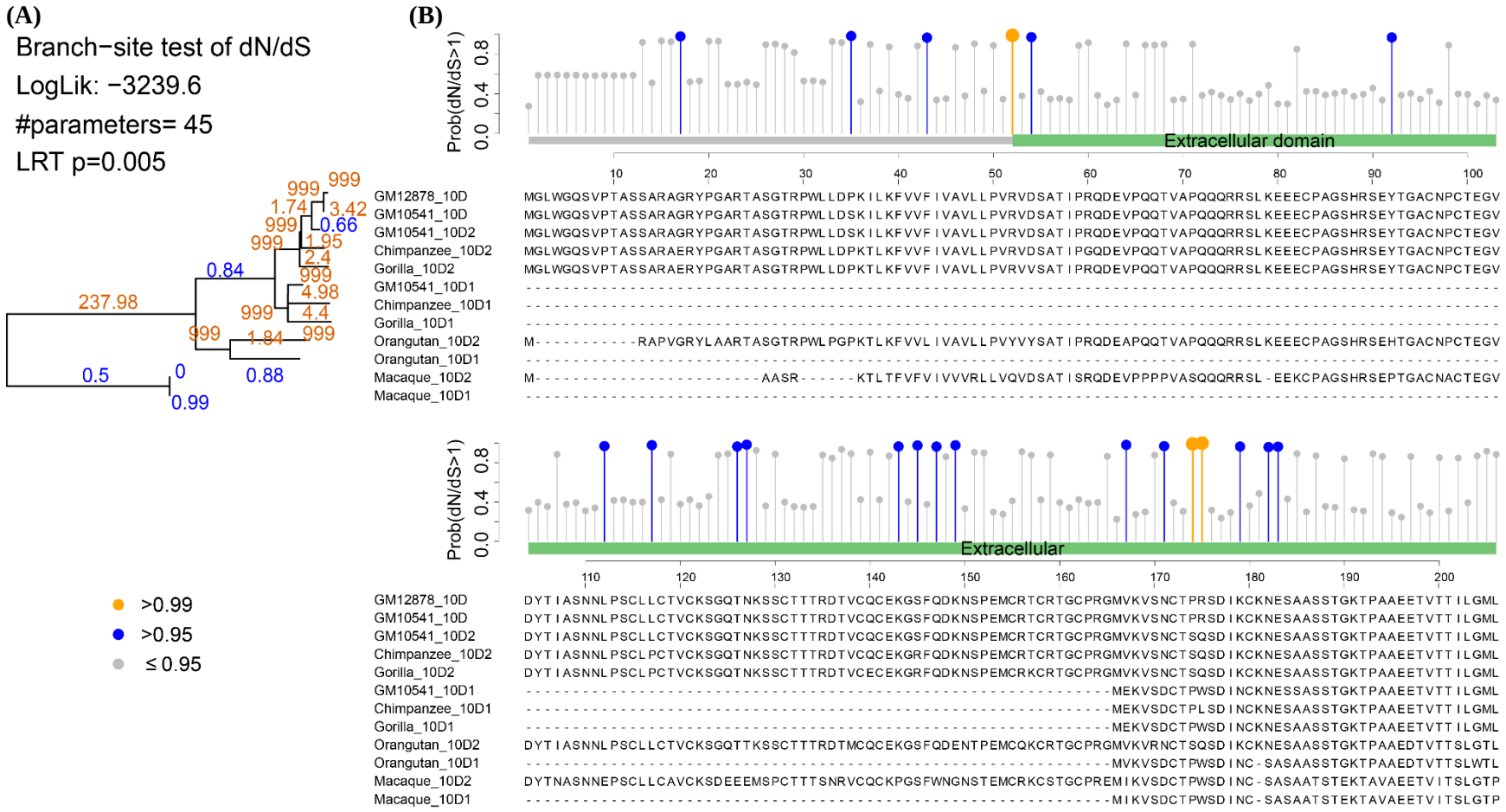


Figure S63. Pairwise *dN/dS* among the ORF sequences of twelve FLNC transcripts in primates. FLNC transcripts of *TNFRSF10D*, *TNFRSF10D*1, and *TNFRSF10D2* were generated from Melanesian (GM10541, CN3), European (GM12878, CN2), and chimpanzee (PanTro, CN4) fibroblast cell lines. Orthologous transcript sequences in gorilla, orangutan, and rhesus macaque were inferred from their BAC sequences. Predicted ORFs were defined as the longest ORF in all frames of individual transcripts. *dN/dS* ratios were estimated using the codeml program in the PAML package (v14.9). 1,000 bootstrap samples of the multiple codon sequence alignment were used to estimate the 95% C.I. Note that PAML reports *dN/dS* = 99 for a branch when no synonymous mutation was inferred (*dS* = 0) along the lineage.


(A)
Branch−site test of dN/dS
LogLik: −3239.6
#parameters= 45
LRT p=0.005
237.98
0.84
0.5
0
0.99
0.88
999
999
1.74
3.42
999
0.66
999
1.95
2.4
999
4.98
999
4.4
999
1.84
999
>0.99
>0.95
≤ 0.95
(B)
Prob(dN/dS>1)
0.0
0.4
0.8
Extracellular domain
10
20
30
40
50
60
70
80
90
100
GM12878_10D
GM10541_10D
GM10541_10D2
Chimpanzee_10D2
Gorilla_10D2
GM10541_10D1
Chimpanzee_10D1
Gorilla_10D1
Orangutan_10D2
Orangutan_10D1
Macaque_10D2
Macaque_10D1
MGLWGQSVPTASSARAGRYPGARTASGTRPWLLDPKILKFVVFIVAVLLPVRVDSATIPRQDEVPQQTVAPQQQRRSLKEEECPAGSHRSEYTGACNPCTEGV
MGLWGQSVPTASSARAGRYPGARTASGTRPWLLDSKILKFVVFIVAVLLPVRVDSATIPRQDEVPQQTVAPQQQRRSLKEEECPAGSHRSEYTGACNPCTEGV
MGLWGQSVPTASSARAGRYPGARTASGTRPWLLDSKILKFVVFIVAVLLPVRVDSATIPRQDEVPQQTVAPQQQRRSLKEEECPAGSHRSEYTGACNPCTEGV
MGLWGQSVPTASSARAERYPGARTASGTRPWLLDPKTLKFVVFIVAVLLPVRVDSATIPGQDEVPQQTVAPQQQRRSLKEEECPAGSHRSEYTGACNPCTEGV
MGLWGQSVPTASSARAERYPGARTASGTRPWLLDPKTLKFVVFIVAVLLPVRVVSATIPRQDEVPQQTVAPQQQRRSLKEEECPAGSHRSEYTGACNPCTEGV
M----------RAPVGRYLAARTASGTRPWLPGPKTLKFVVLIVAVLLPVYVYSATIPRQDEAPQQTVAPQQQRRSLKEEECPAGSHRSEHTGACNPCTEGV
M----------AASR------KTLTFVFVIVVVRLLVQVDSATISRQDEVPPPPVASQQQRRSL-EEKCPAGSHRSEPTGACNACTEGV
Prob(dN/dS>1)
0.0
0.4
0.8
Extracellular
110
120
130
140
150
160
170
180
190
200
GM12878_10D
GM10541_10D
GM10541_10D2
Chimpanzee_10D2
Gorilla_10D2
GM10541_10D1
Chimpanzee_10D1
Gorilla_10D1
Orangutan_10D2
Orangutan_10D1
Macaque_10D2
Macaque_10D1
DYTIASNNLPSCLLCTVCKSGQTNKSSCTTTRDTVCQCEKGSFQDKNSPEMCRTCRTGCPRGMVKVSNCTPRSDIKCKNESAASSTGKTPAAEETVTTILGML
DYTIASNNLPSCLLCTVCKSGQTNKSSCTTTRDTVCQCEKGSFQDKNSPEMCRTCRTGCPRGMVKVSNCTPRSDIKCKNESAASSTGKTPAAEETVTTILGML
DYTIASNNLPSCLLCTVCKSGQTNKSSCTTTRDTVCQCEKGSFQDKNSPEMCRTCRTGCPRGMVKVSNCTSQSDIKCKNESAASSTGKTPAAEETVTTILGML
DYTIASNNLPSCLPCTVCKSGQTNKSSCTTTRDTVCQCEKGRFQDKNSPEMCRTCRTGCPRGMVKVSNCTSQSDIKCKNESAASSTGKTPAAEETVTTILGML
DYTIASNNLPSCLPCTVCKSGQTNKSSCTTTRDTVCECEKGRFQDKNSPEMCRKCRTGCPRGMVKVSNCTSQSDIKCKNESAASSTGKTPAAEETVTTILGML
----------MEKVSDCTPWSDINCKNESSASSTGKTPAAEETVTTILGML
----------MEKVSDCTPLSDINCKNESAASSTGKTPAAEETVTTILGML
----------MEKVSDCTPWSDINCKNESAASSTGKTPAAEETVTTILGML
DYTIASNNLPSCLLCTVCKSGQTTKSSCTTTRDTMCQCEKGSFQDENTPEMCQKCRTGCPRGMVKVRNCTSQSDIKCKNESAASSTGKTPAAEDTVTTSLGTL
----------MVKVSDCTPWSDINC-SASAASSTGKTPAAEDTVTTSLWTL
DYTNASNNEPSCLLCAVCKSDEEEMSPCTTTSNRVCQCKPGSFWNGNSTEMCRKCSTGCPREMIKVSDCTPWSDINC-SASAATSTEKTAVAEETVITSLGTP
----------MIKVSDCTPWSDINC-SASAATSTEKTAVAEETVITSLGTP

