Title: Insights into human genetic variation and population history from 929 diverse genomes

Authors:

Anders Bergström[1,2,*], Shane A. McCarthy[1,3,‡], Ruoyun Hui[3,4,‡], Mohamed A. Almarri[1,‡], Qasim Ayub[1,5,6], Petr Danecek[1], Yuan Chen[1], Sabine Felkel[1,7], Pille Hallast[1,8], Jack Kamm[1,3,9], Hélène Blanché[10,11], Jean-François Deleuze[10,11], Howard Cann[10,†], Swapan Mallick[12,13], David Reich[12,13], Manjinder S. Sandhu[1,14], Pontus Skoglund[2], Aylwyn Scally[3], Yali Xue[1,§], Richard Durbin[1,3,§], Chris Tyler-Smith[1,§,*]

Affiliations:

1. Wellcome Sanger Institute, Hinxton, CB10 1SA, UK
2. The Francis Crick Institute, London, NW1 1AT, UK
3. Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK
4. McDonald Institute for Archaeological Research, University of Cambridge, CB2 3ER, UK
5. Monash University Malaysia Genomics Facility, Tropical Medicine and Biology Multidisciplinary Platform, 47500 Bandar Sunway, Malaysia
6. School of Science, Monash University Malaysia, 47500 Bandar Sunway, Malaysia
7. Institute of Animal Breeding and Genetics, University of Veterinary Medicine Vienna, Vienna, 1210, Austria
8. Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu 50411, Estonia
9. Chan Zuckerberg Biohub, San Francisco, 94158, USA
10. Centre d'Etude du Polymorphisme Humain, Fondation Jean Dausset, 75010 Paris, France
11. GENMED Labex, Paris, France, ANR-10-LABX-0013
12. Department of Genetics, Harvard Medical School, Boston, 02115, USA
13. Broad Institute of Harvard and MIT, Cambridge, 02142, USA
14. Department of Medicine, University of Cambridge, Cambridge, CB2 0QQ, UK

*Correspondence to. Email: ab34@sanger.ac.uk (A.B.); cts@sanger.ac.uk (C.T.-S.)
‡These authors contributed equally to this work.
§These authors contributed equally to this work.
†Deceased.

Abstract:

Genome sequences from diverse human groups are needed to understand the structure of genetic variation in our species and the history of, and relationships between, different populations. We present 929 high-coverage genome sequences from 54 diverse human populations, 26 of which are physically phased using linked-read sequencing. Analyses of these genomes reveal an excess of previously undocumented common genetic variation private to each of southern Africa, central Africa, Oceania and the Americas, but an absence of such variants fixed between major geographical regions. We also find deep and gradual population separations within Africa, contrasting population size histories between hunter-gatherer and agriculturalist groups in the last 10,000 years, and a contrast between single

Neanderthal but multiple Denisovan source populations contributing to present-day human populations.

One Sentence Summary: Genomes from 54 diverse populations expand the genomic record of human diversity and illuminate the history of our species

Main Text:

Genome sequences from diverse human groups can reveal the structure of genetic variation in our species and the history of, and relationships between, different populations. They also provide a framework for the design and interpretation of medical genetics studies. A consensus view of the history of our species includes divergence from the ancestors of the archaic Neanderthal and Denisovan groups 500,000-700,000 years ago, the appearance of anatomical modernity in Africa in the last few hundred thousand years, an expansion out of Africa and the Near East 50,000-70,000 years ago, with a reduction in genetic diversity in the descendant populations, admixture with archaic groups in Eurasia shortly after this and large-scale population growth, migration and admixture following multiple independent transitions from hunter-gatherer to food producing lifestyles in the last 10,000 years (*1*). However, much still remains to be understood about the extent to which population histories differed between continents and regions, and how this has shaped the present-day distribution and structure of genetic variation across the species. Large-scale genome sequencing efforts to date have been restricted to large, metropolitan populations and employed low-coverage sequencing (*2*), while those sampling human groups more widely have mostly been limited to 1-3 genomes per population (*3, 4*). The Human Genome Diversity Project (HGDP)-CEPH panel (*5*) has constituted a key resource to which several iterations of genetic assays have been applied (*3, 6-12*). Here, we present 929 high-coverage genome sequences from 54 geographically, linguistically and culturally diverse populations (Fig. 1A, table S1) from this panel, 142 of which were previously sequenced (*3, 11, 13*).

Genetic variant discovery across diverse human populations

We performed Illumina sequencing to an average coverage of 35x (min: 25x) and mapped reads to the GRCh38 reference assembly. We also used linked-read technology (*14*) to physically resolve the haplotype phase of 26 of these genomes from 13 populations (table S2). By analysing local sequencing coverage across the genome, we identified and excluded nine samples with large-scale alterations in chromosomal copy numbers that likely arose

during lymphoblastoid cell line culturing. The remaining individuals provided high-quality genotype calls (figs. S1, S2, S3). In this set of 929 genomes we identified 67.3 million single-nucleotide polymorphisms (SNPs), 8.8 million small insertions or deletions (indels) and 40,736 copy number variants (CNVs) (*15*). This is nearly as many as the 84.7 million SNPs discovered in 2504 individuals by the 1000 Genomes Project (*2*), reflecting increased sensitivity due to high-coverage sequencing as well as the greater diversity of human ancestries covered by the HGDP-CEPH panel. While the vast majority of the variants discovered by one of the studies but not the other are very low in frequency, the HGDP dataset contains substantial numbers of variants that were not identified by the 1000 Genomes Project but are common or even high-frequency in some populations: ~1 million variants at ≥20%, ~100,000 variants at ≥50% and even ~1000 variants fixed at 100% frequency in at least one sampled population (Fig. 1B). This highlights the importance of anthropologically informed sampling for uncovering human genetic diversity.

The unbiased variant discovery enabled by whole-genome sequencing avoids potential ascertainment biases associated with the pre-defined variant sets used on genotyping arrays. We find that while analyses of the SNPs included on commonly used arrays accurately recapitulate relationships between non-African populations, they sometimes dramatically distort relationships involving African populations (Fig. 1C). Some of the f_4-statistics commonly used to study population history and admixture (*10*) even shift sign when using array SNPs compared to when using all discovered SNPs, thus incorrectly reversing the direction of the implied ancestry relationship (for example: f_4*(BantuKenya,San;Mandenka,Sardinian)* is positive (Z=2.9) using all variants but negative (Z=-3.11) when using commonly employed array sites). A set of 1.3 million SNPs ascertained as polymorphic among three archaic human genomes, mainly reflecting shared ancestral variation (69% of them being polymorphic in Africa), provide more accurate f_4-statistics than the variants on commonly used arrays, as well as more accurate F_{ST} values and cleaner estimates of individual ancestries in model-based clustering analyses (fig. S4), consistent with the theoretical properties of outgroup-ascertained variants (*10*).

Rare variants, largely absent from genotyping arrays, are more likely to derive from recent mutation and can therefore inform upon recently shared ancestry between individuals. The patterns of rare variant sharing across the 929 genomes reveal abundant structure (Fig. 2A), as well as a general pattern of greater between-population rare allele sharing among Eurasian

as opposed to Oceanian and American populations. We do not find a general increase in the power to detect population relationships in the form of non-zero f_4-statistics when using all the discovered SNPs, most of which are rare, compared to using just the ~600,000 variants present on commonly used genotyping arrays (Fig. 1C). However, stratifying *D*-statistics by derived allele frequency can reveal more nuanced views of population relationships (*16*). In the presence of admixture, statistics of the form *D(Chimp,X;A,B)*, quantifying the extent to which the allele frequencies of X are closer to those of A or B, can take different values for variants that have different derived allele frequencies in X. For example, we find that the West African Yoruba have a closer relationship to non-Africans than to the central African Mbuti at high allele frequencies but the opposite relationship at low frequencies (Fig. 2B), suggesting recent gene flow between Mbuti and Yoruba since the divergence of non-Africans. An excess sharing of San with Mandenka relative to Mbuti at low allele frequencies may similarly reflect low amounts of West African-related admixture into San (Fig. 2C) (*17*). The known Denisovan admixture in Oceanian populations manifests itself, without making use of any archaic genome sequences, in a greater affinity of African populations to Eurasians over Oceanians especially at variants that are fixed in Africans (Fig. 2D). In a manner analogous to this, at fixed variants the central African Biaka have much greater affinity to Yoruba than to the Mandenka, another West African population (Fig. 2E), which would be consistent with Mandenka having some ancestry that is basal to other African ancestries (*18*).

The Y chromosome sequences in the dataset recapitulate the well-understood structure of the human Y chromosome phylogeny, but also contain a number of rare lineages of interest (figs. S12, S13). An F* lineage representing the deepest known split in the FT branch that is carried by the vast majority of non-African men was found only once across the 1205 males of the 1000 Genomes Project (*19*). Here, we find it in five out of seven sampled males in the Lahu from Yunnan province in southern China (who also carry high levels of population-specific rare autosomal alleles (Fig. 2A)), pointing to the importance of East Asia for understanding the early dispersal of non-African Y chromosomes, and highlighting how sequencing of diverse human groups can recover genetic lineages that are globally rare.

The extremes of human genetic differentiation

We next studied the extremes of human genetic variation by identifying variants that are private to geographic regions (excluding individuals with likely recent admixture from other

regions, table S4). We find no such private variants that are fixed in a given continent or major region (Fig. 3A-C). The highest frequencies are reached by a few tens of variants present at >70% (and a few thousands at >50%) in each of Africa, the Americas and Oceania. In contrast, the highest frequency variants private to either Europe, East Asia, the Middle East or Central and South Asia reach just 10-30%. This likely reflects greater genetic connectivity within Eurasia owing to culturally driven migrations and admixture in the last 10,000 years, events which did not involve the more isolated populations of the Americas and Oceania (*1*), allowing variation accumulating in the latter to remain private. Even comparing Central and South America, we find variants private to one region but absent from the other reaching >40% frequency. Within Africa, ~1000 variants private to the rainforest hunter-gatherer groups Mbuti and Biaka reach >30%, and the highly diverged San of southern Africa harbour ~100,000 private variants at >30% frequency, ~1000 at >60% and even about 20 that are fixed in our small sample of six individuals.

The vast majority of these geographically restricted variants reflect novel mutations that occurred after, or shortly before, the diversification of present-day groups, with >99% of alleles private to most non-African regions being the derived rather than the ancestral allele (Fig. 3D). Alleles private to Africa, however, include a higher proportion of ancestral alleles, and this proportion increases with allele frequency, reflecting old variants that have been lost outside of Africa. For the same reason, many high frequency private African variants are also found in available Neanderthal or Denisovan genomes (*11, 16, 20*) (Fig. 3E). The fraction of variants private to any given region outside of Africa that are shared with archaic genomes is very low, consistent with most or all gene flow from these archaic groups having occurred before the diversification of present-day non-African ancestries. The exception to this is Oceania, in which at least ~35% of private variants present at ≥20% frequency are shared with the Denisovan genome. Generally, at least ~20% of common (>10% allele frequency) variants that are present outside of Africa but absent inside Africa are shared with and thus likely derive from admixture with Neanderthals and Denisovans (Fig. 3F). The remaining up to ~80% of such common variants are more likely to have derived from novel mutations, which thus have been a stronger force than archaic admixture in introducing novel variants into present-day human populations.

Indel variants private to geographic regions display frequency distributions similar to those of SNPs, although reduced in overall numbers by approximately 10-fold (Fig. 3B). The same is

mostly true of CNVs, with an even greater reduction in overall numbers, except for a slight excess of high-frequency private CNVs in Oceanians over what would be expected on the basis of the number of private Oceanian SNPs (Fig. 3C, fig. S5). Several of these variants are shared with the available Denisovan genome, suggesting that, relative to other variant classes and geographical regions, positive selection may have acted with a disproportionate strength on copy number variants of archaic origin in the history of Oceanian populations.

Effective population size histories

We next examined what present-day patterns of genetic variation can tell us about the past demographic histories of different human populations. The distribution of coalescence times between chromosomes sampled from the same population can be used to infer changes in effective population size over time (*21, 22*). However, resolution in recent times is limited when analysing single human genomes, and haplotype phasing errors can cause artefacts when using multiple genomes (*23, 24*). We therefore applied SMC++ (*24*) which extends this approach to incorporate information from the site frequency spectrum as estimated from a larger number of unphased genomes, enabling inference of effective population sizes into more recent time periods (Fig. 4A). In Europe and East Asia, most populations are inferred to have experienced major growth in the last 10,000 years, but less so in more isolated groups, including the European Sardinians, Basques, Orkney islanders, the southern Chinese Lahu and the Siberian Yakut. In Africa, while the sizes of agriculturalist populations increased over the last 10,000 years, those of the hunter-gatherer groups, Biaka, Mbuti and San, saw no growth or even declined. These findings may reflect a more general pattern of human prehistory, in which hunter-gatherer groups which previously might have been more numerous and widespread decreased in size as agriculturalist groups expanded (*25*).

We also find tentative evidence for population growth in the ancestors of Native Americans coinciding with entry into the American continents ~15 kya (Fig. 4B), mirroring observations of rapid diversification of mitochondrial and Y-chromosome lineages at this time (*26, 27*) but not previously observed with autosomal data. The inference is sensitive to SMC++ parameter settings and likely counteracted by very recent bottlenecks in the Native American groups, but other populations do not display similar histories under these parameter settings (fig. S10). While this finding might be a technical artefact and will require further validation, the inferred growth rate exceeds even those of large European and East Asian populations in the

last 10,000 years, suggesting this could be one of the most dramatic growth episodes in modern human population history.

While informative, these analyses still appear to have limited resolution to infer more fine-scale population size histories during the transitions to agriculture, metal ages and other cultural processes that have occurred during the last 10,000 years. This might require yet larger sample sizes, novel analytical methods that exploit other features of genetic variation (*28*), or both.

The time depth and mode of human population separations

We used the 26 genomes physically phased by linked-read technology to study the time-course of population separations using the MSMC2 method (*22, 29*). As a heuristic approximation to the split time between two populations we take the point at which the estimated rate of coalescence between them is half of the rate of coalescence within them, but we also assess how gradual or extended over time the splits were by comparing the shape of the curves to those obtained by running the method on simulated instant split scenarios without subsequent gene flow. Assuming a mutation rate of 1.25×10^{-8} per base-pair per generation (*30*) and a generation time of 29 years (*31*), our midpoint estimates suggest (Fig. 5A) splits between the two central African rain forest hunter-gatherer groups Mbuti and Biaka ~62 kya, Mbuti and the West African Yoruba ~69 kya, Yoruba and the southern African San ~126 kya and between San and both of Biaka and Mbuti ~110 kya. Non-Africans have separation midpoints from Yoruba ~76 kya, Biaka ~96 kya, Mbuti ~123 kya and, representing the deepest split in the dataset, from San ~162 kya. However, all of these curves are clearly inconsistent with clean splits, suggesting a picture where genetic separations within Africa were gradual and shaped by ongoing gene flow over tens of thousands of years. For example, there is evidence of gene flow between San and Biaka until at least 50 kya, and between each of Mbuti, Biaka and Yoruba until the present day or as recently as the method can infer.

For the deepest splits, there is some evidence of genetic separation dating back to before 300 or even 500 kya, in the sense that even by that time the rate of coalescence between populations still differs from that within populations. The implication of this would be that there lived populations already at this time which have contributed more to some present-day human ancestries than to others. We find that a small degree of such deep structure in

MSMC2 curves might be spuriously caused by batch effects associated with sequencing and genotyping pairs of chromosomes from diploid human samples together, but that such effects are not large enough to fully explain the differences in coalescence rates at these time scales (fig. S7). However, even if this signal reflects actual ancient population structure, its magnitude is such that it would only apply to small fractions of present-day ancestries. An analogy to this is how Neanderthal and Denisovan admixture results in a few percent of non-African ancestries separating from some African ancestries approximately half a million years ago, while most of the ancestry was connected until much more recently. We argue, in the light of such composite ancestries in present-day human populations and the clear deviation of our MSMC2 results from instant split behaviours, that single point estimates are inadequate for describing the timing of early modern human population separations. A more meaningful summary of our results might be that the structure we observe among human populations today formed predominantly during the last 250 ky, with continued genetic contact between all populations during much of this time, but also a small fraction of present-day ancestries retaining traces of structure that is older than this, potentially by hundreds of thousands of years.

We also applied MSMC2 to the history of separation between archaic and modern human populations. While the method relies on phased haplotypes, the high degree of homozygosity of Neanderthals and Denisovans means that it might still perform well despite the absence of phase information for heterozygous sites in these genomes. The midpoint estimates suggest that modern and archaic populations separated 550-700 kya (Fig. 5A), in line with, but potentially slightly earlier than, estimates obtained with other methods (*16, 20*). These results also provide relative constraints on the overall time depth of modern human structure that are independent of the mutation rate we use to scale the results, in the sense that the deepest modern human midpoints are less than one-third of the age of the midpoints of the archaic curves. However, the deep tails of some modern human curves partly overlap a time period when genetic separation from the archaics might still not have been complete. The separation between archaic and modern humans appears more sudden than those between different modern human populations, and only slightly less sudden than expected under an instant split scenario, suggesting a qualitatively different mode of separation between modern and archaic groups than between modern human groups within Africa. While the divergence time between modern human and Neanderthal mitochondrial genomes shows that there is at least some ancestry shared more recently than 500 kya (*32*), these MSMC2 results suggest that

post-split gene flow to and from the archaic groups, likely geographically restricted to Eurasia, overall would have been limited.

Outside of Africa, the time depths of population splits are in line with previous estimates (*3, 4, 22*), with all populations sharing most of their ancestry within the last 70 kya (Fig. 5B). Our analyses of these physically phased genomes do not replicate a previously observed earlier divergence of West Africans from Oceanians than from Eurasians in MSMC analyses (*4, 29*), suggesting those results were caused by some artefact of statistical phasing. Instead, all non-African populations display very similar histories of separation from African populations (fig. S6). Like those within Africa, many curves between non-African populations are more gradual than instant split simulations. However, some curves, including those between the Central American Pima and the South American Karitiana, between Han Chinese and the Siberian Yakut, or between the European Sardinians and the Near Eastern Druze, do not deviate appreciably from those expected under instant splits. This suggests that once modern humans had expanded into the geographically diverse and fragmented continents outside of Africa, populations would sometimes separate suddenly and without much subsequent gene flow.

We also fit simple pairwise split models for the complete set of 1431 population pairs to the site-frequency spectrum using momi2 (*33*), obtaining estimates with high concordance to the MSMC2 midpoints ($r = 0.93$). This much larger set of split time estimates is consistent with present-day populations sharing the majority of their ancestry within the last 200 kya. Using these estimates, we also find that the strength of allele frequency differentiation between populations (F_{ST}) relative to split times is about three times greater outside than inside of Africa (Fig. 5C). This could partly reflect increased rates of drift in some non-African populations, but is likely largely explained by the amplifying effects on F_{ST} of the reduced diversity of these groups following their shared bottleneck event (*34*).

The genetic contribution of archaic hominins to present-day human populations

We estimate an average of 2.4% and 2.1% Neanderthal ancestry in eastern non-Africans and western non-Africans, respectively. We estimate 2.8% (95% confidence interval: 2.1-3.6%) Denisovan ancestry in Papuan highlanders (*15*), substantially lower than the first estimate of 4-6% (*35*) based on less comprehensive modern and archaic data, but only slightly lower than more recent estimates (*11, 36, 37*). The proportion of ancestry that remains in present-day

Oceanian populations after the Denisovan admixture is thus likely not much higher than the amount of Neanderthal ancestry that remains in non-Africans generally.

