Supplementary Materials for

Evolutionary history and adaptation of a human pygmy population of Flores Island, Indonesia

Serena Tucci, Samuel H. Vohr, Rajiv C. McCoy, Benjamin Vernot, Matthew R. Robinson, Chiara Barbieri, Brad J. Nelson, Wenqing Fu, Gludhug A. Purnomo, Herawati Sudoyo, Evan E. Eichler, Guido Barbujani, Peter M. Visscher, Joshua M. Akey*, Richard E. Green*

*Corresponding author. Email: ed@soe.ucsc.edu (R.E.G.); jakey@princeton.edu (J.M.A.)

Published 3 August 2018, *Science* **361**, 511 (2018)
DOI: 10.1126/science.aar8486

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/361/6401/511/suppl/DC1)

Tables S4 to S6 (.xlsx)

SUPPLEMENTARY INFORMATION

Table of Contents

Sample collection and DNA extraction

The Flores samples considered in this study were collected from the village of Rampasasa in the Manggarai District (Flores Island, Nusa Tenggara Timor Province; geographic coordinates 120.5, -8.7). Approval for this study was obtained from the Human Subjects Review Committee at UCSC to Principal Investigator Richard Edward Green (UCSC Institutional Review Board 2196), following local approvals from the elders' committee in Rampasasa, from Dr. Herawati Sudoyo, Deputy Director of Eijkman Institute for Molecular Biology of Jakarta and from the Manggarai District Authority in Ruteng (Flores).

A courtesy visit was made to the village in Fall 2013, during which we gave full explanation of the project aims and sample collection procedure. Upon approval of the elders' committee of Rampasasa and on the community level, samples were collected from 32 healthy adult individuals in collaboration with the Eijkman Institute in Spring 2014. Individuals were sampled randomly among volunteers in the village of Rampasasa. No specific sampling criteria were applied. Informed consent, written in their own language and in English, was obtained from all participants.

Saliva samples were collected using the Oragene DISCOVER (OGR-500) DNA sample collection kits (Genotek, Ottawa, Ontario, Canada). Anthropometric measurements (stature), sex, age, self-reported ancestry information, clan, and language were also collected in the field. DNA was extracted using Qiagen High Molecular Weight Blood and Tissue kit for DNA extraction. See Table S1 for information about samples collected.

Omni 2.5 array genotyping

Data generation and QC

We performed SNP genotyping for all 32 Flores individuals on the Illumina HumanOmni2.5-8 v1.1 BeadChip array. Genotypes were called using the Illumina Genome Studio v2011.1, genotyping Module Version 1.9.4 for a total number of 2,391,739 SNPs.

After discarding 53,260 X-chromosome, 2,246 Y-chromosome and 189 mitochondrial SNPs, 2,336,044 autosomal SNPs were left for analysis (we refer to that as the “OMNI” dataset).

PLINK v1.9 (www.cog-genomics.org/plink/1.9/) (*26*) was used to assess genotyping quality according to the protocol published in (*27*). Samples were checked for outlying heterozygosity (more than 3 standard deviations from the mean) and elevated rates of missing data (genotyping failure rate >3%). The observed heterozygosity rate per individual was calculated and plotted versus the proportion of missing SNPs per individual (Figure S1). Thresholds (dashed lines) were set to ≥ 0.03 for the genotype failure rate and ± 3 standard deviations from the mean for the heterozygosity rate. No sample failed these QCs. Samples were also checked for discordant sex information (mismatches between documented sex and that suggested in the genotyping data) to highlight potential plating errors. No individual with discordant sex was identified.

Identification of related individuals

We used the software KING version 1.4 (*28*) to identify unknown relatedness amongst the Flores individuals. We checked family relationships by estimating the kinship coefficient of 2,336,044 autosomal SNPs, using KING robust algorithm, which allows for the existence of population structure (parameter --kinship). The analysis was performed on the unpruned dataset, following KING documentation. Pairwise relationships were checked between each pair of individuals. Pairs of individuals with estimated kinship coefficients between 0.177 - 0.354, 0.0884 - 0.177 and 0.0442 - 0.0884 were considered first-degree, second-degree, and third-degree relatives, respectively (*28*). We identified 34 pairs of related individuals up to third-degree, which include 7 pairs of parent-offspring (PO), 2 pairs of full-sibling (FS), 6 pairs of second degree, and 19 pairs of third-degree relationships (Figure S2 panel A). To better visualize the results of the relatedness inference obtained using KING, we represented the *NxN* matrix of pairwise relationships between the 32

individuals in a network, using the R package *igraph (*http://igraph.org) (*29*). Figure S2 panel B shows a reconstruction of the family relationships among the Flores individuals. Based on these findings, we discarded 11 related individuals from downstream analysis.

Deconvoluting population affinities in Asia

To focus on population affinities within Asia, we integrated our Flores genotypes with SNP data released by the HUGO Pan-Asian SNP Consortium (*30*), genotyped for Affymetrix Genechip Human Mapping 50K array and available at http://www4a.biotec.or.th/PASNP. Among other populations, the dataset includes ~300 individuals sampled from 15 populations in the Indonesian Archipelago. Of those samples, 17 were collected from the same Rampasasa village (population code "IDRA"), in the context of a previous genetic survey. We used KING to check for the presence of duplicate individuals between our Flores sample and the Pan-Asian dataset. Our analysis confirmed that 6 IDRA individuals overlapped with individuals in our study. After removing duplicates, and close relatives within the Pan-Asian dataset, as reported in (*31*), we merged the dataset with our Flores genotypes. PLINK was used for data management and quality control. Genotyping success rate was set to 95% and MAF to 0.01. The full merged dataset includes 17,035 SNPs genotyped in 1,685 individuals from 74 populations (Table S2). After LD-pruning (--indep-pairwise 50 5 0.4), 15,187 SNPs were left for analysis ("PANASIA" dataset"). To obtain the first, synthetic view of the main patterns of population affinities in Asia, we performed a Principal Component Analysis (PCA) on the PANASIA, using the R package *SNPRelate* (*32*) (Figure S3A). To further refine our analysis, we restricted our dataset to include 1,503 individuals from 64 East and Southeast Asian populations (Figure S3B).

Whole-genome sequencing and filtering

We generated high-coverage whole-genome sequences from approximately 1µg of genomic DNA for 10 Flores samples. We excluded related individuals but sequenced a trio to facilitate haplotype inference (see Table S1; individuals selected for sequencing are marked with an asterisk). Only unrelated individuals were used in downstream analyses. All sequencing work was carried out at the New York Genome Center (NYGC). Sequencing libraries were prepared using TruSeqDNA Nano 350bp kits and 150 bp paired-end reads were obtained on a HiSeq X Ten sequencing platform. Sequenced reads were aligned to the human reference sequence GRCh37 (available from GATK bundle ftp://ftp.broadinstitute.org/bundle/) using the BWA-MEM algorithm (*33*). Duplicates reads were removed using Picard (http://broadinstitute.github.io/picard/). Local realignment around indels and base quality score recalibration were performed using GATK (*34*) to generate the final bam files. Coverage depth was calculated using *bedtools genomecov* (*35*). Each of the genomes was sequenced to a median autosomal depth of 37.8x (range 33 – 49x). QC-passed mapped reads per individual ranged from 654,925,220 to 1,000,842,729 (Table S3).

Variants were called following the GATK3.2-2 pipeline (see GATK "Best Practices" documentation and (*34, 36*) for more details). Calls were obtained on each sample using GATK HaplotypeCaller (default filters were used: Minimum mapping quality = 20, Minimum depth of coverage = 10, Maximum depth = 500, Minimum base quality = 10) and per-sample gVCFs were generated. Subsequently, joint genotyping was performed using GATK GenotypeGVCFs. Variant calls were annotated using *SnpEff 3.4b* (*37*) and *vcftools* (*38*). rsIDs were retrieved from dbSNP Build 141 available at ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b141_GRCh37p13/VCF/All.vcf.gz

All downstream analyses were restricted to SNPs, and unless stated, were carried out using the following filters. First, we identified reliably callable loci from aligned reads with GATKCallableLoci (*34*) considering the following thresholds:

-Minimum base quality of 20

-Minumum mapping quality of 30

-Minimum depth of 10

-Maximum depth equal to the 99.5th percentile of autosomal depth, calculated for each sample.

To ensure overlap among sites for all the individuals, we only considered sites that passed filters in all the individuals. Then, the following sites were masked:

-sites within 5 bp of a short insertion or deletion;

-sites where every individual was heterozygous, following the same approach as in (*39*);

-sites within segmental duplications (*40*) downloaded from:

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/genomicSuperDups.txt.gz-

-sites where at least 18 of 35 overlapping 35-mers from the human reference sequence can be mapped elsewhere with zero or one mismatch as in (*41*)

-sites within a CpG dinucleotide context, as in (*42*)

-sites included in the 1000 Genomes accessibility mask, downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20141020.pilot_mask.whole_genome.bed. This mask was applied in the analyses perfomed in the context of the 1000 Genomes dataset (*43*).

-sites included in the Altai and Denisovan minimal filter mask (*42*), downloaded from: https://bioinf.eva.mpg.de/altai_minimal_filters/

We identified 9,871,310 raw variants which include 8,206,481 SNVs and 1,664,829 INDELs. After applying our callable filters and masking complex regions, 5,335,281 autosomal SNVs were left for the analysis. To assess the quality of our variant calls, we measured the transition/transversion ratio (Ts/Tv) using vcftools --TsTvsummary (*38*). After excluding non-biallelic sites, we found that the Ts/Tv in the entire call set is 2.06, 2.1 for coding and 1.97 for non-coding regions.

To further validate data quality, genotypes obtained from the sequencing data were then compared with those from the HumanOmni2.5-8 v1.1 BeadChip genotyping array, using GATKGenotypeConcordance (*34*). After removing A/T and C/G alleles to prevent flipping-strand problems, INDELs and multiallelic sites, a total of 1,069,991 variants were compared. The correlation between genotypes obtained from the sequencing data and those from the genotyping array was never below 99.8% for any individual.

Haplotype inference

Genotypes were computationally phased using the algorithm implemented in Beagle version 4.0 (*44*), using a panel of 2,504 computationally phased genomes from the 1000 Genomes Project, as reference. Phasing was performed per chromosome for all the autosomes in two steps: (1) we first phased sites in the Flores individuals that were also present in the 1000 Genomes dataset, using the 1000 Genomes sites as the reference panel; (2) sites which were present in the Flores sample but absent in the 1000 Genomes dataset, were then phased using the phased Flores sites as reference panel. Finally, we pooled together these two sets of phased sites. A trio was included in this process as there is a significant number of haplotypes shared among father-mother-child, and this information facilitates phasing of unrelated individuals from the same population. We then merged the fully phased Flores VCFs with the 1000 Genomes phased VCFs. Any sites that were a) unmasked in either the Flores or 1000 Genomes datasets, and b) present in one dataset but absent in the other, were assumed to be homozygous reference.

Integrating WGS with SNP array datasets

We obtained population datasets genotyped on the Affymetrix Human Origins SNP Array (*45-47*) and sequencing data from (*48*). We next merged the Flores sequencing data with the genotyping datasets following the approach as in (*48*). We started by extracting

positions from our multi-VCF that overlap with loci genotyped in the Affymetrix Human Origin SNP Array. Positions not called in the Flores multi-VCF were set as homozygous for the reference allele. Only variant and homozygous sites that passed our callable loci filters were used in the Flores samples (see *Whole-genome sequencing and filtering*). The resulting "Flores callable multi-VCF", filtered for high confidence positions genotyped in the Human Origins Dataset, was converted to PLINK formats using *vcftools* (*38*). The Flores genotypes were then merged with data from the published datasets (see below). We applied the same approach to merge sequencing data from (*48*). We used PLINK version 1.9 for data management (*26*). After removing related individuals up to second degree as identified using KING (*28*) and individuals with a missing genotype rate >3%, our final dataset encompassed 2,507 present day humans from 225 worldwide populations, as well as Neanderthal, Denisovan, and chimpanzee data released in the Human Origin array dataset published in (*45*). After discarding sex-linked, mitochondrial, and multiallelic sites SNPs with ambiguous strand identification and with more than 10% missing call rate, 541,402 autosomal SNPs remained for analysis. This merged and filtered dataset (which we refer to as "WORLD DATASET") was used for the ADMIXTURE analysis (Figure 1C and S4), the archaic principal component analysis (PCA) in Figure 2A and for computing formal tests of admixture. A subset of 769 individuals from 85 populations sampled in East Asia, ISEA and Oceania (see Figure 1A) was further processed to include only individuals with a missing genotype rate less than 3% and SNPs with a missing call rate less than 2% ("SEA DATASET"). We used the SEA DATASET for analysis shown in Figures 1B, S12 and the inbreeding analysis. Population included in the SEA DATASET are listed in Table S4.

Inference of population structure and admixture

To explore Flores genome diversity in the context of genetic variation in East Asia, ISEA and Oceania, we performed a PCA using the SEA DATASET (see above). To avoid the effect of variants in high linkage disequilibrium, we pruned the dataset using PLINK

employing a window of 200 SNPs advanced by 25 SNPs and a r^2 threshold of 0.4 (--indep-pairwise 200 25 0.4), matching parameters previously used (*45*). A PCA was carried out using the R package SNPRelate (*32*) on the pruned SEA DATASET on a total of 168,466 SNPs (Figure 1B).

We used ADMIXTURE (*49*) to infer ancestral clusters in our Flores samples and 225 worldwide populations included in the WORLD DATASET. We pruned our dataset using PLINK following the same approach as above to mitigate the effects of linkage disequilibrium. After pruning, 250,502 SNPs were left for analysis. We ran ADMIXTURE in 10 replicates with different starting points, exploring number of clusters (K) ranging from 2 to 6. The multiple runs were then aligned using the "greedy" algorithm of CLUMPP (*50*) and visualized with the software Distruct (*51*). Figure S4 shows ADMIXTURE results for K from 2 to 6. Our Flores samples are shown in the inset. In Figure 1C of main text we show results for K=6 in a subset of 96 populations selected from the WORLD DATASET.

To quantify the proportion of New Guinean-related ancestry in the Flores population, we used the *F4-ratio statistics*, a formal test of admixture which allows us to infer ancestry proportions in an admixed population, by studying patterns of allele frequency correlations across populations (*52*). We used ADMIXTOOLS (*52*) to compute the *F4-ratio statistics* in the form shown below, as in (*7*). *X* is one of the populations included in the SEA DATASET, excluding New Guinean, Australian and Han (East Asia) populations, as they were used to compute the admixture proportions.

$$\alpha_{NEW\ GUINEAN} = 1 - \frac{f4(Yoruba, Australia;\ X, New\ Guinea)}{f4(Yoruba, Australia;\ \ East\ Asia, New\ Guinea)}$$

Values of $\alpha_{NEW\ GUINEAN}$ (z-*score*>2) are shown for populations in ISEA and Oceania in Figure S5, along with bars representing 2 standard errors of the estimate and on the *x*-axis

in Figure S12. Our ADMIXTURE and *F4-ratio statistics* results are consistent with previous studies based on genome-wide SNP array data (*8, 53*).

We note that although our data represent the first high-coverage genomes from Flores Island, the sampling was performed in a single village and likely does not reflect the diversity of the whole island. Future sampling of the genetic diversity of other villages from the Flores Island would allow comparative studies to be performed.

MSMC analysis

We used an updated version of the multiple sequential Markovian coalescent (MSMC) (*54, 55*), obtained from https://github.com/stschiff/msmc2, to reconstruct the demographic history of the pygmy population or Flores. We performed this analysis on 9 unrelated Flores pygmies, along with 16 previously published high coverage genomes (*42, 56*) downloaded from http://cdna.eva.mpg.de/denisova/. The dataset considered for this analysis include:

- five genomes from African populations (San, Mbuti, Yoruba, Dinka and Mandenka)
- two genomes from European populations (French and Sardinian)
- one genome from a Native American population (Karitiana)
- four genomes from East Asian populations (two Han and two Dai individuals)
- four genomes from populations in Near Oceania (two Papuan and two Australian individuals)

We identified reliably callable loci from aligned reads for each sample, as described in *Whole-genome sequencing and filtering.* We then merged genotype calls for the 16 published genomes and our Flores samples, and considered only sites that passed filters in all the individuals. We masked sites within segmental duplications (*40*) and using a mappability mask (*41*), as also described in *Whole-genome sequencing and filtering*. We phased all genomes using Beagle (*44*), as described in *Haplotype inference*. Sites with ambiguous phasing status in any individual were filtered out from further analysis.

All analyses were performed using default settings, except for the population separation analysis based on four samples where we used --timeSegmentPatter=20*1 to reduce memory and computing time. Results were scaled assuming a mutation rate of $1.25x10^{-8}$ per base pair per generation (*57*) and a generation time of 30 years (*54*).

We inferred population size changes and population separations over time, by randomly selecting one individual from each population (Figure S6 panels A-B). To increase the resolution for populations in East Asia and Oceania, we repeated the analyses using four haplotypes (i.e., two unrelated individuals randomly sampled from Flores, and two individuals from Han, Dai, Papuan, and Australian populations). Four-haplotypes analysis are shown in Figure S6 panels C-D. A deep bottleneck is observed in all non-African populations, as well as in the Flores pygmies, around 60kya, consistent with the time of modern human dispersal into Eurasia, as previously reported (*54, 55, 58*).

We obtained estimates of relative cross coalescence rates (RCCR), which are indicative of the genetic separation between two populations. As expected, the older split is observed between Flores/African, followed by the split between Flores/Eurasian ancestral populations (Figure S6 panel B). Within Eurasian populations, the RCCR curves for pairs of Flores/East Asian versus Flores/European appear separated, consistent with a more recent population history shared with Asian populations. Patterns of divergence of the Flores pygmies from ancestors of East Asian and Oceanic individuals, is more complex, with Flores and Australo-Papuan ancestral populations being closest until about 20kya, and East Asian ancestral showing a lower increase in RCCR going back in time (Figure S6 panel B). However, for more recent times, Flores appears closest to the Dai ancestral population (Figure S6 panel B-D), suggesting recent gene flow from populations of East Asian ancestry.

While we cannot exclude the effect of technical artifacts caused by insufficient phasing quality (*55, 58*) the MSMC results combined with the ADMIXTURE inferences, are consistent with a model in which the Flores pygmies trace most of their ancestry back to a

population close to the ancestors of present day Australo-Papuan populations, and to a more recent admixture event with populations of East Asian ancestry.

Inbreeding analysis

We detected runs of homozygosity across the genome of individuals in the SEA DATASET using PLINK v1.9, following recommendations in (*59*). To minimize calls of spurious ROHs, we removed variants with MAF<0.05 and in LD within a 50 SNP window (--indep 50 5 2), leaving 88,906 SNPs for analysis. Here, we focused on assessing long ROH that likely resulted from recent relatedness in the past ~20 generations. We used a 50 minimum SNP-threshold, allowed for no heterozygote calls, and set other parameters as in (*59*). We identified a total of 9,889 ROHs; the length of ROH ranges from ~1 Mb to 61.3 Mb, with the longest ROH segment found in an individual from the Mengen in New Britain (Oceania) (Figure S7). Within the Flores pygmies we identified a total of 93 ROH segments, ranging from ~1 Mb to 38.7 Mb; the longest ROH was found in individual RPS013, which also harbors the highest cumulative length of ROHs (106.7 Mb) and highest number of segments (15 ROHs). The lowest cumulative ROH length (14.6 Mb) was found in individual RPS015. Both the total number of segments and cumulative ROH length in the Flores samples fall within the range of values seen in other ISEA populations. The median cumulative ROH length in Flores (43.9 Mb) is comparable to median values observed in populations from Borneo, like the Murut (42.6 Mb), Dusun (44.4 Mb) and a sample named "Borneo" (40.6 Mb) (Figure S8).

Y-chromosome and mitochondrial variation

Y-chromosome

Y-chromosome analysis was performed using the variant calls for 5 unrelated males, generated by the NYGC (see *Whole-genome sequencing and filtering*). For each male

individual, we extracted the variant calls using *vcftools* (*38*). To avoid variants that occur in repetitive regions of the chromosome, we only considered variants within the "callable" regions reported by (*60*), while variants were filtered according to stringent criteria (*60*). Haplogroup assignation was performed with in-house scripts searching for all mutations listed in the ISOGG page (release April 21, 2016, revision 7 June 2016). Due to the lack of an established nomenclature system for the Y-chromosome, we referred to different sources. Specifically, haplogroups were defined by 1) their diagnostic SNPs, as suggested by (*61*), 2) by the ISOGG nomenclature and 3) by a combined nomenclature approach employed in a recent publication (*62*).

Four individuals were assigned to haplogroup C1b, (sublineage C1b1a2b1, according to ISOGG, corresponding to diagnostic mutation B465 in a branch referred to as C7b1 reported in (*62*)). Haplogroups within C appear to be the most frequent also in previously published data from Flores (*63*). The fifth Flores individual belongs to haplogroup O3, according to the YCC nomenclature (sub-lineage O2a2b2a2 according to ISOGG diagnostic mutation F706 in a branch referred to as O3i'j, or O-N6 (*62*). The upstream mutation M122 has been also previously found in a sample from Flores (*63*).

mtDNA

In our initial screen of this panel, we generated low-coverage sequencing data for 20 of the 32 Flores individuals. We generated multiplexed shotgun genomic libraries and sequenced them using an Illumina MiSeq sequencer at the University of California Santa Cruz. Of those 20 samples, 12 were pooled and sequenced on a 2x75 paired-end run and 8 of them were pooled and sequenced on a 2x300 paired-end run. Overlapping reads were merged and adapter sequences were trimmed from the reads using SeqPrep (https://github.com/jstjohn/SeqPrep). Reads were mapped to the reference human genome

(hg19) using bwa v0.6.1-r104 (*33*). Samtools v0.1.18 (*64*) was used to sort the reads and remove duplicates. We assembled the mitochondrial genomes for each individual using a reference guided approach. We used the Mapping Iterative Assembler (mia, available at https://github.com/udo-stenzel/mapping-iterative-assembler) (*65*) to generate assemblies by aligning reads from each individual to the mitochondrial reference sequence obtained from hg19 (chrM). From the low-coverage data, we generated assemblies with mean reference coverage between 5.5-27.7x. We examined the bases covering each position in the final iteration and found that the bases at 99% of sequence positions were in 75% agreement or higher.

We used the same strategy to assemble mtDNA sequences from individuals sequenced to high coverage by NYGC. From these data, we generated assemblies with mean coverage between 62.8-95.8x. We examined the bases covering each position in the final iteration, and found that the bases at 99.7% of sequence positions were in 85% agreement or higher.

Consensus sequences were generated for both the high coverage and the low coverage genomes, for a total of 20 individuals sampled. The consensus sequences were merged for the same individuals checking for consistency in the variants found. Four individuals were excluded from the analysis because they were maternally related to the second generation with other individuals. Haplogroup assignment was performed with Haplogrep (*66*) using the nomenclature reported in phylotree.org (*67*).

The maternal haplogroup composition of the Flores pygmies sample, reported in Table S5, with a predominance of haplogroup F1a, is compatible with other data from the island (*63*), but different from the rest of Indonesia (*68*).

A comparative dataset was then assembled with our consensus sequences together with other mtDNA genomes retrieved from the literature for a total of 1,840 individuals (Table S5; *69-77*). Only individual sequences with less than 7 missing sites were considered. A

multiple alignment procedure was performed using MAFFT v7 (http://mafft.cbrc.jp/alignment/software/); the alignment output was then manually checked with Bioedit (www.mbio.ncsu.edu/BioEdit/bioedit.html). From this alignment, 47 populations with a sample size of at least 15 individuals and sampled with non-biased criteria (i.e. not selecting only particular haplogroups of interest) were used for population-based comparisons (at haplogroup and sequence level) (see Table S5 for further details about the dataset used for comparisons).

We visualized F_{ST} distances between populations, included in the comparative dataset, with a Neighbor joining tree (Figure S9 panel A), and displayed the distribution of characteristic haplogroups by means of a Correspondence Analysis (CA) (Figure S9 panel B).

We then computed values of diversity for the full mtDNA genomes dataset, using in house R scripts (Table S5). The presence of very diverse branches influences the nucleotide diversity of the sample, which is the highest amongst the populations from Asia and Oceania considered in the comparative full mtDNA genomes dataset. Similarly, haplotype diversity is also high, with 15 distinct haplotypes over the 16 samples.

The phylogeny of the Flores mitogenomes is visualized with a tree generated using BEAST package v1.8.1 (*78*) (Figure S10 panel A). Runs were performed with two partitions (coding vs. non-coding region), and mutation rates were set to 1.708×10^{-8} substitutions per nucleotide per year for the coding region and 9.883×10^{-8} substitutions per nucleotide per year for the non-coding region, as in (*79*). The best substitution model was determined using jModelTest v2.1.7 (*80*) and resulted in HKY with invariant sites for the non-coding region, and TN93 for the coding region. To determine the best clock model and the best tree model, different runs were performed with BEAST and evaluated with a Bayes Factor analysis (*81*). We tested for constant and skyline tree models, and strict and relaxed clock models. Skyline tree model and Uncorrelated Relaxed clock were the most appropriate to

describe the data (decisive support according to the Bayes Factor analysis (*82*)). Runs were performed with 20 million chains and all the Effective Sample Size (ESS) values retrieved were above 200. The maximum clade credibility was determined using TreeAnnotator and visualized with FigTree (http://tree.bio.ed.ac.uk/software/figtree/). A Bayesian Skyline Plot (BSP) generated from this phylogeny shows that the Flores sample is characterized by a relatively high population size through time (Figure S10 panel B).