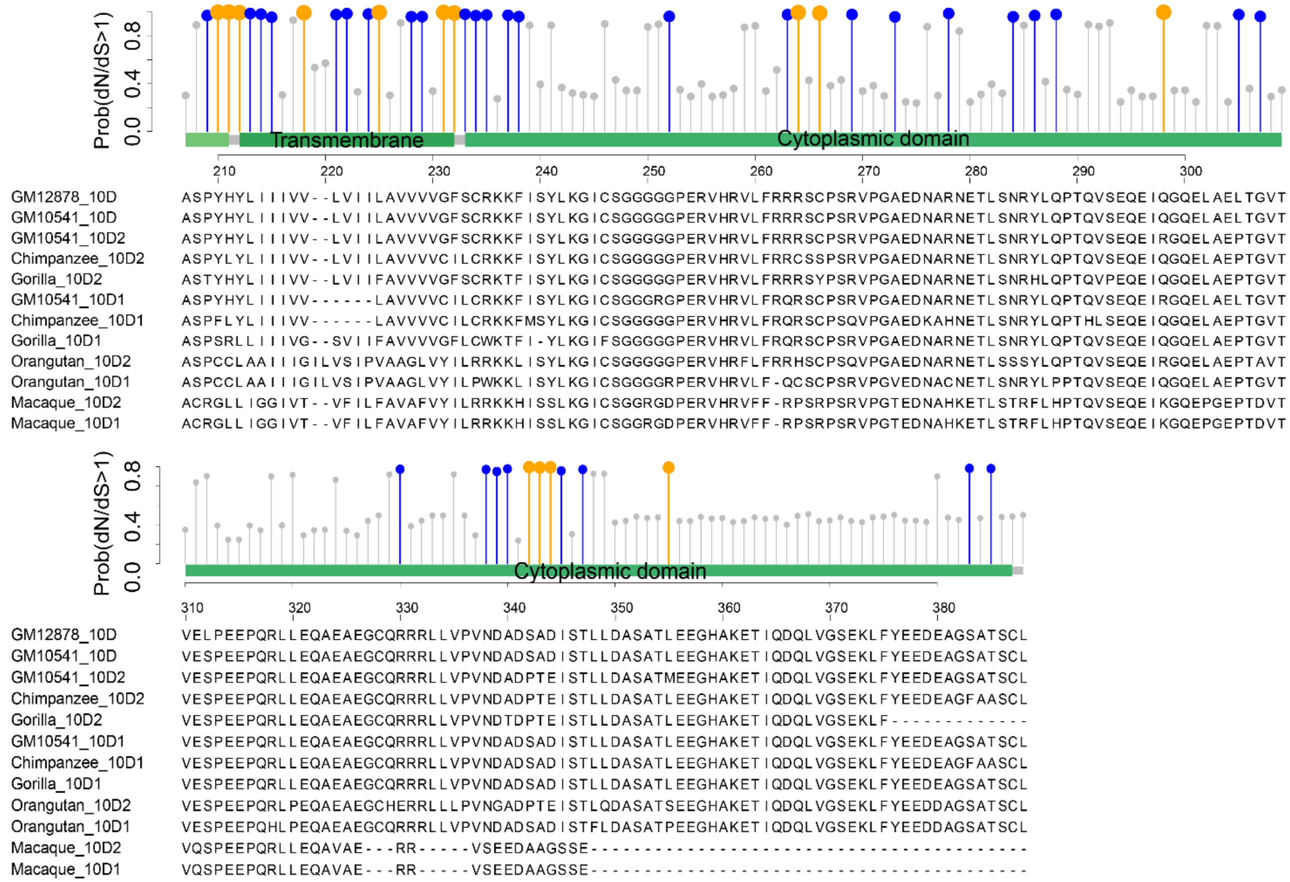


Figure S64. Evidence for positive selection at the clade of *TNFRSF10D1* lineages. Bayesian-based phylogenetic (BEAST) trees and branch supports were inferred for the eight most common ORF sequences from Melanesian (GM10541, CN3), European (GM12878, CN2), chimpanzee (PanTro, CN4), and rhesus macaque (R.macaque) samples. *dN/dS* ratios were estimated using the codeml program in the PAML package (v14.9), and colored orange if they are greater than 1. (**A**) Significant evidence for variable *dN/dS* ratios among the phylogeny. *dN/dS* ratios at the clade of *TNFRSF10D1* lineages are significantly greater than 1, compared to those under the null expectation (*dN/dS* = 1), suggesting the act of positive selection in this clade. Note that PAML reports *dN/dS* = 999 for a branch when no synonymous mutation was inferred (*dS* = 0) along the lineage. (B) The branch-site selection test of PAML identifies a cluster of positively selected sites corresponding to the predicted transmembrane domain of the genes.

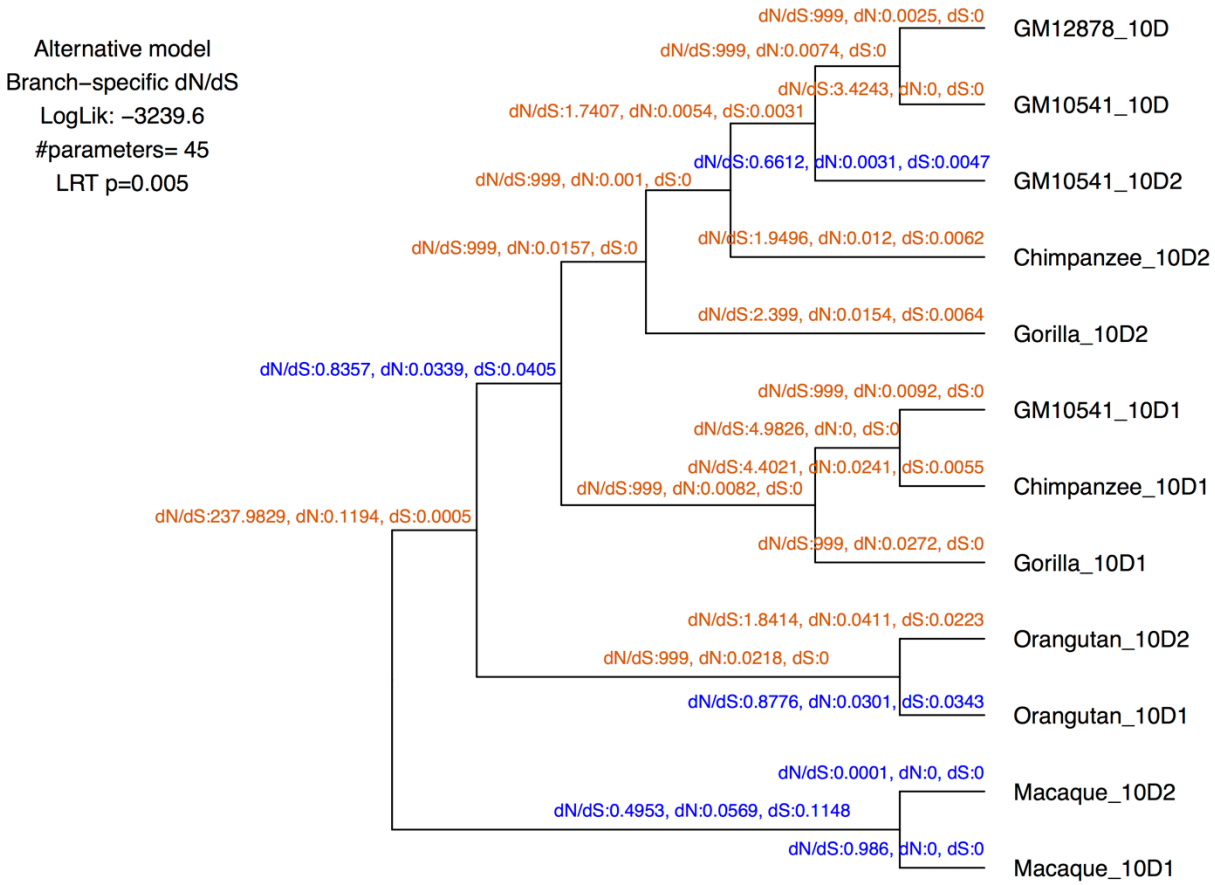


Figure S65. Cladogram of *TNFRSF10D* lineages showing *dN* and *dS* values and *dN/dS* ratios for individual branches. The same *TNFRSF10D* sequences as in **Figure S64** were used to estimate the branch-specific *dN/dS*, *dN*, and *dS* values, shown above each branch, using PAML (v14.9). Orange and blue colors indicate if the test of *dN/dS* ratio ≥1 is significant ($p < 0.05$) or not ($p \geq 0.05$), respectively. A significance test of the free *dN/dS* ratios model was based on a chi-squared likelihood ratio test (d.f. = 1) against the null model of neutral evolution (*dN/dS* = 1). The phylogeny of these *TNFRSF10D* sequences (**Figure S64**, left panel) was inferred using BEAST (v2.5.0). Note that PAML reports *dN/dS* = 99 or 999 for a branch when no synonymous mutation was inferred (*dS* = 0) along the lineage.