We identified Neanderthal and Denisovan segments in non-African genomes using a hidden Markov model (*15*), and studied the diversity of these haplotypes to learn about the structure of these admixture events and whether they involved one or more source populations. For Neanderthals, several lines of evidence are consistent with there having been a single source with no apparent contribution from any additional population which was detectably different in terms of ancestry, geographical distribution or admixture time. Neanderthal segments recovered from modern genomes across the world show very similar distributions along the genome (fig. S18 and table S8) and profiles of divergence to available archaic genomes (fig. S19), and different Neanderthal haplotypes detected at the same location in modern genomes rarely form geographically structured clusters (fig. S23, table S10). The structure of absolute divergence (D_{XY}) in Neanderthal segments between pairs of non-African populations mirrors that in unadmixed segments (Fig. 6A), suggesting a shared admixture event before these populations diverged from each other. A substantial later episode of admixture from Neanderthals into one or more modern populations would have resulted in greater structure (more divergence between some populations) in the Neanderthal segments relative to that in unadmixed segments. Instead, the diversity in unadmixed segments relative to that in Neanderthal segments is higher in western than in eastern non-Africans, perhaps due to gene flow from a source with little or no Neanderthal ancestry into the former (*38*). Although phylogenetic reconstructions indicate that some regions in the genome contain more than 10 different introgressing Neanderthal haplotypes (Fig. 6B, table S9), thus clearly ruling out the scenario of a single contributing Neanderthal individual, the average genetic diversity of admixed Neanderthal sequences is limited (Fig. 6B,C). Coalescent simulations suggest that, genome-wide, as few as 2-4 founding haplotypes are sufficient to produce the observed distribution of haplotype network sizes.

In contrast, Denisovan segments show evidence of a more complex admixture history. Segments in Oceania are distinct from those in East Asia, the Americas and South Asia, as shown by their different distribution along the genome (fig. S18 and table S8), high D_{XY} values (Fig. 6A) and a clear separation in most haplotype networks between these two geographical groups (fig. S24, table S10), corresponding to a deep divergence between the Denisovan source populations. East Asian populations also harbour some Denisovan

segments that are very similar to the Altai Denisovan genome but which are absent from Oceania (fig. S19). This is consistent with the Denisovan ancestry in Oceania having originated from a separate gene flow event not experienced in other parts of the world (*39*). We do not, however, find clear evidence of more than one source in Oceanians (*40*). The more complicated structure of the Denisovan segments in East Asia (and likely also in the Americas and South Asia) is difficult to explain by one or even two admixture events, and may possibly reflect encounters with multiple Denisovan populations by the ancestors of modern humans in Asia. Some Denisovan haplotypes found in Cambodians are somewhat distinct from those in the rest of East Asia with tentative connections to those in Oceania. Overall, these results paint a picture of an admixture history from Denisovan-related populations into modern humans that is substantially more complex than the history of admixture from Neanderthals.

In MSMC2 analyses, we find that non-Africans display clear modes of non-zero cross-coalescence rates with the Vindija Neanderthal in recent time periods (<100 kya), providing an additional line of evidence for the known admixture episode without requiring assumptions about African populations lacking admixture (Fig. 6D, fig. S8). The Denisovan gene flow into Oceanians is also visible in these analyses but is less pronounced and substantially shifted backwards in time (fig. S8), consistent with the introgressing population being highly diverged from the sequenced individual from the Altai mountains. The West African Yoruba also display a Neanderthal admixture signal, similar in shape but much less pronounced than the signal in non-Africans (Fig. 6D, fig S9). Other African populations do not clearly display the same behaviour. These results provide evidence for low amounts of Neanderthal ancestry in West Africa, consistent with previous results based on other approaches (*16, 20*), and we estimate this at 0.18%±0.06% in Yoruba using an f_4-ratio (assuming Mbuti has none). The most likely source for this is West Eurasian admixture (*41*), and assuming a simple linear relationship to Neanderthal ancestry, our estimate implies 8.6%±3% Eurasian ancestry in Yoruba.

While there is an excess of haplotypes deriving from archaic admixture in non-Africans, many single variants present in archaic populations are also present in Africans due to their having segregated in the population ancestral to archaic and modern humans, and some of these variants were subsequently lost in non-Africans due to increased genetic drift. Counting how many of the variants carried in heterozygote state in archaic individuals are segregating

in balanced sets of African and non-African genomes, we find that more Vindija Neanderthal variants survive in non-Africans than in Africans (31.0% vs 26.4%). However, more Denisovan variants survive in Africans (18.9% vs 20.3%). These numbers might change if larger numbers of Oceanian populations were surveyed, but they highlight how the high levels of genetic diversity in African populations mean that, despite having received much less or no Neanderthal and Denisovan admixture, they still retain a substantial, and only partly overlapping (Fig. 3E), subset of the variants which were segregating in late archaic populations.

Discussion

While the number of human genomes sequenced as part of medically motivated genetic studies is rapidly growing into the hundreds of thousands, the number resulting from anthropologically informed sampling to characterize human diversity still remains in the hundreds to low thousands. With the set of 929 genomes from 54 diverse human populations presented here, we greatly extend the number of high-coverage genomes freely available to the research community as part of human global diversity datasets, and substantially expand the catalogue of genetic variation to many underrepresented ancestries. Our analyses of these genomes highlight several aspects of human genetic diversity and history, including the extent and source of geographically restricted variants in different parts of the world, the time depth of separation and extensive gene flow between populations in Africa, a potentially dramatic population expansion following entry into the Americas and a simple pattern of Neanderthal admixture contrasting with a more complex pattern of Denisovan admixture.

One aim of the 1000 Genomes Project (*2*) was to capture most common human genetic variation, which it achieved in the populations included in the study. However, the more diverse HGDP dataset reveals that there are several human ancestries for which this aim was not achieved, and which harbour substantial amounts of genetic variation, some of it common, that so far has been documented poorly or not at all. This is particularly true of Africa and the ancestries represented by the southern African San, and central African Mbuti and Biaka groups. Outside of Africa, Oceanian populations represent one of the major lineages of non-African ancestries and have substantial amounts of private variation, some of it deriving from Denisovan admixture. Any biomedical implications of variants common in these populations but rare or absent elsewhere are unknown, and will remain unknown until

genetic association studies are extended to include these and other currently underrepresented ancestries.

Our analyses demonstrate the value of generating multiple high-coverage whole-genome sequences to characterise variation in a population, compared to genotyping using arrays, sequencing to low-coverage or sequencing just small numbers of genomes. In particular, such an approach enables unbiased variant discovery, including of large numbers of low-frequency variants, and higher resolution assessments of allele frequencies. The experimental phasing of haplotypes using linked-read technology aids analyses of deep human population history and structural variation, and is now becoming a feasible alternative to statistical phasing, especially useful in diverse populations. However, short read sequencing still imposes limitations on the ability to identify more complex structural variation. We expect the application of long-read or linked-read sequencing technologies to large sets of diverse human genomes, combined with de-novo assembly or variation graph (*42*) approaches that are less reliant on the human reference assembly, to unveil these additional layers of human genetic diversity.

While the HGDP genome dataset substantially expands our genomic record of human diversity, it too contains considerable gaps in its geographical, linguistic and cultural coverage. We therefore argue for the importance of continued sequencing of diverse human genomes. Given the scale of ongoing medical and national genome projects, producing high-coverage genome sequences for at least ten individuals from each of the approximately 7000 (*43*) human linguistic groups would now arguably not be an overly ambitious goal for the human genomics community. Such an achievement would represent a scientifically and culturally important step towards diversity and inclusion in human genomics research.

Materials and methods summary

We sequenced DNA, extracted from the lymphoblastoid cell lines of the HGDP-CEPH panel (*5*), on Illumina HiSeq X machines, and incorporated data from a subset of samples that had been previously sequenced (*3, 11*). Reads were mapped to the GRCh38 human reference assembly. We applied per-sample caps on the mapping quality of reads to counteract the effects of low-level index hopping in multiplexed sequencing runs. We analyzed patterns of sequencing coverage along the chromosomes to identify any large-scale copy number

deviations that arose during cell line culturing, and excluded nine samples with such deviations, leaving 929 samples.

Genotypes were called using GATK HaplotypeCaller (*44*) v3.5.0 and filtered by setting to missing any genotype with a GQ (Genotype Quality) or RGQ (Reference Genotype Quality) value equal to or lower than 20, or a DP (depth) value equal to or greater than 1.65 times the genome-wide average coverage for the given sample. We also flagged sites displaying excess heterozygosity and excluded these from analyses. We constructed a genome accessibility mask largely based on the 1000 Genomes Project (*2*) "strict mask" and restricted analyses to these regions. Batch effects between library types observed when analysing unfiltered genotypes were not observed after applying the genotype filters and restricting to the mask, but we cannot rule out that more subtle effects persist. We reassessed the population labels used in the previous literature on the HGDP-CEPH panel to arrive at 54 labels which we use in analyses, along with the seven regional/continental labels previously used (*6*). We constructed 10x Genomics linked-read libraries (*14*) and sequenced these on Illumina HiSeq X machines for 26 of the individuals, from 13 globally representative populations, to physically resolve their haplotype phase and aid analyses of structural variation.

We used ADMIXTOOLS (*10*) v5.0 to compute f_4 and *D*-statistics and EIGENSOFT (*45*) v6.0.1 to compute F_{ST} statistics. To identify variants that are private to geographical regions while avoiding the effects of recent admixture between regions, we used the model-based clustering program ADMIXTURE (*46*) to determine which individuals to use as the ingroup and outgroup for each region.

We used MSMC2 (*22, 29*) to study the time depth and nature of population separations, using the 13 populations for which we had physically phased genomes for two individuals each. By site-frequency spectrum modelling using momi2 (*33*) we also estimated all possible pairwise population divergence times under simple, clean split scenarios. We used SMC++ (*24*) to infer effective population size histories, using all available genomes for a given population. For all demographic analyses, we scaled results using a mutation rate of 1.25×10-8 per site per generation (*30*) and a generation time of 29 years (*31*).

To identify segments in modern human genomes deriving from archaic admixture, we used a Hidden Markov Model trained on simulated haplotypes. The model decodes haplotypes into

archaic or unadmixed on the basis of the allele sharing patterns between sub-Saharan Africans, one or more archaic genomes, and the given genome under examination. We analysed the properties of the inferred haplotypes, including their nucleotide diversity, spatial distributions along the genome and phylogenetic relationships and ages as inferred using haplotype networks.

Detailed descriptions of materials and methods are available in the supplementary materials.

References:

1. R. Nielsen *et al.*, Tracing the peopling of the world through genomics. *Nature* **541**, 302-310 (2017).
2. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
3. S. Mallick *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201-206 (2016).
4. L. Pagani *et al.*, Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538**, 238-242 (2016).
5. H. M. Cann *et al.*, A human genome diversity cell line panel. *Science* **296**, 261-262 (2002).
6. N. A. Rosenberg *et al.*, Genetic structure of human populations. *Science* **298**, 2381-2385 (2002).
7. M. Jakobsson *et al.*, Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998-1003 (2008).
8. J. Z. Li *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104 (2008).
9. W. Shi *et al.*, A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Mol Biol Evol* **27**, 385-393 (2010).
10. N. Patterson *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).
11. M. Meyer *et al.*, A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226 (2012).
12. S. Lippold *et al.*, Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. *Investig Genet* **5**, 13 (2014).
13. M. Raghavan *et al.*, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
14. G. X. Zheng *et al.*, Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**, 303-311 (2016).
15. Materials and methods are available as supplementary materials.
16. K. Prufer *et al.*, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49 (2014).
17. J. K. Pickrell *et al.*, Ancient west Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci U S A* **111**, 2632-2637 (2014).

18. P. Skoglund *et al.*, Reconstructing Prehistoric African Population Structure. *Cell* **171**, 59-71 e21 (2017).
19. G. D. Poznik *et al.*, Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat Genet* **48**, 593-599 (2016).
20. K. Prufer *et al.*, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655-658 (2017).
21. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496 (2011).
22. S. Schiffels, R. Durbin, Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925 (2014).
23. S. Song, E. Sliwerska, S. Emery, J. M. Kidd, Modeling Human Population Separation History Using Physically Phased Genomes. *Genetics* **205**, 385-395 (2017).
24. J. Terhorst, J. A. Kamm, Y. S. Song, Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet* **49**, 303-309 (2017).
25. L. Excoffier, S. Schneider, Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. *Proc Natl Acad Sci U S A* **96**, 10597-10602 (1999).
26. B. Llamas *et al.*, Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci Adv* **2**, e1501385 (2016).
27. T. Pinotti *et al.*, Y Chromosome Sequences Reveal a Short Beringian Standstill, Rapid Expansion, and early Population structure of Native American Founders. *Curr Biol* **29**, 149-157 e143 (2019).
28. S. R. Browning, B. L. Browning, Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *Am J Hum Genet* **97**, 404-418 (2015).
29. A. S. Malaspinas *et al.*, A genomic history of Aboriginal Australia. *Nature* **538**, 207-214 (2016).
30. A. Scally, The mutation rate in human evolution and demographic inference. *Curr Opin Genet Dev* **41**, 36-43 (2016).
31. J. N. Fenner, Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005).
32. C. Posth *et al.*, Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nat Commun* **8**, 16046 (2017).
33. J. Kamm, J. Terhorst, R. Durbin, Y. S. Song, Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. *Journal of the American Statistical Association* **0**, 1-16 (2019).
34. M. Jakobsson, M. D. Edge, N. A. Rosenberg, The relationship between F(ST) and the frequency of the most frequent allele. *Genetics* **193**, 515-528 (2013).
35. D. Reich *et al.*, Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060 (2010).
36. P. Qin, M. Stoneking, Denisovan Ancestry in East Eurasian and Native American Populations. *Mol Biol Evol* **32**, 2665-2674 (2015).
37. B. Vernot *et al.*, Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235-239 (2016).
38. I. Lazaridis *et al.*, Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419-424 (2016).

39. 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 e59 (2018).
40. G. S. Jacobs *et al.*, Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell* **177**, 1010-1021 e1032 (2019).
41. I. Lazaridis *et al.*, Paleolithic DNA from the Caucasus reveals core of West Eurasian ancestry. *bioRxiv 423079*, (2018).
42. E. Garrison *et al.*, Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol* **36**, 875-879 (2018).
43. G. F. Simons, C. D. Fennig, Ethnologue: Languages of the World, Twenty-first edition. (2018).
44. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
45. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).
46. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664 (2009).
47. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
48. G. Tischler, S. Leonard, biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol Med.*, (2014).
49. S. T. Sherry *et al.*, dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311 (2001).
50. G. Jun *et al.*, Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-848 (2012).
51. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993 (2011).
52. A. Bergström *et al.*, A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science* **357**, 1160-1163 (2017).
53. R. E. Handsaker *et al.*, Large multiallelic copy number variations in humans. *Nat Genet* **47**, 296-303 (2015).
54. S. Neph *et al.*, BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919-1920 (2012).
55. O. Delaneau, J. Marchini, C. Genomes Project, C. Genomes Project, Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* **5**, 3934 (2014).
56. P. R. Loh, P. F. Palamara, A. L. Price, Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet* **48**, 811-816 (2016).
57. B. L. Browning, S. R. Browning, Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet* **98**, 116-126 (2016).
58. N. A. Rosenberg *et al.*, Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet* **1**, e70 (2005).
59. P. R. Staab, S. Zhu, D. Metzler, G. Lunter, scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics* **31**, 1680-1682 (2015).

60. G. D. Poznik *et al.*, Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**, 562-565 (2013).
61. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
62. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
63. A. J. Drummond, A. Rambaut, B. Shapiro, O. G. Pybus, Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22**, 1185-1192 (2005).
64. A. J. Drummond, A. Rambaut, BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**, 214 (2007).
65. Q. Fu *et al.*, Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449 (2014).
66. H. Zhao *et al.*, CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006-1007 (2014).
67. P. Skoglund *et al.*, Genetic evidence for two founding populations of the Americas. *Nature* **525**, 104-108 (2015).
68. W. Haak *et al.*, Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207-211 (2015).
69. D. Reich *et al.*, Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* **89**, 516-528 (2011).
70. J. Kelleher, A. M. Etheridge, G. McVean, Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput Biol* **12**, e1004842 (2016).
71. B. Vernot, J. M. Akey, Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017-1021 (2014).
72. S. Sankararaman *et al.*, The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354-357 (2014).
73. H. Li, Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843-2851 (2014).
74. M. Nei, W. H. Li, Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A* **76**, 5269-5273 (1979).
75. I. Lazaridis *et al.*, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409-413 (2014).
76. E. Paradis, pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419-420 (2010).
77. J. Saillard, P. Forster, N. Lynnerup, H. J. Bandelt, S. Norby, mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet* **67**, 718-726 (2000).

Acknowledgments: We thank the sample donors who made this research possible, as well as the CEPH Biobank, Paris, France (BIORESOURCES) at Fondation Jean Dausset-CEPH, for maintaining the cell line resource and distributing DNA. We thank the Wellcome Sanger Institute sequencing facility for generating data, and Susan Fairley and colleagues at the International Genome Sample Resource for incorporating and hosting data. We thank J.

Terhorst, S. Schiffels, R. Handsaker, D. Gurdasani and members of the Tyler-Smith and Durbin groups for useful advice and discussions. **Funding:** A.B., S.A.M., M.A.A, Q.A., P.D., Y.C., S.F., P.H., J.K, M.S.S., Y.X., R.D. and C.T.-S. were supported by Wellcome grants 098051 and 206194, and S.A.M. and R.D. also by Wellcome grant 207492. A.B. and P.S. were supported by the Francis Crick Institute (FC001595) which receives its core funding from Cancer Research UK, the UK Medical Research Council and the Wellcome Trust. P.S. was also supported by the European Research Council (grant no. 852558) and the Wellcome Trust (217223/Z/19/Z). R.H. was supported by a Gates Cambridge scholarship. P.H. was supported by Estonian Research Council Grant PUT1036. D.R. is an Investigator of the Howard Hughes Medical Institute. **Data and materials availability:** Raw read alignments are available from the European Nucleotide Archive under study accession PRJEB6463. Processed per-sample read alignment files are made available by the International Genome Sample Resource at the European Bioinformatics Institute (EMBL-EBI) (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGDP/). The 10x Genomics sequencing data generated for 26 samples are available at the European Nucleotide Archive under study accession PRJEB14173. Genotype calls and other downstream analysis files are available from the Wellcome Sanger Institute (ftp://ngs.sanger.ac.uk/production/hgdp). DNA extracts from the samples in the HGDP-CEPH collection can be obtained from the CEPH Biobank at Fondation Jean Dausset-CEPH in Paris, France (http://www.cephb.fr/en/hgdp_panel.php).

Figure legends:

Figure 1: Genome sequencing and variant discovery in 54 diverse human populations. (A) Geographical origins of the 54 populations from the HGDP-CEPH panel, with the number of sequenced individuals from each in parentheses. (B) Maximum allele frequencies of variants discovered in the HGDP dataset but not in the 1000 Genomes phase 3 dataset, and vice versa. The vertical axis displays the number of variants that have a maximum allele frequency in any single population equal to or higher than the corresponding value on the horizontal axis. To account for higher sampling noise due to smaller population sample sizes in the HGDP dataset, results obtained on versions of the 1000 Genomes dataset down-sampled to match the HGDP sizes are also shown. To conservatively avoid counting variants that are actually present in both datasets but not called in one of them for technical reasons, any variant with a global frequency of >30% in a dataset is excluded. (C) Comparison of Z-

scores from all possible f_4-statistics involving the 54 populations using whole genome sequences and commonly used, ascertained genotyping array sites (*8*). Points are coloured according to the number of African populations included in the statistic.

Figure 2: Insights into population relationships from low-frequency variants. (A) A heatmap of pairwise counts of doubleton alleles (alleles observed exactly twice across the dataset) between all 929 individuals, grouped by population. (B-D) *D*-statistics of the form *D(Chimp,X;A,B)*, stratified by the derived allele frequency in X. Red points correspond to $|Z| > 3$.