The whole comparative dataset of complete mitochondrial sequences available from literature (regardless of sample size or sampling criteria) was considered to trace back connections between single lineages over 72 geographic data points. The limited data from Indonesia might underestimate the amount of population connections for the region. Figure S11 panel A shows connections between individual samples who possess the same haplotype, and therefore share a relatively recent common maternal ancestor. The sample from Flores does not directly share individual sequences with the other data points on the map. Figure S11 panel B displays the connections between haplotypes with F_{ST} distance below 0.0001. Considering almost identical haplotypes extends the time scale of our comparison. In this figure the Flores sample appears to be connected with populations from Sumatra, Taiwan and the Philippines.

In conclusion, the maternal ancestry of the Flores pygmies is characterized by haplogroups common across Southeast Asia and Indonesia (*83*), with a high frequency of F1a, similarly to a previously reported population from Flores (*63*). The population sample appears neither drifted nor isolated. The genetic profile has a primary affinity to continental Asia, but displays more recent connections in terms of haplotype sharing with Taiwan and the Philippines, followed by Near and Remote Oceania.

Inference of archaic hominin ancestry

To explore the relationship between modern humans and archaic hominin species, we performed a PCA on the Altai Neanderthal, Denisovan, and chimpanzee genomes, included in the Human Origin dataset (*45*) and projected present-day humans onto the plane described by the top two principal components. This approach, described in detail in (*84*), allows heterogeneities in similarity between modern humans and archaic humans genomes to be captured (*85, 86*). The PCA was computed using the R package SNPRelate (*32*) on a total of 769 individuals from 85 populations from East Asia, ISEA and Oceania, along with 320 individuals from 28 African populations included in the WORLD DATASET. To facilitate better visualization, we enlarged the central portion of the PCA and plotted the mean values for the top two principal components (PCs) (Figure 2A of main text).

To confirm our inference of genetic similarities between modern humans and archaic hominins, we applied a formal test of admixture. We converted PLINK formats to eigenstrat formats using the "CONVERTF" utility of ADMIXTOOLS v.3 (*52*) and proceed by estimating the proportion of Denisovan ancestry using the *F4-ratio statistic* in the following form, where *X* is a target population in East Asia, ISEA or Oceania:

$$\boldsymbol{\alpha_D} = \frac{f4\ (Yoruba, Altai\ Neandertal; Han\ Chinese, X)}{f4(Yoruba, Altai\ Neandertal; Han\ Chinese, Denisova)}$$

The *F4-ratio statistic* was computed using the “qpF4ratio” software of ADMIXTOOLS, which estimates a standard error using a weighted block jackknife for each estimated quantity (block size set to 5 cM). Values of $\boldsymbol{\alpha_D}$ are shown on the y-axis of Figure S12.

Identifying archaic hominin sequence

To identify archaic sequences in the genomes of the Flores pygmies, we used the statistical framework previously described in (*48*). Namely, we used a three-stage approach

to (1) identify candidate introgressed sequences using the *S** statistics (*87, 88,),* (2) calculate a *p-value* to quantify whether a putative introgressed haplotype matched Denisovan or Neanderthal sequence (*42, 56*) more than expected by chance, and finally (3) refine and probabilistically classify the set of haplotypes. Notably, the first stage, (highlighted in light red in Figure S13), does not rely on the use of ancient DNA sequences from the archaic species. Here, we extended the framework to detect putative introgressed sequences that might be inherited through admixture with extinct hominin species, such as *H. erectus* and *H. floresiensis*, whose prior presence is well documented in Island Southeast Asia but for which there is no genome available. Thus, we categorized the set of haplotypes as Neanderthal, Denisovan, ambiguous and "*unknown*" sequences, with the last category to include haplotypes not introgressed from Neanderthal or Denisovan, but harboring *S** significant haplotypes. A schematic overview of the S* framework used in this study is shown in Figure S13.

To identify Neanderthal and Denisovan introgressed sequences, we used a threshold for archaic match *p-values* to control the FDR ≤5% (i.e., ≤5% of these calls are expected to be non-Neanderthal and non-Denisovan, potentially from other archaic sources and potentially non-introgressed). We proceed by categorizing those haplotypes as Neanderthal, Denisovan, or ambiguous (i.e. haplotypes for which Neanderthal or Denisovan status cannot be confidently distinguished), following the same procedure as in (*48*). The archaic match *p-value* threshold was selected separately for each population. Similarly, we identified putative *unknown* haplotypes using a threshold such that FDR ≤ 5% (i.e., ≤ 5% of these calls are expected to derived from Neanderthal or Denisovan). To be conservative, *unknown* haplotypes were further refined by removing any site overlapping with previously defined Neanderthal, Denisovan, or ambiguous call sets (Figure S14).

Estimating ages of *S haplotypes**

Analysis of pairwise divergence

On average, introgressed haplotypes are expected to have an older time to the most recent common ancestor (TMRCA) compared to non-introgressed genomic regions and to exhibit high levels of divergence. In the specific case of admixture with hominin species, such as *H. erectus* and *H. floresiensis*, which diverged at least 1 My from the common ancestor of Neanderthal, Denisovan and modern humans, we expect to find putative *"floresiensis/erectus"* haplotypes that are more divergent compared to modern humans, Neanderthal and Denisovan haplotypes.

To investigate the presence of truly archaic introgressed haplotypes in our call set, we estimated the TMRCA between the *S** putative introgressed haplotypes and non-introgressed haplotypes. Namely, we computed pairwise divergence between haplotypes identified using the *S** framework (i.e. Neanderthal, Denisovan, ambiguous and unknown categories) and each of the remaining non-introgressed haplotypes. To avoid comparisons between two introgressed haplotypes, haplotypes carrying *S** haplotypes for the same region were excluded in each comparison. For each haplotype category (Neanderthal, Denisovan, ambiguous and unknown) we estimated pairwise divergence using a triangulation approach, which measures the putative introgressed haplotype's "missing" shared substitutions with a non-introgressed haplotype. Importantly, this method of estimating divergence does not rely on the detection of rare alleles on the putative archaic haplotype, which may fail to pass all genotyping filters or may be the result of sequencing error. The resulting divergence estimates were then normalized using the distance from the chimpanzee reference sequence (panTro4) to account for local variation in mutation rate. The hg19/PanTro4 pairwise alignment was downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro4/).

After excluding regions with fewer than 50 informative sites, the deepest divergence found between the *S** segment and the remaining haplotypes in the Flores panel was taken

as the TMRCA for the region. This comparison was repeated for the *S** haplotypes previously identified in 27 Melanesians, 10 East Asians, and 10 Europeans.

Furthermore, to assess how the TMRCA distributions for S* haplotypes differ from those for non-S* haplotypes, we randomly shuffled the S* regions to new genomic regions and estimated the TMRCAs using the same procedure as described above. To avoid placing regions in unassembled or repetitive genomic regions, we shuffled our S* segments using *bedtools shuffle* (*35*) to regions on the same chromosome that passed our filtering parameters.

Figure S15 shows the distributions of the TMRCA (maximum estimated pairwise divergence) computed for each *S** haplotype by classification (Neanderthal, Denisovan, ambiguous, and unknown) and by population. We find that the average estimated TMRCA for the S* haplotypes classified as "unknown" origin was lower compared to TMRCAs for haplotypes classified as Neanderthal, Denisovan, or ambiguous, but higher than average TMRCAs for random genomic regions (in grey). This same TMRCA pattern (i.e. random < unknown < Neanderthal / Denisovan / ambiguous) is observed in all four populations under study. The intermediate position of the distribution of TMRCAs for unknown haplotypes suggest that the unknown category might be enriched for Neanderthal or Denisovan haplotypes (e.g. sequences absent from the current Neanderthal and Denisovan reference genomes, and/or sequences that did not pass our match p-value thresholds), as well as a number of false positive.

Among the S* categories, it is worth noting that the ambiguous haplotypes show TMRCA estimates similar to those for Neanderthal and Denisovan haplotypes, consistent with the inclusion in this ambiguous category of introgressed haplotypes for which the Neanderthal or Denisovan status cannot be confidently distinguished. Further, we noted a striking increase in TMRCAs for Denisovan S* haplotypes in the East Asian sample, relative to Neanderthal and Denisovan haplotypes found in other populations. This observation is

consistent with the hypothesis of containing sequences inherited through admixture with a distinct Denisovan group into East Asians. While confirming the common origin of the Denisovan ancestry in both Melanesian and Flores populations, our analysis also supports the recent finding of a distinct Denisovan admixture event in the ancestors of East Asians (*89*).

ARGweaver analysis

To further explore whether the unknown *S** sequences (i.e. not matching Neanderthal or Denisovan reference genomes) contain a signature of admixture with a divergent hominin lineage, we calculated TMRCA using ARGweaver (*90*) as described in (*91*). The rationale of this analysis is that if the ancestors of the Flores pygmies admixed with an archaic hominin group that diverged with modern humans prior to Neanderthals (e.g. *H. erectus* or *H. floresiensis*), then we would expect some haplotypes in the Flores pygmies to coalesce above haplotypes from other modern humans, and above haplotypes from Neanderthal and Denisovan (Figure S16 panel A). Moreover, we would expect these haplotypes to be enriched in the "unknown" portion of the S* call set, as they would match neither Neanderthal or Denisovan. We have previously observed that the S* call set has ~50% FDR for calling introgressed sequence from archaic groups such as Neanderthals (*48*). Furthermore, we previously demonstrated, through simulations, that the power of *S** increases with divergence time of the "archaic" hominin population (see Figure 3C in *(39)*), suggesting that in the case of introgression with a more divergent hominin lineage, the *unknown* portion would be further enriched for introgressed sequence. If a substantial number of such haplotypes exist, we would expect the unknown portion in Flores pygmies to contain a significant number of haplotypes with older TMRCAs. We then test if the distribution of TMRCAs for the older haplotypes is significantly different between the *unknown* portion and the Neanderthal portion of the *S** call set.

We first calculated the TMRCA of six African genomes (two Yorubans, two San, and two Mbuti; HGDP00927, SS6004475, SS6004471, HGDP0456, HGDP01029, SS6004473 from (*42, 56*)), the Altai Neanderthal and Denisovan genomes (*42, 56*) and our nine unrelated Flores genomes. ARGweaver was run over the above 17 genomes plus chimpanzee (panTro4) in 50kb windows. We computed 5,000 MCMC iterations, with sampling every 20 iterations, starting at iteration 2,000. We then calculated the average TMRCA of each 50kb window from the previous *S** analysis. This TMRCA was calculated for all 17 genomes ("TMRCA(all)") and separately for the six modern humans plus two archaic genomes ("TMRCA (all-Flores)"), i.e., excluding the Flores genomes.

In general, there is a strong correspondence between TMRCA(all) and TMRCA(all-Flores), across all 50kb windows (Figure S16 panel B; Pearson's correlation = 0.998). We examined the differences between these TMRCAs and identified regions where the addition of Flores increases the TMRCA of a region above the coalescence of other modern humans, Neanderthal, and Denisovan (Figure S16 panel C). We can formally test the differences between these distributions, using a kernel density-based test for differences between multidimensional distributions (*92*). The null hypothesis that these TMRCAs are drawn from the same distribution cannot be rejected (p=0.53), suggesting that the set of *unknown S** sequences is not significantly enriched for regions with a different TMRCA distribution in the Flores pygmies genomes. It is possible that even if introgression occurred, the number of surviving introgressed haplotypes may be quite small. This is especially plausible, given the fact that there is some evidence for reproductive barriers between Neanderthals and modern humans (*88, 93*) potentially leading to the rapid removal of introgressed haplotypes shortly after introgression (*94, 95*). This analysis cannot exclude a similar scenario involving low levels of introgression from *H. floresiensis* or other deeply diverged hominin species.

Analysis of copy number variation

We analyzed 45 genomes (10 Indonesian individuals from Flores and 35 Melanesians from (*48*)) for copy-number variation (CNV) using whole genome shotgun sequence read depth detection (WSSD) as previously described in (*96*). Initial QC analysis showed that all Indonesians and most (27/35) Melanesian genomes had >90% of copy number 2 regions correctly called (Figure S17), whereas the remaining (n=8) Melanesian genomes had >85% of CP2 regions correctly called.

We initially focused on a previously described segmental duplication block that introgressed from Denisovan into the ancestors of present-day Papuans ~40 kya (Figure 2C, chromosome 16p12.2; *15*). Sudmant et al. (*15*) described a set of four duplicated loci in this region (A, B, C, D) that are polymorphic among human populations and missing from the GRCh37 and GRCh38 human references. WSSD and paired end read analysis predict that the majority of humans have a duplication block that includes duplications A and C as adjacent segments, but Denisovan and the majority of Papuans have an expanded duplication block that includes A, B, C and D as adjacent segmental duplications. This complex duplication structure is absent from all other previously described human populations.

Here we assessed the chromosome 16p12.2 segmental duplication block using WSSD in a panel of 86 genomes. The panel includes:

- 9 Flores unrelated individuals;
- 49 Oceanic individuals (27 Melanesians from (*48*), and 22 Papuans from the Simons Genome Diversity Project (SGDP; *15*));
- 4 Africans from SGDP;
- 4 Americans from SGDP;
- 15 Eurasians from SGDP;
- 3 ancient humans (*45, 97*);
- the Neanderthal and Denisovan genomes (*42, 56*).

As shown in Figure S18 panel A, structure A/C in the chromosome 16p12.2 duplication, is found in all individuals, while B/D is only observed in the Denisovan, Oceanic and Flores genomes. Using segmental duplication D (Figures 2C and S18 panel B, hg19: chr16:22710749-22782167)(*15*) as a proxy, we estimate that the Denisovan duplication is present at an allele frequency (AF) of 50% in the Flores population and at higher frequency in populations in Oceania (79.6% AF in the 27 unrelated Melanesians from (*48*) and 82.6% AF (n=43), when combined with Papuan individuals from the SGDP). Thus, the Flores population represents, to date, the second modern human group known to carry this large (>220 kbp) segmental duplication block at high frequency.

In an effort to identify additional CNVs specific to the Flores population, we performed digital comparative genomic hybridization (dCGH) comparing the 45 Flores and Melanesian genomes against 17 high quality genomes from SGDP (individuals listed in Table S7 in (*15*)). Because CNV calling is subject to high FDR, we examined inheritance of biallelic CNV calls in the Flores trio (RPS031, RPS018, and RPS020) as a function of size. We found a Mendelian error rate of 4.8% for deletion and 21.3% for duplication calls, which fell to 3.2% and 10.2%, respectively, for calls greater than 10 kbp. Overall, we considered three classes of CNVs genome-wide: biallelic deletions, biallelic duplications, and multiallelic CNVs (Table S6). To identify population-specific CNVs, we searched for variants present in multiple unrelated Indonesians that didn't overlap any SGDP calls from (*15)*, excluding samples that showed evidence of relatedness (RPS020, UV573, UV952, UV927, UV979, UV1266, UV946, UV956, and UV1042), for a total of 9 Indonesians and 27 Papuan samples. A small number of deletions (n=13) and duplications (n=24) were identified that were large (>10 kbp) and present in more than one Flores individual (Table S7).

Considering related samples, we identified an inherited deletion of ~300 kbp present in two Indonesians that intersected *MALRD1* (Figure S19), and putative deletions involving the *ILKAP* (Figure S20), *MYO5C* (Figure S21), and *XIRP2* (Figure S22) genes. With respect

to duplications, we predict a ~150 kbp duplication affecting the *RBM44*, *RAMP1*, and part of *LRRFIP1* (Figure S23), and another duplication that intersected *GOLGA3* and *CHFR* (Figure S24). Only the *XIRP2* and *ILKAP* CNVs were present in multiple unrelated Indonesians, suggesting that most CNVs are rare rather than population-specific.

Scan for recent positive selection

Whole-genome data from the Flores individuals of Rampasasa village (RPS) provides an opportunity to investigate genomic evidence of recent adaptation in this Flores pygmy population. To this end, we applied the population branch statistic (*PBS*) (*23*, *98, 99*) to scan for alleles that are highly differentiated in RPS compared to related populations. *PBS* quantifies the amount of allele frequency change that occurred in the target lineage since its divergence with the reference populations under comparison, and extreme frequency changes may reflect the action of natural selection. To achieve resolution over recent time scales, we used genomes from 27 unrelated Melanesian individuals (PNG) (*48*) as comparison samples and genomes from 103 Han Chinese (CHB) individuals (*43*) as more distantly related outgroup samples.

As selection scans are sensitive to spurious genotype calls, we applied stringent masks to restrict the data to confidently called sites. These sites comprised the intersection of the 1000 Genomes accessibility mask (see ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/accessible_genome_masks/README_20120824_accessibility_mask_bed_files) and sites that passed our filters for the Flores data (see *Whole-genome sequencing and filtering*). We further removed sites for which all genotyped Flores individuals were called heterozygotes, as these are likely to represent mapping errors to repetitive regions. We also removed sites with evidence of strand bias, sites within 5 bp of indels, sites in annotated segmental duplications, as well as sites in regions of low mappability. We note that while only SNPs

were used in the PBS scan, selection targeting indels or structural variants would also be detected if these variants segregate in linkage disequilibrium with tag SNPs in the region.

We then calculated the population branch statistic (*PBS*), which is based on levels of differentiation in allele frequencies across populations, as quantified by the fixation index (F_{ST}) (*100*). F_{ST} is proportional to the evolutionary branch length (*T*) between each pair of populations:

$$T = -\log(1 - F_{ST})$$

For each SNP that segregates in all three populations, *PBS* for the Flores branch is calculated as:

$$PBS = \frac{T_{RPS\times PNG} + T_{CHB\times RPS} - T_{PNG\times CHB}}{2}$$

Per-SNP *PBS* was then averaged in 20 SNP sliding windows with 5 SNP steps.

We identified two genomic regions with extreme signatures of positive selection based on *PBS* (Table S8; Figure S25). One of these regions, falling on chromosome 6, encompasses the human leukocyte antigen (HLA) gene complex. Genes in the HLA play a key role in the adaptive immune system and are known to be subject to strong diversifying selection across human populations (*101, 102*). The strongest *PBS* signal, however, extends over an ~74 Kb region of chromosome 11 that includes *FADS1* and *FADS2*—genes previously implicated as targets of selection in distinct human populations (Figure 3A; *103*). These genes encode fatty acid desaturase (FADS) enzymes that regulate metabolism of long-chain polyunsaturated fatty acids (LC-PUFA), key precursor molecules with diverse biological roles (*104*).

Across worldwide populations, the FADS region contains two major LD blocks—the first encompassing all of *FADS1* and the first half of *FADS2* and the second containing the

latter half of *FADS2* (*105*). While polymorphic across most of Eurasia, a derived haplotype in the first LD block is thought to have swept to near fixation in Africa approximately 85 kya (*21*). This event was hypothesized to have allowed a transition to a more plant-based diet, as the derived haplotype is associated with enhanced conversion of MC-PUFA to LC-PUFA through increased activity of *FADS1*. However, even early-diverging African populations, such as San and Mbuti, appear fixed for derived FADS haplotypes (*58*), raised the intriguing possibility that the sweep may have been complete at the time of the Out-of-Africa migration and that the ancestral haplotype was reintroduced to non-African populations via introgression from an archaic lineage. Refuting this hypothesis, Buckley et al. (*19*) demonstrated that Neanderthal haplotypes cluster most closely with but are highly divergent from the derived haplotypes observed in modern humans, while Denisovan haplotypes cluster most closely with but are highly divergent from ancestral haplotypes observed in modern humans. These data are inconsistent with a simple model of adaptive introgression from a lineage closely related to either archaic species. Instead, the unique topology of divergent trans-specific haplotype clusters suggests potential long-term balancing selection operating at this locus (*19*). Under such a model, functionally distinct haplotypes may be maintained by persistent environmental heterogeneity, which in this case may consist of temporal and geographic variability in diet.

Consistent with a model of recurrent local adaptation on ancient standing variation, derived haplotypes also attain high frequencies in certain South Asian populations, potentially due to selection in response to a vegetarian diet (*22*). Meanwhile, Fumagalli et al. (*23*) demonstrated that ancestral haplotypes are nearly fixed in Greenlandic Inuits, potentially in response to a marine diet rich in omega-3 fatty acids.

Our data add a new chapter to this complex narrative, revealing that the Flores sample is nearly fixed for ancestral haplotypes in LD block 1 (tagged by SNP rs174547) in a pattern consistent with a recent selective sweep (Figure 3B; *106*). This result was

confirmed in the larger Omni 2.5-genotyped sample (*n* = 21 unrelated individuals). The Greenlandic Inuit study (*23*) was based on microarray data and therefore precludes comprehensive comparison to the Flores genome sequences. Nevertheless, all four haplotypes from the genomes of two Greenlandic Inuit individuals published by Raghavan et al. (*107)* match the high-frequency Flores haplotypes described in our study, particularly in LD block 1 (Figure S26). In contrast to the Flores data, however, the Greenlandic Inuit *PBS* scan of Fumagalli et al. (*23*) achieved a maximum in LD block 2, downstream of the Flores PBS signal (Figure S26), providing further evidence that these selection events were likely independent.

Fumagalli et al. (*23*) and Amorim et al. (*108*) reported that the ancestral FADS haplotypes at high frequency in Greenlandic Inuits also segregate at high frequencies in most other Native American populations. This led the authors to hypothesize that the selective sweep may have occurred in Beringia, prior to migration into the Americas and the subsequent differentiation of Native American populations. Analogous to this observation, we found that ancestral haplotypes matching those observed in the Flores sample also segregate at relatively high frequencies in other Southeast Asian populations, including the Chinese Dai in Xishuangbanna, China (CDX) and Kinh in Ho Chi Minh City, Vietnam (KHV) (Figure 3B; *43*). Additional Southeast Asian populations genotyped using the Human Origins SNP chip (*45-47*; see *Integrating WGS with SNP array datasets*) also carry high frequencies of ancestral alleles at SNPs tagging the ancestral haplotypes (Figure 3B, inset; *n* >= 8). This is reflected in the observation that the FADS region also contains the highest scoring window in a genome-wide *PBS* scan in which the Flores population is swapped with the KHV population and compared to Papuan (PNG) and Han Chinese (CHB) reference populations (chr11:61548559-61598288, mean *PBS* = 0.392). Yet the highest scoring window in the region remains in the top 5 x 10^{-4} quantile of genome-wide PBS values in a PBS scan comparing RPS to KHV and CHB (chr11:61548559-61598288, mean *PBS* = 0.977). Taken

together, these observations are consistent with positive selection in an ancestral Southeast Asian population, with drift and additional selection potentially occurring in Flores and other Southeast Asian populations subsequent to divergence.

The high level of divergence between the two haplogroups presents a challenge for identifying the causal variants under selection. SNPs distinguishing the haplogroups, such as rs174547, are strongly associated with circulating levels of fatty acids based on previous GWAS studies (Table S9; *109 -112*), as well as various blood phenotypes based on PheWAS data from the UK Biobank (Table S10; *113*).

As previously described (*19*) data from the GTEx Consortium (*20*) suggest regulatory differences among the haplotypes, which are defined by variants that are strong eQTLs of both *FADS1* and *FADS2*. These associations predict that the Flores individuals have decreased expression of *FADS1*, reducing their capacity to synthesize LC-PUFA from plant-based precursors. Kothapalli et al. (*22*) recently identified the potential causal variant at this locus, demonstrating a regulatory effect of a 22-bp insertion (rs66698963), which they hypothesized to have been the target of positive selection in South Asian populations. Specifically, this region contains binding sites for a sterol regulatory element binding protein (SREBP) and acts as a master switch controlling the expression of *FADS1* and *FADS2*. To determine indel genotypes at rs66698963, the Flores samples were amplified using *FADS2* primers and PCR conditions as described in (*114*). PCR products were run on 2% agarose gels at 100V for 3 hours and compared to a 100 bp DNA Ladder (NEB). Gels were ethidium-bromide stained and visualized under UV light. Samples producing a single 607 bp band were scored as homozygous for the 22 bp deletion, and those with multiple bands were scored as heterozygous. We found that 28 of 31 successfully genotyped individuals were homozygous for the deletion allele. This finding is consistent with a model in which the deletion allele of rs66698963 may have been targeted by selection in ancestors of the Flores population.

In conclusion, we find evidence of population-specific selection spanning the *FADS1* and *FADS2* region—adding to emerging evidence that this region was a recurrent target of selection in diverse human populations, possibly in response to changing diet.