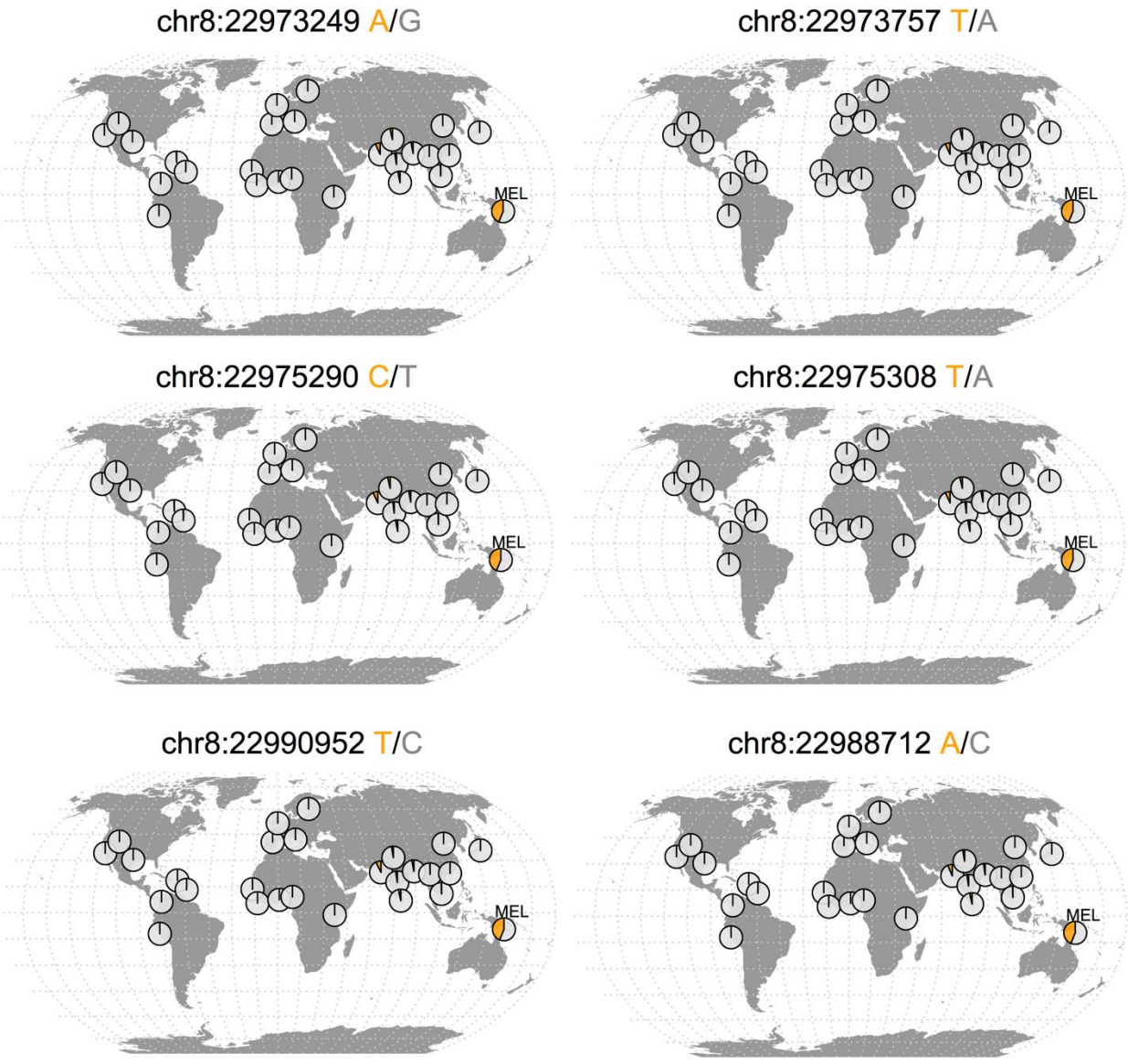


Figure S66. Geographic allele frequency distributions of six $DEL_{MEL\text{-}NDL}$ deletion-tagged SNVs, in addition to rs367585898 shown in Figure 5A. The 1KG populations suggest the deletion variant is geographically restricted to mainly South Asians at low frequencies (<0.07). Note that at all sites, the minor (orange) alleles are the deletion-tagged alleles, and the frequencies of these alleles in the Melanesian samples, are all 0.4375, which also equals the frequency of the deletion alleles (**Table S18**).

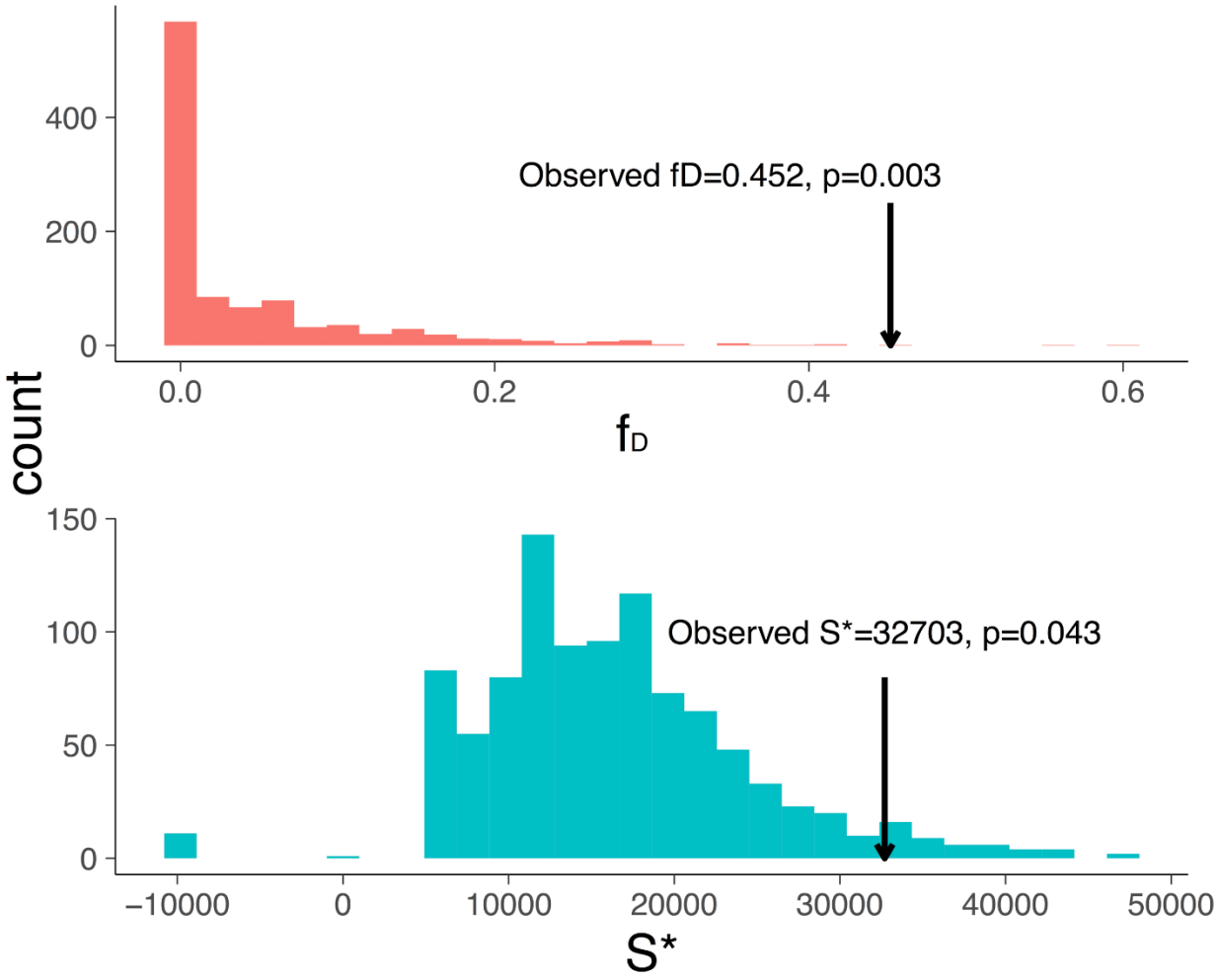


Figure S67. Significant archaic introgression signals in Melanesians at the unique sequence of 18,500 bp at the telomeric side of *TNFRSF10D* that spans the locus of $DEL_{MEL\text{-}NDL}$ (chr8:22,972,880-22,991,380). Two complementary tests, f_D (top) and S^* (bottom), were used to test introgression in Melanesians. In the case of the f_D calculation, the two Neanderthal individuals were used as the archaic reference sequences. Significance levels were determined using coalescent simulations based on 1,000 demographic models (**Methods**).