Figure 3: Counts and properties of geographically private variants. (A-C) Counts of region-specific variants. The vertical axis displays the number of variants private to a given geographical region that have an allele frequency in that region equal to or higher than the corresponding value on the horizontal axis. Shaded areas denote 95% Poisson confidence intervals. (A) SNPs. (B) Indels. (C) CNVs. (D) The fraction of SNPs private to a given region and at a frequency equal to or higher than the corresponding value on the horizontal axis for which the private allele is the derived as opposed to ancestral state. (E) The fraction of SNPs private to a given region and at a frequency equal to or higher than the corresponding value on the horizontal axis for which the private allele is observed in any of three high-coverage archaic genomes. (F) As E, but now counting variants that are present in the given region and absent in Africa, regardless of their frequency elsewhere.

Figure 4: Effective population size histories of 54 diverse populations. (A) Effective population sizes for all populations inferred using SMC++, computed using composite likelihoods across six different distinguished individuals per population. Our ability to infer recent size histories in some South Asian and Middle Eastern populations might be confounded by the effects of recent endogamy. (B) Results for the Native American Karitiana population with varying SMC++ parameter settings. Decreasing the regularization or excluding the last few thousand years from the time period of inference leads to curves displaying massive growth approximately in the period 10 to 20 kya.

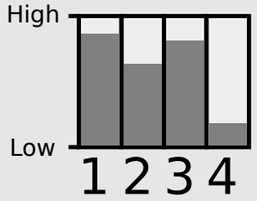
Figure 5: The time depth and mode of population separations. (A) MSMC2 cross-population results for pairs of African populations, including Han Chinese as a representative of non-Africans, as well as between archaic populations and Mbuti as a representative of modern humans. Curves between modern human groups were computed using four physically phased haplotypes per population, while curves between modern and archaic groups were computed using two haplotypes per population and unphased archaic genomes. The results of simulated histories with instantaneous separations at different time points are displayed in the background in alternating yellow and grey curves. (B) MSMC2 cross-population results, as in A, for pairs of non-African populations. (C) Split times estimated under simple, sudden pairwise split models using momi2 for all possible pairs among the 54 populations against F_{ST}, a measure of allele frequency differentiation. The plot does not include Native American populations, as we could not obtain reliable momi2 fits for these.

Figure 6: Archaic haplotypes in modern human populations. (A) Nucleotide divergence D_{XY} within segments deriving from archaic admixture and within other segments in non-African populations. (B) The mean number of archaic founding haplotypes estimated by constructing maximum likelihood trees for each archaic segment identified in present-day non-Africans, and then determining the number of ancestral branches in the tree at the approximate time of admixture (2000 generations ago). (C) The distribution of estimated ages of archaic haplotype networks in the present-day human population. The distribution is compared to results obtained in simulations performed with different numbers of archaic founding haplotypes. (D) MSMC2 cross-population results for African (two individual curves per population) and selected non-African (one individual curve per population) against the Vindija Neanderthal, zooming in on the signal of Neanderthal genome flow in modern human genomes (note the highly reduced range of the vertical axis).

Supplementary materials:

Materials and Methods
Figs. S1 to S26
Tables S1 to S11
References (47-77)

Amounts of different classes of genetic variation

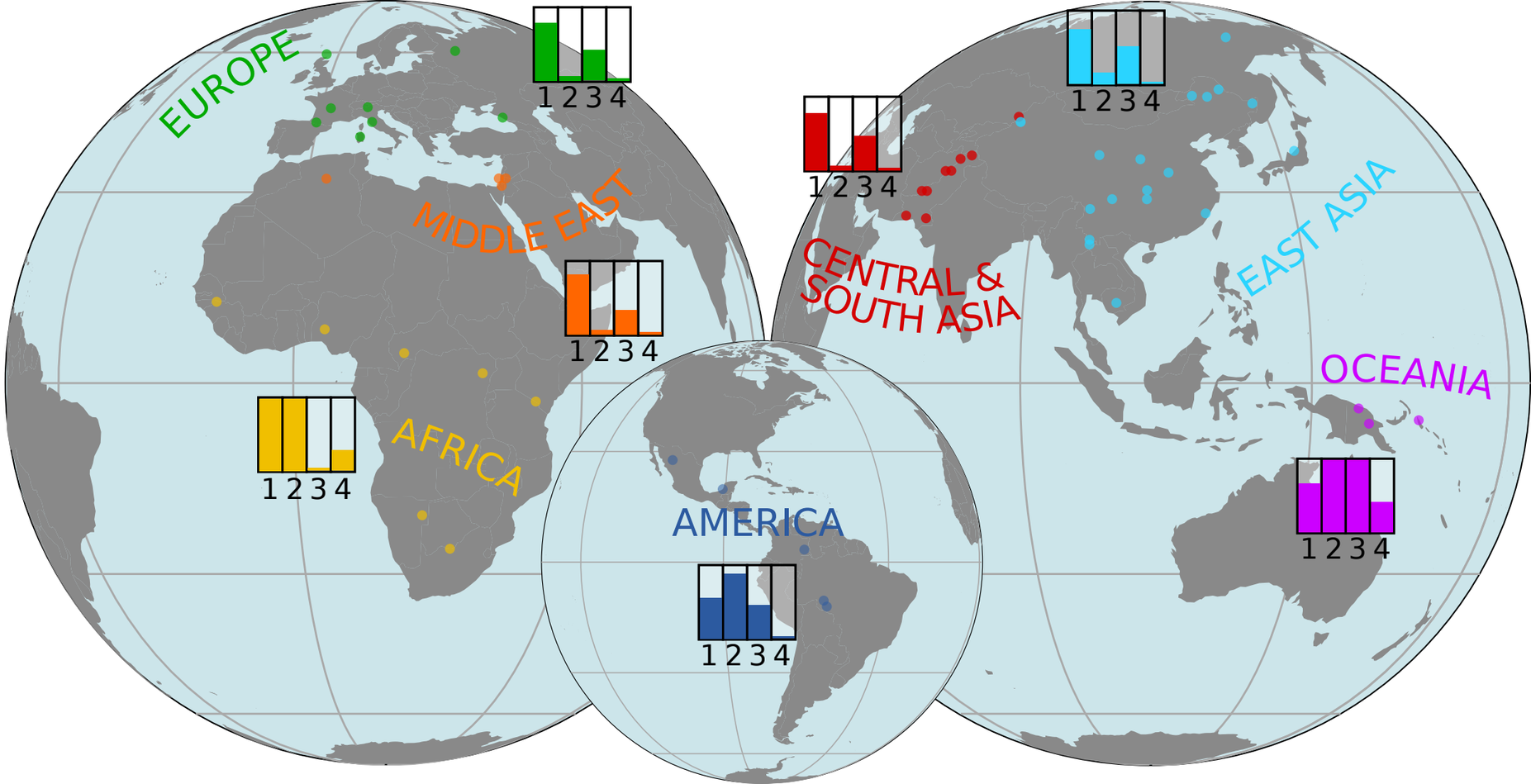


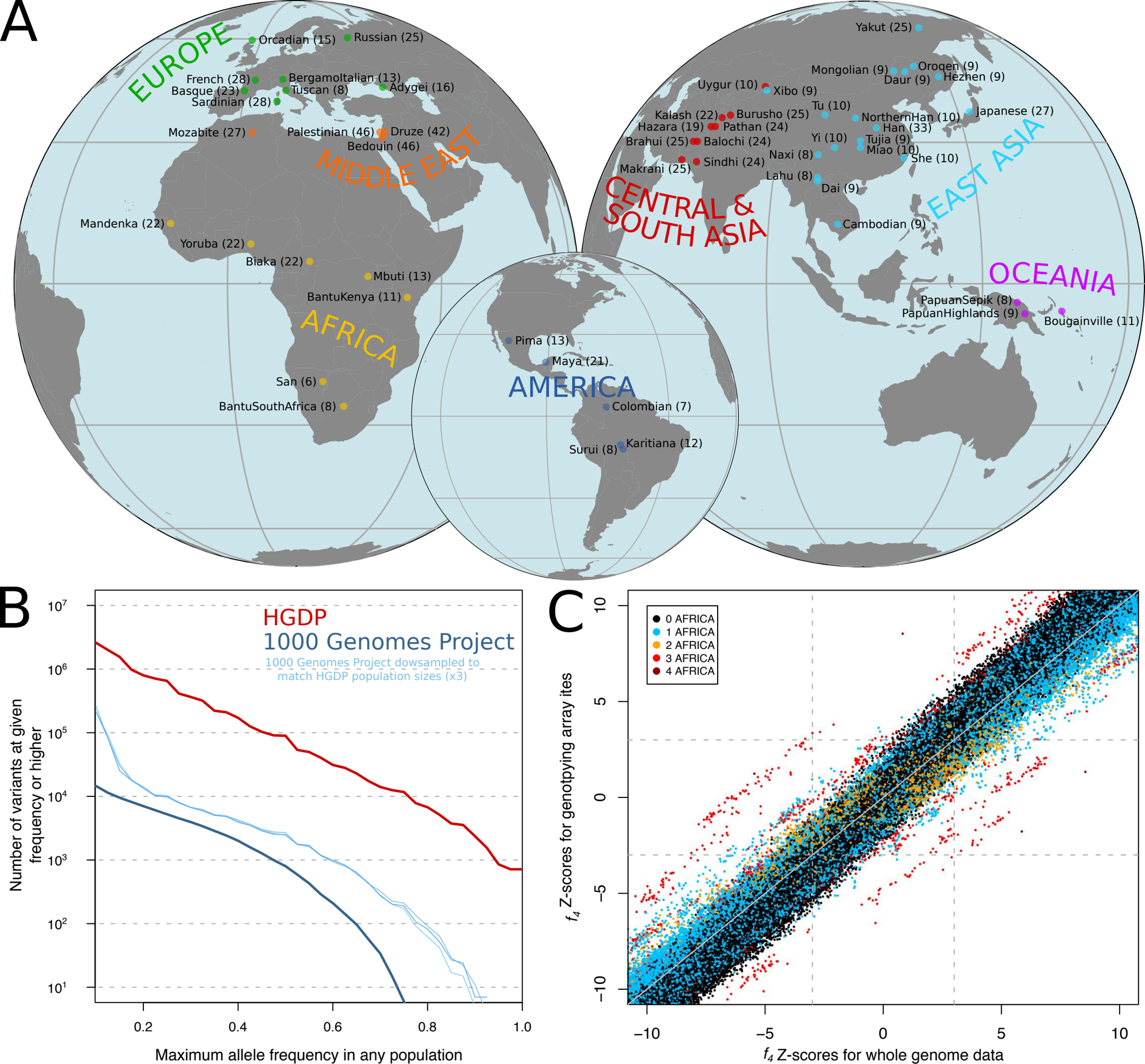
1. Total variation
Relative to highest value (Africa)

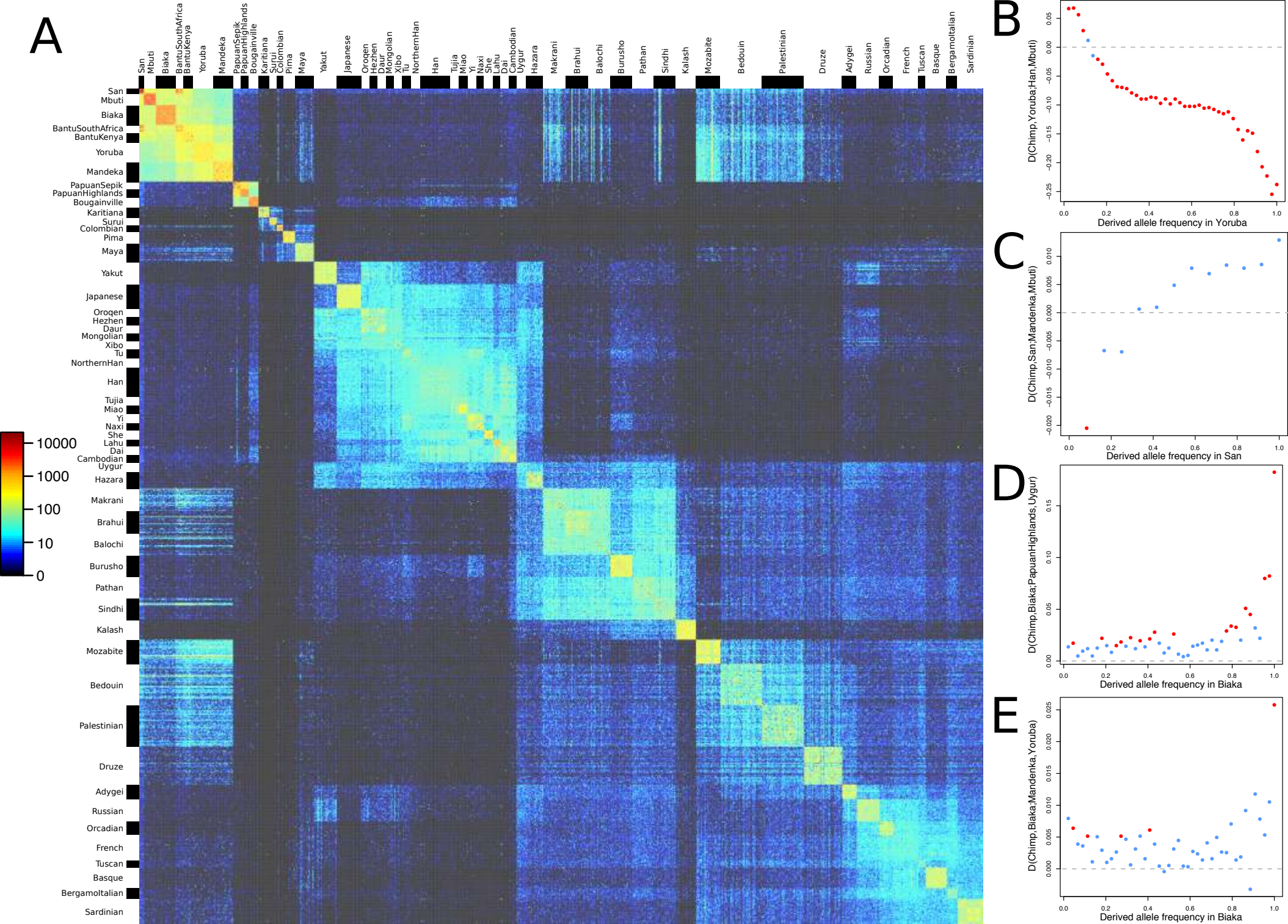
2. Private, common variation
Relative to highest value (Africa/Oceania)

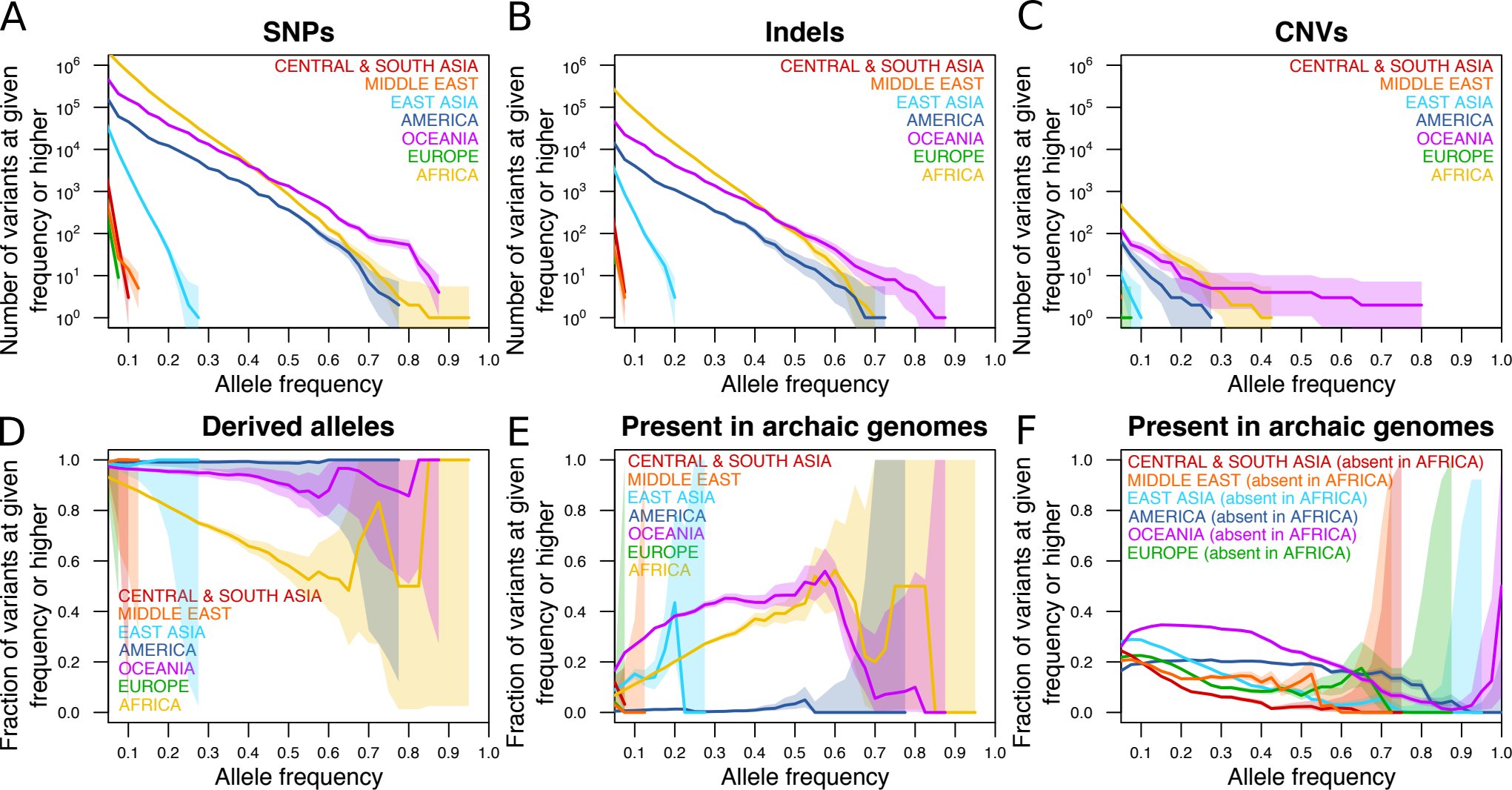
3. Variation deriving from archaic admixture
Relative to highest value (Oceania)