Inference of polygenic selection for reduced stature

Quality control of genotype data

We used data from the Flores Omni 2.5 genotypes, generated in this study (see *Omni 2.5 array genotyping*), and from the Human Genome Diversity Panel (HGDP). Alleles were aligned to the forward strand using the Genotype Harmonizer software (*115*) and the 1000 Genomes phased version 5 genotypes as reference set. We downloaded the 1000 Genomes Omni 2.5 genotypes (which are aligned on the forward strand, and available at http://ftp.1000genomes.ebi.ac.uk/ALL.chip.omni_broad_sanger_combined.20140818.snps.genotypes.vcf.gz) and selected individuals from CHS (Southern Han Chinese), CDX (Dai in Xishuangbanna) and KHV (Kinh in Viet Nam) populations. The three sets of genotypes were then independently checked for strand alignment, alleles, position, reference assignments and frequency differences from the 1000 Genomes sequence data using the Genotype Harmonizer.

We removed any loci with minor allele frequency (MAF) differences greater than 0.2 among populations and greater than 0.2 as compared to the reference. We also removed palindromic SNPs with frequency differences to the reference greater than 0.4, identified any further strand flips, and removed loci of MAF less than 1%.

To maximize the number of overlapping SNPs between the HGDP data and the other Omni genotyped data, we then imputed the HGDP genotypes to 1000 Genomes reference panel and repeated the quality control of the imputed data, using the 1000 Genomes sequence data as the reference (details about the imputation pipeline can be found at https://github.com/CNSGenomics/impute-pipe) (*116, 117*). We then combined the genotype

data together, selecting overlapping SNP loci with less than 5% missing data and individuals from the HGDP data from the Melanesian (Bougainville Island), Papua New Guinean and Cambodian populations. This left 448,068 SNPs and 453 individuals from the 1000 Genomes CHS (*n*=153), CDX (*n* =100) and KHV (*n* =121) populations, the Human Genetic Diversity Panel Melanesian (*n* =19), Papua New Guinea (*n* =17), and Cambodian (*n* =11) populations, and the Flores (*n* =32) population. Our approach examines the combined effect of polygenic selection at multiple segregating common loci. Thus, while this extensive quality control will likely remove loci that show a very high level of true differentiation among populations, we prefer a conservative approach where we ensure that a limited number (<5,000) of extremely differentiated SNPs do not influence the results we present here.

Height-associated loci

We extracted 456,426 participants of European ancestry from the UK Biobank (*113*). Ancestry was inferred using a two-stage approach. The first stage consisted in projecting each study participant onto the first two genotypic principal components (PC) calculated from HapMap 3 SNPs genotyped in 2,504 participants of the 1,000 Genomes Project (ftp://ftp.ncbi.nlm.nih.gov/hapmap/). We then used five super-populations (European, African, East Asian, South Asian and Admixed) as reference, and assigned each participant to the closest population. Distance was defined as the posterior probability under a bivariate Gaussian distribution of each participant to belong to one of the five super-populations. This method generalizes the k-means method and considers the shape the reference cluster to improve the clustering. Vectors of means and *2×2* variance-covariance matrices were calculated for each super-population and we used a uniform prior. We used SNPs with an imputation quality score above 0.3 and hard-called the genotypes with a posterior probability larger than 0.9. We then kept for analysis SNPs in the Haplotype Reference Consortium (HRS) (*118*) with call rate >0.95, a minor allele frequency >0.0001, and Hardy-Weinberg

test p-value larger than 10^{-6}; this left 16,653,239 SNPs for analysis. We ran a genome-wide association study of height in the 456,426 individuals using mixed linear model association (MLMA) testing implemented in BOLT 2.2 software (*119*) assuming an infinitesimal model. We used ~700,000 HapMap 3 SNPs (LD pruned for SNPs with $r^2 > 0.9$) as model SNPs in our analysis. Height was adjusted for age, genotypic batches and 10 PCs calculated using ~78,000 genotyped SNPs pre-selected by the UK Biobank quality control team for principal component analysis (details about UK Biobank data quality control can be found at www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf) .

PCs were calculated using flashPCA (*120*). This analysis controls for population stratification and yields SNP effects that are unbiased of population stratification (*119, 121*). We then used the PLINK clumping approach to select a genome wide set of SNPs in approximate linkage equilibrium (LE: LD $R^2 < 0.05$ within a 2 Mb window).

Analysis of population genetic differentiation

Selection on loci associated with height has been reported before in European populations (*24*, *122*) including selection for reduced height on the island population of Sardinia (*123*). We followed the approach in (*24*) to estimate population genetic differentiation of height in the 453 individuals from the 7 populations in Southeast Asia and Oceania. Briefly, we created a genetic predictor of height from the SNP loci identified in the MLMA analysis, we estimate the population mean value of the predictor and the population genetic variation in a Bayesian linear model and compared our estimates to the values from a null quantitative genetic model of multivariate population differentiation. The null model was constructed, as in (*24*). Briefly, the SNP effects were randomized across all SNPs 1,000 times, and 1,000 genetic predictors were created in the Flores island and neighbouring population samples. By keeping the effect sizes consistent but attributing these effects across SNPs at random,

the genetic predictors generated reflect the action of genetic drift. Second, each set of genetic predictors was standardized to a z-score and used as a response variable in the Bayesian mixed-effects model outlined in equation (2.2) of (*24*). This provided 1,000 estimates of the population genetic variance and population means under drift; these values are displayed in the figures as the estimates from the neutral model. Third, the across-population sample covariance matrix of these 1,000 estimates was calculated, which provided an estimate of the expected population-level covariance in phenotype under drift. We then used a Mahalanobis distance statistic to provide a measure of the relative deviation of our predicted population-level means from their multivariate theoretical expectations under drift (see Supplementary Note in (*24*)). This calculation provided the test statistic used to compare our predicted estimates to the expected values under drift. As both the drift profile and trait profile scores were transformed to a z-score, this comparison was on the same standard deviation scale. This tests whether the difference in predicted genetic value of Flores compared to all of the other neighbouring population was larger than expected given the maximum amount of variation in a genetic predictor that could be created across populations under our neutral model.

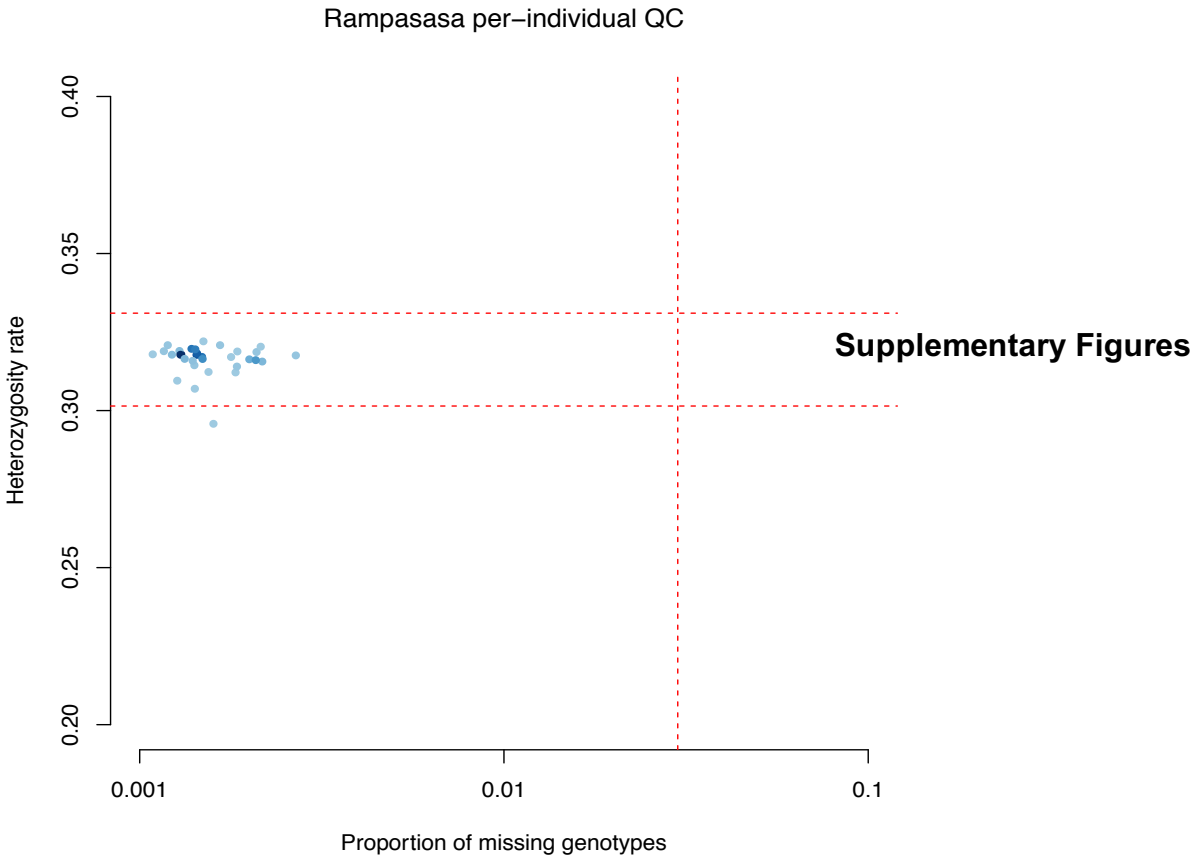
We then compared our estimated population genetic differentiation to the null using a multivariate chi-squared test. Figure 4A shows a lower predicted genetic value for the Flores population as compared to their neighbouring populations. It also shows that this difference is significantly larger than expected under neutrality (a significant departure from a neutral model). A lower predicted value for Flores occurs because, on average, the common loci comprising the genetic predictor are differentiated in a direction that is consistent with the direction of their effects on height, which in turn creates differences among populations in a genetic predictor. So in this example, the predictor is lower because height increasing loci occur less frequently than in surrounding populations, and less frequently than expected under drift.

We then calculated allele frequency differences between the Flores population and the 1000 Genomes populations using the software GCTA (http://cnsgenomics.com/software/gcta/) (*124*) and tested for a relationship between effect size and allele frequency difference at 4,000 loci of smallest p-value from the MLMA analysis. Figure 4B is in essence a comparison of F_{ST} between Flores and neighbouring populations at trait-associated loci (*y*-axis) and the effect size at those loci from a European reference population (*x*-axis). It shows an association between F_{ST} and both the strength and direction of an allele's effect on height: the larger the effect size of the increasing allele, the more differentiated its frequency is from neighbouring populations, in a directional manner (lower frequency).

Finally, we used a linear model to test for an association between the genetic predictor of height and the phenotype recorded in the Flores data (Figure 4C). The linear model assumes that each data point is independent, which is not the case in the Flores data because of the presence of close relatives who share genetic and environmental effects.

As we simply wished to determine whether the predictor was associated with Flores phenotypic value, we provide an approximate standard error for the proportion of phenotypic variance explained by the regression that we obtain from simulation. For the simulation, we created a set of data with 32 individuals that contained 10 pairs of close relatives. For simplicity, we assumed the 10 pairs were first-degree relatives, whose phenotypic covariance at a genetic predictor $= 0.5var(\hat{g}) + var(c)$ with $var(\hat{g})$ the genetic variance and $var(c)$ the common or shared environment effects within families. Assuming the covariance of first-degree relatives is 0.5 at a genetic predictor obtained from a phenotype of variance 1, then $var(c) = 0.5 - 0.5var(\hat{g})$, and assuming $var(\hat{g}) = 0.085$, which is the phenotypic variance explained, then $var(c) = 0.458$. We assumed the phenotypic covariance among unrelated individuals was zero and simulated a phenotype, $\boldsymbol{y} = \hat{\boldsymbol{g}} + \boldsymbol{r} + \boldsymbol{e}$, where $\boldsymbol{y}, \hat{\boldsymbol{g}}, \boldsymbol{r}$ and $\boldsymbol{e}$ (the individual-level environmental values), are normally distributed

random variables with variances 1, 0.085, 0.458, and 0.457 respectively. We calculated the standard error of the proportion of phenotypic variance explained by $\hat{\boldsymbol{g}}$ in a linear model across 1,000 simulated datasets.



Supplementary Figures

Figure S1. **Genotype failure rate vs. heterozygosity across all 32 individuals in the OMNI Dataset.** Shading indicates sample density and dashed lines denote QC thresholds.

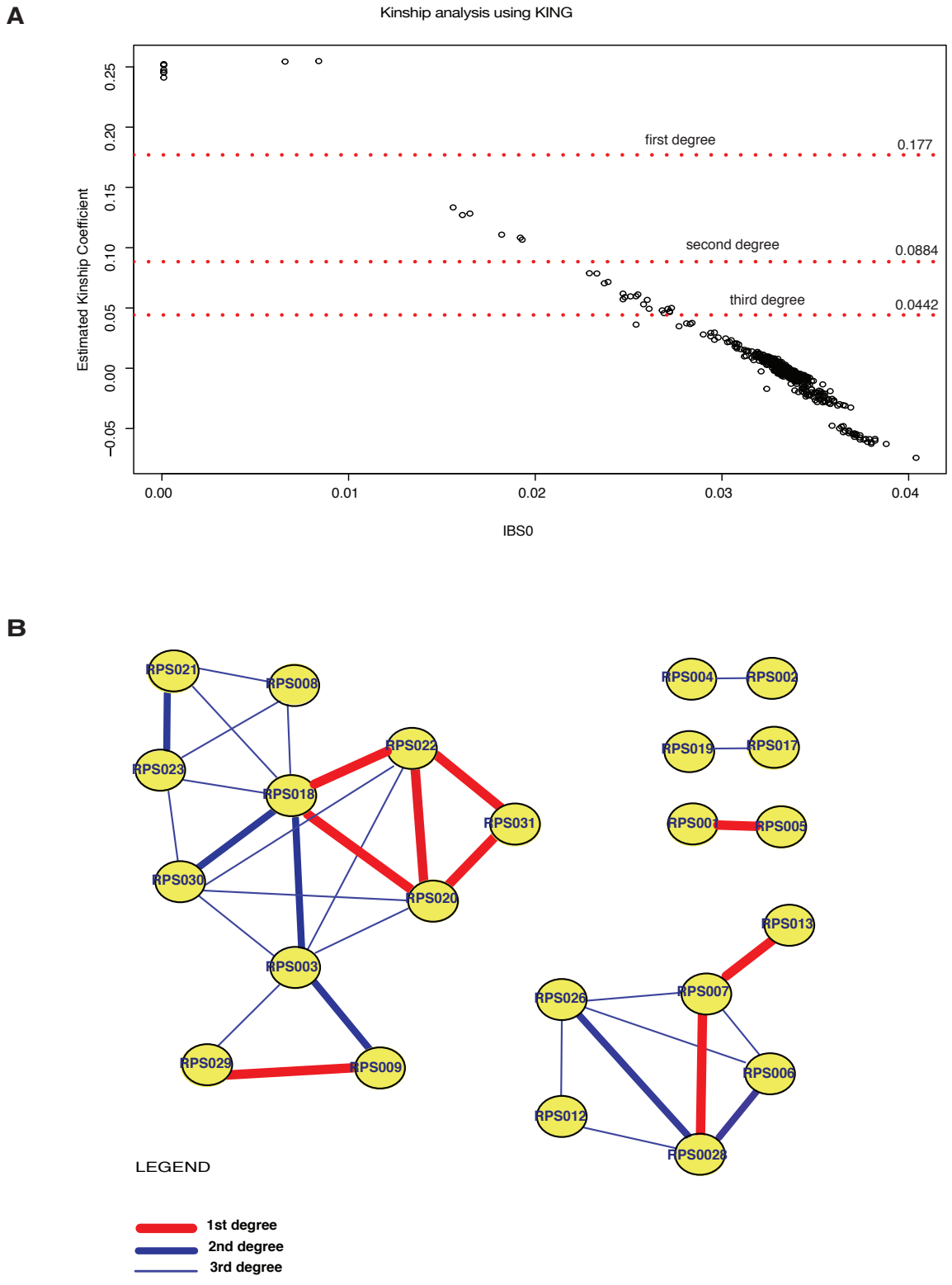


Figure S2. Relatedness analysis. (**A**) Kinship coefficient estimated using KING is plotted against the proportion of zero IBS-sharing. Dashed red lines correspond to the thresholds for first, second and third degree relationships, respectively, based on KING documentation. (**B**) Network of family relationships as inferred in our analysis. Vertices represent individuals. Edge colors and thickness represent relationships inferred using KING, as specified in the legend.

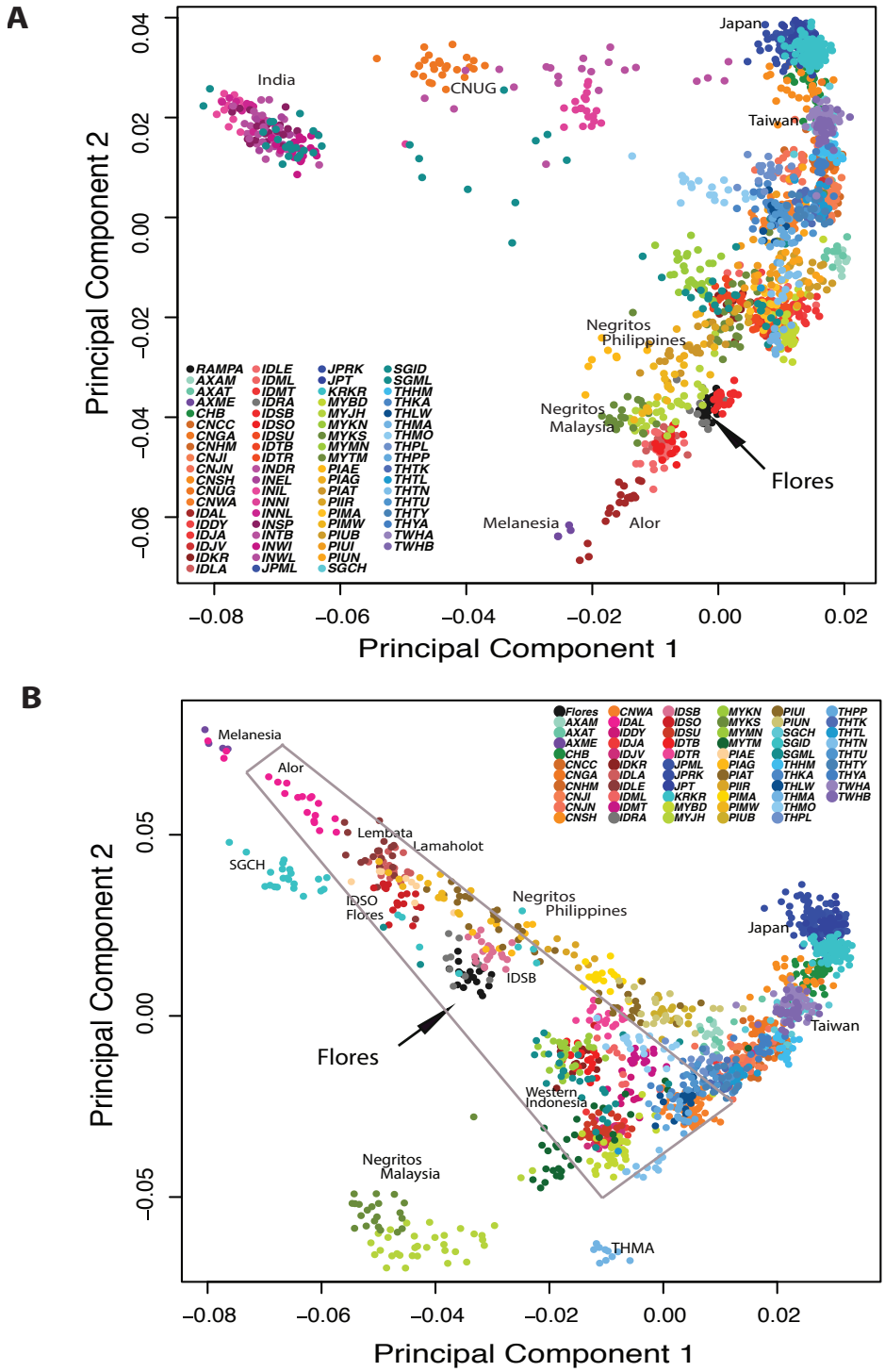


Figure S3. Flores in the context of Asian genetic diversity. Principal Component Analysis performed on (**A**) 74 populations from South Asia, East Asia, ISEA and Oceania and (**B**) on a subset of 64 populations from East Asian, ISEA and Oceania, included in the PANASIA Dataset. Indonesian populations (in reds) as shown inside the polygon. The Flores individuals under study are shown in black.

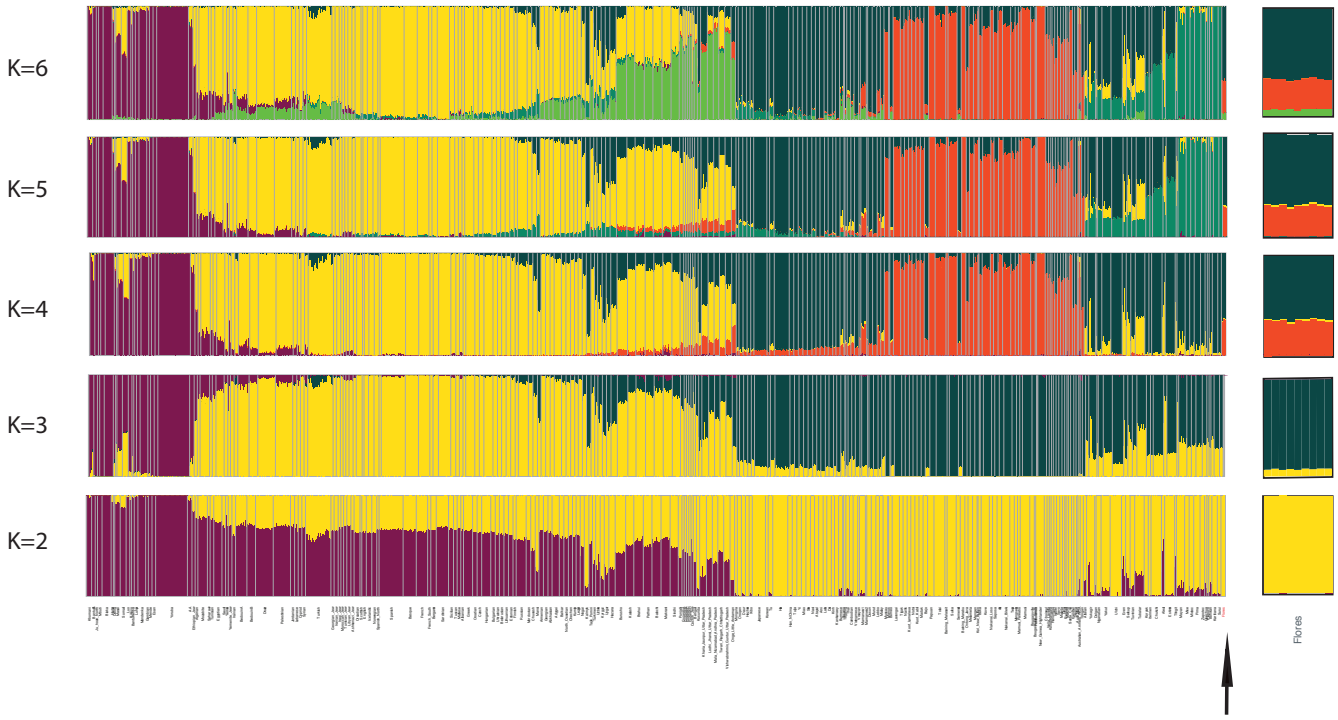


Figure S4. ADMIXTURE analysis of 2,507 individuals from 225 wolrdwide populations. We refer to this panel as the "WORLD DATASET", and show results for K from 2 to 6. Flores samples are shown in the inset.