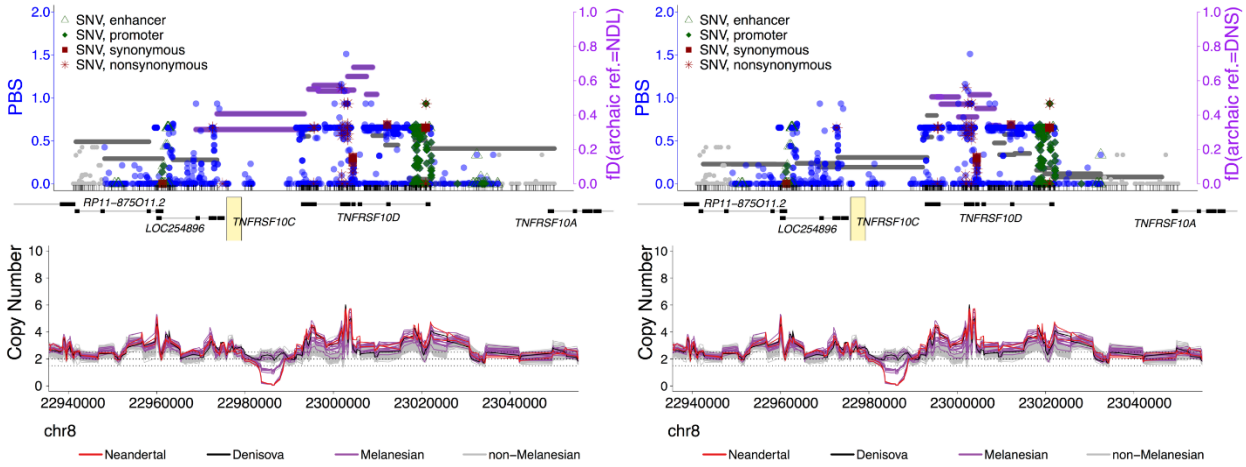


Figure S68. Significant introgression signals (f_D statistic) at the deletion locus ($DEL_{MEL\text{-}NDL}$) using Neanderthals (NDL) as the archaic reference (left panel). Note that the signals of introgression become mostly insignificant when the Denisovan individual was used as the archaic reference, but with some signals at the DUP_{10D} variant locus due to PSVs shared between Neanderthals and Denisovan samples (right panel).

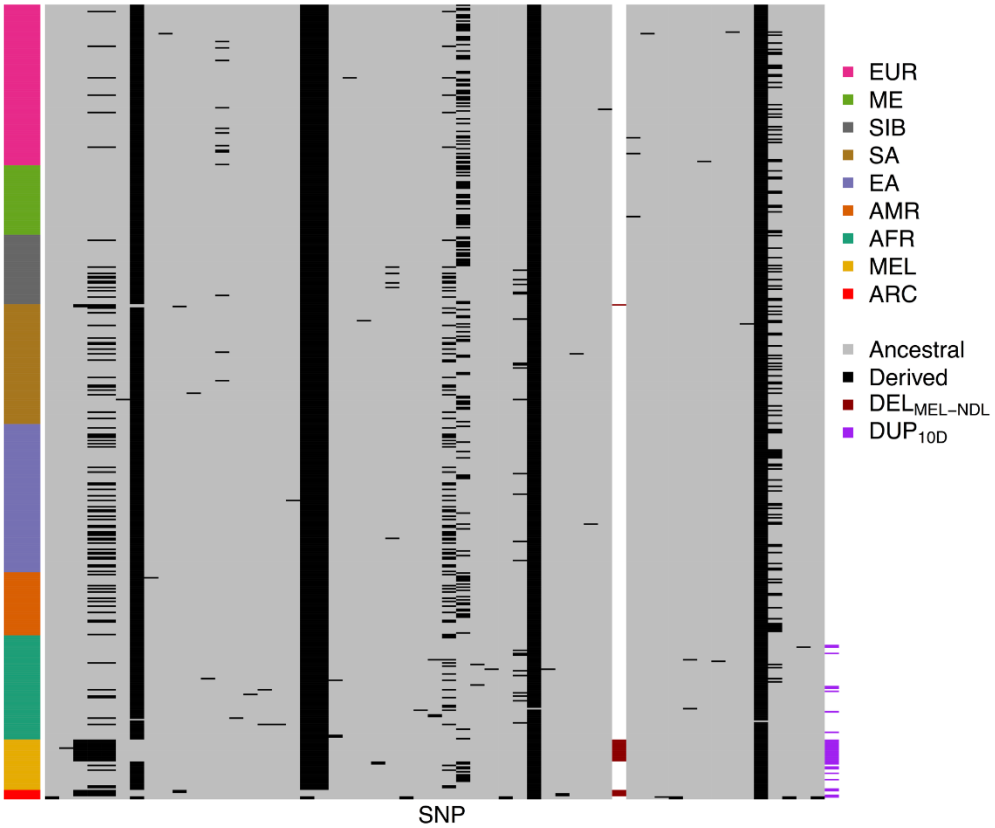


Figure S69. Unique deletion-linked haplotype observed in Melanesians (n=14), Neanderthals (n=4), and South Asians (Punjabi, n=1). Haplotypes of 56 SNVs, along with the two bi-allelic variants, $DEL_{MEL\text{-}NDL}$ (dark red) and DUP_{10D} (purple), for chr8:22972880-22991380.

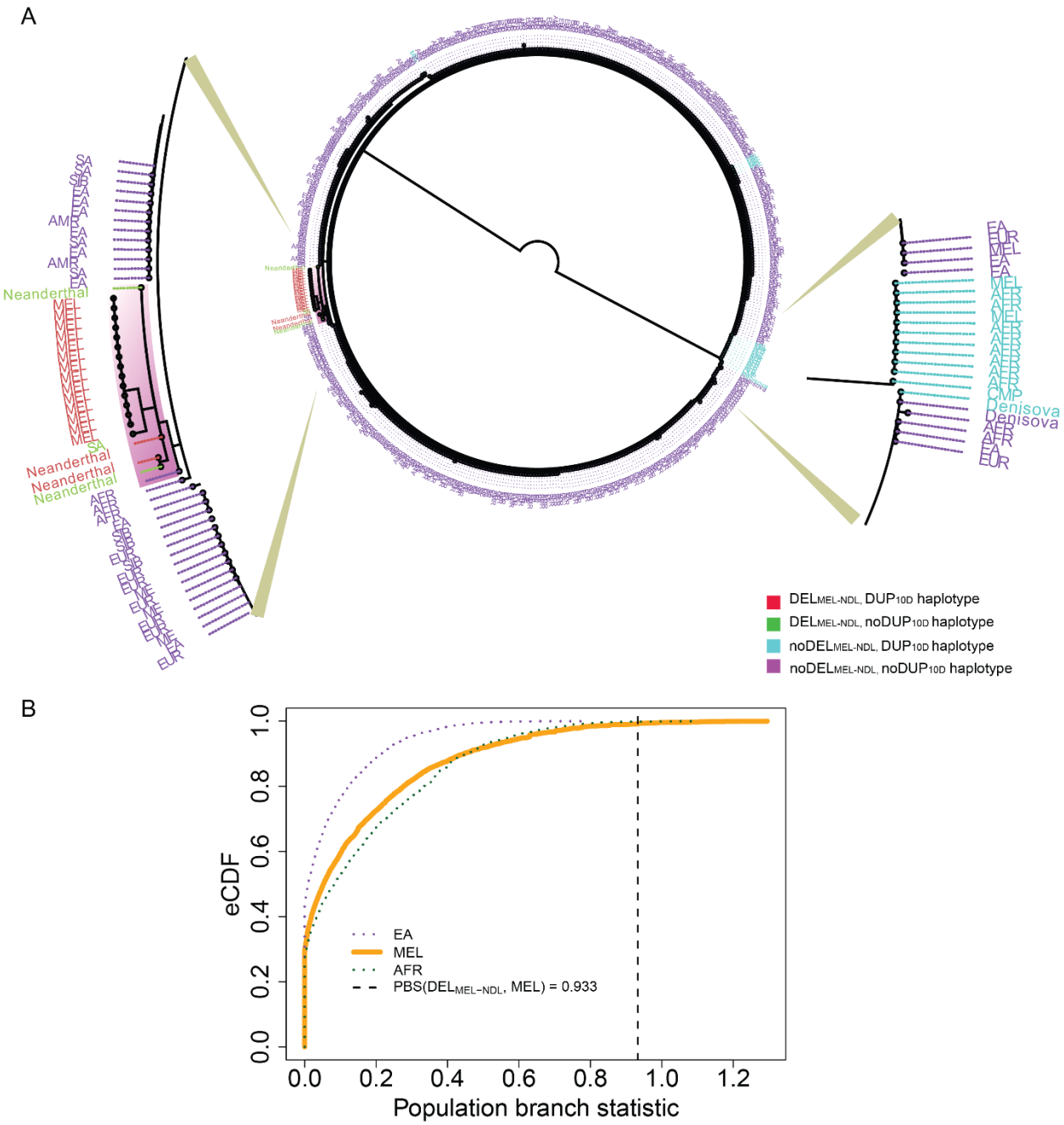


Figure S70. A single clade for deletion-linked haplotypes and evidence for ongoing positive selection. (**A**) The maximum likelihood phylogeny (top, log likelihood = -21578.402) inferred using RAxML (v.8.2.10) and the putative CN2 sequences (chr8:22972880-22991380). (**B**) Cumulative distribution function for *PBS* of AFR, EA, and MEL. The *PBS* distributions were generated using SNVs from the coalescent simulations for the putative CN2 regions around the $DEL_{MEL\text{-}NDL}$ variant, conditional on frequencies within 30% of the observed frequency of the deletion variant in MEL (0.4375; 30% range: 0.306–0.568). The numbers of SNVs in AFR, EA, and MEL are 8,229, 7,684, and 7,580, respectively. Compared with all three parametric bootstrap distributions, it is unlikely to observe a *PBS* value as large as 0.933 under the null demographic models given the age of the variant ($p < 0.0082$).