4. Private, common variation shared with archaic genomes
Fraction of private variation

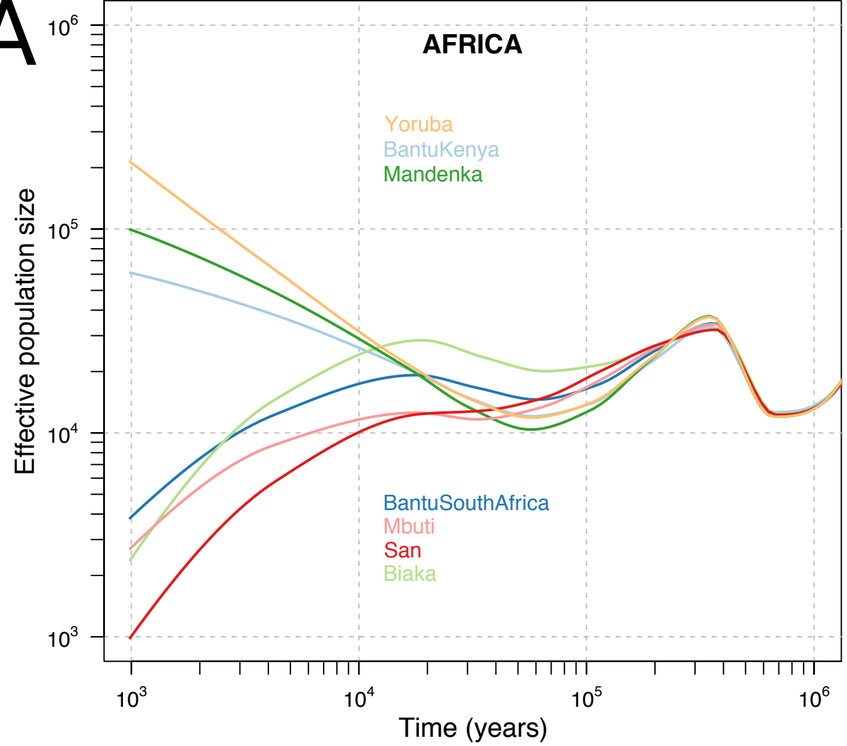


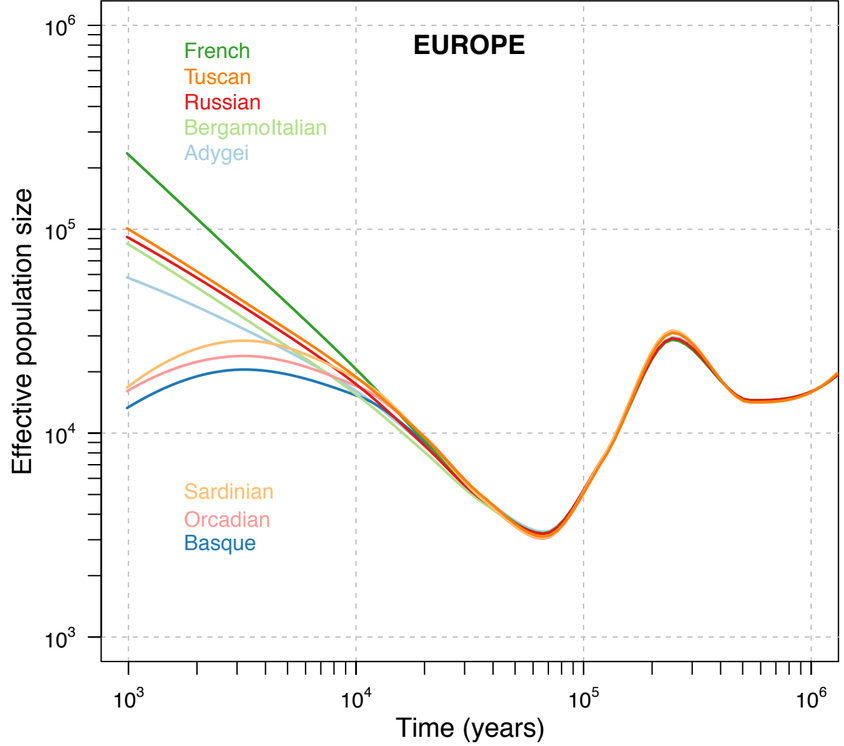

A
EUROPE
Orcadian (15)
Russian (25)
French (28)
BergamoItalian (13)
Basque (23)
Tuscan (8)
Adygei (16)
Sardinian (28)
Mozabite (27)
Palestinian (46)
Druze (42)
Bedouin (46)
MIDDLE EAST
Mandenka (22)
Yoruba (22)
Biaka (22)
Mbuti (13)
BantuKenya (11)
AFRICA
San (6)
BantuSouthAfrica (8)
Pima (13)
Maya (21)
AMERICA
Colombian (7)
Surui (8)
Karitiana (12)
Yakut (25)
Mongolian (9)
Oroqen (9)
Daur (9)
Hezhen (9)
Uygur (10)
Xibo (9)
Tu (10)
Japanese (27)
Kalash (22)
Burusho (25)
NorthernHan (10)
Hazara (19)
Pathan (24)
Han (33)
Brahui (25)
Balochi (24)
Yi (10)
Tujia (9)
Naxi (8)
Miao (10)
She (10)
Sindhi (24)
Makrani (25)
Lahu (8)
CENTRAL & SOUTH ASIA
EAST ASIA
Dai (9)
Cambodian (9)
OCEANIA
PapuanSepik (8)
PapuanHighlands (9)
Bougainville (11)
B
HGDP
1000 Genomes Project
1000 Genomes Project dowsampled to match HGDP population sizes (x3)
Number of variants at given frequency or higher
Maximum allele frequency in any population
C
0 AFRICA
1 AFRICA
2 AFRICA
3 AFRICA
4 AFRICA
f_4 Z-scores for genotpying array ites
f_4 Z-scores for whole genome data

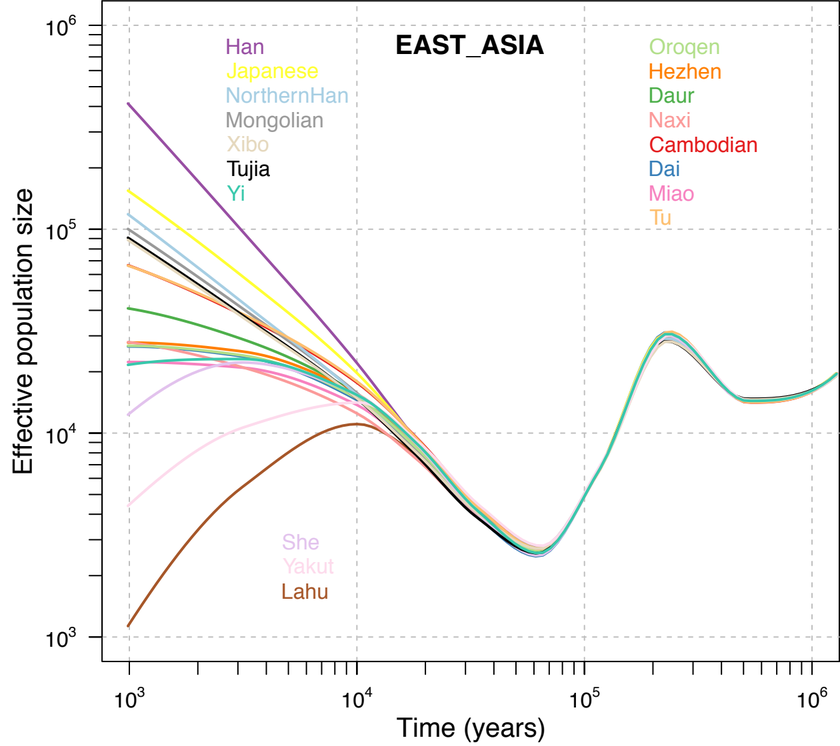

A
San
Mbuti
Biaka
BantuSouthAfrica
BantuKenya
Yoruba
Mandeka
PapuanSepik
PapuanHighlands
Bougainville
Karitiana
Surui
Colombian
Pima
Maya
Yakut
Japanese
Oroqen
Hezhen
Daur
Mongolian
Xibo
Tu
NorthernHan
Han
Tujia
Miao
Yi
Naxi
She
Lahu
Dai
Cambodian
Uygur
Hazara
Makrani
Brahui
Balochi
Burusho
Pathan
Sindhi
Kalash
Mozabite
Bedouin
Palestinian
Druze
Adygei
Russian
Orcadian
French
Tuscan
Basque
Bergamoltalian
Sardinian
10000
1000
100
10
0
B
D(Chimp,Yoruba;Han,Mbuti)
Derived allele frequency in Yoruba
C
D(Chimp,San;Mandenka,Mbuti)
Derived allele frequency in San
D
D(Chimp,Biaka;PapuanHighlands,Uygur)
Derived allele frequency in Biaka
E
D(Chimp,Biaka;Mandenka,Yoruba)
Derived allele frequency in Biaka

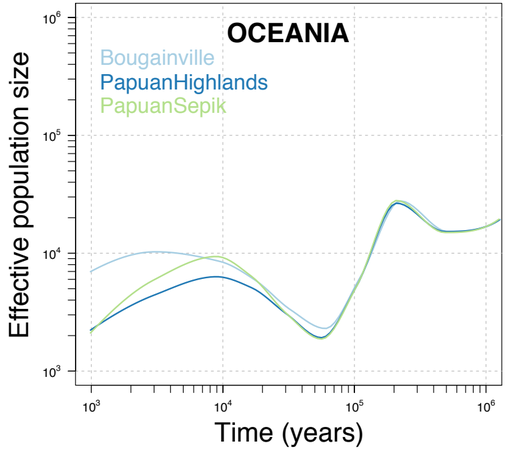

A
SNPs
B
Indels
C
CNVs
D
Derived alleles
E
Present in archaic genomes
F
Present in archaic genomes
Number of variants at given frequency or higher
Fraction of variants at given frequency or higher
Allele frequency
CENTRAL & SOUTH ASIA
MIDDLE EAST
EAST ASIA
AMERICA
OCEANIA
EUROPE
AFRICA
CENTRAL & SOUTH ASIA (absent in AFRICA)
MIDDLE EAST (absent in AFRICA)
EAST ASIA (absent in AFRICA)
AMERICA (absent in AFRICA)
OCEANIA (absent in AFRICA)
EUROPE (absent in AFRICA)

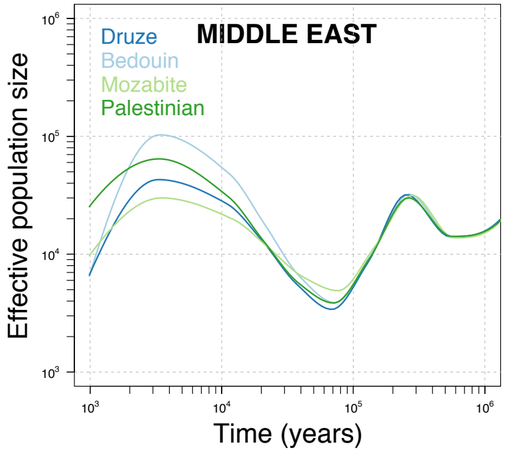
A

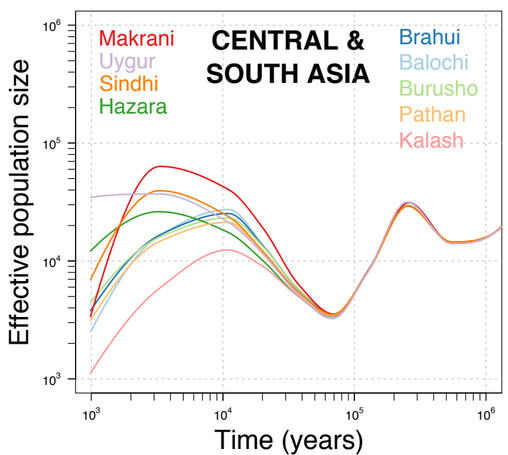

AFRICA
Yoruba
BantuKenya
Mandenka
BantuSouthAfrica
Mbuti
San
Biaka
Effective population size
Time (years)

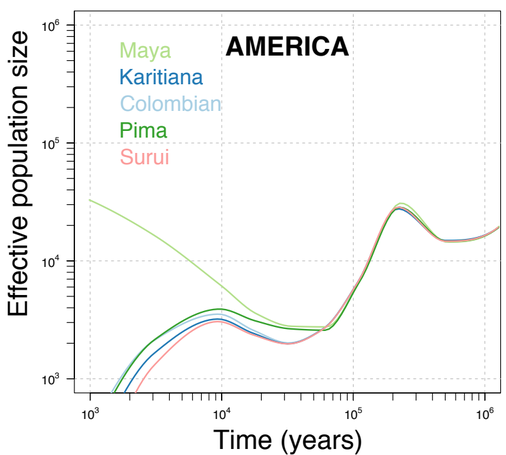

EUROPE
French
Tuscan
Russian
BergamoItalian
Adygei
Sardinian
Orcadian
Basque
Effective population size
Time (years)


EAST_ASIA
Han
Japanese
NorthernHan
Mongolian
Xibo
Tujia
Yi
Oroqen
Hezhen
Daur
Naxi
Cambodian
Dai
Miao
Tu
She
Yakut
Lahu
Effective population size
Time (years)

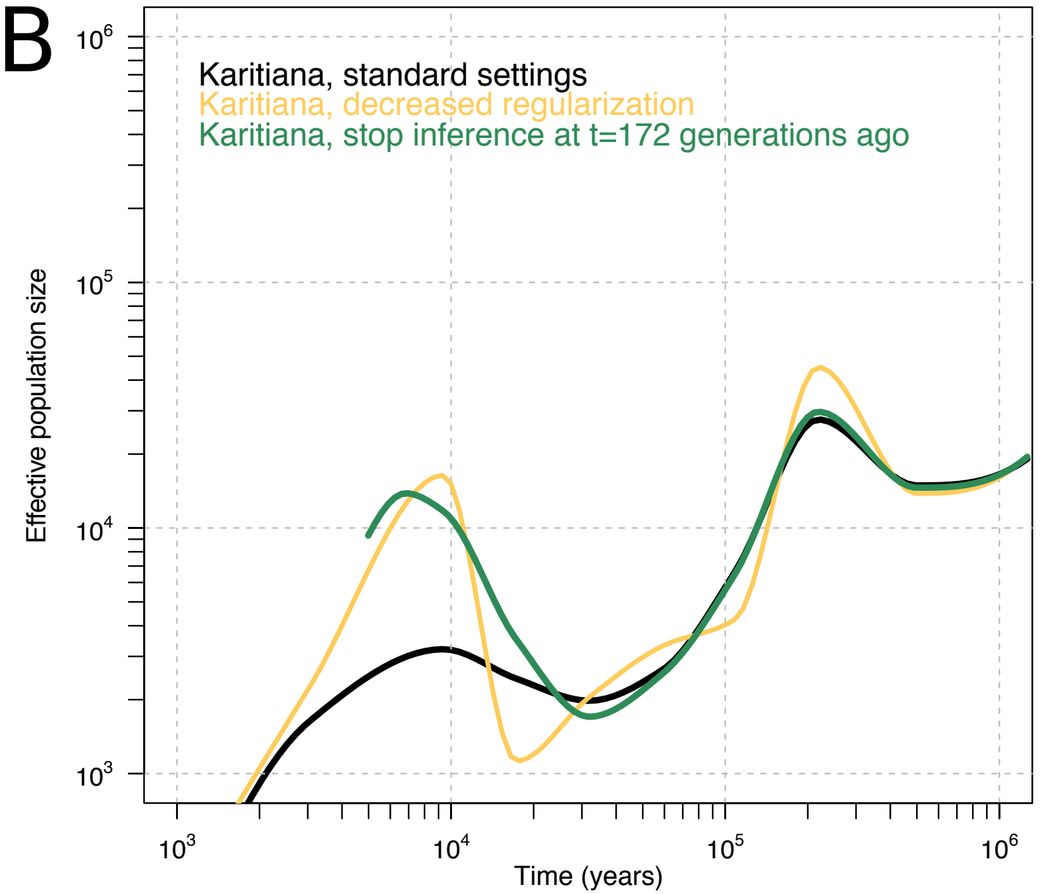

OCEANIA
Bougainville
PapuanHighlands
PapuanSepik
Effective population size
Time (years)

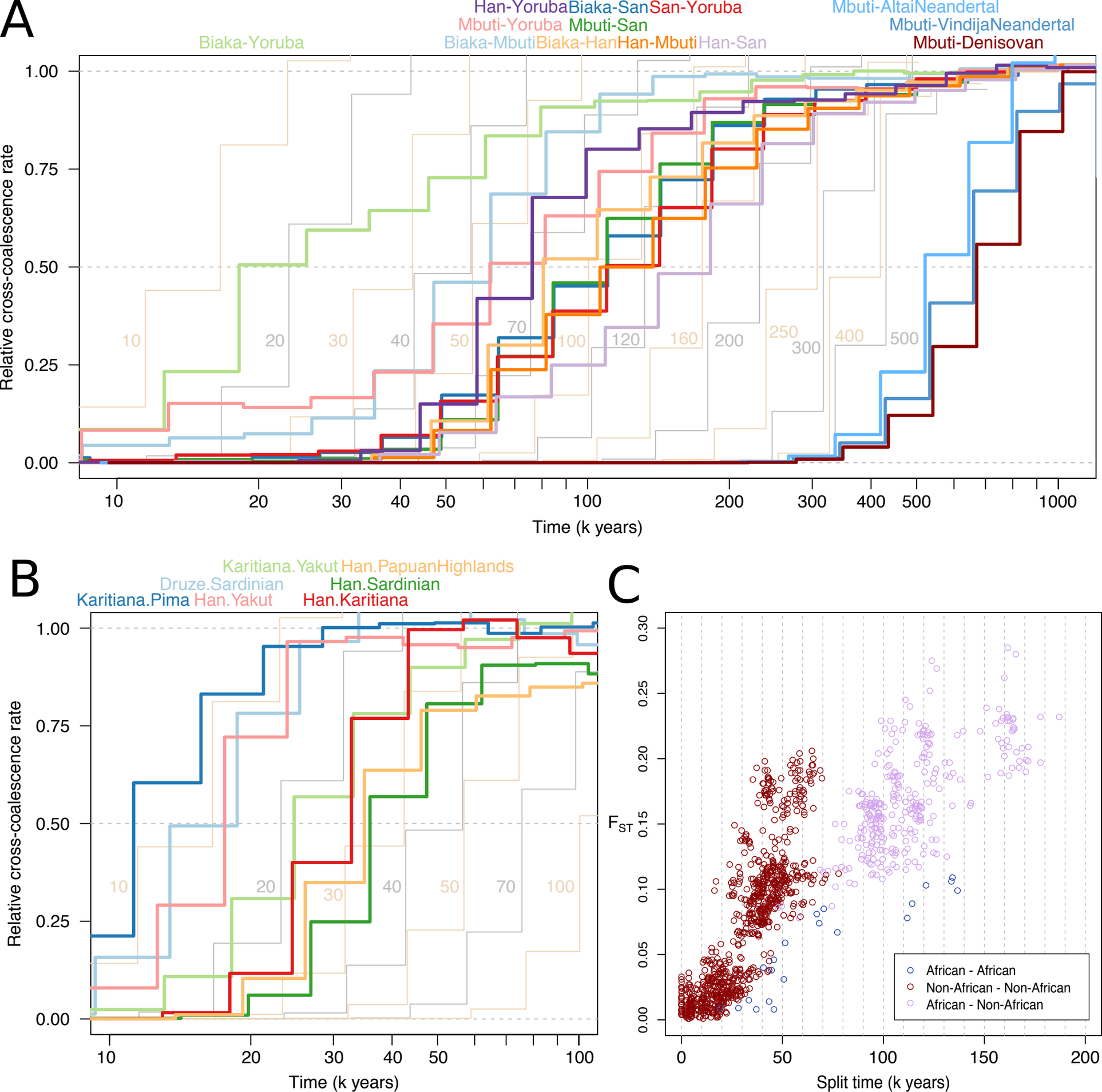

MIDDLE EAST
Druze
Bedouin
Mozabite
Palestinian
Effective population size
Time (years)


CENTRAL &
SOUTH ASIA
Makrani
Uygur
Sindhi
Hazara
Brahui
Balochi
Burusho
Pathan
Kalash
Effective population size
Time (years)