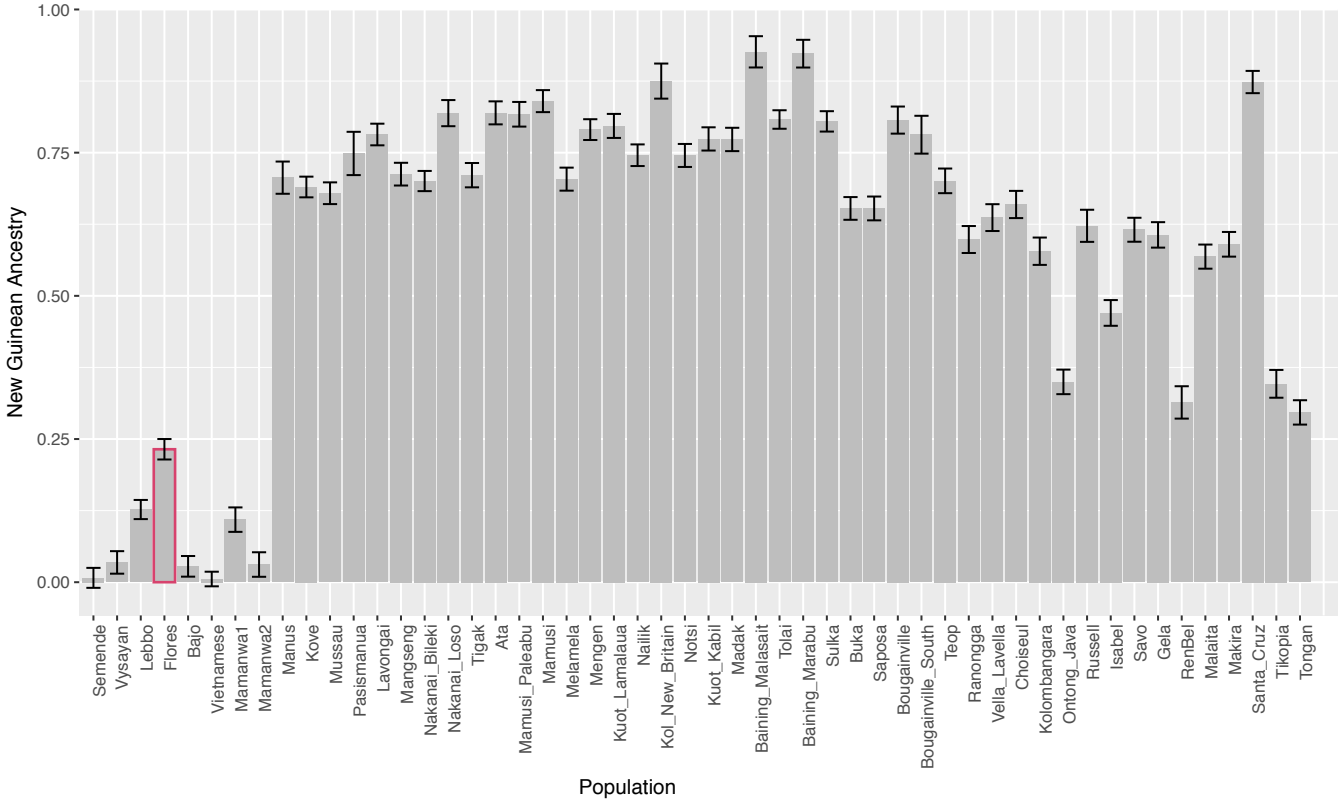


Figure S5. Proportion of New Guinean ancestry in populations in ISEA and Oceania, estimated using the *F4*-ratio statistics. Populations are ordered based on increasing longitude. Only values with z-score > 2 were reported. Bars represent 2 standard errors. The Flores pygmy population is highlighted in red.

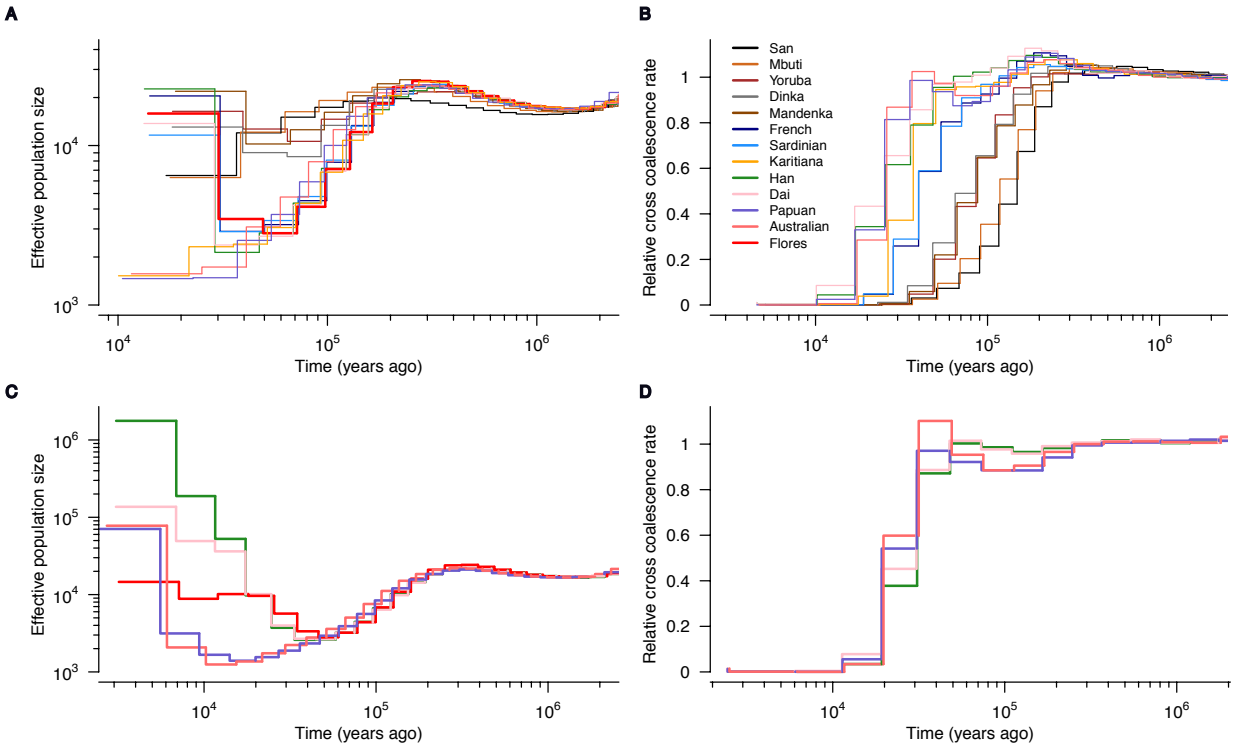


Figure S6. MSMC analysis of effective population sizes. Estimates of effective population sizes and relative cross coalescence rate as a function of time in thousands of years ago (ka) between Flores pygmies and previously published diploid high coverage genomes, estimated using MSMC, from two-haplotypes (**A** and **B**), and from four-haplotypes (**C** and **D**).

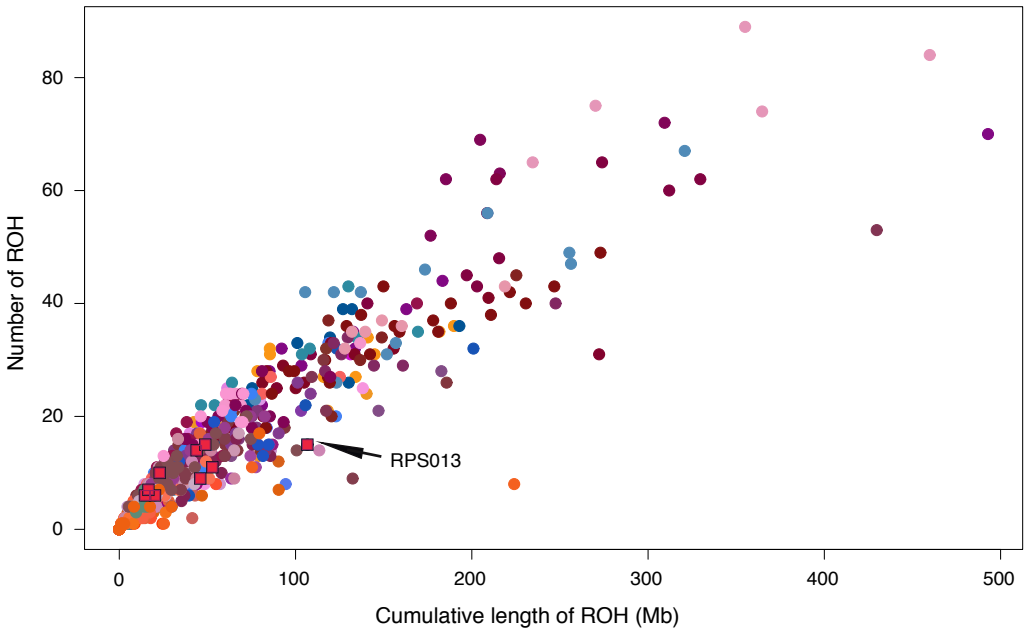


Figure S7. Distribution of runs of homozygosity. Individual patterns of runs of homozygosity in the Flores pygmies and populations included in the SEA DATASET. Colors code corresponds to colors in the map in Fig. 1A in main text, specifically orange shades=East Asia, blue shades=Island Southeast Asia, purple shades=Near Oceania, pink shades=Remote Oceania.

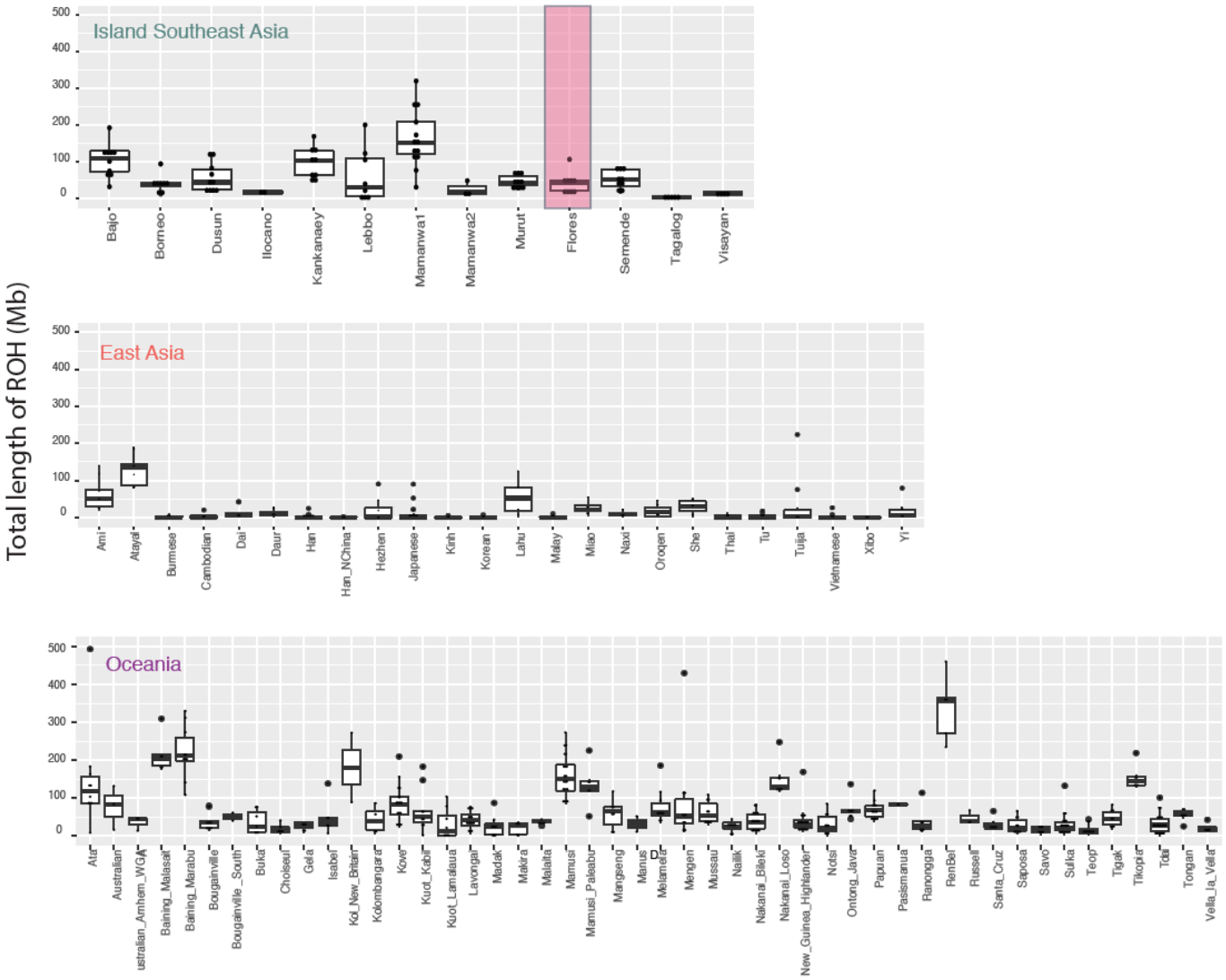


Figure S8. Patterns of long ROH across populations in Island Southeast Asia, East Asia and Oceania. Data are shown as boxplots representing the distribution of cumulative length of ROH over all individuals. The Flores population is highlighted in red.

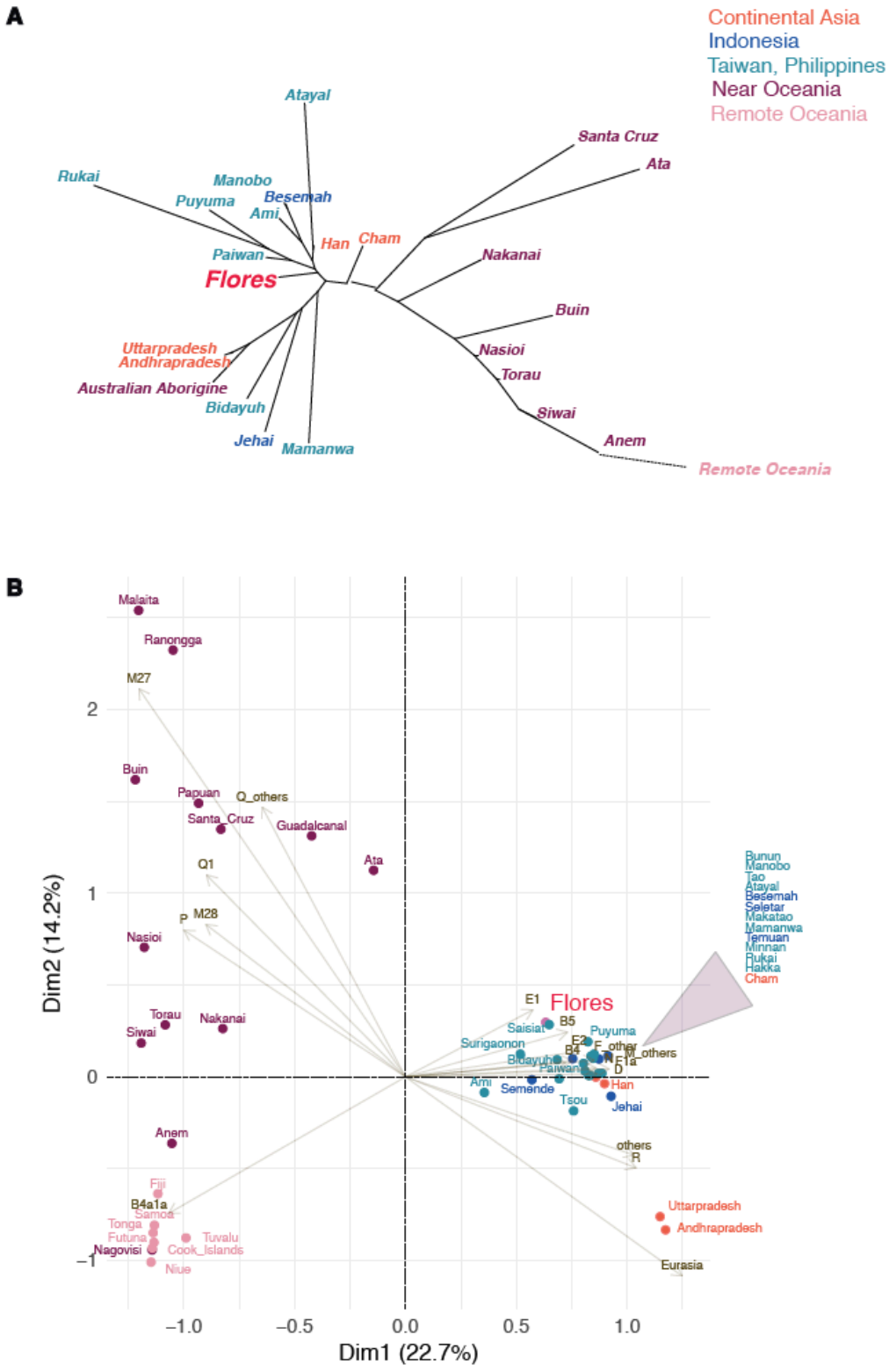
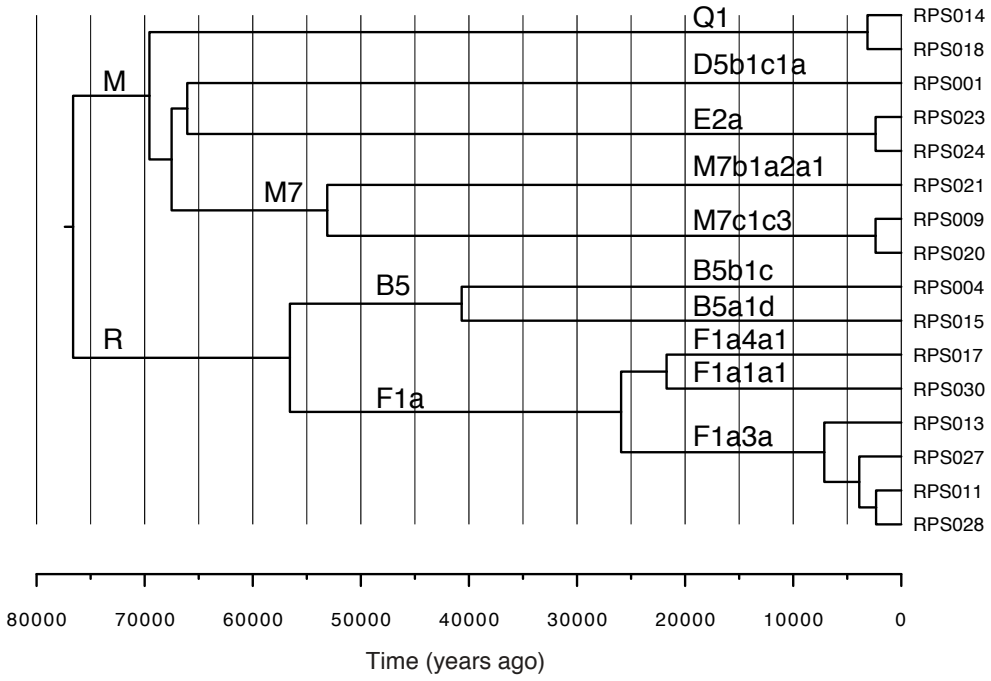


Figure S9. mtDNA variation. (**A**) Neighbor-joining tree from genetic distances estimated from mitochondrial whole genome sequences, computed in R with package ape (https://cran.r-project.org/web/packages/ape/index.html). (**B**) Haplogroup variation in 48 populations from Asia, Southeast Asia and Oceania, visualized by a CA plot. The plot is generated in R with the package factoextra (https://cran.r-project.org/web/packages/factoextra/index.html). The sample from Australia is excluded for its outlier haplogroup composition.

A



B

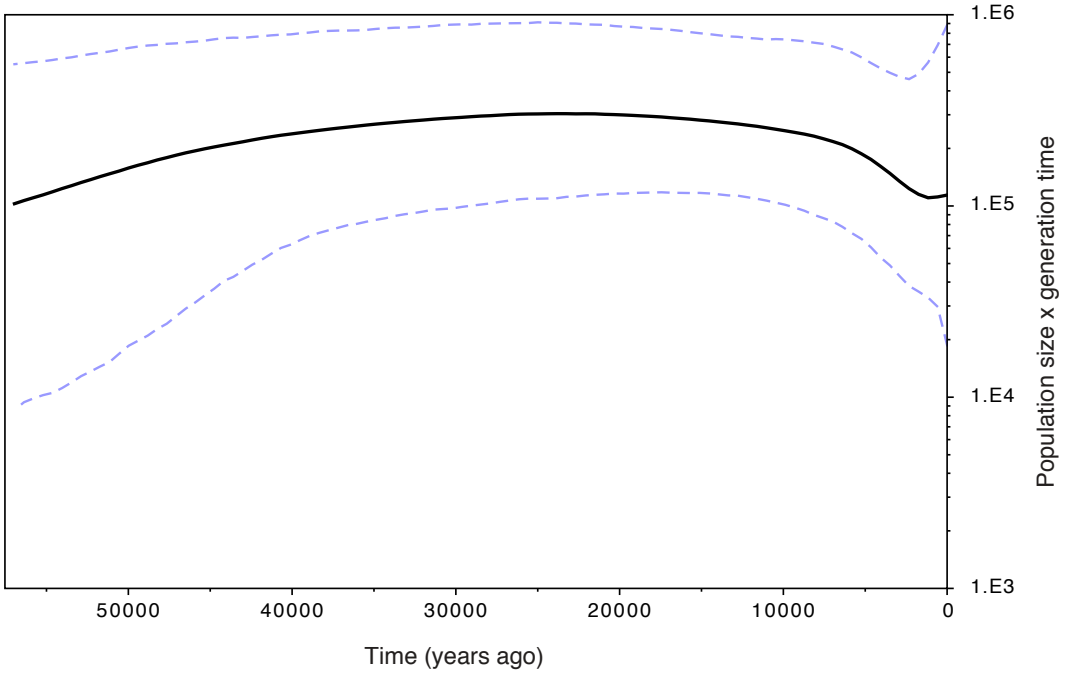


Figure S10. BEAST phylogenetic analysis. (**A**) Tree of the 16 mtDNA sequences from Flores. Haplogroups and sublineages are indicated on branches; (**B**) Bayesian Skyline Plot (BSP) for the same 16 sequences.

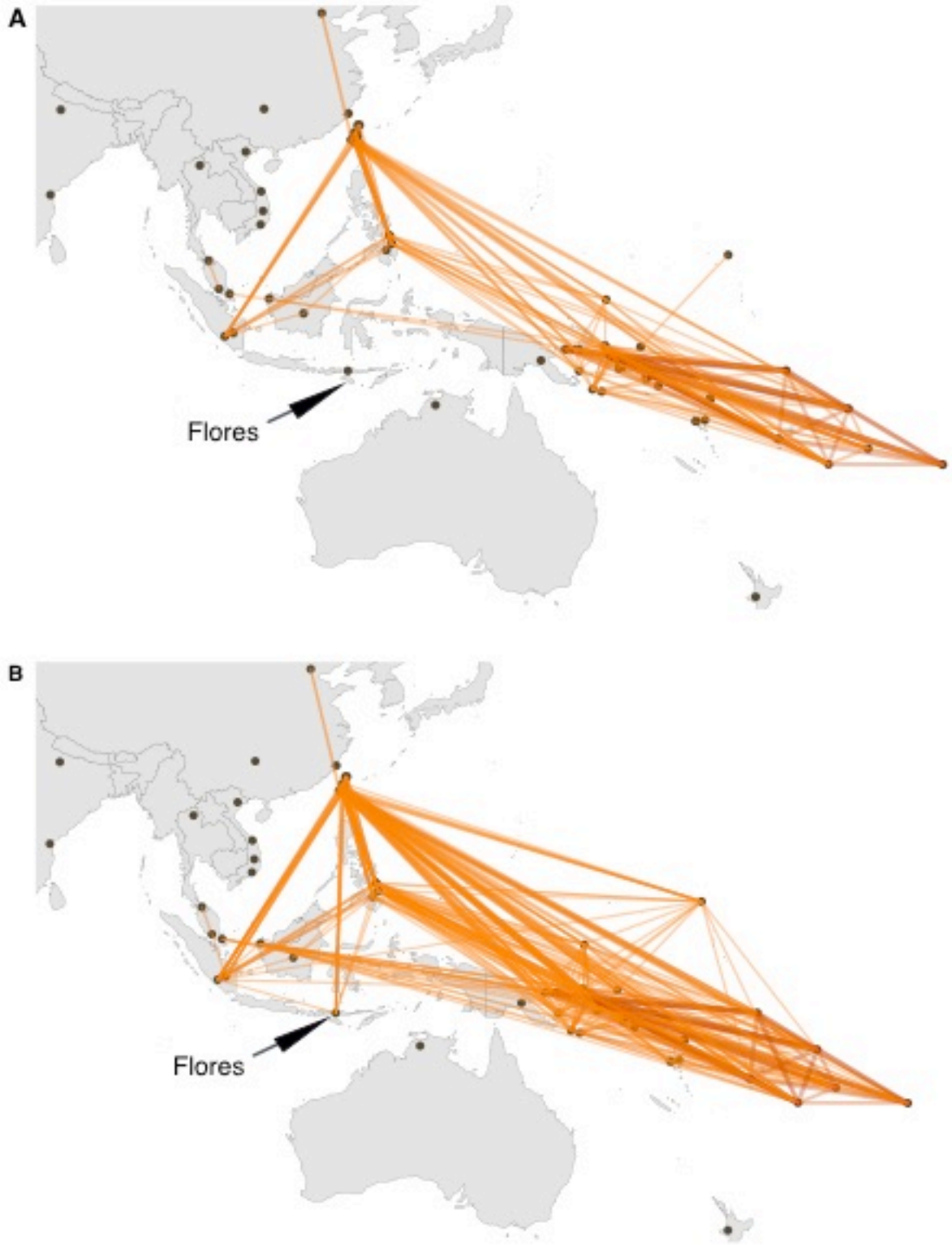


Figure S11. Haplotype sharing between sampled locations. Maps depicting patterns of mtDNA haplotype sharing at a continental scale. Thin yellow lines indicate the lowest levels of exchange (from just a single pair of individuals sharing an identical or similar haplotype); thick orange lines indicate highest sharing. (**A**) Sharing of identical haplotypes; (**B**) Sharing of similar haplotypes whose F_{ST} distance is less than 0.0001. Maps were generated in R with packages maps (https://cran.r-project.org/web/packages/maps/index.html) and geosphere (https://cran.r-project.org/web/packages/geosphere/index.html).