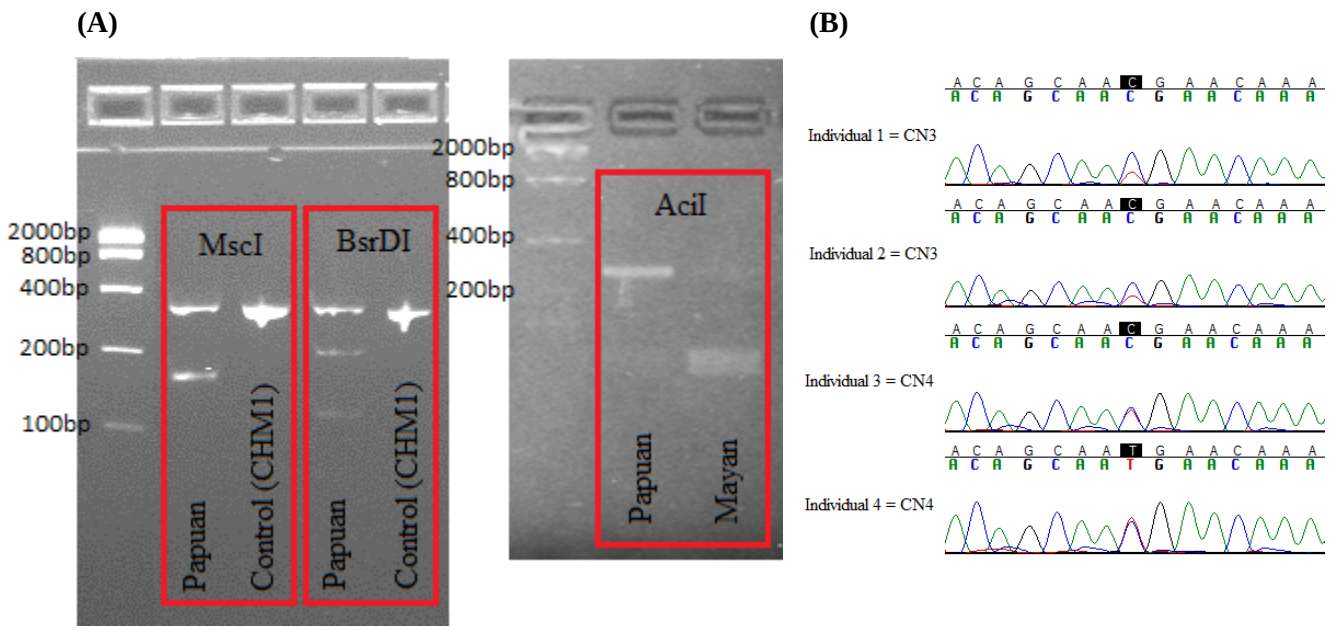


Figure S71. PCR experiments for genotyping DUP_{16p12} polymorphism. (A) Gel images of the restriction digests used to test for the chromosome 16 duplication. The ladder appears to the left with base pair sizes listed. The two enzymes, which cut the alternative haplotype, are shown with cuts present in the Papuan (~150 bp in MscI and ~100 and ~200 bp bands in BsrDI) and no cut in the control (bands at ~300 bp). AciI was tested with a Mayan sample as the control and shows cuts at ~150 bp and no cuts in the Papuan. (B) Sanger traces of four blood-derived Melanesian DNA samples at site 22768213 showing the variable peak height of the alternate T allele versus the reference C allele. The first two samples have half the peak height for the T allele indicating that they have only one copy of the duplication (CN3).

Supplementary Tables

All supplementary tables are provided in separate EXCEL files and available as online tables.

Table S1. Copy number genotypes of 5,135 CNVs in 249 modern human, 3 archaic hominin, and 72 nonhuman primate genomes. CNVs were called in a discovery panel of 20 genomes, including 17 modern and 3 archaic humans, using the dCGH CNV discovery method as described in the main text. A read-depth-based approach was used to genotype these CNVs.

Table S2. Copy number genotypes of 402 hominin-specific CNVs in 249 modern human, 3 archaic hominin, and 72 nonhuman primate genomes. These CNVs were derived from the 5,135 CNV genotypes from Table S1 using the approach described in the main text and the Methods section in the Supplementary Materials.

Table S3. 368,256 CNVs identified from 266 SGDP samples using five different callers. Note that only Genome STRiP and dCGH produce multi-allelic CNV (mCNV) calls.

Table S4. Validation of CNVs using single-nucleotide polymorphism (SNP) microarray. Array-based CNV calls were generated using Illumina 2.1M SNP microarrays for 123 samples. For the purpose of validation, only CNV calls with more than 5 or 10 SNP probes were considered. Eight out of the 123 samples were removed from this analysis because they suffer large background noise as determined in (*9*). We examined variants that have <50% segmental duplication (SegDup) in their content.

Table S5. Numbers of population-stratified CNVs as reported by three summary statistics. The p-value of observing the number of stratified CNVs in each population was estimated through 10,000 non-parametric simulations of permuting the CN estimates for the 19,211 CNVs.

Table S6. Primary three population demographic models for Africans, East Asians, and Melanesians evaluated using ∂a∂i (Gutenkunst et al. 2009). Assume a generation time of 29 years and mutation rate $\mu=0.5\times10^{-9}$ per site per year (*61*). N is effective population size; T is the time of a demographic event (years); ms is symmetric migration rate, while $m_{A\text{-}B}$ is the asymmetric migration rate from population B to A (per chromosome per generation). The parameter P_{flip} models the proportion of variants with ancestral state misidentification. * indicates the best-fit model in this table and # indicates a model whose optimization does not converge.

Table S7. Maximum likelihood parameter estimates and confidence intervals for the two best-fit demographic models of the population trio: AFR-EA-MEL (Figure S6). Parameter estimates are calculated using the mutation rate of 0.5×10^{-9} per base per year (*61*) and a generation time of 29 years. N is effective population size; T is the time of a demographic event (years); ms is symmetric migration rate, while mA-B is the asymmetric migration rate from population B to A (per chromosome per generation). The parameter P_{flip} models the proportion of variants with ancestral state misidentification. 95% confidence intervals (C.I.) were estimated using Godambe information matrix (Materials and Methods).

Table S8. Uniform (unif) priors for demographic parameters relevant to events prior to the anatomically modern humans. Time in years and population sizes are the number of individuals.

Table S9. CNV candidates of selection in Melanesians. Melanesian-stratified CNVs were identified using Dmedian, VST, and MWU tests. Selection signals were inferred using SNVs from the flanking diploid sequences of candidate CNVs and the population branch statistics (*PBS*) and their significances (p-values) are determined using coalescent simulations. Asterisks (*) indicate introgression signals in close proximity.

Table S10. Top CNV candidates with archaic introgression signals in Melanesians. Introgression signals were inferred based on the f_D statistics calculated using Neanderthal (NDL) and Denisovan (DNS) as archaic reference genomes, separately (Methods). P-values for selection and introgression scans are calculated using coalescent simulations. The two CNVs, indicated with †, represent a single 383 kbp duplication found specifically in the Melanesian and Denisovan samples.

Table S11. CN estimates of DUPchr16p12 for an independent set of 242 blood-derived Melanesian DNA samples using genotypes from Sanger sequencing.

Table S12. Experimental CN estimates for cell line samples. We validated CN estimates in three Melanesian and one European ancestry cell lines. These cell lines are fibroblast derived. The details of the FISH and PCR experiments can be found in Supplementary Note: Materials and Methods.

Table S13. Fosmid and BAC clones in FISH experiments.

Table S14. Summary statistics for the de novo SD assembly for the DUP16p12 Melanesian duplication variant. A Melanesian SDA contig was generated using an iteration approach with SDA (*41*). Each iteration used a backbone from the previous iteration to extend the assembly. The paralogous sequence variant (PSV) agreement is computed based on the overlap (left and right ends of the contig) between two assemblies from two consecutive iterations. The final ~383 kbp Melanesian duplication contig was resolved after the seventh iteration.

Table S15. Divergence between Melanesian DUP16p12 sequences and their closest related lineages. Melanesian sequences were sampled along the DUP16p12 sequences on the Melanesian contig (Figure S22). Homologous sequences to the Melanesian sequences were pulled from the ancestral (16p12.2, KV880768.1) and insertion loci (16p11.2, GRCh37) of DUP16p12, and 7 BAC contigs from Nuttle et al. (*15*).

Table S16. Increasing the number of potential sequences (>10 kbp) for unequal crossover and potential susceptibility to rearrangement at 16p11.2 in the Melanesian contig. Sequences (>10 kbp) from the duplication block distal to the 500 kbp autism-critical region in GRCh37 and the assembled Melanesian contig were aligned against the those proximal to the critical region in GRCh37 at 16p11.2. Only pairs with >95% identity were shown. The additional NPIPB duplication sequence (MelanesianContig:1084327-1164934) in the Melanesian contig with >95% identity to that in the duplication block proximal to autism-critical region provides additional genetic material for potential unequal crossover events.

Table S17. PCR validation of the DELMEL-NDL deletion for 1000 Genomes Project (1KG) genomes. 64 putative $DEL_{MEL\text{-}NDL}$ carriers from the 1KG (Phase 3) were identified using a read-depth-based genotyper. Validations of these samples were performed using an in-house designed PCR assay described in the main text. Note that we did not perform a PCR assay for HG01308 due to the lack of DNA for that sample.

Table S18. Allele frequencies of seven tagging alleles to the $DEL_{MEL\text{-}NDL}$ deletion variant in SGDP samples. The seven tagging SNVs are in nearly complete linkage disequilibrium with the deletion allele (r^2>0.9 and D'>0.9) in the SGDP data set. Note that in all three Neanderthals, the last four SNVs were no-calls due to lack of sequence coverage. Also note that due to low sequence coverage, we did not call CNVs for the Mezmaiskaya Neanderthal.