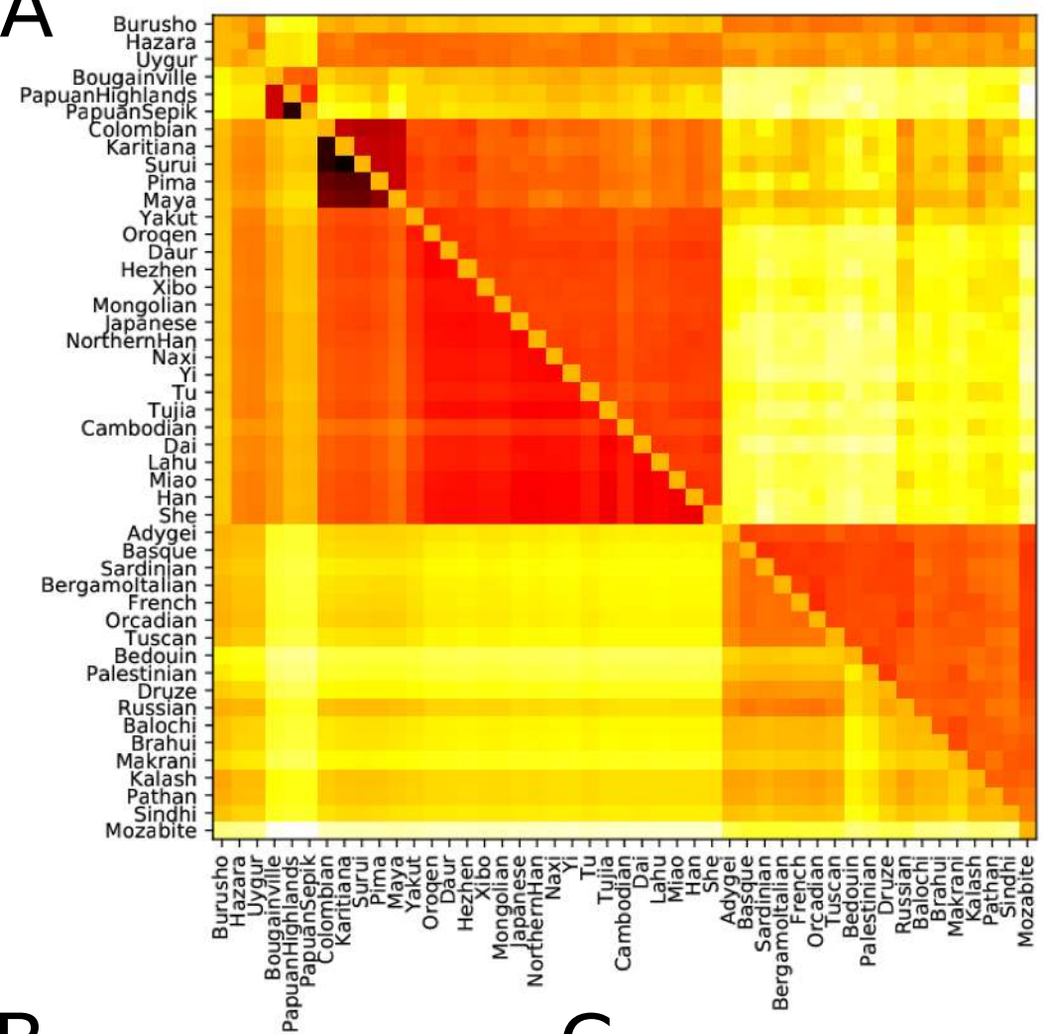

AMERICA
Maya
Karitiana
Colombian
Pima
Surui
Effective population size
Time (years)

B


Karitiana, standard settings
Karitiana, decreased regularization
Karitiana, stop inference at t=172 generations ago
Effective population size
Time (years)

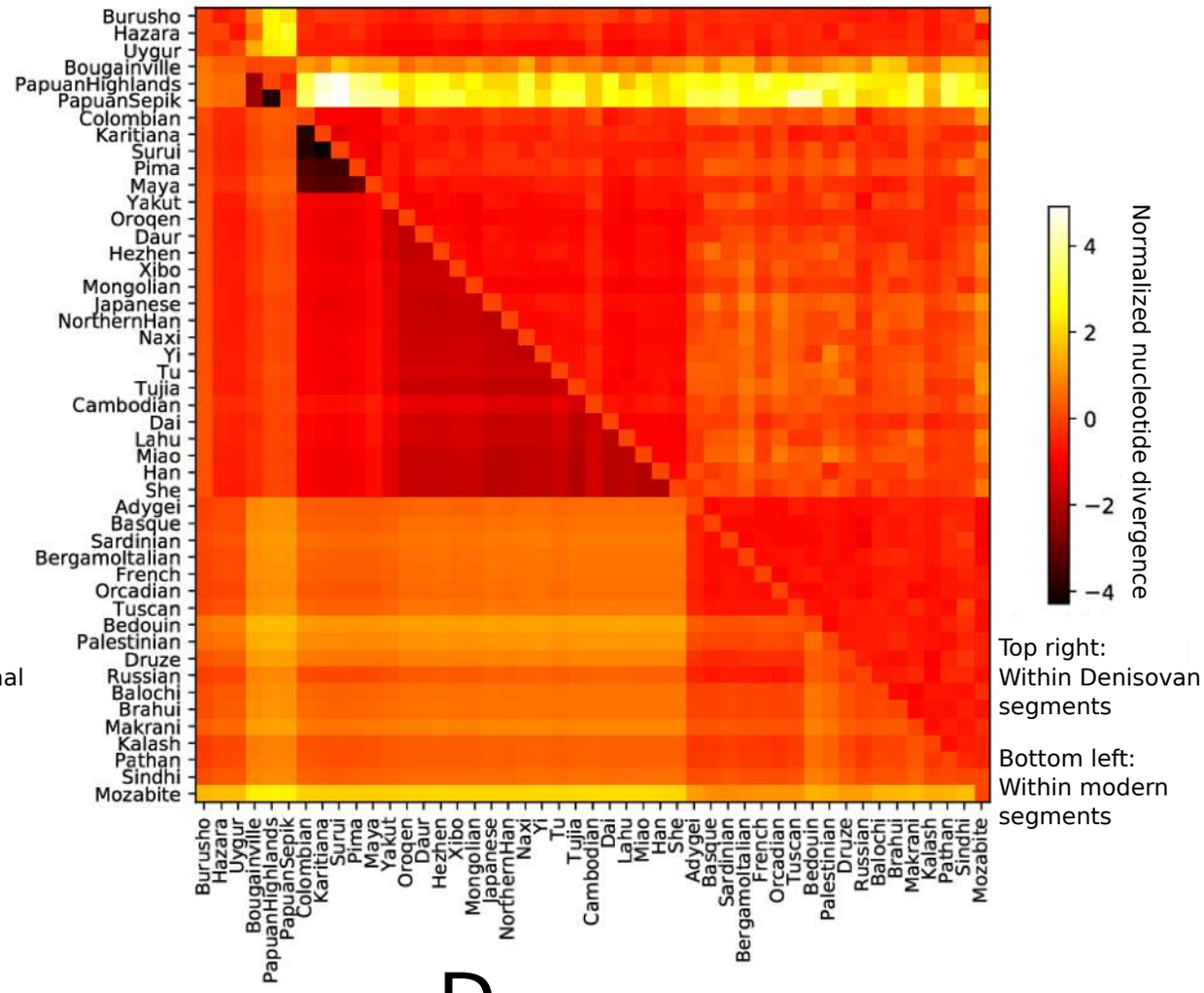

A
Biaka-Yoruba
Han-YorubaBiaka-SanSan-Yoruba
Mbuti-Yoruba Mbuti-San
Biaka-MbutiBiaka-HanHan-MbutiHan-San
Mbuti-AltaiNeandertal
Mbuti-VindijaNeandertal
Mbuti-Denisovan
Relative cross-coalescence rate
Time (k years)
B
Karitiana.Yakut Han.PapuanHighlands
Druze.Sardinian Han.Sardinian
Karitiana.Pima Han.Yakut Han.Karitiana
Relative cross-coalescence rate
Time (k years)
C
F_{ST}
African - African
Non-African - Non-African
African - Non-African
Split time (k years)

A



Top right:
Within Neanderthal segments

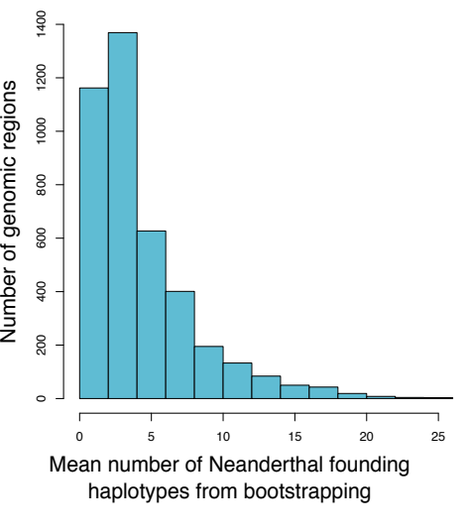
Bottom left:
Within modern segments



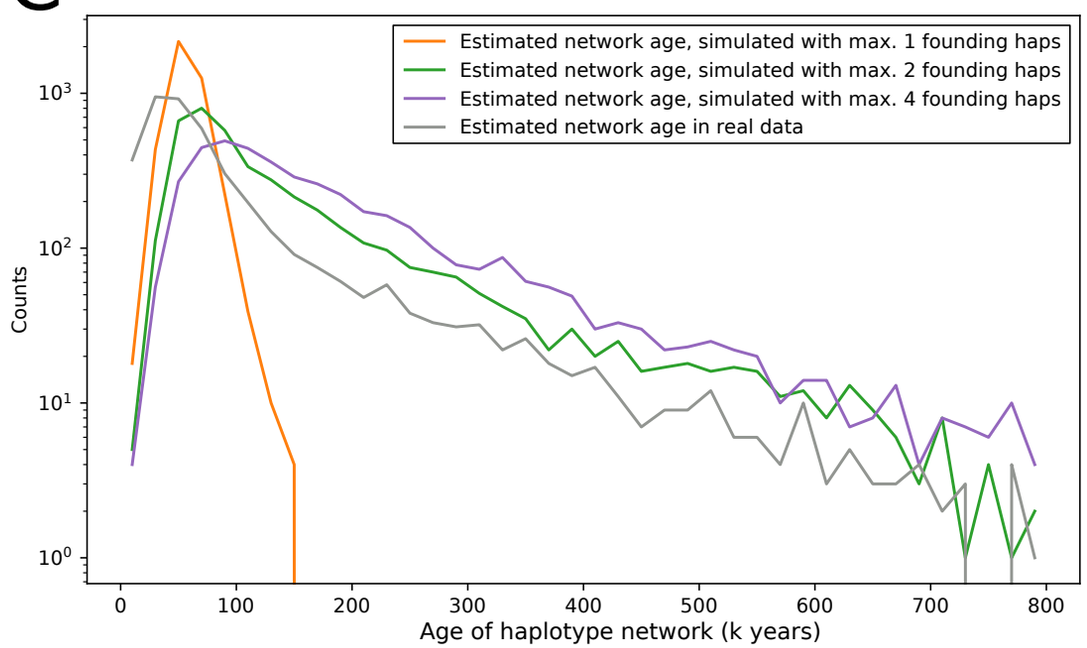
Top right:
Within Denisovan segments

Bottom left:
Within modern segments

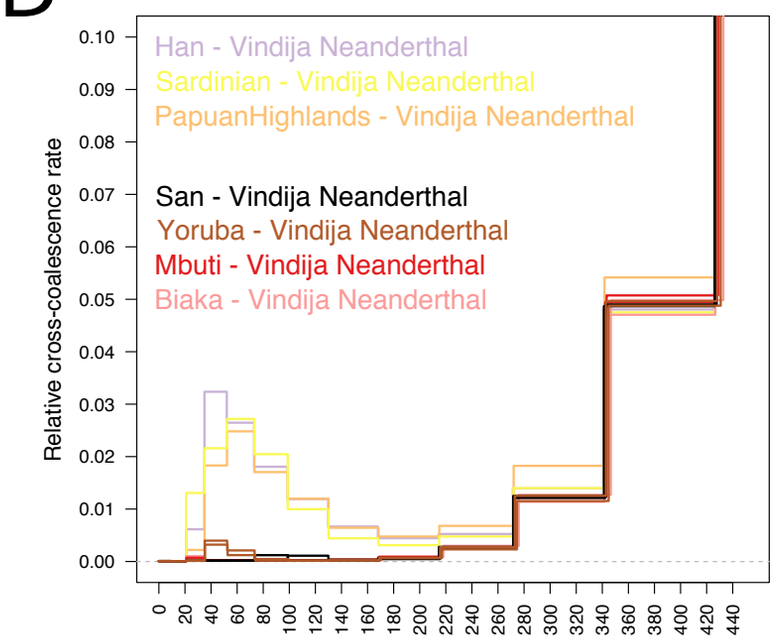
B



C



D



Supplementary Materials for

Insights into human genetic variation and population history from 929 diverse genomes

Anders Bergström[*], Shane A. McCarthy[‡], Ruoyun Hui[‡], Mohamed A. Almarri[‡], Qasim Ayub, Petr Danecek, Yuan Chen, Sabine Felkel, Pille Hallast, Jack Kamm, Hélène Blanché, Jean-François Deleuze, Howard Cann[†], Swapan Mallick, David Reich, Manjinder S. Sandhu, Pontus Skoglund, Aylwyn Scally, Yali Xue[§], Richard Durbin[§], Chris Tyler-Smith[§,*]

*Corresponding author. Email: ab34@sanger.ac.uk (A.B.); cts@sanger.ac.uk (C.T.-S.)
‡These authors contributed equally to this work
§These authors contributed equally to this work
†Deceased

This PDF file includes:

Materials and Methods

Sequencing and read processing

DNA extracted from lymphoblastoid cell lines was shipped from the CEPH-Biobank at Fondation Jean Dausset-CEPH laboratory in Paris, and sequenced at the Wellcome Sanger Institute in Hinxton, United Kingdom. PCR libraries were constructed for an initial batch of samples (PCR-free library construction was not available for large-scale production at the institute at this time), while PCR-free libraries were constructed for the rest. Libraries were sequenced on Illumina HiSeq X machines, on single lanes for the PCR libraries and multiplexed across 12 lanes for the PCR-free libraries, producing paired-end reads of length 2×151 base-pairs (bp), a mean insert size of 447 bp and a mean coverage of 35.0X. 14 of these libraries were described and used for population genetic analyses in a previous publication (*13*).

Reads were processed through the automated pipeline of the Wellcome Sanger Institute sequencing facility, mapping to the GRCh38 reference assembly (GRCh38_full_analysis_set_plus_decoy_hla.fa) using bwa mem version 0.7.12 (*47*) with the –T 0 parameter. Tools from the biobambam package (*48*) were used to trim adaptor sequences from reads prior to mapping (bamadapterclip) and to mark duplicate reads after mapping (bamstreamingmarkduplicates). We performed no further post-processing of the alignments, e.g. base quality recalibration or local indel realignment.

Data previously generated on a subset of the samples from the HGDP-CEPH panel as part of the Simons Genome Diversity Project (*3*) (hereafter referred to as "SGDP") was also incorporated into the project (paired-end reads of length 2×100 bp, mean insert size of 310 bp, mean coverage of 42.4x). Data from another earlier publication (*11*) (hereafter referred to as "Meyer") was obtained from ENA experiment accession number SRX103808, de-multiplexed allowing for one nucleotide mismatch in the barcode sequence and then also incorporated (paired-end read lengths of 94+100 bp or 95+101 bp, mean insert size of 264 bp, mean coverage of 28.7).

Reads from the SGDP and Meyer libraries were processed as the Sanger libraries above, except the trimadap tool (https://github.com/lh3/trimadap) was used for adapter trimming, the –T 0 parameter to bwa mem was not applied, and the bwa-postalt.js post-processing script was run after mapping.

Sample quality control, some of which is described in more detail below, included assessments of overall sequencing coverage, read error rates, genotype concordance to published genotype array data and cell-line chromosomal artefacts. For a few samples in the panel we had more than one library from different sources (i.e. Sanger, SGDP or Meyer, and PCR or PCR-free), and in each such case choose to include the library with the highest array genotype concordance, but excepting that rule in a few cases to avoid any of the 54 populations in the panel having an atypical composition of libraries from these different sources (for example, having more than the typical two SGDP libraries), when possible. After quality control and sample exclusions, the final set of 929 libraries contained 649 Sanger PCR-free, 152 Sanger PCR, 111 SGDP PCR-free, 9 SGDP PCR and 8 Meyer PCR libraries (table S1).

Previously published array genotypes available on the panel were lifted over from hg18 (*8*) and GRCh37 (*10*) to the GRCh38 assembly using the NCBI Remap tool and through looking up rs numbers in dbSNP (*49*).

Capping of mapping qualities

We noticed that the genotypes called from many of the Sanger PCR-free libraries displayed higher rates of discordance with array genotypes from the same individuals than the other sample sets, including the Sanger PCR libraries. To test if this could be due to cross-sample contamination caused by index hopping in the multiplexed sequencing runs performed for the PCR-free libraries, we ran the VerifyBamID tool (*50*) to estimate a per-library contamination rate (the "FREEMIX" estimate). These estimates were higher for many of the Sanger PCR libraries (in the highest cases 1-2%), and correlated strongly with the array discordance rate, strongly suggesting this was the cause of the reduced genotype accuracy (fig. S1A).

To reduce the impact of the index hopping on genotype accuracy, we applied a per-sample cap on the mapping qualities (MAPQ) of reads as a function of the contamination estimate (https://github.com/mcshane/capmq):

$$\max(MAPQ) = \max\left(20{,}10 \cdot \log_{10}\frac{1}{FREEMIX}\right)$$

The rationale behind this is that if e.g. 0.1% of the reads are contaminant and do not actually derive from the given individual, for any given read we cannot, regardless of how well it has aligned to the reference genome, have more than 99.9% confidence that it reflects the genome sequence of the individual. We thus express that uncertainty by lowering the mapping quality of any read above this confidence level, in this example corresponding to a mapping quality of 30. However, we do not set the cap for any sample at lower than 20.

Applying these sample-specific mapping quality caps substantially improved the array discordance rate for the Sanger PCR-free libraries, bringing them into the same range as non-multiplexed libraries (fig. S1B). It also slightly improved the rate for some other libraries. One library displayed a very high contamination rate (~4%) and a high array discordance even after capping mapping qualities, and was therefore marked as failing quality control and excluded from analyses.

Cell line copy number alterations

To identify any large-scale chromosomal copy number changes having occurred during culturing of the HGDP-CEPH lymphoblastoid cell lines from which DNA was obtained, we studied the mapped read coverage along the chromosomes of each sample. We calculated the coverage at approximately 300,000 single positions across the genome, and then plotted rolling means of these normalized by the genome-wide median. We visually inspected the plots for each sample and identified local deviations from the expected normalized copy number. As the cell line is a population of cells, the magnitude of the coverage deviation will be proportional to the fraction of the cells carrying the copy number alteration and can thus fall anywhere within the continuous range between the expected (1) and the most extreme possible values (1.5 in the case of a gain and 0.5 in the case of loss).

We marked 66 libraries as displaying at least one instance of mild deviation in coverage along a substantial stretch of a chromosome or major deviation in a megabase-sized stretch. Most events affected entire chromosomes or chromosome arms, but some smaller events were also observed. An especially large number of whole-chromosome gains were observed on chromosomes 9 and 12. We considered 9 of these 66 libraries to have deviations that were too extreme, informed by a computational experiment described below, and marked these as failing quality control. This included one sample (HGDP01097, Tujia) which did not display any copy number alterations, but which we discovered was completely homozygous along the entire length of chromosome 1 (we confirmed this was also the case in previously published genotype array data), most likely reflecting a cell line uniparental disomy event (though we cannot rule out that this reflects an actual germline uniparental disomy event in the sample donor). While several of the identified cases of unexpected copy number involved the sex chromosomes, particularly loss of chromosome Y in males and loss of chromosome X in females (fig. S2B), we did not exclude any sample on the basis of sex chromosome coverage alone. One self-reported male individual displayed what is likely an actual XXY karyotype rather than a cell line alteration.

We performed a computational experiment to assess the effects of coverage deviations on the accuracy of genotype calling at small nucleotide variants and determine whether it was appropriate to retain samples with mild deviations in our final dataset. We took the reads from the haploid chromosome X of two male individuals (HGDP00262 and HGDP00251, both Pathan) and subsampled these in varying proportions to generate synthetic diploid datasets corresponding to varying levels of coverage imbalance between the two chromosomes. We then called genotypes and determined how often the genotype called from a given imbalanced dataset matched that called from the perfectly balanced dataset at heterozygote SNPs (copy number changes will not impact the calling of homozygous genotypes). We found that copy number gains of a chromosome, even extreme ones, have only very minor effects on genotyping accuracy on high-coverage data (fig. S2A), consistent with the variant calling algorithm being able to tolerate some fluctuation from the expected balanced proportions of allele observations. However, we found that copy number losses result in larger effects on genotype accuracy. We therefore considered copy number losses as particularly strong reasons for sample exclusion.