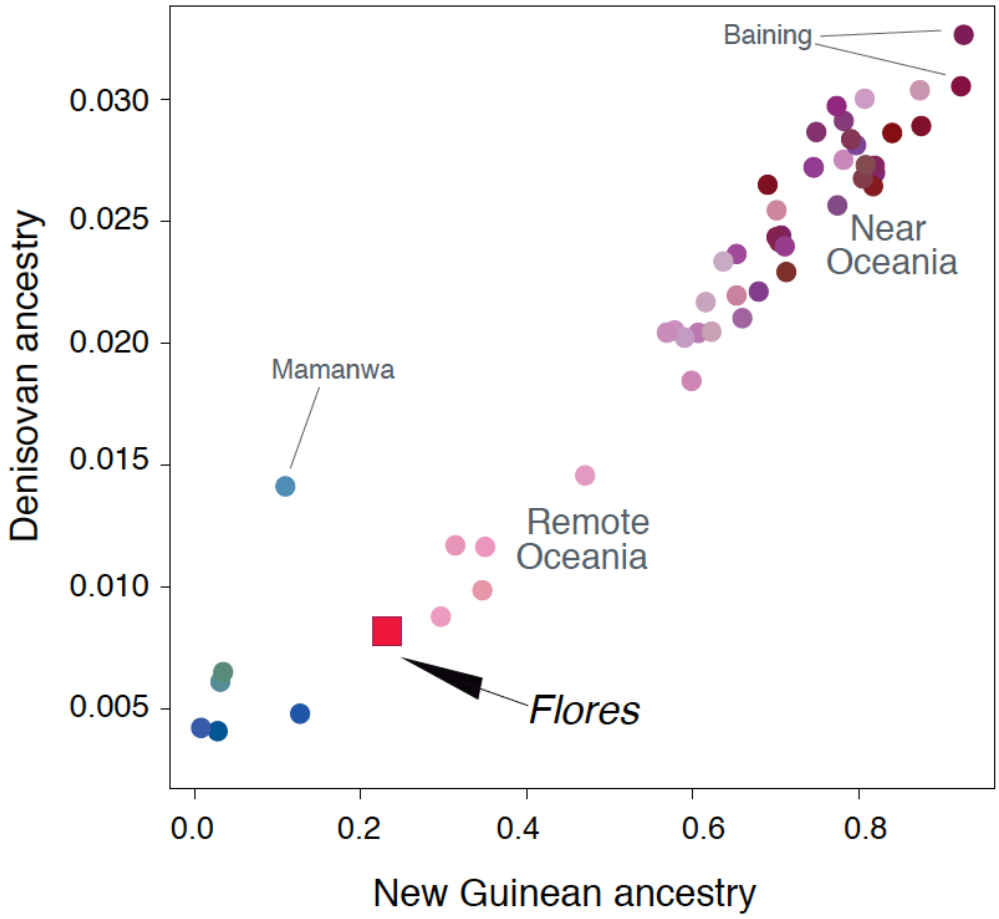


Figure S12. Denisovan ancestry, found in populations east of Wallace's line, is correlated with estimated New Guinean ancestry.

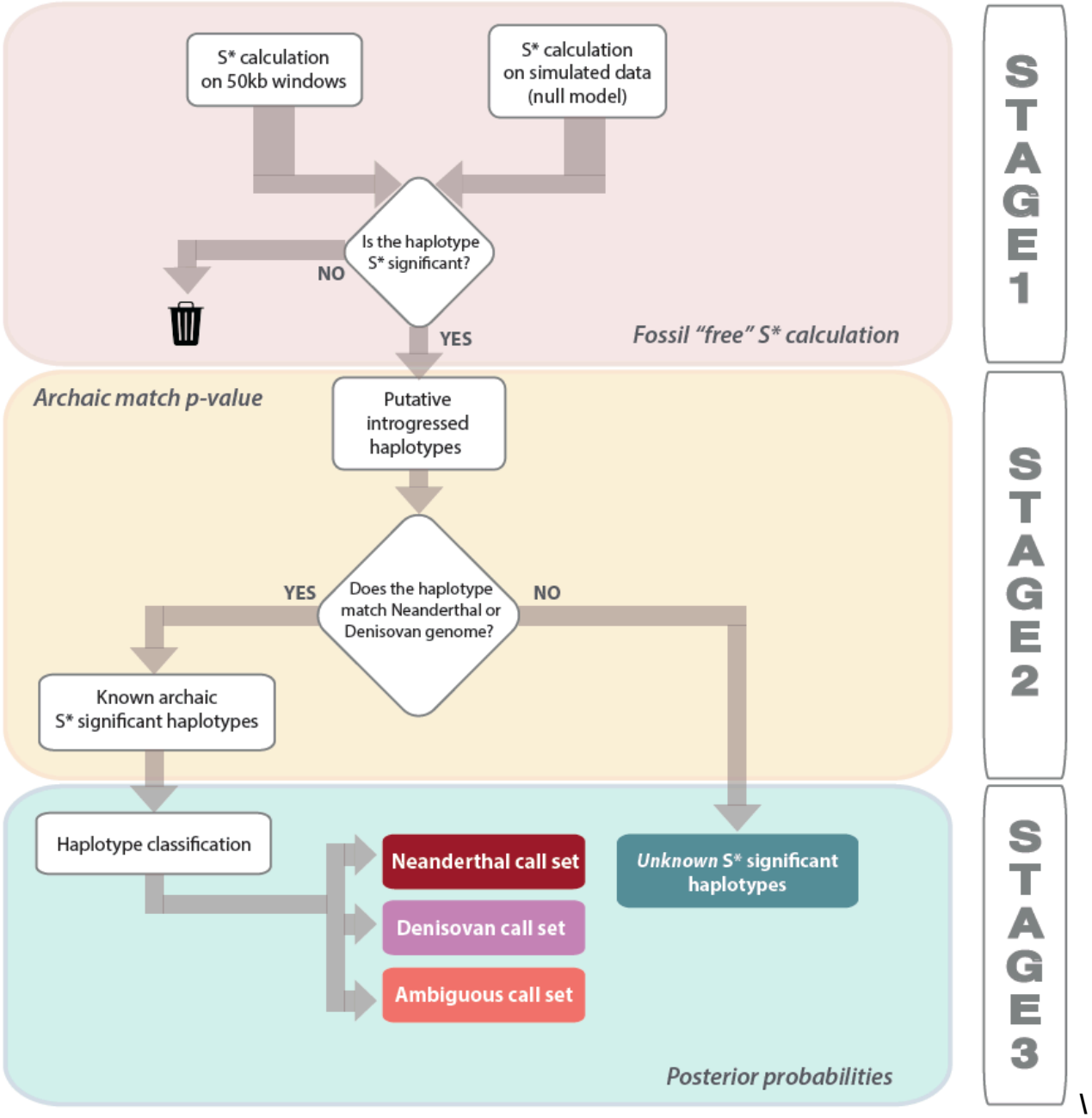


Figure S13. A schematic overview of the three-stages S* statistical framework. Adapted from Figure S3 in (*88*).

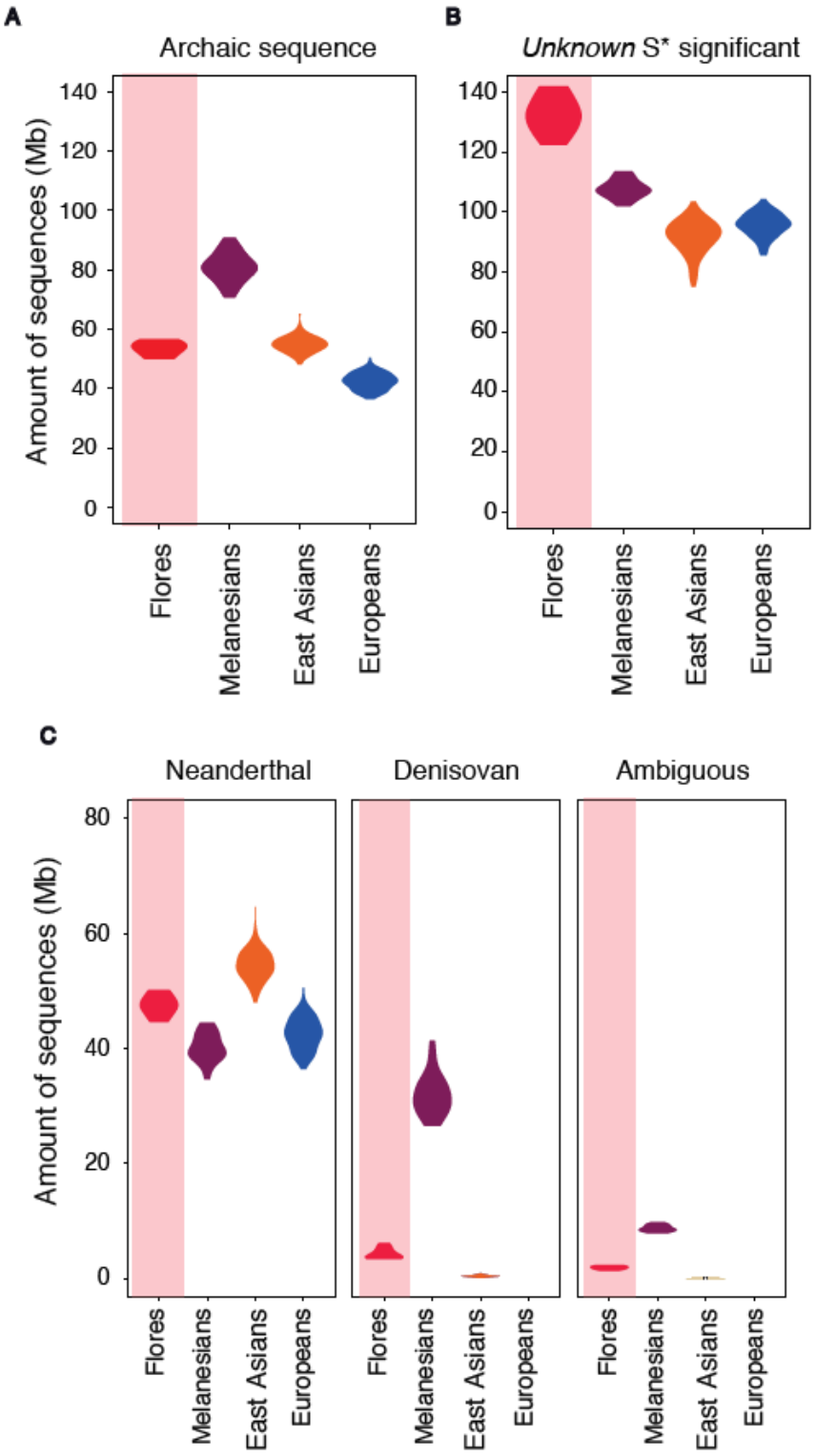


Figure S14. Amount of significant S* sequence in each population. (**A**) Amount of total archaic introgressed sequence and (**B**) Amount of unknown S* significant sequence identified in each population. (**C**) Amount of archaic introgressed sequence, classified as Neanderthal, Denisovan and ambiguous, identified in each population. The Flores sample is highlighted in red.

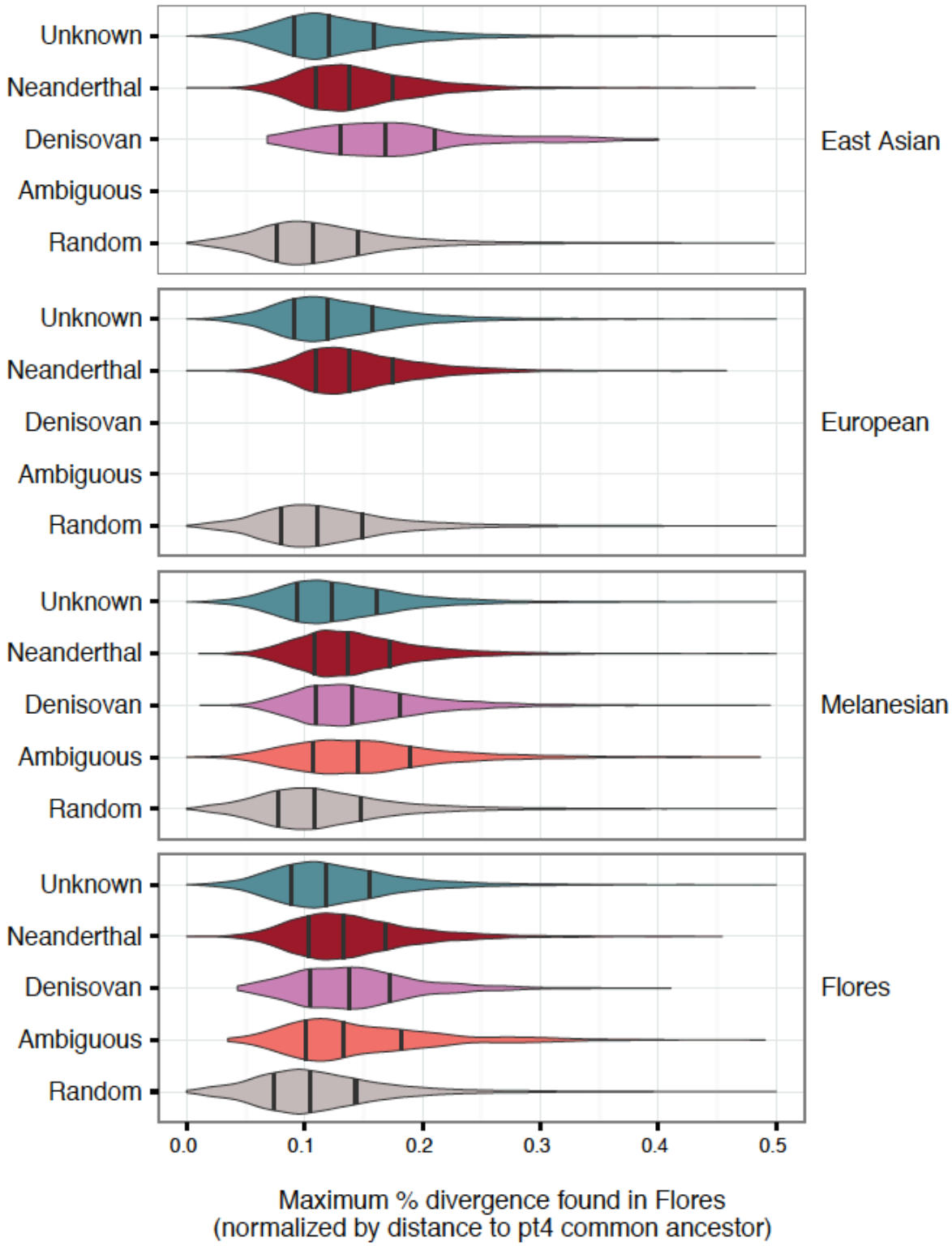


Figure S15. Distribution of the maximum pairwise divergence (TMRCA) estimated between S* significant haplotypes (Neanderthal, Denisovan, ambiguous and unknown) and non-S* haplotypes (random). Pairwise sequence divergence was estimated by triangulation. Vertical lines mark the first, second, and third quartiles.

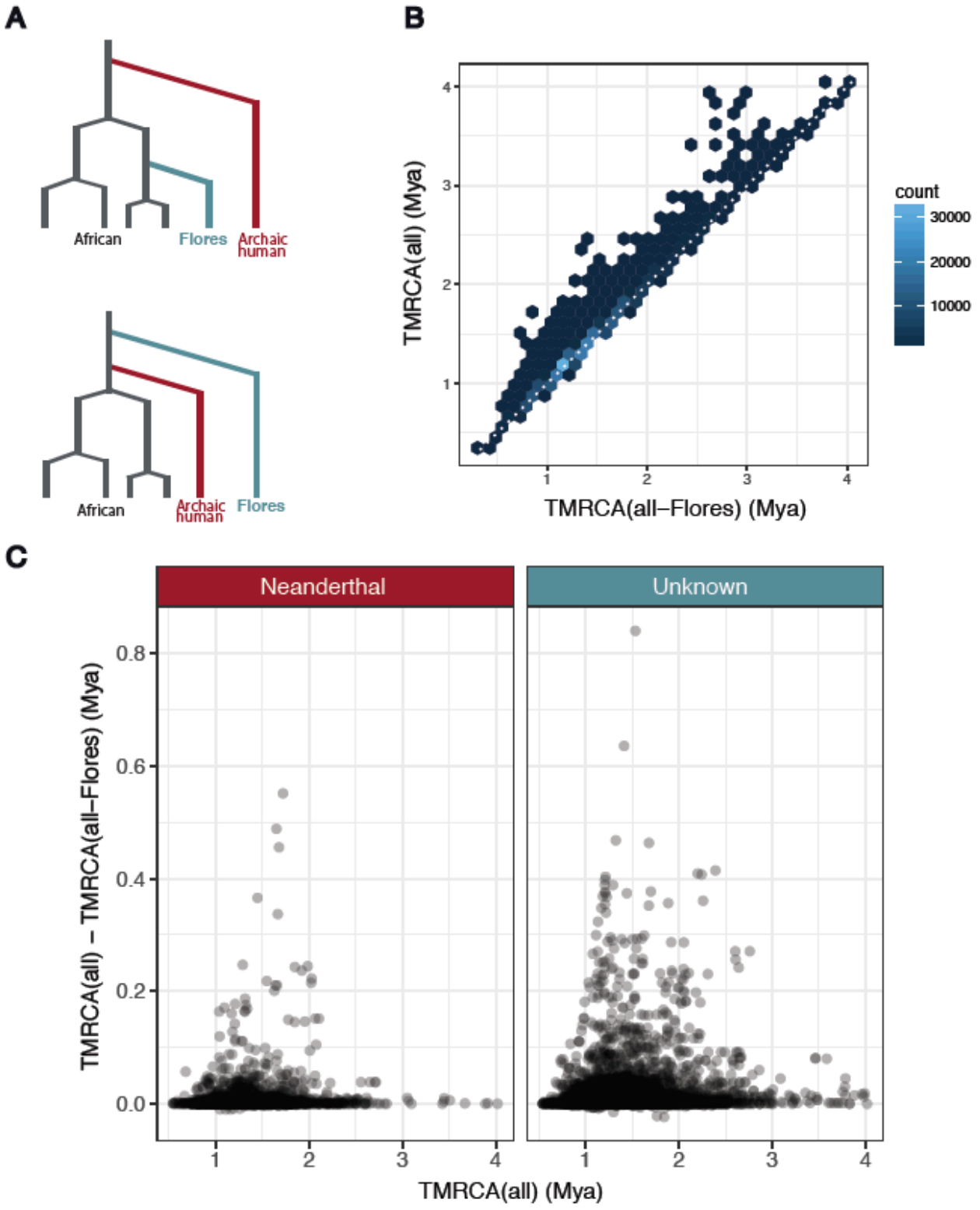


Figure S16. Estimates of TMRCA for significant S* haplotypes. (**A**) Schematic of the two models being compared. If the Flores genomes contain ancestry from archaic hominin group that diverged prior to Neanderthals (e.g. *H. floresiensis* or *H. erectus*), then we would expect some haplotypes in the Flores pygmies to coalesce above haplotypes from other modern humans ("African"), and above haplotypes from Neanderthal and Denisovan ("Archaic human"). (**B**) TMRCA estimates obtained excluding the Flores genomes "TMRCA(all-Flores)", versus TMRCA estimates obtained from all 17 genomes; (**C**)

Identifying regions where the addition of Flores increases the TMRCA of a region above the coalescence of other modern humans. Comparisons of TMRCAs for Neanderthal and unknown haplotypes estimated on all 17 individuals (Flores genomes included).

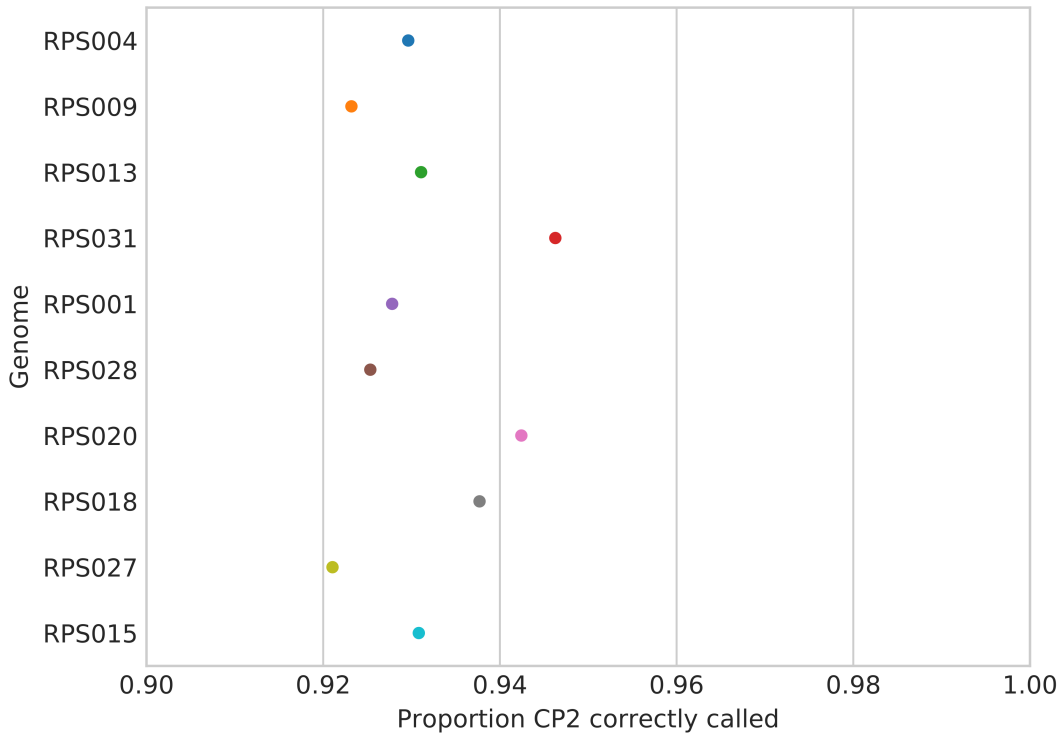


Figure S17. **Proportion of copy number 2 regions called correctly per Flores genome.** Thresholds above >90% are typically used for discovery while thresholds above >80% are useful for genotyping purposes (*15*).

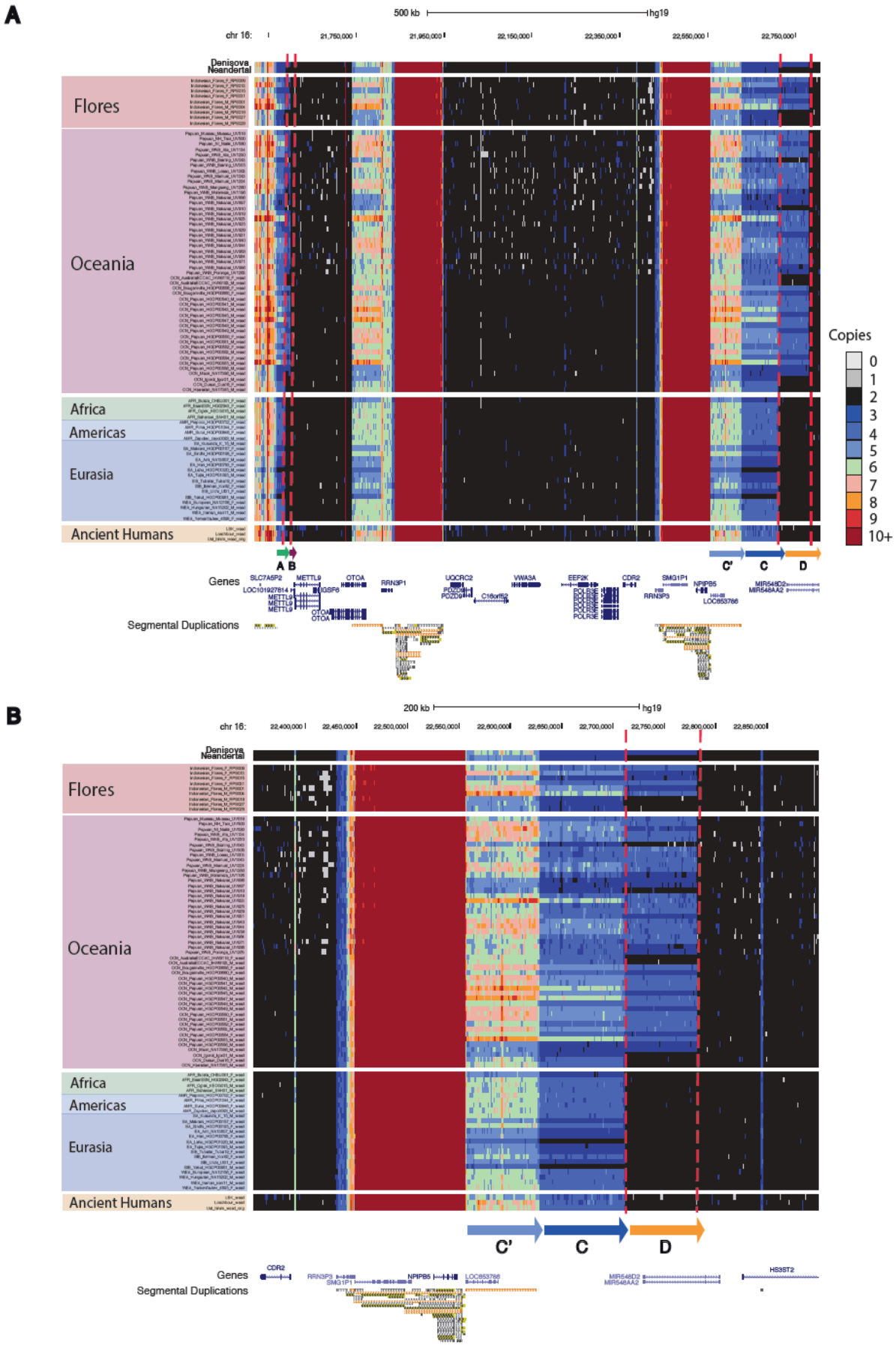


Figure S18. Heatmap representation of the chromosome 16p12.2 segmental duplication block (~1Mbp) in a panel of worldwide populations. (**A**) UCSC genome

browsershot (*125*) of WSSD in a panel of 86 genomes, which includes 9 Flores unrelated individuals, 49 Oceanic (27 genomes from (*48*), and 22 from the SGDP panel), 4 African, 4 American, 15 Eurasian, 3 ancient humans (*45, 97*), the Denisovan and Neanderthal genomes (*42, 56*). Each row represents the estimated copy number in 1kbp window for single individual. Arrows correspond to the positions and orientation of the duplications. The chromosome 16p12.2 duplication, described in (*15*) includes duplicated loci A, B, C and D. Structure A/C is found in all individuals, while B/D is only observed in the Denisovan, Oceanic and Flores genomes. See heatmap color key for copy number. (**B**) Analogous of panel A but "zoomed" in order to show the Denisovan D duplication (>220 kbp) in further detail. Copy number greater than two (4 and 3 for light and dark blue, respectively) in region D (far right dCGH call) indicates presence of the duplication. See color key for number of copies.