Table S19. Sites shared between Melanesian and Denisovan samples at the chromosome 16p12 duplication locus. Selected sites (arbitrary site number assigned) with position on chr16 (GRCh37) and haplotype found in Papuans and Melanesians from the island of Bougainville and other control individuals, which match the reference. *Sites that were used for further restriction digest tests.

Table S20. Primers for the PCR assay to genotype copy number status of the 383 kbp Melanesian-Denisovan-specific duplication at its ancestral locus of chromosome 16p12.2. Primers designed to amplify approx. 300 bp regions around the site listed in Table S17. PCR protocol used a standard PCR master mix and standard amplification protocol with 35 cycles of 95° denaturation for 45 seconds, annealing at 55° for 30 seconds, and extension at 72° for 45 seconds.

Table S21. Primers used in the PCR assays to genotype three additional selective CNV candidates shown in Figure S11.

References and Notes

1. S. A. Tishkoff, F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, P. Deloukas, Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**, 31–40 (2007). doi:10.1038/ng1946 Medline

2. X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, H. Zheng, T. Liu, W. He, K. Li, R. Luo, X. Nie, H. Wu, M. Zhao, H. Cao, J. Zou, Y. Shan, S. Li, Q. Yang, P. Asan, P. Ni, G. Tian, J. Xu, X. Liu, T. Jiang, R. Wu, G. Zhou, M. Tang, J. Qin, T. Wang, S. Feng, G. Li, J. Huasang, J. Luosang, W. Wang, F. Chen, Y. Wang, X. Zheng, Z. Li, Z. Bianba, G. Yang, X. Wang, S. Tang, G. Gao, Y. Chen, Z. Luo, L. Gusang, Z. Cao, Q. Zhang, W. Ouyang, X. Ren, H. Liang, H. Zheng, Y. Huang, J. Li, L. Bolund, K. Kristiansen, Y. Li, Y. Zhang, X. Zhang, R. Li, S. Li, H. Yang, R. Nielsen, J. Wang, J. Wang, Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010). doi:10.1126/science.1190371 Medline

3. M. Fumagalli, I. Moltke, N. Grarup, F. Racimo, P. Bjerregaard, M. E. Jørgensen, T. S. Korneliussen, P. Gerbault, L. Skotte, A. Linneberg, C. Christensen, I. Brandslund, T. Jørgensen, E. Huerta-Sánchez, E. B. Schmidt, O. Pedersen, T. Hansen, A. Albrechtsen, R. Nielsen, Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343–1347 (2015). doi:10.1126/science.aab2319 Medline

4. M. A. Ilardo, I. Moltke, T. S. Korneliussen, J. Cheng, A. J. Stern, F. Racimo, P. de Barros Damgaard, M. Sikora, A. Seguin-Orlando, S. Rasmussen, I. C. L. van den Munckhof, R. Ter Horst, L. A. B. Joosten, M. G. Netea, S. Salingkat, R. Nielsen, E. Willerslev, Physiological and genetic adaptations to diving in sea nomads. *Cell* **173**, 569–580.e15 (2018). doi:10.1016/j.cell.2018.03.054 Medline

5. F. L. Mendez, J. C. Watkins, M. F. Hammer, A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am. J. Hum. Genet.* **91**, 265–274 (2012). doi:10.1016/j.ajhg.2012.06.015 Medline

6. E. Huerta-Sánchez, X. Jin, Z. Asan, Z. Bianba, B. M. Peter, N. Vinckenbosch, Y. Liang, X. Yi, M. He, M. Somel, P. Ni, B. Wang, X. Ou, J. Huasang, J. Luosang, Z. X. Cuo, K. Li, G. Gao, Y. Yin, W. Wang, X. Zhang, X. Xu, H. Yang, Y. Li, J. Wang, J. Wang, R. Nielsen, Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014). doi:10.1038/nature13408 Medline

7. F. Racimo, D. Gokhman, M. Fumagalli, A. Ko, T. Hansen, I. Moltke, A. Albrechtsen, L. Carmel, E. Huerta-Sánchez, R. Nielsen, Archaic Adaptive Introgression in TBX15/WARS2. *Mol. Biol. Evol.* **34**, 509–524 (2017). Medline

8. D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, M. E. Hurles; Wellcome Trust Case Control Consortium, Origins and functional impact of copy

number variation in the human genome. *Nature* **464**, 704–712 (2010). doi:10.1038/nature08516 Medline

9. P. H. Sudmant, S. Mallick, B. J. Nelson, F. Hormozdiari, N. Krumm, J. Huddleston, B. P. Coe, C. Baker, S. Nordenfelt, M. Bamshad, L. B. Jorde, O. L. Posukh, H. Sahakyan, W. S. Watkins, L. Yepiskoposyan, M. S. Abdullah, C. M. Bravi, C. Capelli, T. Hervig, J. T. S. Wee, C. Tyler-Smith, G. van Driem, I. G. Romero, A. R. Jha, S. Karachanak-Yankova, D. Toncheva, D. Comas, B. Henn, T. Kivisild, A. Ruiz-Linares, A. Sajantila, E. Metspalu, J. Parik, R. Villems, E. B. Starikovskaya, G. Ayodo, C. M. Beall, A. Di Rienzo, M. F. Hammer, R. Khusainova, E. Khusnutdinova, W. Klitz, C. Winkler, D. Labuda, M. Metspalu, S. A. Tishkoff, S. Dryomov, R. Sukernik, N. Patterson, D. Reich, E. E. Eichler, Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015). doi:10.1126/science.aab3761 Medline

10. P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, J. O. Korbel; 1000 Genomes Project Consortium, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015). doi:10.1038/nature15394 Medline

11. P. A. Audano, A. Sulovari, T. A. Graves-Lindsay, S. Cantsilieris, M. Sorensen, A. E. Welch, M. L. Dougherty, B. J. Nelson, A. Shah, S. K. Dutcher, W. C. Warren, V. Magrini, S. D. McGrath, Y. I. Li, R. K. Wilson, E. E. Eichler, Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675.e19 (2019). doi:10.1016/j.cell.2018.12.019 Medline

12. G. H. Perry, N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain, R. Misra, N. P. Carter, C. Lee, A. C. Stone, Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007). doi:10.1038/ng2123 Medline

13. Y. Xue, D. Sun, A. Daly, F. Yang, X. Zhou, M. Zhao, N. Huang, T. Zerjal, C. Lee, N. P. Carter, M. E. Hurles, C. Tyler-Smith, Adaptive evolution of UGT2B17 copy-number variation. *Am. J. Hum. Genet.* **83**, 337–346 (2008). doi:10.1016/j.ajhg.2008.08.004 Medline

14. R. J. Hardwick, L. R. Machado, L. W. Zuccherato, S. Antolinos, Y. Xue, N. Shawa, R. H. Gilman, L. Cabrera, D. E. Berg, C. Tyler-Smith, P. Kelly, E. Tarazona-Santos, E. J. Hollox, A worldwide analysis of beta-defensin copy number variation suggests recent selection of a high-expressing DEFB103 gene copy in East Asia. *Hum. Mutat.* **32**, 743–750 (2011). doi:10.1002/humu.21491 Medline

15. X. Nuttle, G. Giannuzzi, M. H. Duyzend, J. G. Schraiber, I. Narvaiza, P. H. Sudmant, O. Penn, G. Chiatante, M. Malig, J. Huddleston, C. Benner, F. Camponeschi, S. Ciofi-Baffoni, H. A. F. Stessman, M. C. N. Marchetto, L. Denman, L. Harshman, C. Baker, A. Raja, K. Penewit, N. Janke, W. J. Tang, M. Ventura, L. Banci, F. Antonacci, J. M. Akey, C. T. Amemiya, F. H. Gage, A. Reymond, E. E. Eichler, Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536**, 205–209 (2016). doi:10.1038/nature19075 Medline

16. S. Lindeberg, P. Nilsson-Ehle, B. Vessby, Lipoprotein composition and serum cholesterol ester fatty acids in nonwesternized Melanesians. *Lipids* **31**, 153–158 (1996). doi:10.1007/BF02522614 Medline

17. J. Flint, A. V. S. Hill, D. K. Bowden, S. J. Oppenheimer, P. R. Sill, S. W. Serjeantson, J. Bana-Koiri, K. Bhatia, M. P. Alpers, A. J. Boyce, D. J. Weatherall, J. B. Clegg, High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature* **321**, 744–750 (1986). doi:10.1038/321744a0 Medline

18. P. T. Katzmarzyk, W. R. Leonard, Climatic influences on human body size and proportions: Ecological adaptations and secular trends. *Am. J. Phys. Anthropol.* **106**, 483–503 (1998). doi:10.1002/(SICI)1096-8644(199808)106:4<483:AID-AJPA4>3.0.CO;2-K Medline