DNA from these cell lines has previously been used to generate a large amount of data using a variety of technologies, and such data might potentially have been affected by the chromosomal alterations we identify here. To test to which extent a previously published and widely used array genotype dataset from the panel (*8*) was affected, we asked if the total sequencing coverage of a chromosome in our whole-genome sequencing data correlated with heterozygosity of the array genotypes. We found that increased coverage of a chromosome is associated with reduced heterozygosity in the array genotypes (fig. S2C), consistent with imbalanced allele counts leading to undercalling of heterozygotes. However, the effect is not very dramatic and thus is unlikely to have had much impact on population genetic analyses.

Genotype calling and filtering

We identified and genotyped SNPs and small indels using GATK HaplotypeCaller (*44*) version 3.5.0, applying genotype priors without bias towards the reference allele through the "--input_prior 0.001 --input_prior 0.4995" arguments (*3*), the "--pcr_indel_model NONE" argument for the PCR-free libraries and the "--includeNonVariantSites" argument to include monomorphic sites in the output VCF files.

We wished to apply filters that are equally stringent for variant sites as for non-variant sites. Any filter that applies to variant sites but not to non-variant sites, e.g. GATK's Variant Quality Score Recalibration, comes with a risk of introducing a bias against variants and thereby introduce skews into various population genetic analyses that rely on the balance between these two classes of sites. We applied filters to the genotype calls using the GQ ("Genotype Quality") and RGQ ("Reference Genotype Quality") annotations that are produced by GATK for variant sites and non-variant sites, respectively. The GATK software outputs these as separate annotations and does not guarantee that they are comparable, but we performed a computational experiment to test how these annotations behave in practice. We performed single-sample calling for a given sample, and from these calls extracted the genotype annotations at the ~60 million sites which during the joint calling of the whole panel had been called as polymorphic SNPs but with a homozygous reference ("0/0") genotype for the given sample (meaning some other sample in the panel carried an alternative allele at this site). This thus gives us a set of sites where the genotypes for the given sample has been evaluated by GATK once as non-variant sites and assigned RGQ values, and once as variant sites and assigned GQ values. For the sample HGDP01377, out of the 61,028,821 surveyed sites, GQ != RGQ at 6,650,620 sites (10.9%), but |RGQ – GQ| > 1 only at 23,372 sites (0.038%), suggesting most differences are just due to numerically rounding off the values into different consecutive integer bins. These comparisons are slightly complicated by multi-allelic sites in the joint calls – restricting to bi-allelic SNPs, |RGQ – GQ| > 1 at 0 sites. We performed the same experiment for two additional samples (HGDP00995 and HGDP01377) and obtained similar results. Thus, at least on this dataset called with non-reference biased priors, GQ and RGQ behave extremely similarly in practice and we thus take the same threshold applied to these annotations as providing equally stringent filtering for variant and non-variant sites.

We proceeded to filter the genotypes as follows. For each sample, we set any genotype to missing if it had a GQ or RGQ value equal to or lower than 20, or a coverage ("DP" annotation) equal to or greater than 1.65 times the genome-wide average coverage for the sample.

We also computed two site level annotations: GATK's Variant Quality Score Recalibration (VQSR) and excess heterozygosity. VQSR was run on the unfiltered genotypes, for SNPs with the hapmap_3.3.hg38.vcf.gz, 1000G_omni2.5.hg38.vcf.gz and 1000G_phase1.snps.high_confidence.hg38.vcf.gz variant sets from the GATK GRCh38 bundle for training and the former two for truth sets, and using the QD, MQRankSum, ReadPosRankSum, FS and MQ annotations as features. VQSR was run for indels with the Mills_and_1000G_gold_standard.indels.hg38.vcf.gz variant set for training and truth sets, and using the FS, ReadPosRankSum, InbreedingCoeff, MQRankSum and QD annotations as features. After annotating the VCF with the obtained VQSLOD values, another annotation "VQSRMODE" was added to each site to indicate whether it was evaluated in the SNP or indel mode, as at multi-allelic site this information might otherwise not always be retained when filtering out one of the alleles or subsetting to a sample set in which the evaluated allele is not present. The excess heterozygosity "ExcHet" annotation was calculated on a per-allele basis using the bcftools (*51*) fill-tags plugin. We then marked any site with a SNP VQSR score below -8.3929 or indel VQSR score below -1.0158 with the "LOW_VQSLOD" filter tag, and any site harbouring an allele with an ExcHet value equal to or larger than 60 (corresponding to an excess heterozygosity p-value of 10^{-6}) with the "ExcHet" filter tag. For

downstream analyses, we did not make use of the VQSR scores for filtering, unless noted, but we excluded sites tagged with excess heterozygosity from all analyses.

We wished to restrict certain analyses to bi-allelic SNPs. Rather than excluding the entire site in case a third allele and/or an indel allele is present, which would mean that an e.g. an additional allele observed in a single individual would lead to the exclusion of a site harbouring two otherwise informative and perfectly usable common alleles, for most analyses we instead masked out alleles. If an allele is lower in frequency than two other alleles at the same site, or if the allele is an indel allele, we excluded the allele and set the genotype of any individual carrying a copy of the allele to missing, allowing us to retain the site and the two most common SNP alleles.

We annotated SNP variants with two ancestral allele tags: one using the allele predicted by the Ensembl 8 primates EPO alignments (cc21_ensembl_compara_86), and one using the Chimpanzee allele in the GRCh38-PanTro4 alignment from the UCSC genome browser.

We noticed that in population genetic analyses (e.g. principal component analyses) of the unfiltered genotype calls, there were noticeable batch effects between the libraries of different sources, primarily between Sanger and SGDP libraries but also to a smaller extent between PCR and PCR-free libraries. The genotype filtering described above reduces these effects, as does increasingly stringent VSQLOD thresholds, and when applying the accessibility mask described above the effects are not discernible. In the final genotypes used for analyses there is thus no batch effect visible (fig. S3). However, we still urge any users of the data to be aware of the possibility that some sensitive analyses could still be affected by these effects, particularly if not applying the accessibility mask.

Overview of small variant callset

Our filtered variant call set across 929 samples, excluding sites labelled with excess heterozygosity, contained 75,310,370 variant sites. This included 67,325,692 SNPs and 8,797,538 indels. 3,085,457 of all variants were multi-allelic, and 855,977 of SNP sites (1.3%) were multi-allelic. The transition/transversion ratio was 1.88. 29,279,819 SNPs (43.5%) and 3,431,934 indels (39.0%) were singletons (the alternative allele observed only in one copy across the individuals).

We compared the identified variants with those identified by the 1000 Genomes Project (*2*) (the 20170504 GRCh38 lift-over version). While both datasets contain millions of variant alleles that are not found in the other dataset, the vast majority of these are very low frequency, including many singletons, and a simple count of overlapping versus unique variants does not reveal to which extent either of the datasets contains variants that might be of higher frequency in particular populations. We therefore calculated, for each variant present in one dataset but not the other, its maximum allele frequency in any population. To avoid cases where a variant is actually present in the sequenced individuals in both datasets but absent from one of the VCFs because of technical issues in variant calling and/or lift-over, we excluded 1000 Genomes variants that did not have the "GRCH37_38_REF_STRING_MATCH" tag (indicating that the reference allele string matches between GRCh37 and GRCh38), and we excluded from both datasets any variant that had a global allele frequency across the dataset of 30% or higher (the reasoning being that it would be very unlikely that such a globally common variant would not be sampled by both of these datasets).

The HGDP dataset contains a larger number of populations and these populations have smaller sample sizes than in the 1000 Genomes dataset, leading to higher variance in the per-population allele frequency estimates and thus potentially resulting in an upwards bias when finding the maximum allele frequency across populations. To enable a fair comparison, we down sampled the 1000 Genomes dataset to resemble the HGDP dataset: for each of the 26 1000 Genomes populations, two random subsets were constructed with sizes determined by sampling without replacement from the set of HGDP population sizes. The distribution of maximum allele frequencies across these size-matched 1000 Genomes populations for variants not present in HGDP shifted upwards, but variant numbers still remained much lower than the number of variants private to HGDP. We performed this down-sampling three times, and the results were very similar across the three replicates.

gVCF construction

Having genotype calls at monomorphic sites is very valuable, as they are needed in certain population genetic analyses and also as they make appropriate merging of different datasets possible. We produced VCFs containing every called site, but these files are very large and therefore challenging to distribute and store. We therefore constructed per-sample gVCFs ("genomic" VCFs) in which consecutive sites with homozygous reference genotypes and similar confidence level are grouped into single, block VCF records. We grouped such sites into four classes of blocks, in which the GQ and RGQ annotations are collapsed into a single GQ annotation:

Block definition	**GT field**	**FILTER tag**
DP ≥ mean(DP) × 1.65	set to missing	EXCESS_DP
GQ>60 & DP < mean(DP) × 1.65	unchanged	.
GQ>20 & GQ<=60 & DP < mean(DP) × 1.65	unchanged	.
GQ<=20	set to missing	GQ20

For each block record, only the DP and GQ genotype annotations are retained and used to represent the minimum value across all sites contained within the block. If a site with a non-reference genotype fails any filter, we set the genotype to missing but still retain the alternative allele and all the genotype annotation fields, such that it's always possible to restore the genotype called by GATK if desired. For sites with non-reference genotypes we also carried over a few site-level annotations from the joint variants VCF (ExcHet, VQSLOD, VQSRMODE).

These gVCF files constitute compact representations of the genome sequences of single individuals, while retaining most of the information of relevance to assess the uncertainty of a given genotype call and allow for custom filtering. The mean size of these files across the 929 samples is 1.32 GB (standard deviation = 0.48, min = 0.25, max = 3.25), with the variation largely explained by differences in sample coverage ($r_{coverage,file\ size}$ = -0.79).

Accessibility mask

Rather than variant level filtering, for most analyses we instead relied on an accessibility mask. This mask was constructed on the basis of the 1000 Genomes Project's strict mask for GRCh38 (20160622 version), which is based on coverage and mapping quality patterns in the 1000 Genomes Project dataset. From this mask, we also subtracted any regions of the

primary GRCh38 assembly which have alternative loci or patch scaffolds, as defined by the NCBI assembly resource for the GCA_000001405.15_GRCh38 entry. As we performed variant calling on GRCh38 read alignments that had not been post-processed to adjust the mapping qualities in an alt-aware manner, genotypes in these regions are likely not reliable. We also subtracted all sites tagged with excess heterozygosity, considering them unsuitable for genotyping from Illumina reads. The resulting masks leaves approximately 73% of the primary assembly for analyses, and unless otherwise noted our analyses are restricted to this mask. While we find that restricting to this mask without further site-level filtering appears suitable for the population genetics analyses we perform here, other types of analyses using this dataset might benefit from alternative filtering strategies, e.g. using the VQSR scores or other annotations. In particular, analyses of variants of potential functional, medical or selection relevance might benefit from interrogating variants also in the approximately 27% of the genome that falls outside of this mask.

10x Genomics sequencing and haplotype phasing

We selected 26 samples from 13 populations to process with the 10x Genomics Chromium technology (*14*), producing linked reads with long-range physical information that enable haplotype phasing. We selected 13 of the 54 populations representing key ancestries. Within each of these populations, we tried to fulfill several criteria when selecting which two samples to process: an absence of any large-scale chromosomal copy number alterations; a typical ancestry profile with respect to their population as assessed through principal component and model-based clustering analyses; no evidence of high relatedness between the two individuals; and male individuals if possible, to obtain Y chromosome data.

DNA quality and molecule length distributions were assessed using the Agilent TapeStation, and Chromium libraries were then constructed for the 26 selected samples and sequenced on single lanes of Illumina HiSeqX machines (2×151 bp reads, average coverage 30.2X) (table S2). Four of these libraries were described and used for population genetic analyses in a previous publication (*52*). The resulting barcoded reads were processed using the 10x Genomics Long Ranger software version 2.1.2 with genotype calling through GATK HaplotypeCaller 3.5.0, to obtain phased VCFs for each individual. In order to maintain only a single set of genotype calls for these 26 individuals for which we had calls both from the standard Illumina data and from the Chromium data, we lifted over the haplotype phase information at heterozygous sites from the latter to the former. Any genotype where the two calls disagreed were set to unphased, and variants within phase blocks that contained only one variant were also set to unphased. The resulting VCF files thus contained the unaltered genotypes that we called from the standard Illumina data, with only haplotype phase information obtained from the Chromium experiments.

Structural variation calling

We called copy number variants using GenomeSTRiP v2.00 (*53*) using default parameters. We initially ran the algorithm jointly on all 965 available libraries, including libraries not passing quality control for short variant calling, and including the Meyer libraries. However, we found the quality of resulting calls for the Meyer libraries to be low and re-ran the algorithm excluding them.

As we have a number of duplicate samples prepared using PCR and PCR-free libraries for quality control purposes, we ran the algorithm twice for these, separately for each library

preparation set, in each case together with the rest of the dataset. We found that more variants were called for PCR-based libraries compared to PCR-free libraries. We also found that samples prepared with PCR libraries had a larger number of shared heterozygous calls that are missing from the PCR-free libraries, suggesting these are artefactual calls. We subsequently excluded variants with excessive heterozygosity as computed by bcftools v1.9 (ExcHet < 0.0001) separately for each library preparation and sequencing location set (i.e. SGDP PCR, SGDP PCR-free, Sanger PCR and Sanger PCR-free). For the SGDP PCR samples we used ExcHet < 0.05 as this set had only 9 samples.

We investigated potential cell-line artefacts in further details by analysing coverage across the genome for each sample, as CNV calling might be more sensitive to such artefacts than short variant calling. From the 929 samples included in the SNP analysis, we excluded the Meyer samples as well as 10 samples that displayed evidence of alterations across multiple chromosomes. We also masked regions in 74 samples that showed more limited putative alterations, meaning we retain the samples but we did not consider variants called for them within these masked regions. This resulted in a VCF file with 911 samples in which GenomeSTRiP called 50,474 CNVs.

We examined the calls and found cases where the algorithm splits variants into multiple shorter entries which are not always overlapping. This a known behaviour of the GenomeSTRiP CNV pipeline, and it seems to occur when there are variants with different copy numbers across different individuals within a sub-segment of a larger variant. This issue can also occur if a low-quality variant is found within a larger CNV. To address these issues and be able to more accurately estimate the total number of identified CNVs in our dataset, we merged high quality (CNQ > 12) calls that have same diploid copy number and are within 50 kb of each other, for each sample individually. For the X-chromosome we performed this separately for male and female samples. At this point, we observed that one sample (HGDP01254) had an elevated number of variants compared to the rest of the samples. Closer inspection showed that these calls had relatively low genotype quality. To be conservative, we excluded this sample from downstream analysis, leaving 910 individuals. All variants were then merged using bedmap v2.4.35 (*54*) based on 100% overlap. This resulted in 39,634 autosomal variants and 1,102 variants on the X-chromosome.

Statistical haplotype phasing

As we only had 10x Genomics experimental phasing data for a subset of samples, we also performed statistical phasing of the whole panel. We restricted statistical phasing to biallelic SNPs (by masking out third or higher alleles and indel alleles, rather than excluding entire sites) lacking filter tags. To take advantage of a large external reference panel without having to exclude variants not present in the reference panel, we applied a scaffold-based method implemented as a genotype calling mode in SHAPEIT2 (*55*). We first obtained the scaffold by phasing the genomes using Eagle (v2.3.2) (*56*) after three PBWT iterations with 4,956 genomes from the African Genome Resources (https://www.apcdr.org/) in the reference panel (this includes the full 1000 Genomes Project panel). Subsequently we divided the unphased chromosomes into windows each spanning 2,400 SNPs with 200 overlapping ones between adjacent windows. Beagle 4.1 (*57*) was run with the -gtgl option to produce genotype probabilities in each window. Based on the genotype probabilities, the variants were then mapped onto the phased scaffold with the --call and --input-scaffold options in SHAPEIT2. To avoid under- and overflow errors (as described in (*3*)), we ran SHAPEIT2 in the same windows as in the previous step instead of assigning windows by physical lengths,

and further enlarged the windows when such errors still occurred on rare occasions. Missing genotypes in the unphased input dataset were reset to missing in the phased dataset for consistency.

We evaluated the accuracy of the statistical phasing by comparing the inferred haplotypes to those obtained for the samples sequenced with 10x Genomics linked reads. If adjacent phase-resolved heterozygous sites in the same phasing block in the 10x genome formed a different haplotype structure than that in the statistically phased genome, we considered it a switch error. Table S3 lists the switch error rates on chromosome 1, measured as the total number of switch errors divided by the total number of possible switches (namely the number of heterozygous sites minus one). When singleton alleles are excluded, the highest switch error rates of around 1.2% are found in the San and Papuan populations, which roughly corresponds to on average one switch error per 160 kB. We therefore do not expect phasing errors to have a detectable impact on downstream analysis targeting haplotypes substantially shorter than this.

Metadata curation

There is some inconsistency in the population labels used in the prior literature on the HGDP-CEPH samples. We reviewed the population labels, aiming to adhere as much as possible to the labels provided in the official CEPH sample documentation but also to define the most scientifically useful groupings with labels that are ethnically, linguistically and geographically appropriate. After this review, we arrived at 54 population labels. While most differences to previously used labels involve only minor spelling variations, a few cases involve more notable changes. We comment on some of the population labels and the motivation behind our choices, as well as any notes on the coordinates used to indicate geographical origins, here:

- **BantuSouthAfrica** and **BantuKenya**: These have sometimes been collapsed into a single "Bantu" label, however we use two labels as they are substantially separated both geographically and genetically ($F_{ST} \approx 0.008$). The South African samples have sometimes been further subdivided into "South-Western" and "South-Eastern" sets or into individual language groups, however to retain a decent sample size for the population we do not make use of these further subdivisions.

- **Colombian**: These have sometimes been subdivided into two individual language groups (Piapoco and Curripaco), however to retain a decent sample size we do not subdivide these samples (the CEPH documentation also does not describe which of the samples are from which of the two language groups).

- **Han** and **NorthernHan**: While representing individuals from the same ethno-linguistic group, given the large number of Han individuals in the panel we made use of the sometimes utilized second label for the set sampled in northern China, for which we assign new geographical coordinates based on documentation from the original sample collection.

- **Mongolian**: "Mongola" has been used, however we believe this reflects a spelling error. The geographical coordinates reported for this population in (*58*) differs slightly from those in the CEPH documentation – we believe the latter are the correct coordinates.

- **Bougainville**: “Melanesian” or “NAN Melanesian” (NAN=Non-Austronesian language) has been used for this group from Bougainville Island, but is overly generic. The name of a specific language group has sometimes been used, but we do not make use of this as it is not part of the CEPH documentation.

- **PapuanSepik** and **PapuanHighlands**: The prior literature has used the single label “Papuan” for these samples from Papua New Guinea, however it has been shown that they consist of two genetically highly distinct ($F_{ST} \approx 0.03$) subsets, one with affinities to populations in the eastern highlands and one with affinities to populations in the Sepik river region of the northern lowlands of New Guinea (*52*), and we thus separate them into two separate labels. For the PapuanSepik population, we use the geographical coordinates provided in the CEPH documentation, which are consistent with a Sepik region location. For the PapuanHighlands population, we assign new coordinates on the basis of the results reported in (*52*).

- **San**: “Ju|’hoan North” has been used, but this label is not used in the CEPH documentation and so we use the “San” label, even if it is more generic.