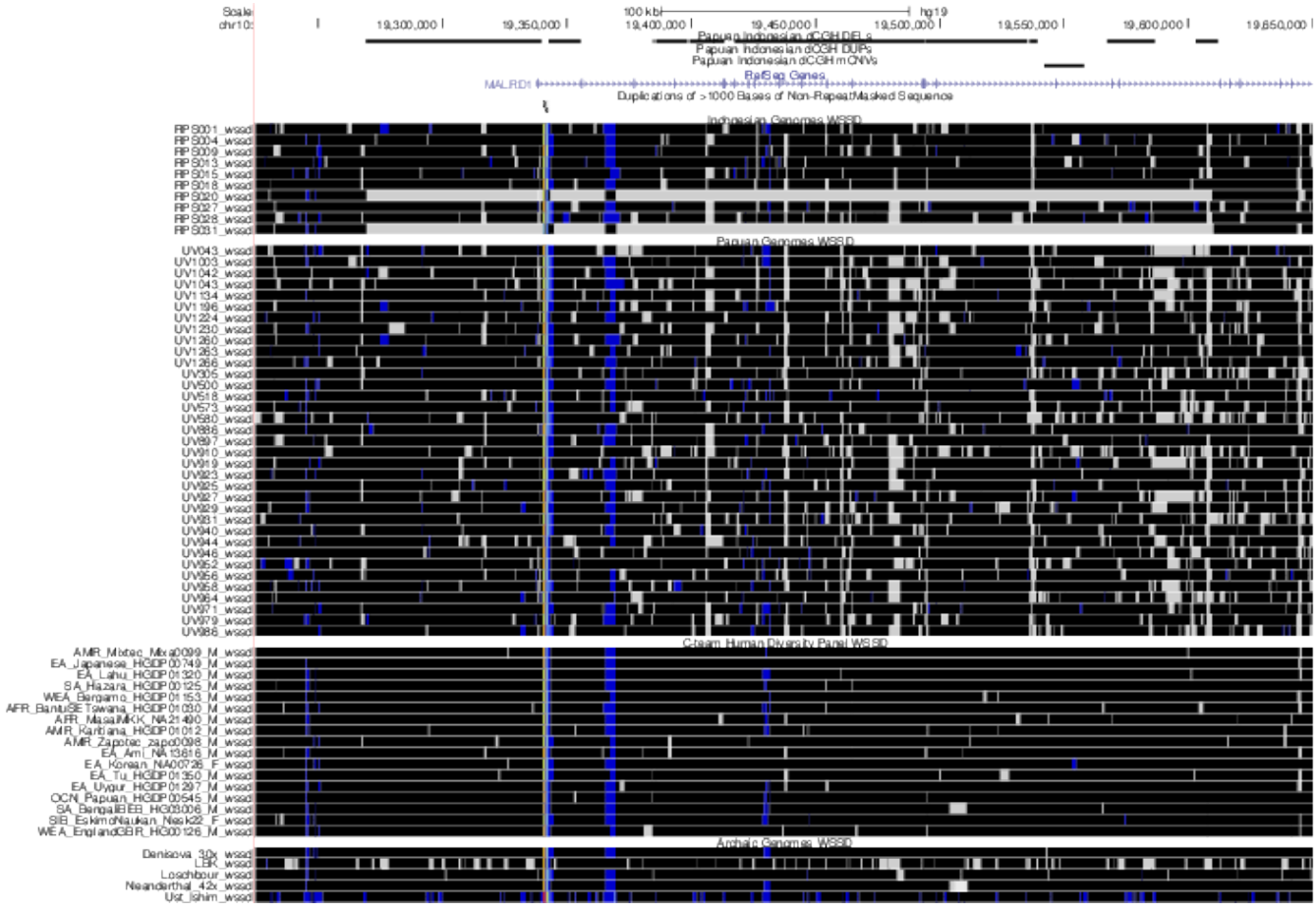


Figure S19. UCSC browser shot showing a maternally inherited ~340 kbp deletion (grey bar) of *MALRD1* absent from all SGDP and Melanesian genomes.

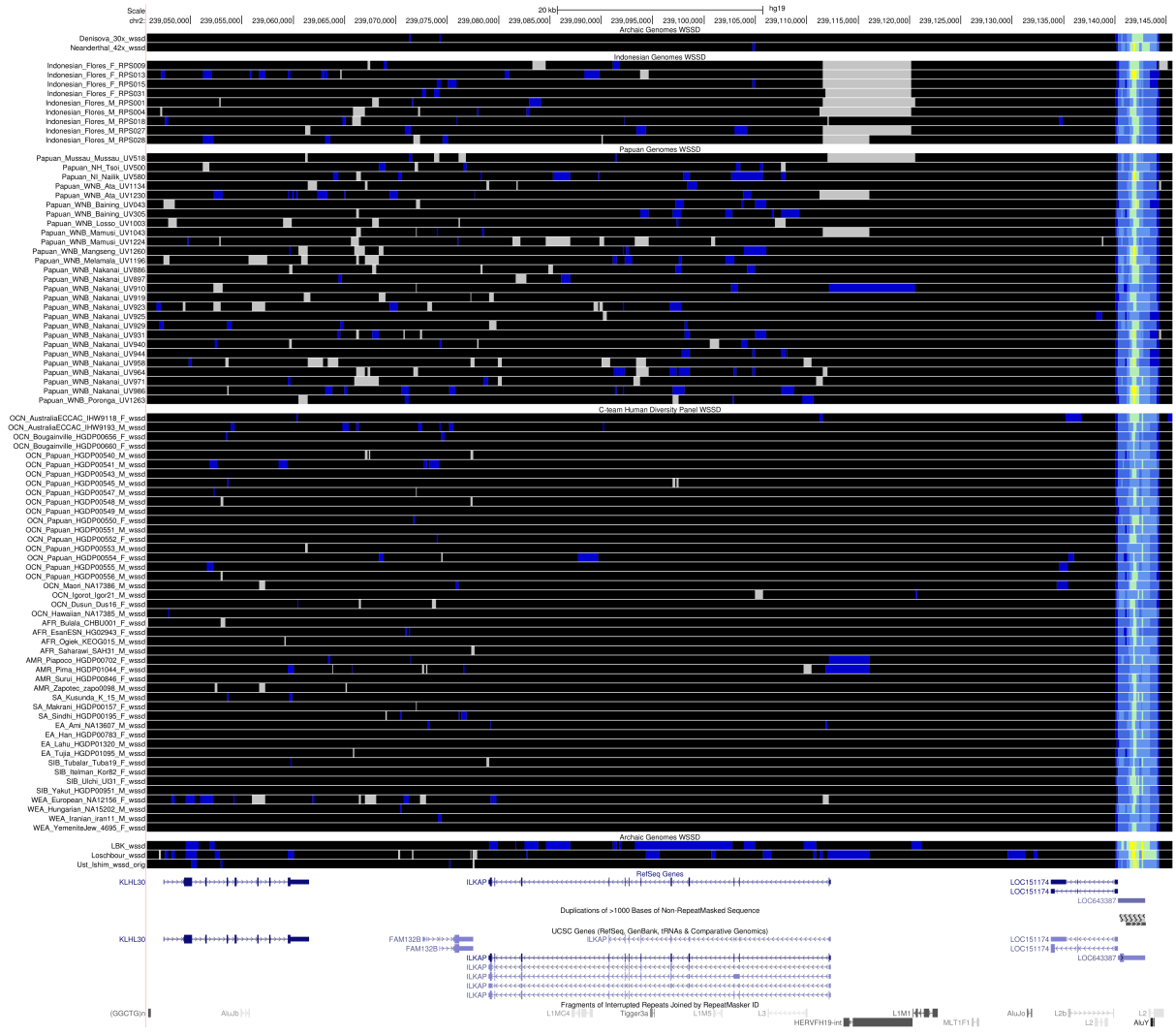


Figure S20. UCSC browsershot of a ~9 kbp deletion present in 9 Flores individuals and 2 Melanesians that intersects *ILKAP and a HERV element*.

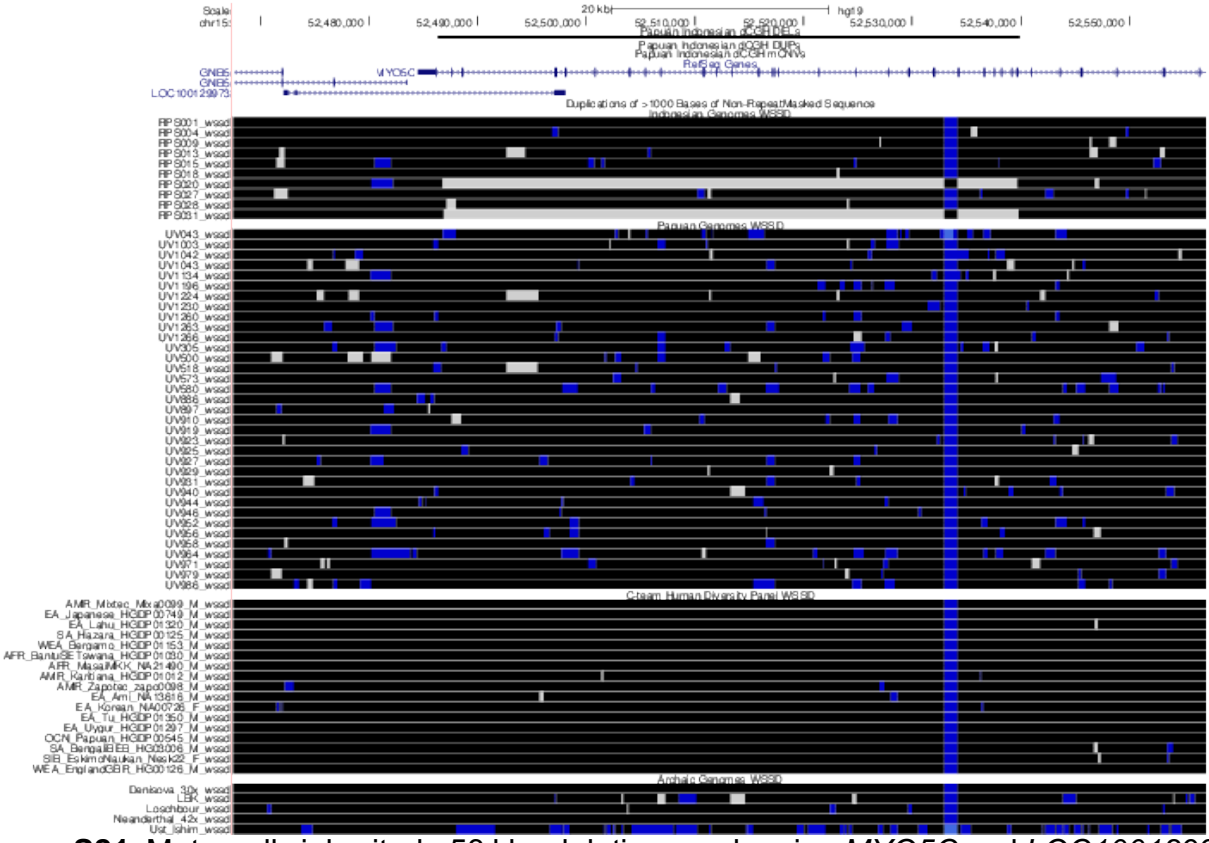


Figure S21. Maternally inherited ~53 kbp deletion overlapping *MYO5C* and *LOC100129973*.

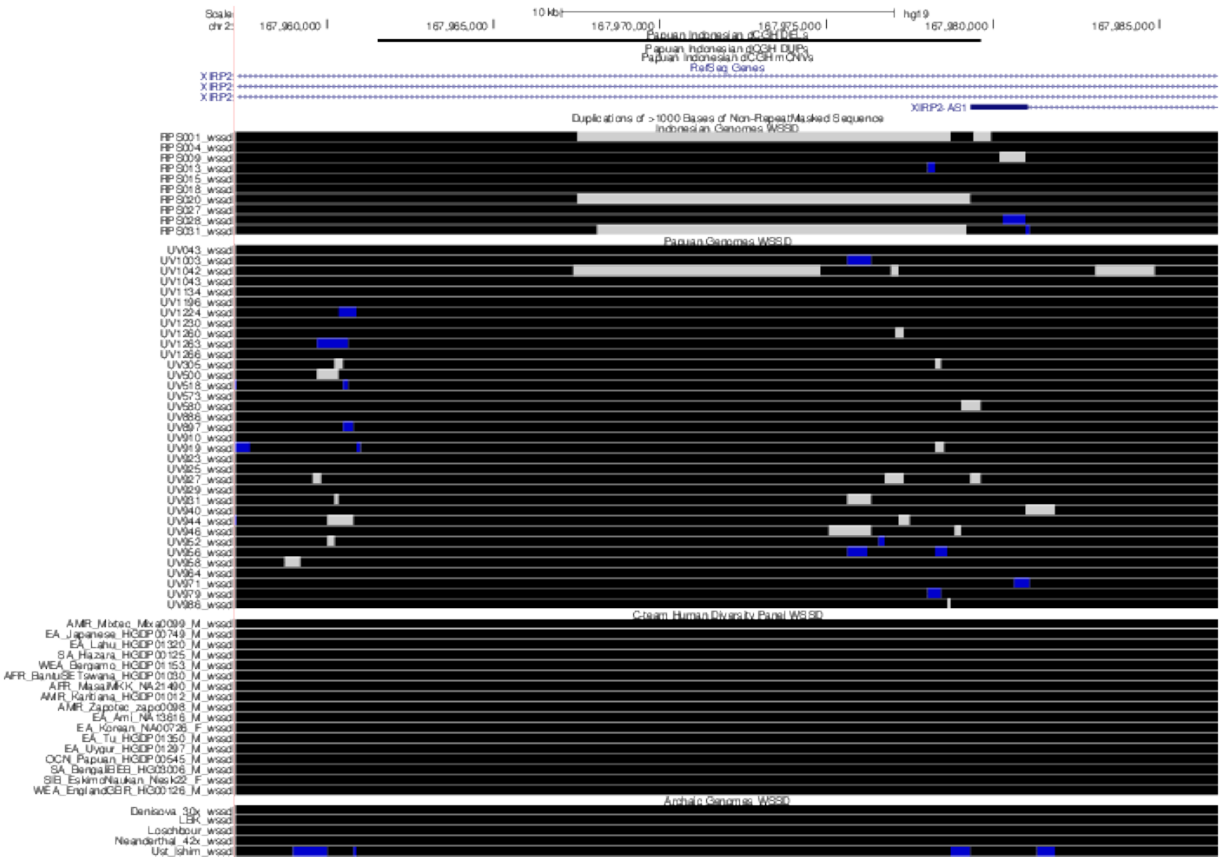


Figure S22. ~11 kbp deletion that intersects *XIRP2* present in 3 Flores genomes including two unrelated samples (inherited by RPS020 from RPS031, also found in RPS001).

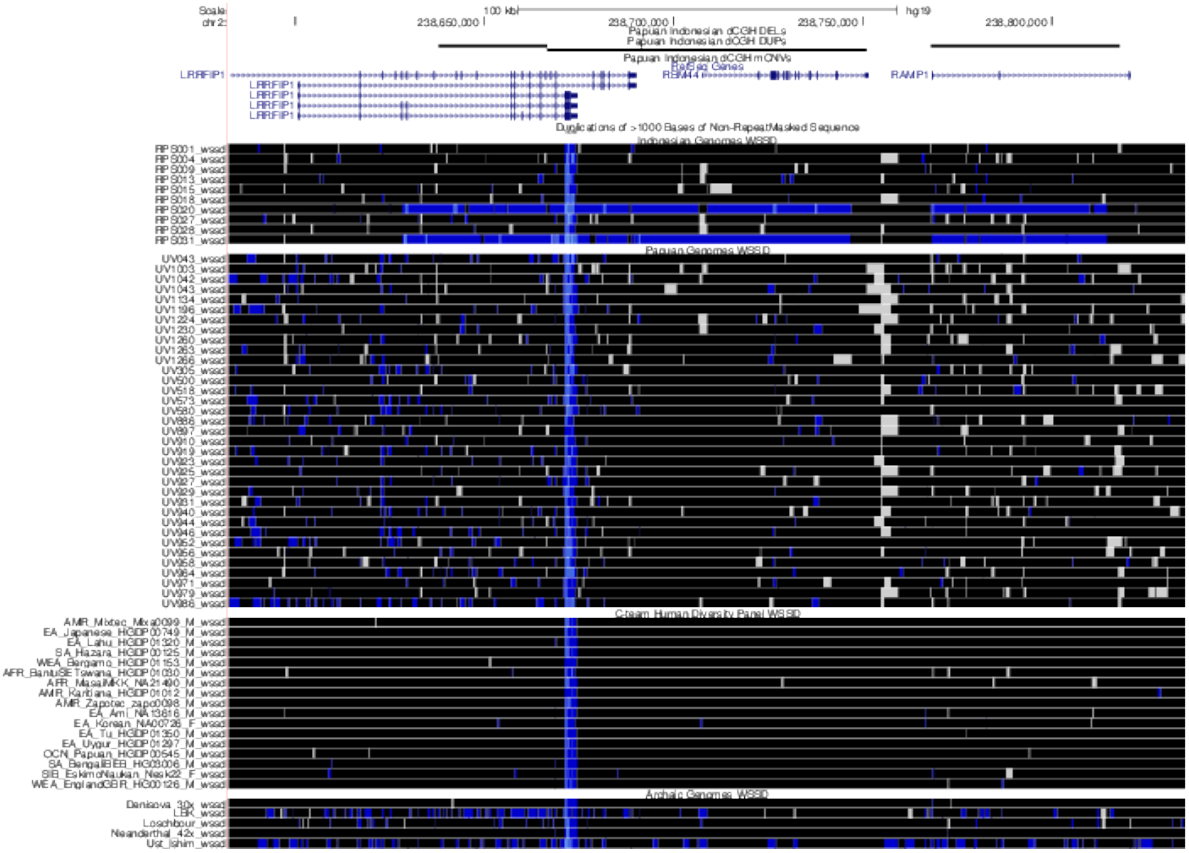


Figure S23. Inherited ~185 Kbp duplication (blue horizontal bar) of *RBM44, RAMP1*, and portion of *LRRFIP1* in two Flores genomes.

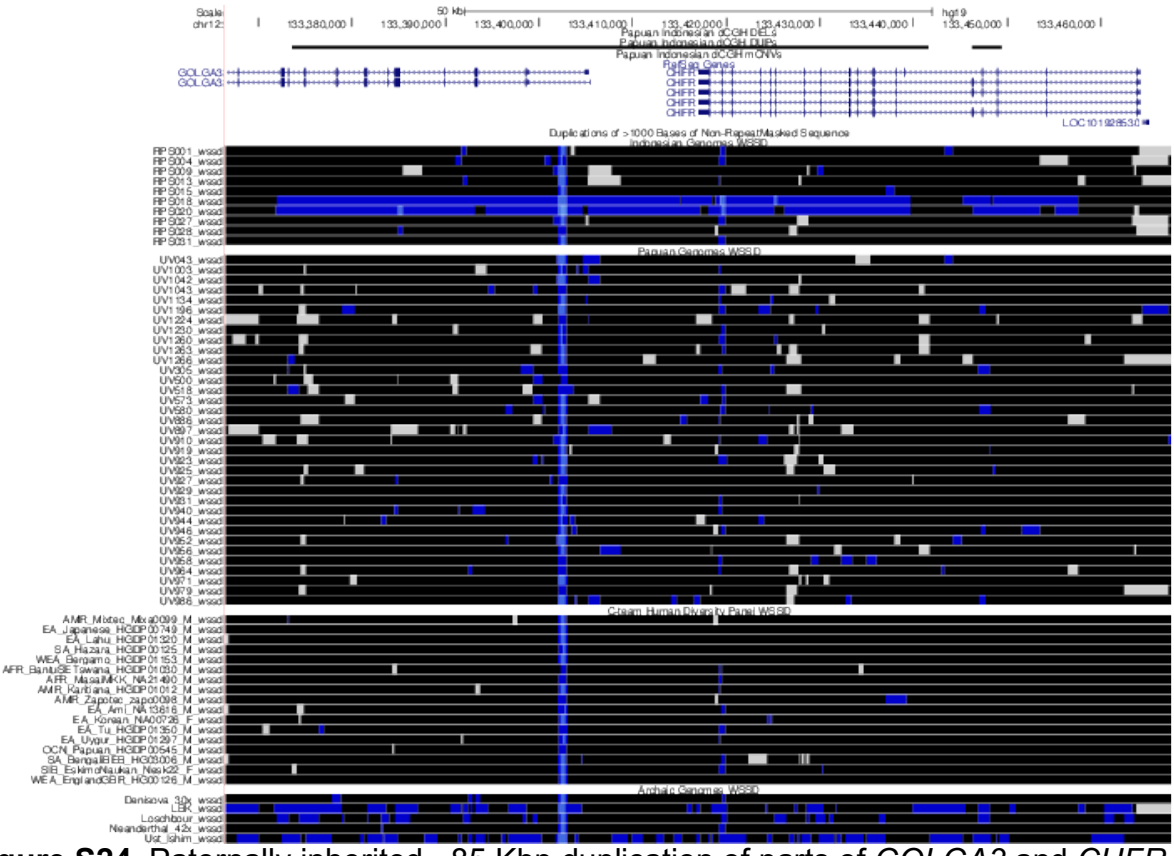


Figure S24. Paternally inherited ~85 Kbp duplication of parts of *GOLGA3* and *CHFR*.

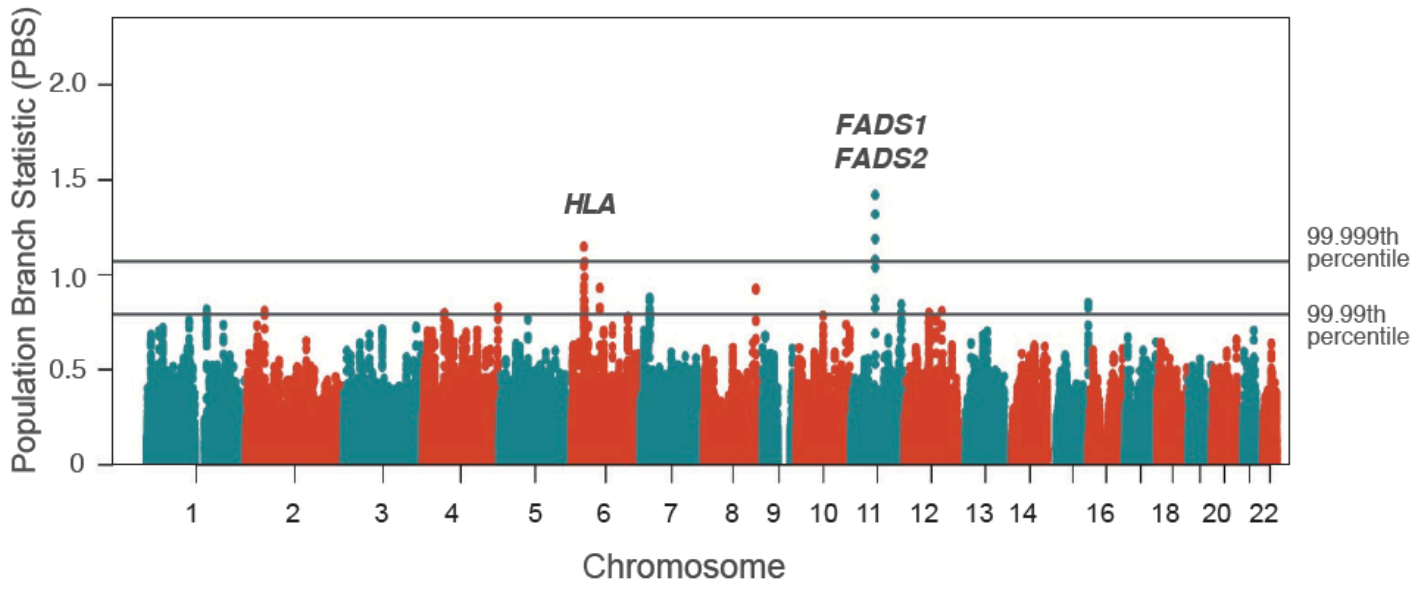


Figure S25. Manhattan plot of PBS values in sliding windows across the genome. The two grey lines indicate the top 0.01 and 0.001 PBS percentiles, respectively.