19. A. S. Malaspinas, M. C. Westaway, C. Muller, V. C. Sousa, O. Lao, I. Alves, A. Bergström, G. Athanasiadis, J. Y. Cheng, J. E. Crawford, T. H. Heupink, E. Macholdt, S. Peischl, S. Rasmussen, S. Schiffels, S. Subramanian, J. L. Wright, A. Albrechtsen, C. Barbieri, I. Dupanloup, A. Eriksson, A. Margaryan, I. Moltke, I. Pugach, T. S. Korneliussen, I. P. Levkivskyi, J. V. Moreno-Mayar, S. Ni, F. Racimo, M. Sikora, Y. Xue, F. A. Aghakhanian, N. Brucato, S. Brunak, P. F. Campos, W. Clark, S. Ellingvåg, G. Fourmile, P. Gerbault, D. Injie, G. Koki, M. Leavesley, B. Logan, A. Lynch, E. A. Matisoo-Smith, P. J. McAllister, A. J. Mentzer, M. Metspalu, A. B. Migliano, L. Murgha, M. E. Phipps, W. Pomat, D. Reynolds, F.-X. Ricaut, P. Siba, M. G. Thomas, T. Wales, C. M. Wall, S. J. Oppenheimer, C. Tyler-Smith, R. Durbin, J. Dortch, A. Manica, M. H. Schierup, R. A. Foley, M. M. Lahr, C. Bowern, J. D. Wall, T. Mailund, M. Stoneking, R. Nielsen, M. S. Sandhu, L. Excoffier, D. M. Lambert, E. Willerslev, A genomic history of Aboriginal Australia. *Nature* **538**, 207–214 (2016). doi:10.1038/nature18299 Medline

20. A. Bergström, S. J. Oppenheimer, A. J. Mentzer, K. Auckland, K. Robson, R. Attenborough, M. P. Alpers, G. Koki, W. Pomat, P. Siba, Y. Xue, M. S. Sandhu, C. Tyler-Smith, A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science* **357**, 1160–1163 (2017). doi:10.1126/science.aan3842 Medline

21. D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. F. 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). doi:10.1038/nature09710 Medline

22. P. Skoglund, C. Posth, K. Sirak, M. Spriggs, F. Valentin, S. Bedford, G. R. Clark, C. Reepmeyer, F. Petchey, D. Fernandes, Q. Fu, E. Harney, M. Lipson, S. Mallick, M. Novak, N. Rohland, K. Stewardson, S. Abdullah, M. P. Cox, F. R. Friedlaender, J. S.

Friedlaender, T. Kivisild, G. Koki, P. Kusuma, D. A. Merriwether, F.-X. Ricaut, J. T. S. Wee, N. Patterson, J. Krause, R. Pinhasi, D. Reich, Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**, 510–513 (2016). doi:10.1038/nature19844 Medline

23. B. Vernot, S. Tucci, J. Kelso, J. G. Schraiber, A. B. Wolf, R. M. Gittelman, M. Dannemann, S. Grote, R. C. McCoy, H. Norton, L. B. Scheinfeldt, D. A. Merriwether, G. Koki, J. S. Friedlaender, J. Wakefield, S. Pääbo, J. M. Akey, Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239 (2016). doi:10.1126/science.aad9416 Medline

24. K. Prüfer, C. de Filippo, S. Grote, F. Mafessoni, P. Korlević, M. Hajdinjak, B. Vernot, L. Skov, P. Hsieh, S. Peyrégne, D. Reher, C. Hopfe, S. Nagel, T. Maricic, Q. Fu, C. Theunert, R. Rogers, P. Skoglund, M. Chintalapati, M. Dannemann, B. J. Nelson, F. M. Key, P. Rudan, Ž. Kućan, I. Gušić, L. V. Golovanova, V. B. Doronichev, N. Patterson, D. Reich, E. E. Eichler, M. Slatkin, M. H. Schierup, A. M. Andrés, J. Kelso, M. Meyer, S. Pääbo, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017). doi:10.1126/science.aao1887 Medline

25. 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). doi:10.1126/science.1224344 Medline

26. 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. F. 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). doi:10.1038/nature12886 Medline

27. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300

genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016). doi:10.1038/nature18964 Medline

28. J. Prado-Martinez, P. H. Sudmant, J. M. Kidd, H. Li, J. L. Kelley, B. Lorente-Galdos, K. R. Veeramah, A. E. Woerner, T. D. O'Connor, G. Santpere, A. Cagan, C. Theunert, F. Casals, H. Laayouni, K. Munch, A. Hobolth, A. E. Halager, M. Malig, J. Hernandez-Rodriguez, I. Hernando-Herraez, K. Prüfer, M. Pybus, L. Johnstone, M. Lachmann, C. Alkan, D. Twigg, N. Petit, C. Baker, F. Hormozdiari, M. Fernandez-Callejo, M. Dabad, M. L. Wilson, L. Stevison, C. Camprubí, T. Carvalho, A. Ruiz-Herrera, L. Vives, M. Mele, T. Abello, I. Kondova, R. E. Bontrop, A. Pusey, F. Lankester, J. A. Kiyang, R. A. Bergl, E. Lonsdorf, S. Myers, M. Ventura, P. Gagneux, D. Comas, H. Siegismund, J. Blanc, L. Agueda-Calpena, M. Gut, L. Fulton, S. A. Tishkoff, J. C. Mullikin, R. K. Wilson, I. G. Gut, M. K. Gonder, O. A. Ryder, B. H. Hahn, A. Navarro, J. M. Akey, J. Bertranpetit, D. Reich, T. Mailund, M. H. Schierup, C. Hvilsom, A. M. Andrés, J. D. Wall, C. D. Bustamante, M. F. Hammer, E. E. Eichler, T. Marques-Bonet, Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013). doi:10.1038/nature12228 Medline

29. Materials and methods are available as supplementary materials.

30. S. H. Martin, J. W. Davey, C. D. Jiggins, Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257 (2015). doi:10.1093/molbev/msu269 Medline

31. V. Tillander, S. E. H. Alexson, D. E. Cohen, Deactivating Fatty Acids: Acyl-CoA Thioesterase-Mediated Control of Lipid Metabolism. *Trends Endocrinol. Metab.* **28**, 473–484 (2017). doi:10.1016/j.tem.2017.03.001 Medline

32. J. M. Kidd, T. L. Newman, E. Tuzun, R. Kaul, E. E. Eichler, Population stratification of a common APOBEC gene deletion polymorphism. *PLOS Genet.* **3**, e63 (2007). doi:10.1371/journal.pgen.0030063 Medline

33. P. An, R. Johnson, J. Phair, G. D. Kirk, X.-F. Yu, S. Donfield, S. Buchbinder, J. J. Goedert, C. A. Winkler, APOBEC3B deletion and risk of HIV-1 acquisition. *J. Infect. Dis.* **200**, 1054–1058 (2009). doi:10.1086/605644 Medline

34. N. J. Smith, T. R. Fenton, The APOBEC3 genes and their role in cancer: Insights from human papillomavirus. *J. Mol. Endocrinol.* **62**, R269–R287 (2019). doi:10.1530/JME-19-0011 Medline

35. Y. Y. Qi, X. J. Zhou, F. J. Cheng, P. Hou, L. Zhu, S. F. Shi, L. J. Liu, J. C. Lv, H. Zhang, DEFA gene variants associated with IgA nephropathy in a Chinese population. *Genes Immun.* **16**, 231–237 (2015). doi:10.1038/gene.2015.1 Medline

36. K. Mohajeri, S. Cantsilieris, J. Huddleston, B. J. Nelson, B. P. Coe, C. D. Campbell, C. Baker, L. Harshman, K. M. Munson, Z. N. Kronenberg, M. Kremitzki, A. Raja, C. R. Catacchio, T. A. Graves, R. K. Wilson, M. Ventura, E. E. Eichler, Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region. *Genome Res.* **26**, 1453–1467 (2016). doi:10.1101/gr.211284.116 Medline

37. K. M. Steinberg, F. Antonacci, P. H. Sudmant, J. M. Kidd, C. D. Campbell, L. Vives, M. Malig, L. Scheinfeldt, W. Beggs, M. Ibrahim, G. Lema, T. B. Nyambo, S. A. Omar, J.-M. Bodo, A. Froment, M. P. Donnelly, K. K. Kidd, S. A. Tishkoff, E. E. Eichler, Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat. Genet.* **44**, 872–880 (2012). doi:10.1038/ng.2335 Medline

38. H. Stefansson, A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson, J. Barnard, A. Baker, A. Jonasdottir, A. Ingason, V. G. Gudnadottir, N. Desnica, A. Hicks, A. Gylfason, D. F. Gudbjartsson, G. M. Jonsdottir, J. Sainz, K. Agnarsson, B. Birgisdottir, S. Ghosh, A. Olafsdottir, J.-B. Cazier, K. Kristjansson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, A. Kong, K. Stefansson, A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005). doi:10.1038/ng1508 Medline

39. F. Antonacci, J. M. Kidd, T. Marques-Bonet, B. Teague, M. Ventura, S. Girirajan, C. Alkan, C. D. Campbell, L. Vives, M. Malig, J. A. Rosenfeld, B. C. Ballif, L. G. Shaffer, T. A. Graves, R. K. Wilson, D. C. Schwartz, E. E. Eichler, A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat. Genet.* **42**, 745–750 (2010). doi:10.1038/ng.643 Medline

40. L. A. Weiss, Y. Shen, J. M. Korn, D. E. Arking, D. T. Miller, R. Fossdal, E. Saemundsen, H. Stefansson, M. A. R. Ferreira, T. Green, O. S. Platt, D. M. Ruderfer, C. A. Walsh, D. Altshuler, A. Chakravarti, R. E. Tanzi, K. Stefansson, S. L. Santangelo, J. F. Gusella, P. Sklar, B.-L. Wu, M. J. Daly; Autism Consortium, Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008). doi:10.1056/NEJMoa075974 Medline