- **BergamoItalian**: “North Italian” has been used, but we include the Bergamo origin of these samples in the label to clarify its distinction from the Tuscan population that is also part of the panel.

f-statistics analyses

We calculated f_4 and *D*-statistics using the ADMIXTOOLS package version 5.0 (*10*). As the ADMIXTOOLS programs used excessive memory when attempting to use all variants to calculate large numbers of statistics, e.g. the 948,753 f_4-statistics corresponding to all possible relationships among the 54 populations, we ran ADMIXTOOLS separately on 5 Mb blocks across the genome and then performed our own block jackknifing across the per-block estimates. We verified on a small subset of statistics that the obtained f_4 values and Z-scores were highly correlated to those calculated one by one directly by ADMIXTOOLS. F_{ST} was calculated using EIGENSOFT version 6.0.1 (*45*).

Effects of variant ascertainment on population genetic analyses

The ideal class of variants for analyses that rely on genetic drift are those that were polymorphic in the shared ancestral population of the given populations under study. One approach to approximate this ancestral polymorphism is to ascertain variants in an outgroup. We ascertained SNPs that are polymorphic among three high-coverage archaic human genomes: the Altai Neanderthal, the Vindija Neanderthal and the Denisovan genome. Within the accessibility mask, this resulted in 2,809,464 variants. Out of these, 1,350,097 were also polymorphic within the set of 929 modern humans. Out of these, 931,790 (69.0%) were polymorphic also among Africans only (101 individuals, excluding three BantuKenya individuals displaying some non-African ancestry components in ADMIXTURE runs), demonstrating that the presence of these variants in present-day modern human populations is to a large extent explained by them being present in the shared ancestral population of modern and archaic humans, and only to a smaller extent a consequence of archaic admixture into non-Africans.

We compared results of various population genetic analyses obtained on all the discovered variants and the above outgroup ascertained set to sets of variants present on commonly used genotyping arrays: the Illumina 650K (or "Li 2008") array (*8*), the Humans Origins array (*10*) and the Illumina Multi-Ethnic Global Array ("MEGA"). We lifted over the site lists of these arrays to GRCh38 using the NCBI Remap tool. While most of the samples in the HGDP-CEPH panel have been typed on the former two arrays, rather than using those datasets directly we extracted the genotype calls made from our whole-genome sequencing data on the sites present on the arrays, such that genotypes at any given site are held constant and only the sites used differ. We found:

- f_4-statistics calculated using array sites are highly correlated to those calculated using all variants or outgroup ascertained variants, overall. However, some statistics, especially those involving African populations, deviate in the array sites as a result of ascertainment bias (Fig. 1C, fig. S4A). The overall magnitude of the f_4-statistics differs systematically between array sites and all sites, but this just reflects the overall larger number of variants included in the latter, many of which will not be polymorphic among the set of four populations used in a given statistic and therefore just contributes to smaller allele frequency differences on average.

- Estimates of individual ancestry components using ADMIXTURE (*46*) are cleaner when using outgroup ascertained sites, with less "leaking" of a given regional ancestry component into individuals outside that region (fig. S4B). For example, running ADMIXTURE on k=5 using the Li 2008 sites, individuals from Central and South Asian populations are assigned 5-10% of the Oceanian component, but this is substantially reduced with outgroup ascertained sites. Similarly, it reduces the fraction of Native American component in Europeans.

- F_{ST} is generally slightly overestimated when using array sites, especially so when comparing African to non-African populations (fig. S4C). The Pearson correlations between the F_{ST} values estimated using whole-genome sequencing data (excluding singletons) and those estimated using different sets of ascertained sites were, for the full set of F_{ST} values: $r_{archaic\text{-}ascertained} = 0.99976$, $r_{Li\ 2008} = 0.99726$, $r_{Human\ Origins} = 0.99726$, $r_{MEGA} = 0.99923$. When restricting to F_{ST} values between one African and one non-African populations, the correlations were: $r_{archaic\text{-}ascertained} = 0.99883$, $r_{Li\ 2008} = 0.99492$, $r_{Human\ Origins} = 0.99730$, $r_{MEGA} = 0.99692$.

Region specific variants

We studied the number and frequency distributions of variant alleles that are private to a particular region of the world, meaning alleles that have counts of zero across all individuals outside the given region. For these analyses, we did not restrict only to variants falling within the accessibility mask, with the rationale that technical errors are unlikely to result in genotype distributions that correlate perfectly with geographical labels. To reduce the effects of recent admixture between regions, we excluded individuals displaying evidence of such admixture in model-based clustering analyses. We ran ADMIXTURE (*46*), on 1,350,097 bi-allelic SNPs ascertained as polymorphic among three high-coverage archaic human genomes, with five ancestry components. The components obtained corresponded well to five continental-level ancestries: sub-Saharan African, West Eurasian, East Eurasian, Native American and Oceanian. We used the estimated per-individual ancestry proportions together with the population and region level metadata to define which individuals should be included

when counting variants private to a given region (the "ingroup") and which individuals should be used to ascertain the allele count of zero (the "outgroup") (table S4). Our rationale when defining these criteria was that, if we are looking for variants that are found in region A but not region B, it is more important to exclude individuals from B with recent admixture from A than vice versa – the inclusion of the former might cause otherwise private A variants to be observed in B individuals and therefore not be identified as private, while the inclusion of the latter will only lead to underestimation of the allele frequencies of private variants in A. We calculated 95% Poisson confidence intervals around the private variant counts using the R function poisson.test. In addition to the cumulative displays, fig. S5A displays the counts of region-specific SNPS in a non-cumulative fashion.

When analysing CNVs, we set any individual CNV genotypes with GQ < 20 to missing and excluded variants with >15% missingness across individuals (calculated in the VCF after filtering but before merging of adjacent variants). To further conservatively avoid overcounting single CNVs that have been called as multiple adjacent entries, we subsequently merged variants with similar allele frequencies and the same copy number lying within 25 kb of each other and only report the variant with the lowest genotype missingness. If merged variants have the same missingness but slightly different allele frequencies, we calculated and report the average frequency.

To assess whether the high number of high-frequency private Oceanian CNVs could be expected by sampling noise alone or representing an enrichment indicative of positive selection, we randomly sampled 123 private Oceanian SNPs (matching the number of private Oceanian CNVs, with minimum allele count > 2) 1000 times and compared the frequency distributions in these random samples to the observed CNV distribution. We observed 12 sample sets with a variant having an equal or higher frequency than the single most frequent CNV (at a frequency of 82.14%). However, among these 12 sets, 10 sets had only one variant at >50% frequency and two sets had two variants at >50%, whereas in the observed CNV distribution we find four variants at >50% frequency. The observed CNV distribution is thus highly unlikely to be the result of sampling noise (fig. S5B).

MSMC2 split time analyses

Pairwise population separation histories were studied using the MSMC2 software (*22, 29*). Input files were prepared from genotypes on all called sites, including non-variant sites, and with haplotype phase obtained from the 10x Genomics Chromium experiments we had performed on two individuals each from 13 populations. MSMC2 v2.0.2 was run on eight haplotypes (four from each of the two populations) with the "--skipAmbiguous" argument to skip unphased segments of the genome. Results were scaled to real time by applying a mutation rate of 1.25×10^{-8} per site per generation and a generation time of 29 years.

Clean split scenarios were simulated using scrm (*59*) version 1.7.2, using a mutation rate of 1.25×10^{-8} per site per generation, a generation time of 29 years, a recombination rate of 1.12×10^{-8} per site per generation, an initial N_e of 20,000, a doubling of N_e between 300 and 1000 kya and simulating 14 chromosomes each of size 150 Mbp (for a total genome size of 2.1 Gbp, similar to the size of the empirical genomes restricted to the accessibility mask). The average heterozygosity of the resulting sequences was 0.986 per kbp, similar to human genomes of African ancestry. The sequences were analysed with MSMC2 as above, except the --fixedRecombination parameter was applied (running without this resulted in non-monotonic curves for many simulated histories).

We performed additional MSMC2 runs to test if the deep structure observed in our empirical results, i.e. that the cross-coalescence curves remain below 1 even several hundreds of thousands of years ago, could be caused by batch effects associated with sequencing and processing the two haploid genomes from a diploid human sample together. Any process that could cause the genotypes of these two genomes to appear artificially similar, from somatic or cell-line loss of heterozygosity events to under-calling of heterozygous genotypes during data processing, could cause the time to coalescence between these two genomes to appear lower than that to genomes from another individual. We ran MSMC2 on four haplotypes, but in each of the two populations selecting one haplotype from two different individuals, using the "-I" command line syntax of MSMC2 v2.1.1 on input files prepared with eight haplotypes, e.g. "-I 0,2", "-I 4,6" and "0-4,0-6,2-4,2-6" for the within first population, within second population and between population runs, respectively. We thus use four haplotypes sequenced in four different individuals, and should avoid any batch effects of the type described above. The curves obtained from these runs in many cases tend towards slightly higher relative cross-coalescence values in deep time periods then when using haplotypes sequenced in the same individual, and more so when involving two non-African populations, but they still display largely the same behaviour (fig. S7). While it's possible that the magnitude of the deep structure exhibited in our results could thus be slightly exaggerated by technical artefacts of this nature, they are unlikely to be responsible for the whole effect.

Application of MSMC2 to archaic genomes

We also ran MSMC2 on pairs of modern human and archaic populations. We used the high-coverage Altai Neanderthal, Vindija Neanderthal and Denisovan genomes, left all heterozygous genotypes in these as unphased, and ran MSMC2 as above except using only two haplotypes per population (we also performed runs without the "–skipAmbiguous" argument but found this made little difference to the results). While the method relies on phased genotypes for the between population coalescence rate estimation, the low heterozygosity of the archaic genomes means that many segments of these genomes will be homozygous and thus by necessity phased, which might be the reason we obtain seemingly sensible results.

We also performed simulation experiments to validate the robustness of running MSMC2 with one archaic genome lacking haplotype phase information. We used the basic parameters described above for the clean split simulations, and used scrm to simulate a history approximately mirroring the divergence and admixture between modern humans and Neanderthals: a divergence between two populations at 500 kya, followed (forward in time) by a ten-fold reduction in effective size of the second population at 400 kya, then followed by 2% gene flow from the second, low-diversity population into the first population at 50 kya. We then stripped the haplotype phase from the genome sampled from the second population and ran MSMC2. The results accurately recapitulate the simulated history, both the initial divergence and the later gene flow peak, and the stripping of the haplotype phase from the genome sampled from the low-diversity population does not substantially affect the results (fig. S9A).

We plotted the results naively as above, but also made plots where we attempted to correct for the fact that the archaic genomes are from ancient remains that stopped accumulating mutations at their time of death several tens of thousands of years ago. MSMC2 reports the within and the between population coalescence rates separately, and we could thus adjust

these as a function of the sample age before calculating the relative cross-coalescence rates. We shifted back the within population coalescence rates by adding a new time segment between 0 and the sample age with a rate of 0 and then adding the sample age to all existing time segments. We similarly shifted back the between populations coalescence rates by half the age of the sample age. We used sample ages of 122,000 years for the Altai Neanderthal, 52,000 years for the Vindija Neanderthal and 72,000 years for the Denisovan genome (*20*). Overall, we found that these sample age adjustments did not have substantial effects on the results, especially not the timing of the separation between modern and archaic humans on the order of 500 kya. They did however shift backwards in time the archaic admixture signal in non-African genomes such that it peaks around 80-120 kya, rather than 40-80 kya, when using the Vindija Neanderthal genome – this might actually more accurately reflect the timing of the event, as the introgressing Neanderthal population has some level of divergence from the Vindija individual (*20*) such that the coalescence events with the introgressed haplotypes will be older. The curves obtained when using the Altai Neanderthal are shifted backwards in time relative to when using the Vindija Neanderthal, reflecting how the former is further diverged from the introgressing source.

We observe the MSMC2 Neanderthal gene flow signal in all non-African genomes and to a much-reduced degree in Yoruba. The most likely explanation for these results is non-African gene flow carrying Neanderthal ancestry into West Africa (*41*). The other African populations do not display the same behaviour, with the tiny deviations from zero relative cross-coalescence rate in recent time periods, especially so when shifting back rates by sample age (including when using the Denisovan genome), likely not being distinguishable from technical noise. We do not observe any clear recent gene flow signal when running the Denisovan genome against African, Eurasian or Native American genomes (a tiny increase in the Yakut is difficult to distinguish from technical noise). However, when running Denisovan against Oceanian genomes (PapuanHighlands and PapuanSepik) a subtle upwards shift in the curve is visible roughly in the time span between 140 and 400 kya (moving back by ~40 kya when shifting rates by the Denisovan sample age). This very likely reflects the Denisovan gene flow in these populations, with the large backwards shift in time of the signal reflecting how the sequenced Denisovan from the Altai mountains is highly diverged from the population that contributed to Oceanians (*16*), such that the coalescence events with the introgressed haplotypes in Oceanians are quite old.

The Neanderthal gene flow signal is observed in both of the analysed Yoruba individuals, which provides some reassurance that it is a genuine signal. To further evaluate the robustness of the signal, we generated 50 separate block bootstrap replicate datasets for the Yoruba versus Vindija Neanderthal case by sampling 5 Mb chunks from across the genome, and ran MSMC2 on each of these. The Neanderthal gene flow signal is consistently observed across the bootstrap replicates, demonstrating that is it for example very unlikely to be driven by some technical artefact in a small subset of the genome (fig. S9B).

The application of MSMC2 to archaic genomes thus provides an additional line of evidence for the known archaic admixture in non-Africans, as well as for small amounts of Neanderthal ancestry in West Africans. We note that a conceptually similar approach, identifying haplotypes in non-Africans with low absolute divergence to archaic genomes, was used previously (*16*). Importantly, while methods relying on allele frequency correlations, e.g. *D* or f_4-statistics, produce results that are always relative between pairs of modern human populations, these MSMC2 results involve just one modern human genome at the time without the need for any baseline assumptions. These results thus allow us to say not only

that most sub-Saharan African groups have less Neanderthal ancestry than non-Africans, but also that they most likely in an absolute sense have very little Neanderthal ancestry, e.g. in the case of Mbuti a level that, within the limits of resolution of the method, likely is compatible with no Neanderthal ancestry at all.

Site frequency spectrum models with momi2

We used the momi2 software (*33*), which fits models to the site-frequency spectrum, to estimate pairwise split times between all 1431 combinations of the 54 populations, assuming a simple clean split without subsequent gene flow and a mutation rate of 1.25×10^{-8} per site per generation, using ancestral allele information from the Ensembl EPO alignments, and with confidence intervals obtained through bootstrapping across 500 genomic blocks. While the clean split assumption will be unrealistic in many cases, as evidenced by our MSMC2 results, the overall correlation between these momi2 split time estimates and the MSMC2 midpoint estimates is quite high ($r = 0.93$), suggesting the former will provide a decent approximation of the latter without requiring phased haplotypes. However, we also observed that the momi2 estimates are affected by the sample size of a population, and that they clearly greatly underestimated split times involving Native American populations (for example, some Native American against East Asian split estimates at just a few thousand years), perhaps as an artefact of the low recent effective size of these populations. We therefore do not place much emphasis on any particular single estimates. More elaborate models incorporating multiple populations, post-split gene flow and archaic admixture in the history of non-African populations have been shown to provide more accurate split time estimates (*33*).

Effective population size histories

We used smc++ v1.12.1 (*24*) to estimate effective population size histories for each of the 54 populations separately. Input files were prepared from genotypes on all called sites, masking out indel alleles and any third or further minor alleles at multi-allelic sites by setting the genotype of any individual carrying such alleles to missing. In addition to sites falling outside the accessibility mask, sites not present in the input VCFs were masked out from the analyses. smc++ requires one individual in the run to be specified as the "distinguished" individual, forming the basis of the coalescence time element of the inference, with the remaining individuals only contributing allele frequency information. For each population, we established a ranking for which individuals to use the distinguished individual, firstly prioritizing individuals not listed as ancestry outliers relative to their population by (*10*), secondly prioritizing Sanger PCR-free, then Sanger PCR, then SGDP PCR-free and then SGDP PCR-free libraries, and thirdly prioritizing higher sequencing coverage.

We ran smc++ assuming a mutation rate of 1.25×10^{-8} per site per generation, inferring effective population size up until 34 generations (approximately 1000 years assuming a generation time of 29 years) ago and otherwise default settings. For each run we prepared six alternative input files with different individuals as the distinguished individual and then gave all of these as input, resulting in composite likelihood results. To evaluate the variability of the inferred effective population size curves, we also generated 50 separate block bootstrap replicate datasets for each population, in each case by sampling 5 Mb chunks from across the genome, and ran the inference on each of these (fig. S10A).

Some populations display substantial decreases in inferred effective population size in the last ~5000 years, but we suspect that in many cases this reflects very recent endogamy or bottlenecks, the effects of which are being spread out over a larger time interval by the inference. We also noticed that under some parameter settings, in particular when decreasing the strength of the regularization (allowing more flexible curves to be fit, but also with a greater risk of overfitting) or when stopping the inference at 172 instead of 34 generations, Native American groups are inferred to have experienced dramatic growth approximately in the period between 20 and 10 kya. This is not observed when running on the parameter settings above, but we speculate that this could be a real signal which is counteracted in the inference by the strong bottlenecks experienced in the very recent time by the analysed Native American groups. Rapid population growth in this time period, coinciding with the initial peopling of the American continents, would be consistent with observations from the mitochondrial and Y-chromosomal phylogenies, which both display dramatic star-like behaviour starting around 15 kya indicative of rapid population expansion at this time. While the large degree of variability in the inferred curves between SMC++ parameter settings means we cannot be highly confident that the inferred growth reflects real growth rather than an artefact, we do not observe the same behaviour in other populations under the same settings (fig. S10B,C), suggesting that at least the phenomenon is a function of Native American genomes specifically rather than simply inherent to the parameter settings themselves.

For the 13 populations for which we had physically phased genomes, we also inferred effective population sizes histories using MSMC2. We ran MSMC2 on the two diploid genomes (four haplotypes) from each of these populations with the "--skipAmbiguous" argument to skip unphased segments of the genome. The MSMC2 results obtained on these physically phased genomes were largely concordant with the SMC++ results (fig. S11), confirming broad observations including recent declines in the African hunter-gatherer groups Mbuti, Biaka and San.

Y chromosome analyses

The haploid genotype calls for 603 males (here excluding the Meyer libraries which were of lower quality) across 10.3 Mb of accessible regions on the Y chromosome (*60*), lifted over to GRCh38 using the UCSC liftOver tool, were extracted using bedtools v2.22.0 (*61*). Within these regions, sites were filtered out if they contained indels, had missing genotype calls in more than 5% of the samples, had genotype qualities below 30 in more than 5% of the samples or had coverage above twice or below a third of the sample mean for more than 5% of the samples. This left a final set of 52,032 variant and 10,016,322 invariant sites.

The haplogroup of each sample was predicted using the yHaplo software (https://github.com/23andMe/yhaplo) after substituting the marker coordinates in the relevant input files to correspond to the GRCh38 assembly. An initial maximum likelihood phylogenetic tree was constructed using RAxML (v8.2.10) (*62*) with the GTRCAT substitution model with the set of 52,032 variant sites and then used as a starting tree for dating with BEAST v1.7.2 (*63, 64*). Markov chain Monte Carlo samples were based on 11,000,000 generations, logging every 1,000 generations. The first 10% of generations were discarded as burn-in. Eight independent runs were combined using LogCombiner. A constant-sized coalescent tree prior, the HKY substitution model, accounting for site heterogeneity (gamma) and a strict clock with a substitution rate of 0.76×10^{-9} (95% confidence interval: 0.67×10^{-9} to 0.86×10^{-9}) single nucleotide mutations per bp per year (*65*)

was used. A prior with a normal distribution based on the 95% confidence interval of the substitution rate was applied. Only the variant sites were used, but the number of invariant sites was defined in the BEAST xml file. A summary tree was produced using TreeAnnotator v1.8.1. The final tree (fig. S12, fig. S13) was visualised using the FigTree software (http://tree.bio.ed.ac.uk/software/figtree/).