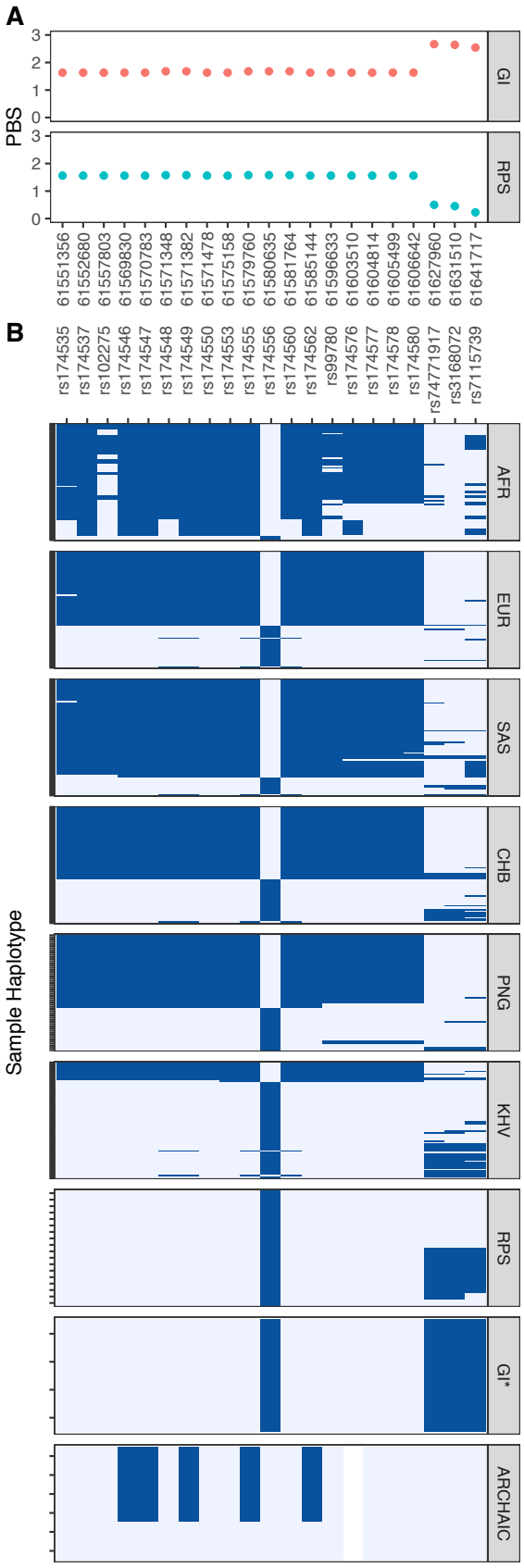


Figure S26. Per-SNP PBS results and haplotype visualization in the FADS region of chromosome 11, stratified by population. All SNPs meeting filtering criteria in all intersected modern human datasets were included, allowing for filtered data in the archaic genomes. (**A**) PBS scan results from Greenlandic Inuits (GI vs. CHB and CEU) (*23*) are plotted in the upper panel, while PBS scan results from Flores (RPS vs. PNG and CHB) are plotted in the lower panel. Strong frequency differentiation is observed over the entire region

in the Greenlandic Inuit scan, but the greatest signal is concentrated in the downstream region encompassing the latter half of *FADS2* and part of *FADS3* (chr11:61627960-61641717). The Flores population is polymorphic in this region, with unremarkable PBS scores, but similarly exhibits strong frequency differentiation in the upstream region encompassing *FADS1* and the first half of *FADS2*. (**B**) Visual haplotypes stratified by population. Columns distinguish SNPs (same as upper panel), while rows distinguish individual sample haplotypes. Dark blue indicates derived alleles, polarized with respect to the chimpanzee genome, while light blue indicates ancestral alleles. Sample sizes vary substantially across population, and individual haplotypes are therefore denoted with *y*-axis tick marks. AFR, EUR, and SAS refer to African, European, and South Asian super-populations from the 1000 Genomes Project, respectively, from each of which 50 haplotypes were randomly sampled. CHB and KHV refer to the Han Chinese in Beijing, China and Kinh in Ho Chi Minh City, Vietnam populations from the 1000 Genomes project, respectively, from each of which 50 haplotypes were also randomly sampled. PNG refers to the Melanesian samples from (*48*). RPS refers to the Flores sample, while GI* refers to Greenlandic Inuit samples from (*107*) ARCHAIC refers to Neanderthal and Denisovan genomes, with Altai (*42*) and Vindija Neanderthal (*126*) sequences plotted above the Denisovan (*56*) sequences. While the archaic data are not phased, the genomes were homozygous at all SNPs depicted here.

Supplementary Tables

Table S1. Samples collected in this study. Asterisks (*) denote samples that were later selected for whole genome sequencing.

Sample ID	sex	Stature (cm)
RPS001*	M	135
RPS002	M	175
RPS003	M	152
RPS004*	M	145
RPS005	F	138
RPS006	M	155
RPS007	M	150
RPS008	M	155
RPS009*	F	132
RPS010	M	153
RPS011	M	148
RPS012	M	155
RPS013*	F	138
RPS014	M	150
RPS015*	F	132
RPS016	M	153
RPS017	M	142
RPS018*	M	136
RPS019	M	153
RPS020*	M	143
RPS021	F	137
RPS022	M	147
RPS023	F	146
RPS024	F	145
RPS025	F	140
RPS026	F	140
RPS027*	M	130
RPS028*	M	139
RPS029	M	143
RPS030	F	152
RPS031*	F	133
RPS032	F	150

Table S2. Populations included in the PANASIA Dataset.

PANASIA HUGO CONSORTIUM DATASET

Pop Code	Source	Language	Language family	n
AXAM	Taiwan_Affymetrix	Ami	Austronesian	10
AXAT	Taiwan_Affymetrix	Atayal	Austronesian	10
AXME	Melanesian_Affymetrix	Naasioi	Indo-Pacific (East Papuan)	4
CHB	China_HapMap	Chinese	Sino-Tibetan	45
CNCC	China	Zhuang,	Northern Tai-Kedai	24
CNGA	China	Cantonese	Sino-Tibetan	30
CNHM	China	Hmong	Miao-Yao	19
CNJI	China	Jiamao	Tai-Kedai	31
CNJN	China	Jinuo	Sino-Tibetan (Loloish)	26
CNSH	China	Mandarin	Sino-Tibetan	21
CNUG	China	Uyghur	Altaic	26
CNWA	China	Wa	Austro-Asiatic	48
IDAL	Indonesia	Alorese	Austronesian	19
IDDY	Indonesia	Dayak	Austronesian	12
IDJA	Indonesia	Javanese	Austronesian	34
IDJV	Indonesia	Javanese	Austronesian	19
IDKR	Indonesia	Karo	Batak Karo Austronesian	17
IDLA	Indonesia	Lamaholot	Austronesian	20
IDLE	Indonesia	Lembata	Austronesian	19
IDML	Indonesia	Malay	Austronesian	11
IDMT	Indonesia	Mentawai	Austronesian	15
IDRA	Indonesia	Manggarai	Austronesian	9
IDSB	Indonesia	Kambera	Austronesian	20
IDSO	Indonesia	Manggarai	Austronesian	18
IDSU	Indonesia	Sunda	Austronesian	25
IDTB	Indonesia	Toba	Batak Toba Austranesian	20
IDTR	Indonesia	Toraja	Austronesian	20
INDR	India	Telugu,	Kannada Dravidian	23
INEL	India	Bengali	Indo-European	10
INIL	India	Hindi	Indo-European	12
INNI	India	Pahari	Indo-European	20
INNL	India	Hindi	Indo-European	13
INSP	India	Hindi	Indo-European	20
INTB	India	Spiti	Sino-Tibetan	23
INWI	India	Bhili	Indo-European	23
INWL	India	Marathi	Indo-European	12
JPML	Japan	Japanese	Altaic	71
JPRK	Japan	Okinawan	Altaic	43
JPT	Japan_HapMap	Japanese	Altaic	44
KRKR	Korea	Korean	Altaic	90
MYBD	Malaysia	Jagoi	Austronesian	41
MYJH	Malaysia	Jehai	Austro-Asiatic	34
MYKN	Malaysia	Malay	Austronesian	15
MYKS	Malaysia	Kensiu	Austro-Asiatic	22
MYMN	Malaysia	Malay	Austronesian	16
MYTM	Malaysia	Temuan	Austronesian	30
PIAE	Philippines	Ayta	Austronesian	8
PIAG	Philippines	Agta	Austronesian	8
PIAT	Philippines	Ati	Austronesian	22
PIIR	Philippines	Iraya	Austronesian	9
PIMA	Philippines	Manobo	Austronesian	17
PIMW	Philippines	Mamanwa	Austronesian	17
PIUB	Philippines	Ilocano	Austronesian	20
PIUI	Philippines	Visaya, Chabakano	Austranesian	20
PIUN	Philippines	Tagalog	Austronesian	19
SGCH	Singapore	Mandarin	Sino-Tibetan	30
SGID	Singapore	Tamil	Dravidian	30
SGML	Singapore	Malay	Austronesian	30
THHM	Thailand	Hmong	Hmong-Mien	19
THKA	Thailand	Karen	Sino-Tibetan	20
THLW	Thailand	Lawa	Austro-Asiatic	16
THMA	Thailand	Mlabri	Austro-Asiatic	10
THMO	Thailand	Mon	Austro-Asiatic	18
THPL	Thailand	Paluang	Austro-Asiatic	17
THPP	Thailand	Blang	Austro-Asiatic	17
THTK	Thailand	Tai Khuen	Tai-Kadai	17
THTL	Thailand	Lue	Tai-Kadai	18
THTN	Thailand	Mal	Austro-Asiatic	13
THTU	Thailand	Tai Yuan	Tai-Kadai	20
THTY	Thailand	Tai Yong	Tai-Kadai	18
THYA	Thailand	Iu Mien	Hmong-Mien	17
TWHA	Taiwan	Hakka	Sino-Tibetan	48
TWHB	Taiwan	Minnan	Sino-Tibetan	32

Table S3. Sequencing coverage depth statistics for 10 Flores samples. Individual "RPS020" was removed in analyses of unrelated individuals.

Sample ID	Mean depth (genome)	Median depth (Autosomes)	99.5th Percentile	Mapped reads
RPS001	31.1	33	62	654,925,220
RPS004	32.4	34	69	681,860,460
RPS009	33.3	34	70	702,133,586
RPS013	37.4	38	75	786,877,329
RPS015	36.6	38	71	770,583,648
RPS018	37.8	40	75	795,182,382
RPS020	42.2	44	82	889,949,803
RPS027	32.2	34	67	678,706,518
RPS028	32.5	34	66	684,725,484
RPS031	47.5	49	92	1,000,842,729

Table S4. Populations included in the SEA Dataset and proportion of Denisovan ancestry. See accompanying excel spreadsheet

Table S5. Summary of mtDNA haplogroup diversity. See accompanying excel spreadsheet.

Table S6: Full Melanesian and Flores CNVs table. CNVs predicted using dCGH comparing 35 Melanesian and 10 Flores genomes (including related individuals) against 17 genomes from the SGDP panel. Values of -1 indicate no call. See accompanying excel spreadsheet.

Table S7: Melanesian and Flores CNV summary. Deletions and duplications predicted using dCGH comparing 36 genomes (Melanesian+Flores) against 17 SGDP genomes. CNVs are filtered by size (>10 kb), frequency (i.e. if observed in 2 or more unrelated Flores individuals or 6 or more unrelated Melanesians) and whether they were previously observed in the SGDP (SGDP Negative).

Group (n)	SV type	Count	Count >10 kbp	Frequent	>10 kbp and frequent	SGDP negative count	SGDP negative and >10 kbp count	SGDP negative and frequent	SGDP negative, >10Kbp and frequent
All (36)	DEL	3245	503						
	DUP	3808	270						
	mCNV	1465	684						
Flores (9)	DEL	971	188	489	88	469	68	125	13
	DUP	894	87	276	32	833	69	247	24
Melanesians (27)	DEL	2967	434	489	75	2335	292	161	9
	DUP	3126	209	392	32	2989	184	353	25

Table S8. Genomic windows in the upper tail of mean PBS scores. Only the top-scoring window for every unique PBS peak is reported for all peaks in top 0.01 percentile of genome-wide PBS scores.

Chr	Start Position	End Position	Mean PBS	Empirical P-value
11	61549025	61601872	1,421	2×10^{-6}
6	30788195	30794004	1,148	6×10^{-6}
6	32581193	32581782	1,067	1×10^{-5}
6	71037769	71053500	0,931	2×10^{-5}
8	130492991	130522508	0,928	2×10^{-5}
7	24248620	24259852	0,880	3×10^{-5}
15	101499930	101507525	0,852	4×10^{-5}
11	126883297	126891327	0,845	4×10^{-5}
4	188402828	188439963	0,830	5×10^{-5}
1	150388385	150457876	0,817	7×10^{-5}
2	45961676	45971253	0,810	7×10^{-5}
12	92930471	92950033	0,809	7×10^{-5}
12	60639416	60681383	0,802	8×10^{-5}
4	54635203	54646237	0,800	8×10^{-5}

Table S9. NHGRI-EBI GWAS catalog hits for high-PBS haplotype-tagging SNP rs174547.

Phenotype	PMID	First Author	Year	Journal	P-value
plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid)	24823311	Guan W	2014	Circ Cardiovasc Genet	3×10^{-971}
linoleic acid: polyunsaturated fatty acids ratio (LA/PUFA)	22286219	Kettunen J	2012	Nat Genet	8×10^{-262}
phosphatidylcholine (PC) diacyl (aa) C36:3 / PC aa C36:4	20037589	Illig T	2009	Nat Genet	7×10^{-179}
lysophosphatidylcholine acyl C20:4	26068415	Draisma HH	2015	Nat Commun	2×10^{-175}
phosphatidylcholine diacyl C38:4	26068415	Draisma HH	2015	Nat Commun	9×10^{-172}

Table S10. UK Biobank Gene ATLAS PheWAS hits for high-PBS haplotype-tagging SNP rs174547. Only hits with *p-values* less than Bonferroni-adjusted threshold of 6.4×10^{-5} are reported. Signs of beta values indicate the effects of the ancestral 'C' allele.

Trait	Beta	P-value
Red blood cell (erythrocyte) count	0,0104	$< 5 \times 10^{-324}$
Red blood cell (erythrocyte) distribution width	-0,0386	$< 5 \times 10^{-324}$
Platelet count	1,9752	$< 5 \times 10^{-324}$
Mean platelet (thrombocyte) volume	-0,0378	$< 5 \times 10^{-324}$
Haemoglobin concentration	0,0264	1.79×10^{-37}
Haematocrit percentage	0,0605	2.32×10^{-23}
Eosinophill count	-0,0028	1.46×10^{-22}
Mean sphered cell volume	-0,1016	1.55×10^{-22}
Platelet crit	0,0008	1.02×10^{-18}
Mean corpuscular volume	-0,0754	2.66×10^{-18}
Mean reticulocyte volume	-0,1326	4.07×10^{-17}
Monocyte percentage	0,0458	1.11×10^{-15}
Standing height	-0,0924	4.12×10^{-15}
Eosinophill percentage	-0,0305	7.87×10^{-15}
Neutrophill count	-0,0221	1.64×10^{-13}
asthma	-0,0052	4.94×10^{-13}
White blood cell (leukocyte) count	-0,0285	1.03×10^{-10}
Mean corpuscular haemoglobin concentration	0,0145	3.42×10^{-10}
Reticulocyte count	0,0005	1.40×10^{-9}
Platelet distribution width	0,0064	1.47×10^{-9}
High light scatter reticulocyte count	0,0001	1.97×10^{-9}
thyroid problem (not cancer)	-0,0031	3.21×10^{-9}
K80 Cholelithiasis	0,0025	6.24×10^{-9}
hypothyroidism/myxoedema	-0,0028	7.33×10^{-9}
K80-K87 Disorders of gallbladder, biliary tract and pancreas	0,0027	1.18×10^{-8}
Sleep duration	0,0137	1.96×10^{-8}
high cholesterol	-0,0040	3.20×10^{-8}
J45 Asthma	-0,0030	1.08×10^{-7}
Number of self-reported non-cancer illnesses	-0,0214	1.13×10^{-7}
Salt added to food	0,0095	5.27×10^{-7}
Sitting height	-0,0361	5.76×10^{-7}
Number of treatments/medications taken	-0,0280	9.39×10^{-7}
E03 Other hypothyroidism	-0,0020	1.70×10^{-6}
J40-J47 Chronic lower respiratory diseases	-0,0030	1.75×10^{-6}
Neutrophill percentage	-0,0858	1.81×10^{-6}
Lymphocyte percentage	0,0713	4.09×10^{-6}
venous thromboembolic disease	-0,0016	7.83×10^{-6}
J33 Nasal polyp	-0,0009	1.69×10^{-5}
Reticulocyte percentage	0,0084	1.85×10^{-5}
deep venous thrombosis (dvt)	-0,0013	3.90×10^{-5}
Skin colour	-0,0050	4.33×10^{-5}
E00-E07 Disorders of thyroid gland	-0,0018	5.70×10^{-5}

References and Notes

1. C. C. Swisher III, G. H. Curtis, T. Jacob, A. G. Getty, A. Suprijo, Widiasmoro, Age of the earliest known hominids in Java, Indonesia. *Science* **263**, 1118–1121 (1994). doi:10.1126/science.8108729 Medline

2. C. C. Swisher III, W. J. Rink, S. C. Antón, H. P. Schwarcz, G. H. Curtis, A. Suprijo, Widiasmoro, Latest Homo erectus of Java: Potential contemporaneity with *Homo sapiens* in southeast Asia. *Science* **274**, 1870–1874 (1996). doi:10.1126/science.274.5294.1870 Medline

3. P. Brown, T. Sutikna, M. J. Morwood, R. P. Soejono, E. Jatmiko, E. W. Saptomo, R. A. Due, A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature* **431**, 1055–1061 (2004). doi:10.1038/nature02999 Medline

4. M. J. Morwood, R. P. Soejono, R. G. Roberts, T. Sutikna, C. S. M. Turney, K. E. Westaway, W. J. Rink, J. X. Zhao, G. D. van den Bergh, R. A. Due, D. R. Hobbs, M. W. Moore, M. I. Bird, L. K. Fifield, Archaeology and age of a new hominin from Flores in eastern Indonesia. *Nature* **431**, 1087–1091 (2004). doi:10.1038/nature02956 Medline

5. T. Sutikna, M. W. Tocheri, M. J. Morwood, E. W. Saptomo, R. D. Jatmiko, R. D. Awe, S. Wasisto, K. E. Westaway, M. Aubert, B. Li, J. X. Zhao, M. Storey, B. V. Alloway, M. W. Morley, H. J. Meijer, G. D. van den Bergh, R. Grün, A. Dosseto, A. Brumm, W. L. Jungers, R. G. Roberts, Revised stratigraphy and chronology for Homo floresiensis at Liang Bua in Indonesia. *Nature* **532**, 366–369 (2016). doi:10.1038/nature17179 Medline

6. M. Aubert, A. Brumm, M. Ramli, T. Sutikna, E. W. Saptomo, B. Hakim, M. J. Morwood, G. D. van den Bergh, L. Kinsley, A. Dosseto, Pleistocene cave art from Sulawesi, Indonesia. *Nature* **514**, 223–227 (2014). doi:10.1038/nature13422 Medline

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

8. S. Xu, I. Pugach, M. Stoneking, M. Kayser, L. Jin; HUGO Pan-Asian SNP Consortium, Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 4574–4579 (2012). doi:10.1073/pnas.1118892109 Medline

9. G. D. van den Bergh, Y. Kaifu, I. Kurniawan, R. T. Kono, A. Brumm, E. Setiyabudi, F. Aziz, M. J. Morwood, Homo floresiensis-like fossils from the early Middle Pleistocene of Flores. *Nature* **534**, 245–248 (2016). doi:10.1038/nature17999 Medline

10. A. Brumm, G. M. Jensen, G. D. van den Bergh, M. J. Morwood, I. Kurniawan, F. Aziz, M. Storey, Hominins on Flores, Indonesia, by one million years ago. *Nature* **464**, 748–752 (2010). doi:10.1038/nature08844 Medline

11. T. Verhoeven, Proto-Negrito in den Grotten auf Flores. *Anthropos* **53**, 229–232 (1958).

12. T. Jacob, E. Indriati, R. P. Soejono, K. Hsü, D. W. Frayer, R. B. Eckhardt, A. J. Kuperavage, A. Thorne, M. Henneberg, Pygmoid Australomelanesian Homo sapiens skeletal remains

from Liang Bua, Flores: Population affinities and pathological abnormalities. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 13421–13426 (2006). doi:10.1073/pnas.0605563103 Medline

13. See supplementary materials.

14. A. Cooper, C. B. Stringer, Did the Denisovans cross Wallace's Line? *Science* **342**, 321–323 (2013). doi:10.1126/science.1244869 Medline

15. 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

16. J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, H. Parkinson, The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017). doi:10.1093/nar/gkw1133 Medline

17. C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, R. Collins, UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* **12**, e1001779 (2015). doi:10.1371/journal.pmed.1001779 Medline

18. K. Ye, F. Gao, D. Wang, O. Bar-Yosef, A. Keinan, Dietary adaptation of FADS genes in Europe varied across time and geography. *Nat. Ecol. Evol* **1**, 0167 (2017). doi:10.1038/s41559-017-0167 Medline

19. M. T. Buckley, F. Racimo, M. E. Allentoft, M. K. Jensen, A. Jonsson, H. Huang, F. Hormozdiari, M. Sikora, D. Marnetto, E. Eskin, M. E. Jørgensen, N. Grarup, O. Pedersen, T. Hansen, P. Kraft, E. Willerslev, R. Nielsen, Selection in Europeans on fatty acid desaturases associated with dietary changes. *Mol. Biol. Evol.* **34**, 1307–1318 (2017). doi:10.1093/molbev/msx103 Medline

20. GTEx Consortium, Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017). Medline

21. R. A. Mathias, W. Fu, J. M. Akey, H. C. Ainsworth, D. G. Torgerson, I. Ruczinski, S. Sergeant, K. C. Barnes, F. H. Chilton, Adaptive evolution of the FADS gene cluster within Africa. *PLOS ONE* **7**, e44926 (2012). doi:10.1371/journal.pone.0044926 Medline

22. K. S. D. Kothapalli, K. Ye, M. S. Gadgil, S. E. Carlson, K. O. O'Brien, J. Y. Zhang, H. G. Park, K. Ojukwu, J. Zou, S. S. Hyon, K. S. Joshi, Z. Gu, A. Keinan, J. T. Brenna, Positive Selection on a Regulatory Insertion-Deletion Polymorphism in FADS2 Influences

Apparent Endogenous Synthesis of Arachidonic Acid. *Mol. Biol. Evol.* **33**, 1726–1739 (2016). doi:10.1093/molbev/msw049 Medline

23. 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

24. M. R. Robinson, G. Hemani, C. Medina-Gomez, M. Mezzavilla, T. Esko, K. Shakhbazov, J. E. Powell, A. Vinkhuyzen, S. I. Berndt, S. Gustafsson, A. E. Justice, B. Kahali, A. E. Locke, T. H. Pers, S. Vedantam, A. R. Wood, W. van Rheenen, O. A. Andreassen, P. Gasparini, A. Metspalu, L. H. Berg, J. H. Veldink, F. Rivadeneira, T. M. Werge, G. R. Abecasis, D. I. Boomsma, D. I. Chasman, E. J. C. de Geus, T. M. Frayling, J. N. Hirschhorn, J. J. Hottenga, E. Ingelsson, R. J. F. Loos, P. K. E. Magnusson, N. G. Martin, G. W. Montgomery, K. E. North, N. L. Pedersen, T. D. Spector, E. K. Speliotes, M. E. Goddard, J. Yang, P. M. Visscher, Population genetic differentiation of height and body mass index across Europe. *Nat. Genet.* **47**, 1357–1362 (2015). doi:10.1038/ng.3401 Medline

25. J. Yang, A. Bakshi, Z. Zhu, G. Hemani, A. A. E. Vinkhuyzen, S. H. Lee, M. R. Robinson, J. R. B. Perry, I. M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, T. Esko, L. Milani, R. Mägi, A. Metspalu, A. Hamsten, P. K. E. Magnusson, N. L. Pedersen, E. Ingelsson, N. Soranzo, M. C. Keller, N. R. Wray, M. E. Goddard, P. M. Visscher; LifeLines Cohort Study, Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015). doi:10.1038/ng.3390 Medline

26. C. C. Chang, C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015). doi:10.1186/s13742-015-0047-8 Medline

27. C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, K. T. Zondervan, Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010). doi:10.1038/nprot.2010.116 Medline

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

29. G. Csardi, T. Nepusz, The igraph software package for complex network research, *InterJournal*, *Complex Systems* 1695 (2006); http://igraph.org.