41. M. R. Vollger, P. C. Dishuck, M. Sorensen, A. E. Welch, V. Dang, M. L. Dougherty, T. A. Graves-Lindsay, R. K. Wilson, M. J. P. Chaisson, E. E. Eichler, Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019). doi:10.1038/s41592-018-0236-3 Medline

42. M. E. Johnson, L. Viggiano, J. A. Bailey, M. Abdul-Rauf, G. Goodwin, M. Rocchi, E. E. Eichler, Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001). doi:10.1038/35097067 Medline

43. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007). doi:10.1093/molbev/msm088 Medline

44. B. P. Coe, H. A. F. Stessman, A. Sulovari, M. R. Geisheker, T. E. Bakken, A. M. Lake, J. D. Dougherty, E. S. Lein, F. Hormozdiari, R. A. Bernier, E. E. Eichler, Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019). doi:10.1038/s41588-018-0288-4 Medline

45. G. Pan, J. Ni, G. Yu, Y. F. Wei, V. M. Dixit, TRUNDD, a new member of the TRAIL receptor family that antagonizes TRAIL signalling. *FEBS Lett.* **424**, 41–45 (1998). doi:10.1016/S0014-5793(98)00135-5 Medline

46. A. Scally, J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero, A. Hobolth, T. Lappalainen, T. Mailund, T. Marques-Bonet, S. McCarthy, S. H. Montgomery, P. C. Schwalie, Y. A. Tang, M. C. Ward, Y. Xue, B. Yngvadottir, C. Alkan, L. N. Andersen, Q. Ayub, E. V. Ball, K. Beal, B. J. Bradley, Y. Chen, C. M. Clee, S. Fitzgerald, T. A. Graves, Y. Gu, P. Heath, A. Heger, E. Karakoc, A. Kolb-Kokocinski, G. K. Laird, G.

Lunter, S. Meader, M. Mort, J. C. Mullikin, K. Munch, T. D. O'Connor, A. D. Phillips, J. Prado-Martinez, A. S. Rogers, S. Sajjadian, D. Schmidt, K. Shaw, J. T. Simpson, P. D. Stenson, D. J. Turner, L. Vigilant, A. J. Vilella, W. Whitener, B. Zhu, D. N. Cooper, P. de Jong, E. T. Dermitzakis, E. E. Eichler, P. Flicek, N. Goldman, N. I. Mundy, Z. Ning, D. T. Odom, C. P. Ponting, M. A. Quail, O. A. Ryder, S. M. Searle, W. C. Warren, R. K. Wilson, M. H. Schierup, J. Rogers, C. Tyler-Smith, R. Durbin, Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012). doi:10.1038/nature10842 Medline

47. D. A. Pollard, V. N. Iyer, A. M. Moses, M. B. Eisen, Widespread discordance of gene trees with species tree in Drosophila: Evidence for incomplete lineage sorting. *PLOS Genet.* **2**, e173 (2006). doi:10.1371/journal.pgen.0020173 Medline

48. M. Florio, M. Heide, A. Pinson, H. Brandl, M. Albert, S. Winkler, P. Wimberger, W. B. Huttner, M. Hiller, Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *eLife* **7**, e32332 (2018). doi:10.7554/eLife.32332 Medline

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

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

51. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009). doi:10.1101/gr.094052.109 Medline

52. M. L. Dougherty, J. G. Underwood, B. J. Nelson, E. Tseng, K. M. Munson, O. Penn, T. J. Nowakowski, A. A. Pollen, E. E. Eichler, Transcriptional fates of human-specific segmental duplications in brain. *Genome Res.* **28**, 1566–1576 (2018). doi:10.1101/gr.237610.118 Medline

53. Z. N. Kronenberg, I. T. Fiddes, D. Gordon, S. Murali, S. Cantsilieris, O. S. Meyerson, J. G. Underwood, B. J. Nelson, M. J. P. Chaisson, M. L. Dougherty, K. M. Munson, A. R. Hastie, M. Diekhans, F. Hormozdiari, N. Lorusso, K. Hoekzema, R. Qiu, K. Clark, A. Raja, A. E. Welch, M. Sorensen, C. Baker, R. S. Fulton, J. Armstrong, T. A. Graves-Lindsay, A. M. Denli, E. R. Hoppe, P. Hsieh, C. M. Hill, A. W. C. Pang, J. Lee, E. T. Lam, S. K. Dutcher, F. H. Gage, W. C. Warren, J. Shendure, D. Haussler, V. A. Schneider, H. Cao, M. Ventura, R. K. Wilson, B. Paten, A. Pollen, E. E. Eichler, High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018). doi:10.1126/science.aar6343 Medline

54. P. H. Sudmant, J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig, A. Tsalenko, N. Sampas, L. Bruhn, J. Shendure, E. E. Eichler; 1000 Genomes Project, Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010). doi:10.1126/science.1197005 Medline

55. Z. N. Kronenberg, E. J. Osborne, K. R. Cone, B. J. Kennedy, E. T. Domyan, M. D. Shapiro, N. C. Elde, M. Yandell, Wham: Identifying structural variants of biological consequence. *PLOS Comput. Biol.* **11**, e1004572 (2015). doi:10.1371/journal.pcbi.1004572 Medline

56. R. M. Layer, C. Chiang, A. R. Quinlan, I. M. Hall, LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014). doi:10.1186/gb-2014-15-6-r84 Medline

57. T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, J. O. Korbel, DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012). doi:10.1093/bioinformatics/bts378 Medline

58. R. E. Handsaker, V. Van Doren, J. R. Berman, G. Genovese, S. Kashin, L. M. Boettger, S. A. McCarroll, Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015). doi:10.1038/ng.3200 Medline

59. F. Hach, F. Hormozdiari, C. Alkan, F. Hormozdiari, I. Birol, E. E. Eichler, S. C. Sahinalp, mrsFAST: A cache-oblivious algorithm for short-read mapping. *Nat. Methods* **7**, 576–577 (2010). doi:10.1038/nmeth0810-576 Medline

60. A. J. Coffman, P. H. Hsieh, S. Gravel, R. N. Gutenkunst, Computationally Efficient Composite Likelihood Statistics for Demographic Inference. *Mol. Biol. Evol.* **33**, 591–593 (2016). doi:10.1093/molbev/msv255 Medline

61. A. Scally, The mutation rate in human evolution and demographic inference. *Curr. Opin. Genet. Dev.* **41**, 36–43 (2016). doi:10.1016/j.gde.2016.07.008 Medline

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

63. P. Hsieh, K. R. Veeramah, J. Lachance, S. A. Tishkoff, J. D. Wall, M. F. Hammer, R. N. Gutenkunst, Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. *Genome Res.* **26**, 279–290 (2016). doi:10.1101/gr.192971.115 Medline

64. K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E.

Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, J. Stewart; International HapMap Consortium, A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007). doi:10.1038/nature06258 Medline

65. K. E. Langergraber, K. Prüfer, C. Rowney, C. Boesch, C. Crockford, K. Fawcett, E. Inoue, M. Inoue-Muruyama, J. C. Mitani, M. N. Muller, M. M. Robbins, G. Schubert, T. S. Stoinski, B. Viola, D. Watts, R. M. Wittig, R. W. Wrangham, K. Zuberbühler, S. Pääbo, L. Vigilant, Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15716–15721 (2012). doi:10.1073/pnas.1211740109 Medline

66. R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. González, M. Gratacòs, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, M. E. Hurles, Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006). doi:10.1038/nature05329 Medline

67. P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, E. S. Lander, Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002). doi:10.1038/nature01140 Medline

68. 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). doi:10.1086/521987 Medline

69. R. Thomson, J. K. Pritchard, P. Shen, P. J. Oefner, M. W. Feldman, Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 7360–7365 (2000). doi:10.1073/pnas.97.13.7360 Medline

70. A. Stamatakis, RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006). doi:10.1093/bioinformatics/btl446 Medline

71. R. Bouckaert, J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, A. J. Drummond, BEAST 2: A software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **10**, e1003537 (2014). doi:10.1371/journal.pcbi.1003537 Medline

72. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi:10.1093/molbev/mst010 Medline

73. M. Suyama, D. Torrents, P. Bork, PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34** (suppl. 2), W609–W612 (2006). doi:10.1093/nar/gkl315 Medline

74. K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007). doi:10.1093/molbev/msm092 Medline

75. V. Lulla, A. M. Dinan, M. Hosmillo, Y. Chaudhry, L. Sherry, N. Irigoyen, K. M. Nayak, N. J. Stonehouse, M. Zilbauer, I. Goodfellow, A. E. Firth, An upstream protein-coding region in enteroviruses modulates virus infection in gut epithelial cells. *Nat. Microbiol.* **4**, 280–292 (2019). doi:10.1038/s41564-018-0297-1 Medline

76. Z. Yang, Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573 (1998). doi:10.1093/oxfordjournals.molbev.a025957 Medline

77. P. Lichter, S. A. Ledbetter, D. H. Ledbetter, D. C. Ward, Fluorescence in situ hybridization with Alu and L1 polymerase chain reaction probes for rapid characterization of human chromosomes in hybrid cell lines. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 6634–6638 (1990). doi:10.1073/pnas.87.17.6634 Medline

78. B. S. Weir, C. C. Cockerham, Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358–1370 (1984). Medline

79. F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989). Medline

80. F. Antonacci, J. M. Kidd, T. Marques-Bonet, M. Ventura, P. Siswara, Z. Jiang, E. E. Eichler, Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.* **18**, 2555–2566 (2009). doi:10.1093/hmg/ddp187 Medline