30. M. A. Abdulla, I. Ahmed, A. Assawamakin, J. Bhak, S. K. Brahmachari, G. C. Calacal, A. Chaurasia, C.-H. Chen, J. Chen, Y.-T. Chen, J. Chu, E. M. C. Cutiongco-de la Paz, M. C. A. De Ungria, F. C. Delfin, J. Edo, S. Fuchareon, H. Ghang, T. Gojobori, J. Han, S.-F. Ho, B. P. Hoh, W. Huang, H. Inoko, P. Jha, T. A. Jinam, L. Jin, J. Jung, D. Kangwanpong, J. Kampuansai, G. C. Kennedy, P. Khurana, H.-L. Kim, K. Kim, S. Kim, W.-Y. Kim, K. Kimm, R. Kimura, T. Koike, S. Kulawonganunchai, V. Kumar, P. S. Lai, J.-Y. Lee, S. Lee, E. T. Liu, P. P. Majumder, K. K. Mandapati, S. Marzuki, W. Mitchell, M. Mukerji, K. Naritomi, C. Ngamphiw, N. Niikawa, N. Nishida, B. Oh, S. Oh, J.

Ohashi, A. Oka, R. Ong, C. D. Padilla, P. Palittapongarnpim, H. B. Perdigon, M. E. Phipps, E. Png, Y. Sakaki, J. M. Salvador, Y. Sandraling, V. Scaria, M. Seielstad, M. R. Sidek, A. Sinha, M. Srikummool, H. Sudoyo, S. Sugano, H. Suryadi, Y. Suzuki, K. A. Tabbada, A. Tan, K. Tokunaga, S. Tongsima, L. P. Villamor, E. Wang, Y. Wang, H. Wang, J.-Y. Wu, H. Xiao, S. Xu, J. O. Yang, Y. Y. Shugart, H.-S. Yoo, W. Yuan, G. Zhao, B. A. Zilfalil; HUGO Pan-Asian SNP Consortium; Indian Genome Variation Consortium, Mapping human genetic diversity in Asia. *Science* **326**, 1541–1545 (2009). doi:10.1126/science.1177074 Medline

31. X. Yang, S. Xu; HUGO Pan-Asian SNP Consortium; Indian Genome Variation Consortium, Identification of close relatives in the HUGO Pan-Asian SNP database. *PLOS ONE* **6**, e29502 (2011). doi:10.1371/journal.pone.0029502 Medline

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

33. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010). doi:10.1093/bioinformatics/btp698 Medline

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

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

36. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011). doi:10.1038/ng.806 Medline

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

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

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

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

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

42. 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

43. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). Medline

44. 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

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

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

47. 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

48. 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

49. 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

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

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

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

53. M. Lipson, P.-R. Loh, N. Patterson, P. Moorjani, Y.-C. Ko, M. Stoneking, B. Berger, D. Reich, Reconstructing Austronesian population history in Island Southeast Asia. *Nat. Commun.* **5**, 4689 (2014). doi:10.1038/ncomms5689 Medline

54. S. Schiffels, R. Durbin, Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014). doi:10.1038/ng.3015 Medline

55. 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

56. 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

57. A. Scally, R. Durbin, Revising the human mutation rate: Implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012). doi:10.1038/nrg3295 Medline

58. 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

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

60. G. D. Poznik, B. M. Henn, M.-C. Yee, E. Sliwerska, G. M. Euskirchen, A. A. Lin, M. Snyder, L. Quintana-Murci, J. M. Kidd, P. A. Underhill, C. D. Bustamante, Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**, 562–565 (2013). doi:10.1126/science.1237619 Medline

61. M. van Oven, A. Van Geystelen, M. Kayser, R. Decorte, M. H. D. Larmuseau, Seeing the wood for the trees: A minimal reference phylogeny for the human Y chromosome. *Hum. Mutat.* **35**, 187–191 (2014). doi:10.1002/humu.22468 Medline

62. M. Karmin, L. Saag, M. Vicente, M. A. Wilson Sayres, M. Järve, U. G. Talas, S. Rootsi, A.-M. Ilumäe, R. Mägi, M. Mitt, L. Pagani, T. Puurand, Z. Faltyskova, F. Clemente, A. Cardona, E. Metspalu, H. Sahakyan, B. Yunusbayev, G. Hudjashov, M. DeGiorgio, E.-L. Loogväli, C. Eichstaedt, M. Eelmets, G. Chaubey, K. Tambets, S. Litvinov, M. Mormina, Y. Xue, Q. Ayub, G. Zoraqi, T. S. Korneliussen, F. Akhatova, J. Lachance, S. Tishkoff, K. Momynaliev, F.-X. Ricaut, P. Kusuma, H. Razafindrazaka, D. Pierron, M. P. Cox, G. N. N. Sultana, R. Willerslev, C. Muller, M. Westaway, D. Lambert, V. Skaro, L. Kovačevic, S. Turdikulova, D. Dalimova, R. Khusainova, N. Trofimova, V. Akhmetova, I. Khidiyatova, D. V. Lichman, J. Isakova, E. Pocheshkhova, Z. Sabitov, N. A. Barashkov, P. Nymadawa, E. Mihailov, J. W. T. Seng, I. Evseeva, A. B. Migliano, S. Abdullah, G. Andriadze, D. Primorac, L. Atramentova, O. Utevska, L. Yepiskoposyan, D. Marjanovic´, A. Kushniarevich, D. M. Behar, C. Gilissen, L. Vissers, J. A. Veltman, E. Balanovska, M. Derenko, B. Malyarchuk, A. Metspalu, S. Fedorova, A. Eriksson, A. Manica, F. L. Mendez, T. M. Karafet, K. R. Veeramah, N. Bradman, M. F. Hammer, L. P. Osipova, O. Balanovsky, E. K. Khusnutdinova, K. Johnsen, M. Remm, M. G. Thomas, C. Tyler-Smith, P. A. Underhill, E. Willerslev, R. Nielsen, M. Metspalu, R. Villems, T.

Kivisild, A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res.* **25**, 459–466 (2015). doi:10.1101/gr.186684.114 Medline

63. S. Mona, K. E. Grunz, S. Brauer, B. Pakendorf, L. Castrì, H. Sudoyo, S. Marzuki, R. H. Barnes, J. Schmidtke, M. Stoneking, M. Kayser, Genetic admixture history of Eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis. *Mol. Biol. Evol.* **26**, 1865–1877 (2009). doi:10.1093/molbev/msp097 Medline

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

65. A. W. Briggs, J. M. Good, R. E. Green, J. Krause, T. Maricic, U. Stenzel, C. Lalueza-Fox, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, R. Schmitz, V. B. Doronichev, L. V. Golovanova, M. de la Rasilla, J. Fortea, A. Rosas, S. Pääbo, Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**, 318–321 (2009). doi:10.1126/science.1174462 Medline

66. H. Weissensteiner, D. Pacher, A. Kloss-Brandstätter, L. Forer, G. Specht, H.-J. Bandelt, F. Kronenberg, A. Salas, S. Schönherr, HaploGrep 2: Mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–W63 (2016). doi:10.1093/nar/gkw233 Medline

67. M. van Oven, PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Sci. International. Genet. Suppl. Ser.* **5**, e392–e394 (2015). doi:10.1016/j.fsigss.2015.09.155

68. C. Hill, P. Soares, M. Mormina, V. Macaulay, D. Clarke, P. B. Blumbach, M. Vizuete-Forster, P. Forster, D. Bulbeck, S. Oppenheimer, M. Richards, A mitochondrial stratigraphy for island southeast Asia. *Am. J. Hum. Genet.* **80**, 29–43 (2007). doi:10.1086/510412 Medline

69. A. T. Duggan, B. Evans, F. R. Friedlaender, J. S. Friedlaender, G. Koki, D. A. Merriwether, M. Kayser, M. Stoneking, Maternal history of Oceania from complete mtDNA genomes: Contrasting ancient diversity with recent homogenization due to the Austronesian expansion. *Am. J. Hum. Genet.* **94**, 721–733 (2014). doi:10.1016/j.ajhg.2014.03.014 Medline

70. E. D. Gunnarsdóttir, M. R. Nandineni, M. Li, S. Myles, D. Gil, B. Pakendorf, M. Stoneking, Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. *Nat. Commun.* **2**, 228 (2011). doi:10.1038/ncomms1235 Medline

71. M. Ingman, U. Gyllensten, Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. *Genome Res.* **13**, 1600–1606 (2003). doi:10.1101/gr.686603 Medline

72. A. M.-S. Ko, C.-Y. Chen, Q. Fu, F. Delfin, M. Li, H.-L. Chiu, M. Stoneking, Y.-C. Ko, Early Austronesians: Into and out of Taiwan. *Am. J. Hum. Genet.* **94**, 426–436 (2014). doi:10.1016/j.ajhg.2014.02.003 Medline

73. S. Lippold, H. Xu, A. Ko, M. Li, G. Renaud, A. Butthof, R. Schröder, M. Stoneking, Human paternal and maternal demographic histories: Insights from high-resolution Y chromosome and mtDNA sequences. *Investig. Genet.* **5**, 13 (2014). doi:10.1186/2041-2223-5-13 Medline

74. M. G. Palanichamy, C. Sun, S. Agrawal, H.-J. Bandelt, Q.-P. Kong, F. Khan, C.-Y. Wang, T. K. Chaudhuri, V. Palla, Y.-P. Zhang, Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: Implications for the peopling of South Asia. *Am. J. Hum. Genet.* **75**, 966–978 (2004). doi:10.1086/425871 Medline

75. M.-S. Peng, H. H. Quang, K. P. Dang, A. V. Trieu, H.-W. Wang, Y.-G. Yao, Q.-P. Kong, Y.-P. Zhang, Tracing the Austronesian footprint in Mainland Southeast Asia: A perspective from mitochondrial DNA. *Mol. Biol. Evol.* **27**, 2417–2430 (2010). doi:10.1093/molbev/msq131 Medline

76. E. D. Gunnarsdóttir, M. Li, M. Bauchet, K. Finstermeier, M. Stoneking, High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Res.* **21**, 1–11 (2011). doi:10.1101/gr.107615.110 Medline

77. T. A. Jinam, L.-C. Hong, M. E. Phipps, M. Stoneking, M. Ameen, J. Edo, N. Saitou; HUGO Pan-Asian SNP Consortium, Evolutionary history of continental southeast Asians: "early train" hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. *Mol. Biol. Evol.* **29**, 3513–3527 (2012). doi:10.1093/molbev/mss169 Medline

78. A. J. Drummond, M. A. Suchard, D. Xie, A. Rambaut, Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012). doi:10.1093/molbev/mss075 Medline

79. P. Soares, L. Ermini, N. Thomson, M. Mormina, T. Rito, A. Röhl, A. Salas, S. Oppenheimer, V. Macaulay, M. B. Richards, Correcting for purifying selection: An improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* **84**, 740–759 (2009). doi:10.1016/j.ajhg.2009.05.001 Medline

80. D. Darriba, G. L. Taboada, R. Doallo, D. Posada, jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012). doi:10.1038/nmeth.2109 Medline

81. G. Baele, W. L. S. Li, A. J. Drummond, M. A. Suchard, P. Lemey, Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Mol. Biol. Evol.* **30**, 239–243 (2013). doi:10.1093/molbev/mss243 Medline

82. R. E. Kass, A. E. Raftery, Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995). doi:10.1080/01621459.1995.10476572

83. T. Kivisild, Maternal ancestry and population history from whole mitochondrial genomes. *Investig. Genet.* **6**, 3 (2015). doi:10.1186/s13323-015-0022-2 Medline

84. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLOS Genet.* **2**, e190 (2006). doi:10.1371/journal.pgen.0020190 Medline

85. D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. Johnson, T. Maricic, J. M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. E. Eichler, M. Stoneking, M. Richards, S. Talamo,

M. V. Shunkov, A. P. Derevianko, J. J. Hublin, J. Kelso, M. Slatkin, S. Pääbo, Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010). doi:10.1038/nature09710 Medline

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

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

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

89. S. R. Browning, B. L. Browning, Y. Zhou, S. Tucci, J. M. Akey, Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173**, 53–61.e9 (2018). doi:10.1016/j.cell.2018.02.031 Medline

90. M. D. Rasmussen, M. J. Hubisz, I. Gronau, A. Siepel, Genome-wide inference of ancestral recombination graphs. *PLOS Genet.* **10**, e1004342 (2014). doi:10.1371/journal.pgen.1004342 Medline

91. M. Kuhlwilm, I. Gronau, M. J. Hubisz, C. de Filippo, J. Prado-Martinez, M. Kircher, Q. Fu, H. A. Burbano, C. Lalueza-Fox, M. de la Rasilla, A. Rosas, P. Rudan, D. Brajkovic, Ž. Kucan, I. Gušic, T. Marques-Bonet, A. M. Andrés, B. Viola, S. Pääbo, M. Meyer, A. Siepel, S. Castellano, Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**, 429–433 (2016). doi:10.1038/nature16544 Medline

92. T. Duong, B. Goud, K. Schauer, Closed-form density-based framework for automatic detection of cellular morphology changes. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8382–8387 (2012). doi:10.1073/pnas.1117796109 Medline

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

94. K. Harris, R. Nielsen, The Genetic Cost of Neanderthal Introgression. *Genetics* **203**, 881–891 (2016). doi:10.1534/genetics.116.186890 Medline

95. I. Juric, S. Aeschbacher, G. Coop, The Strength of Selection against Neanderthal Introgression. *PLOS Genet.* **12**, e1006340 (2016). doi:10.1371/journal.pgen.1006340 Medline

96. P. H. Sudmant, J. Huddleston, C. R. Catacchio, M. Malig, L. W. Hillier, C. Baker, K. Mohajeri, I. Kondova, R. E. Bontrop, S. Persengiev, F. Antonacci, M. Ventura, J. Prado-Martinez, T. Marques-Bonet, E. E. Eichler; Great Ape Genome Project, Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* **23**, 1373–1382 (2013). doi:10.1101/gr.158543.113 Medline

97. Q. Fu, H. Li, P. Moorjani, F. Jay, S. M. Slepchenko, A. A. Bondarev, P. L. F. Johnson, A. Aximu-Petri, K. Prüfer, C. de Filippo, M. Meyer, N. Zwyns, D. C. Salazar-García, Y. V. Kuzmin, S. G. Keates, P. A. Kosintsev, D. I. Razhev, M. P. Richards, N. V. Peristov, M. Lachmann, K. Douka, T. F. G. Higham, M. Slatkin, J.-J. Hublin, D. Reich, J. Kelso, T. B.

Viola, S. Pääbo, Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014). doi:10.1038/nature13810 Medline

98. M. D. Shriver, G. C. Kennedy, E. J. Parra, H. A. Lawson, V. Sonpar, J. Huang, J. M. Akey, K. W. Jones, The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum. Genomics* **1**, 274–286 (2004). doi:10.1186/1479-7364-1-4-274 Medline

99. 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

100. B. S. Weir, C. C. Cockerham, Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984). Medline

101. P. Parham, T. Ohta, Population biology of antigen presentation by MHC class I molecules. *Science* **272**, 67–74 (1996). doi:10.1126/science.272.5258.67 Medline

102. P. J. Norman, J. A. Hollenbach, N. Nemat-Gorgani, L. A. Guethlein, H. G. Hilton, M. J. Pando, K. A. Koram, E. M. Riley, L. Abi-Rached, P. Parham, Co-evolution of human leukocyte antigen (HLA) class I ligands with killer-cell immunoglobulin-like receptors (KIR) in a genetically diverse population of sub-Saharan Africans. *PLOS Genet.* **9**, e1003938 (2013). doi:10.1371/journal.pgen.1003938 Medline

103. R. J. Pruim, R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines, T. P. Gliedt, M. Boehnke, G. R. Abecasis, C. J. Willer, LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010). doi:10.1093/bioinformatics/btq419 Medline

104. C. Glaser, J. Heinrich, B. Koletzko, Role of FADS1 and FADS2 polymorphisms in polyunsaturated fatty acid metabolism. *Metabolism* **59**, 993–999 (2010). doi:10.1016/j.metabol.2009.10.022 Medline

105. A. Ameur, S. Enroth, A. Johansson, G. Zaboli, W. Igl, A. C. Johansson, M. A. Rivas, M. J. Daly, G. Schmitz, A. A. Hicks, T. Meitinger, L. Feuk, C. van Duijn, B. Oostra, P. P. Pramstaller, I. Rudan, A. F. Wright, J. F. Wilson, H. Campbell, U. Gyllensten, Genetic adaptation of fatty-acid metabolism: A human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *Am. J. Hum. Genet.* **90**, 809–820 (2012). doi:10.1016/j.ajhg.2012.03.014 Medline

106. J. H. Marcus, J. Novembre, Visualizing the geography of genetic variants. *Bioinformatics* **33**, 594–595 (2017). Medline

107. M. Raghavan, M. DeGiorgio, A. Albrechtsen, I. Moltke, P. Skoglund, T. S. Korneliussen, B. Grønnow, M. Appelt, H. C. Gulløv, T. M. Friesen, W. Fitzhugh, H. Malmström, S. Rasmussen, J. Olsen, L. Melchior, B. T. Fuller, S. M. Fahrni, T. Stafford Jr., V. Grimes, M. A. P. Renouf, J. Cybulski, N. Lynnerup, M. M. Lahr, K. Britton, R. Knecht, J. Arneborg, M. Metspalu, O. E. Cornejo, A.-S. Malaspinas, Y. Wang, M. Rasmussen, V. Raghavan, T. V. O. Hansen, E. Khusnutdinova, T. Pierre, K. Dneprovsky, C. Andreasen, H. Lange, M. G. Hayes, J. Coltrain, V. A. Spitsyn, A. Götherström, L. Orlando, T.

Kivisild, R. Villems, M. H. Crawford, F. C. Nielsen, J. Dissing, J. Heinemeier, M. Meldgaard, C. Bustamante, D. H. O'Rourke, M. Jakobsson, M. T. P. Gilbert, R. Nielsen, E. Willerslev, The genetic prehistory of the New World Arctic. *Science* **345**, 1255832 (2014). doi:10.1126/science.1255832 Medline

108. C. E. G. Amorim, K. Nunes, D. Meyer, D. Comas, M. C. Bortolini, F. M. Salzano, T. Hünemeier, Genetic signature of natural selection in first Americans. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 2195–2199 (2017). doi:10.1073/pnas.1620541114 Medline

109. T. Illig, C. Gieger, G. Zhai, W. Römisch-Margl, R. Wang-Sattler, C. Prehn, E. Altmaier, G. Kastenmüller, B. S. Kato, H. W. Mewes, T. Meitinger, M. H. de Angelis, F. Kronenberg, N. Soranzo, H. E. Wichmann, T. D. Spector, J. Adamski, K. Suhre, A genome-wide perspective of genetic variation in human metabolism. *Nat. Genet.* **42**, 137–141 (2010). doi:10.1038/ng.507 Medline

110. J. Kettunen, T. Tukiainen, A.-P. Sarin, A. Ortega-Alonso, E. Tikkanen, L.-P. Lyytikäinen, A. J. Kangas, P. Soininen, P. Würtz, K. Silander, D. M. Dick, R. J. Rose, M. J. Savolainen, J. Viikari, M. Kähönen, T. Lehtimäki, K. H. Pietiläinen, M. Inouye, M. I. McCarthy, A. Jula, J. Eriksson, O. T. Raitakari, V. Salomaa, J. Kaprio, M.-R. Järvelin, L. Peltonen, M. Perola, N. B. Freimer, M. Ala-Korpela, A. Palotie, S. Ripatti, Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* **44**, 269–276 (2012). doi:10.1038/ng.1073 Medline

111. W. Guan *et al*., Genome-Wide Association Study of Plasma N6 Polyunsaturated Fatty Acids within the CHARGE Consortium. *Circ. Cardiovasc. Genet.* **7**, 321–331 (2014). doi:10.1161/CIRCGENETICS.113.000208 Medline

112. H. H. M. Draisma, R. Pool, M. Kobl, R. Jansen, A.-K. Petersen, A. A. M. Vaarhorst, I. Yet, T. Haller, A. Demirkan, T. Esko, G. Zhu, S. Böhringer, M. Beekman, J. B. van Klinken, W. Römisch-Margl, C. Prehn, J. Adamski, A. J. M. de Craen, E. M. van Leeuwen, N. Amin, H. Dharuri, H.-J. Westra, L. Franke, E. J. C. de Geus, J. J. Hottenga, G. Willemsen, A. K. Henders, G. W. Montgomery, D. R. Nyholt, J. B. Whitfield, B. W. Penninx, T. D. Spector, A. Metspalu, P. E. Slagboom, K. W. van Dijk, P. A. C. 't Hoen, K. Strauch, N. G. Martin, G. B. van Ommen, T. Illig, J. T. Bell, M. Mangino, K. Suhre, M. I. McCarthy, C. Gieger, A. Isaacs, C. M. van Duijn, D. I. Boomsma, Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat. Commun.* **6**, 7208 (2015). doi:10.1038/ncomms8208 Medline

113. O. Canela-Xandri, K. Rawlik, A. Tenesa, An atlas of genetic associations in UK Biobank. bioRxiv 176834 [Preprint]. 18 August 2017. https://doi.org/10.1101/176834.

114. H. T. Reardon, J. Zhang, K. S. D. Kothapalli, A. J. Kim, W. J. Park, J. T. Brenna, Insertion-deletions in a FADS2 intron 1 conserved regulatory locus control expression of fatty acid desaturases 1 and 2 and modulate response to simvastatin. *Prostaglandins Leukot. Essent. Fatty Acids* **87**, 25–33 (2012). doi:10.1016/j.plefa.2012.04.011 Medline

115. P. Deelen, M. J. Bonder, K. J. van der Velde, H.-J. Westra, E. Winder, D. Hendriksen, L. Franke, M. A. Swertz, Genotype harmonizer: Automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7**, 901 (2014). doi:10.1186/1756-0500-7-901 Medline

116. B. N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genet.* **5**, e1000529 (2009). doi:10.1371/journal.pgen.1000529 Medline

117. A. L. Williams, N. Patterson, J. Glessner, H. Hakonarson, D. Reich, Phasing of many thousands of genotyped samples. *Am. J. Hum. Genet.* **91**, 238–251 (2012). doi:10.1016/j.ajhg.2012.06.013 Medline

118. Haplotype Reference Consortium, A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016). doi:10.1038/ng.3643 Medline

119. P.-R. Loh, G. Tucker, B. K. Bulik-Sullivan, B. J. Vilhjálmsson, H. K. Finucane, R. M. Salem, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, N. Patterson, A. L. Price, Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015). doi:10.1038/ng.3190 Medline

120. G. Abraham, Y. Qiu, M. Inouye, FlashPCA2: Principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017). doi:10.1093/bioinformatics/btx299 Medline

121. J. Yang, N. A. Zaitlen, M. E. Goddard, P. M. Visscher, A. L. Price, Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014). doi:10.1038/ng.2876 Medline

122. M. C. Turchin, C. W. K. Chiang, C. D. Palmer, S. Sankararaman, D. Reich, J. N. Hirschhorn; Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat. Genet.* **44**, 1015–1019 (2012). doi:10.1038/ng.2368 Medline

123. M. Zoledziewska, C. Sidore, C. W. K. Chiang, S. Sanna, A. Mulas, M. Steri, F. Busonero, J. H. Marcus, M. Marongiu, A. Maschio, D. Ortega Del Vecchyo, M. Floris, A. Meloni, A. Delitala, M. P. Concas, F. Murgia, G. Biino, S. Vaccargiu, R. Nagaraja, K. E. Lohmueller, N. J. Timpson, N. Soranzo, I. Tachmazidou, G. Dedoussis, E. Zeggini, S. Uzzau, C. Jones, R. Lyons, A. Angius, G. R. Abecasis, J. Novembre, D. Schlessinger, F. Cucca; UK10K consortium; Understanding Society Scientific Group, Height-reducing variants and selection for short stature in Sardinia. *Nat. Genet.* **47**, 1352–1356 (2015). doi:10.1038/ng.3403 Medline

124. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011). doi:10.1016/j.ajhg.2010.11.011 Medline

125. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, D. Haussler, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). doi:10.1101/gr.229102 Medline

126. 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