Global estimates of archaic ancestry proportions

Genotypes for the high-coverage Vindija Neanderthal (*20*), Altai Neanderthal (*16*) and Denisovan (*11*) individuals on the GRCh37 reference assembly, along with corresponding filter files, were obtained from ftp.eva.mpg.de/neandertal/. The genomic coordinates were lifted over to GRCh38 using CrossMap v0.2.5 (*66*).

To estimate global proportions of archaic ancestry, we constructed a VCF which contained genotypes for the high-coverage Vindija Neanderthal, Altai Neanderthal and Denisovan genomes not just at sites that are polymorphic among modern human genomes, but also including sites that are monomorphic among modern humans but polymorphic among the entire set of modern and these three archaic genomes. We used these files, restricted to the accessibility mask, to estimate:

- *The proportion of Neanderthal ancestry in Eurasians*: We used an f_4-ratio, assuming a baseline of no Neanderthal ancestry in Mbuti, calculated using popstats (*67*), with the "--informative" option to include only variants polymorphic on both sides of the f_4-statistic:

$$\frac{f_4(Chimpanzee, Altai\ Neanderthal; X, Mbuti)}{f_4(Chimpanzee, Altai\ Neanderthal; Vindija\ Neanderthal, Mbuti)}$$

  The statistic gives mean estimates of 1.18% for Middle Eastern, 2.09% for Central & South Asian, 2.14% for European, 2.26% for American and 2.24% for East Asian populations. This statistic will not provide accurate estimates for Oceanian populations because of their high Denisovan ancestry. The highest population estimate is for the Chinese Xibo population at 2.41% (95% CI: 2.12 – 2.70%). Within Africa, the estimates are 0.048% (95% CI: -0.061 – 0.16%) for Biaka, 0.13% (95% CI: -0.025 – 0.28%) for San, 0.16% (95% CI: 0.031 – 0.29%) for Mandenka, 0.16% (95% CI: 0.035 – 0.29%) for BantuKenya, 0.16% (95% CI: 0.035 – 0.29%) for BantuSouthAfrica and 0.18% (95% CI: 0.05 - 0.30%) for Yoruba.

- *The proportion of Denisovan ancestry in Oceanians*, in two different ways:
  1) Using an f_4-ratio, assuming a similar level of Neanderthal ancestry in PapuanHighlands as in Han (*36*), calculated using popstats with the "--informative" option to include only variants polymorphic on both sides of the statistic:

$$\frac{f_4(Mbuti, Vindija\ Neanderthal; Han, PapuanHighlands)}{f_4(Mbuti, Vindija\ Neanderthal; Han, Denisovan)}$$

     This gave an estimate of 2.86% (95% CI: 2.08 - 3.64%).

  2) Using the qpAdm method (*68*) in ADMIXTOOLS version 5.0 (*10*). A model with PapuanHighlands as target, Karitiana and Denisovan as sources, and Yoruba,

Altai Neanderthal, Chimpanzee, Vindija Neanderthal, Mbuti and Biaka as outgroups gives an estimate of 2.70% (95% CI: 2.11% - 3.29%), with a model fit p-value of 0.1892. The rationale behind this approach is to simply model PapuanHighlands as a two-way mixture between Denisovan and non-African ancestry, and we here use Karitiana as a representative of the latter. It's likely that Karitiana has a small but non-zero amount of Denisovan ancestry, which would bias the estimate downwards, but given the very low amount of this ancestry the bias should be very small. Using a West Eurasian population with likely zero levels of Denisovan ancestry in place of Karitiana in the model would not be ideal as the lower levels of Neanderthal ancestry in such populations compared to Oceanians would lead to bias.

If we take the midpoint of these two estimates and conservatively take the largest confidence intervals, we obtain an ancestry fraction of 2.78% (CI: 2.08 - 3.64)

In summary, these estimates of the proportion of Denisovan ancestry in Oceanians are largely consistent with previous *f*-statistics based estimates, but lower than some of the highest estimates: 4.8% (*35*), 3.0% (*11*), 3.5% (*69*), 3.4% (*36*), 3.2% (*37*). The first of these estimates was made using a low-coverage Denisovan genome sequence and with lower quality modern data, and is thus likely not very reliable. A proportion of around or, as suggested by our new estimates here, even slightly below 3.0%, thus seems probable.

- *The proportion of Neanderthal ancestry in Oceanians:* Due to the high levels of both Neanderthal and Denisovan ancestry in Oceanians, and the partially shared drift between these two archaic groups, it is difficult to obtain an unbiased estimate of Neanderthal ancestry in Oceanians using a simple f_4-ratio. To get around this, we tried to estimate both Denisovan and Neanderthal ancestry jointly using qpAdm. We tried a model with PapuanHighlands as target, Denisovan, Vindija Neanderthal and Yoruba as sources and Chimp, Altai Neanderthal, Mbuti and San as outgroups, but this did not fit the data ($p = 4\times10^{-36}$), likely due to the model not accurately representing the relationships among the African groups and the position of Yoruba as a proxy for non-African ancestry. An East African population would likely be a better proxy, however the HGDP dataset does not contain a suitable population. We therefore tried this analysis on the Simons Genome Diversity Project (*3*) dataset which contains the Dinka population. A model with Papuan as target, Dinka, Vindija Neanderthal and Denisovan as sources and Chimpanzee, Altai Neanderthal, Yoruba, Mende, Biaka and Mbuti as outgroups fits the data ($p = 0.061$) and gives estimates of 2.0% (95% CI: 1.41 - 2.59%) Neanderthal ancestry and 3.4% (95% CI: 2.42 - 4.38%) Denisovan ancestry in Papuans. While the exact estimates might still be sensitive to slight model violations, these results suggest that the level of Neanderthal ancestry in Papuans is similar to the levels in other non-Africans.

Implementation of a Hidden Markov Model for archaic haplotype detection

A hidden Markov model (HMM) was used to detect introgressed segments from the Neanderthal and Denisovan populations in modern human genomes. The source code is available at https://github.com/ryhui/hmmarc. The HMM decodes segments of haploid genomes (thus requiring phased haplotypes) into two hidden states: unadmixed (0) and archaic (1). We summarize the observed data by the pattern of allele sharing between a panel

of sub-Saharan African genomes, the haploid genome under examination, and a panel of one or more archaic genomes. Only informative sites where a derived allele is shared between two groups and absent in the third are considered by the HMM. For convenience, the three informative combinations are encoded as emission types 1, 2 and 3 (table S5). If a genetic segment entered the modern human population from an archaic hominin population relatively recently, it should share more derived variants with the archaic genomes. Since sub-Saharan Africa is assumed to have no or very small amounts of Neanderthal and Denisovan ancestry, an allele shared between African and non-African genomes would most likely have arisen on a lineage within the modern human population. Incomplete lineage sorting in modern segments might cause some African lineages to coalesce first with the archaic genomes, thus sharing a derived allele unseen in a non-African genome. Such observations are less likely to occur in the archaic segments due to the small effective size of the archaic populations.

Transitions are only allowed between informative sites. Because the distance between informative sites is not constant, the transition probabilities are updated on-the-fly using genetic distance:

$$T = \begin{bmatrix} 1-(1-e^{-dt})\cdot\alpha & (1-e^{-dt})\cdot\alpha \\ (1-e^{-dt})\cdot(1-\alpha) & 1-(1-e^{-dt})\cdot(1-\alpha) \end{bmatrix}$$

Where $T_{ij}(i,j \in \{0,1\})$ is the probability to transit from state i to state j, t the time since admixture, d the genetic distance either extrapolated from a genetic map or, in the absence of such a map, using the physical distance and a constant per-site recombination rate. The admixture proportion α also defines the initial state probabilities $\pi = (1-\alpha, \alpha)$. The full model is specified by t, α and the emission probabilities matrix E.

Model training

Since the HMM is designed to work with a single archaic source at the time, we simulated 20 haploid genomes under the demographic model in (fig. S14) using msprime (*70*), and obtained the maximum likelihood estimation (MLE) of model parameters based on the true underlying state of the genetic segments (table S6). We fixed t at the true value of 2,000 but noted that varying it between 1,000 and 4,000 while keeping the other parameters unchanged has little influence on the decoding results. Applying this model to simulated data recovered 90.74% of the true archaic segments, with a false discovery rate of 3.68%.

We also explored the Baum-Welch algorithm and numerical likelihood optimization to train the HMM, however both training methods sometimes produced a minor state that absorbs type 2 emissions. It appears difficult to establish an archaic and a modern state from unsupervised training.

Comparison with published methods

In a scenario where only one archaic source of introgression is concerned, we compared the performance of the HMM in detecting Neanderthal segments with the S* method (*37, 71*), which searches for long haplotypes in linkage disequilibrium unseen in the African panel, and a method using a conditional random field (CRF) (*72*), which examines allele sharing, haplotype divergence and local recombination rate. Since the implementation of neither method is publicly available, we compared the result of running CRF, S* and the informative-site-only HMM on the same set of genomes instead. The Neanderthal segments detected by CRF and S* in individuals from 1000 Genomes Project were downloaded from the authors' websites. The HMM was then run on chromosome 1 of the 544 individuals that were included in both studies and the Viterbi sequences were obtained. To be consistent with

the other two methods, only the high-coverage Altai Neanderthal genome (*16*) was used in the archaic panel. The highest agreement is between HMM and S* (fig. S15A), where the shared regions constitute 72.84% of the total material recovered by the HMM and 82.43% of that recovered by S*. It is worth noting that although the S* score itself does not rely on the archaic genome, the reported segments in (*37*) have undergone subsequent filtering based on their match score to the archaic genomes. The other pairwise comparisons between methods only show around 40% reciprocal overlap. The overlap patterns are consistent across the seven Eurasian populations analyzed.

Under a different metric, a segment detected by one method is treated as a match if at least half of it is also reported by another method. Although the HMM does not preferentially detect or miss segments of particular lengths compared to the two other methods, segments detected by the HMM but not by the other methods tend to be shorter (fig. S15B). Segments shorter than 50 kB constitute 91.22% of those not detected by S* and 82.37% of those not detected by CRF. Most segments longer than 50 kB are reported by all three methods.

Distinguishing sources of archaic segments

Since Neanderthal and Denisovan segments coexist in many non-African genomes, we also tested various methods to distinguish between them. We launched another set of simulations following the demographic model in fig. S16, where a non-African population received 3% gene flow from the Neanderthal population, followed by 1% from the Denisovan population. The two-state HMM can be extended to include a Neanderthal state, a Denisovan state and a modern state, but this model showed very high false discovery rate (i.e. mislabeling unadmixed segments as archaic) on simulated data. Instead, we ran the two-state HMM twice with the same model parameters, first with two Neanderthal genomes in the archaic panel, then with the Denisovan genome. The posterior probabilities of the archaic state at each informative site from both runs, p_N and p_D, were used to assign them into the following categories:

- If $p_N \leq 0.5$ and $p_D \leq 0.5$, tagged as "modern";
- If $p_N > p_D$ and $p_D < 0.8$, tagged as "Neanderthal";
- If $p_D > p_N$ and $p_N < 0.8$, tagged as "Denisovan";
- If $p_N \geq 0.8$ and $p_D \geq 0.8$, tagged as "ambiguous archaic".

If a site only showed up as informative regarding the Neanderthal run or the Denisova run, the missing p_D or p_N was extrapolated linearly by physical distance from adjacent sites. The Neanderthal, Denisovan and ambiguous archaic segments were identified by linking neighbouring sites in the same category.

We found that this approach reduces the false discovery rate in simulations to around 0.05, at the cost of pooling a large proportion of archaic segments into the ambiguous category. The probability of labeling true Neanderthal segments as Denisovan is also below 0.05, and even lower the other way around. The criteria for assigning the categories based on p_N and p_D can also be adjusted according to the needs of downstream analyses, reflecting a trade-off between type 1 and type 2 errors. In analyses involving the haplotypes of Neanderthal and Denisovan segments, we used a more stringent set of criteria to obtain a "strict" set of archaic segments:

- If $p_N \geq 0.8$ and $p_D < 0.5$, tagged as "Neanderthal";
- If $p_D \geq 0.8$ and $p_N < 0.5$, tagged as "Denisovan".

In simulated data, this reduced the proportion of true Neanderthal segments to 0.016 among predicted Denisovan segments, and the proportion of true Denisovan segments to 0.0013 among predicted Neanderthal segments.

Detecting Neanderthal and Denisovan segments in modern populations

The two-state HMM was run twice on the 929 phased HGDP genomes with a genetic map to obtain the posterior probabilities of being in the Neanderthal and Denisova state at all informative sites. All 104 genomes from sub-Saharan Africa were included in the African panel, but when extracting observations, we allowed the archaic allele to reach a maximum frequency of 0.01 in this panel to allow for small amounts of archaic ancestry within Africa. Two high-coverage Neanderthal genomes, one from Denisova cave (*16*) and the other from Vindija cave (*20*), were used in the archaic panel in the Neanderthal run whilst the Altai Denisovan genome (*11*) was used in the Denisova run. In addition to the HGDP accessibility mask, we also included a low complexity regions mask (*73*). All sites that do not pass the masks were ignored as non-informative in the HMM runs. We used the ancestral sequences from Ensembl EPO alignment to determine the ancestral state; and in case of unknown sites in this panel, assumed the genotype in chimpanzee (Pan_tro 3.0) to be ancestral. Only sites that are polymorphic in the HGDP dataset were retained after merging. In effect, this leaves out derived sites shared by all modern human genomes but not the archaics, which are type 2 emissions (table S5) that supports assigning the modern state. However, such sites should be very rare in the genome, as derived alleles shared by all modern humans will be older than 200k years.

We obtained two sets of archaic segments following the first (hereafter the "basic" set) and the second (hereafter the "strict" set) criteria described in the previous section. In practice, the "basic" set assigns less material to the ambiguous category than in simulation studies, such that the actual distinguishing power is expected to be greater than in the results in the simulations.

Robustness of HMM to parameter misspecification

To explore the impact of parameter misspecification, we altered the model in table S6 in the following ways: 1. Halving and doubling π_1, the proportion of archaic ancestry; 2. Halving and doubling t, the time of admixture; 3. Increasing and decreasing each entry in the emission matrix by 10% (unless the new value exceeds 1), while scaling the other two entries in the same row proportionally such that each row sums to 1. We compared archaic segments detected on chromosome 1 using these altered models to the segments obtained using the original model in 10 randomly-drawn individuals from each geographical region. Table S7 shows the proportion of Neanderthal and Denisovan segments that are still detected following the "basic" criteria. The concordance is high except when we drastically reduce $E_{1,2}$ (the probability of emitting a variant absent in the African panel and shared with the archaic genome when the true state is archaic) from 0.9981 to 0.8983, which correspondingly makes the model over 50 times more permissive to observing variants shared between the African panel and the archaic genomes, or between the African panel and the genome under study. Even in this case, the concordance only falls below 0.8 in regions with very low level of archaic ancestry (notably Denisovan ancestry in Europe and the Middle East).

Geographical distribution of archaic ancestry

Fig. S17A compares the average amount of Neanderthal, Denisovan and ambiguous segments identified in the HGDP genomes (from the "basic" set) by geographical region, and fig. S17B shows the mean and standard deviation of the amount of Neanderthal and Denisovan segments identified in each population. The length of masked regions was excluded in all plots.

Very few archaic segments are detected in sub-Saharan African populations, but this is expected on technical grounds alone as our method conditions on allele frequencies in these populations. In accordance with previous studies, the amount of Neanderthal ancestry is higher in East Asia and the Americas than in Europe and the Middle East. The highest amount of Neanderthal ancestry is found in Oceania, but this is likely an artefact caused by some misclassified Denisovan segments. No prominent differences are observed between populations within the same geographic regions (fig. S17B). The intra-population variance is higher in Middle Eastern populations (especially Mozabite and Bedouin), likely reflecting recent admixture between sources with different levels of Neanderthal ancestry (e.g. African and West Eurasian). Denisovan segments are most abundant in Oceania (fig. S17A). They are also detectable at much lower levels in East Asia, the Americas and Central and South Asia, but negligible in Europe and the Middle East. Within Oceania, the Bougainville population has less Denisovan ancestry than the two populations from New Guinea (fig. S17B), consistent with dilution due to Southeast Asian admixture in the former.

All archaic segments in the "strict" set were pooled by geographical region to obtain maps of archaic ancestry frequencies along the genome. Fig. S18 depicts the distribution of Neanderthal and Denisovan segments along chromosome 1 as an example.

The difference between Oceania and other non-African populations appears more pronounced in the distribution of Denisovan than Neanderthal segments (fig. S18). To address this more formally, we quantified the length of overlapping genomic regions throughout the genome covered by at least two archaic segments between pairs of geographical regions, regardless of the genotypes in the segments (table S8). $P(A|B)$ here denotes the probability that a genomic region being observed in geographical region A, conditioned on it being observed in geographical region B. The geographical structure is stronger in the genomic distribution of Denisovan segments with less overlapping between America, East Asia, Central/South Asia and Oceania. Denisovan segments might have been lost through genetic drift more often or sampled less frequently than Neanderthal segments because of their low frequency in most populations; however, despite similar amount of Neanderthal and Denisovan ancestry in Oceania, $P(\text{Oceania}|\text{non-Oceania})$ is also lower for Denisovan segments than for Neanderthal regions. This indicates that the local landscapes of Denisovan ancestry across the genome is less similar between Oceanian and Eurasian populations than what the Neanderthal landscapes are. Neanderthal ancestry therefore probably results from a less complicated admixture history than Denisovan ancestry.

Divergence of archaic segments to archaic genomes

All Neanderthal and Denisovan segments from the "strict" set in each genome were compared with the Altai Neanderthal, Vindija Neanderthal and Denisovan genomes. To recover archaic-private variants that were not present in files produced by merging variants-only modern VCFs with all-sites archaic VCFs, we assumed that all modern sites passing the strict mask but not present in the VCF files carried the reference allele; if these sites also pass the respective archaic mask and appear in the archaic all-sites files with alternative alleles, they also contribute to the counts of differences.

The overall patterns of divergence to the Neanderthal genomes are almost identical across all six geographical regions (fig. S19A), while the patterns of divergence to the Denisovan genomes follow visibly different shapes in East Asia and Oceania: the points in Oceania form a well-defined single cluster; in East Asia, the pattern appears more noisy, with additional

segments displaying low divergence to the Altai Denisovan genome (divergence less than ~0.0001) that are absent from the Oceanian populations. This component is also potentially visible in America and Central/South Asia. We quantified the similarity in archaic divergence distributions between pairs of regions using the statistic *D* from the Kolmogorov-Smirnov test, after downsampling Neanderthal and Denisovan segments in each geographical region to match the minimum number (1980), to avoid any effects arising from unequal total amounts of segments. These tests confirm that Oceanian populations are less similar to East Asian and American populations in the divergence of Denisovan segments to the Altai Denisovan genome..:

$D_{America\_EastAsia} = 0.069$
$D_{America\_Oceania} = 0.152$
$D_{EastAsia\_Oceania} = 0.180$

… than in the divergence of Neanderthal segments to the Vindija Neanderthal genome:

$D_{America\_EastAsia} = 0.032$
$D_{America\_Oceania} = 0.027$
$D_{EastAsia\_Oceania} = 0.032$

Thus in pairwise comparisons between East Asians, Americans and Oceanians, distributions of Neanderthal segment divergence from the Vindija Neanderthal are similar in all three cases, whereas distributions of Denisovan segment divergence from the Altai Denisovan show much more similarity between East Asia and America than between either population and Oceania. This strongly suggests a distinct mix of introgressed ancestry in Oceanian Denisovan segments, but not Neanderthal segments.

These results corroborate the finding from (*39*) of an additional pulse of Denisovan gene flow into East Asia. However, it is also worth noting that similarity in relation to known archaic genomes does not guarantee that the source is the same, since different source populations might show identical relationships to a given archaic individual. Based on evidence from nucleotide diversity and haplotype networks described below, we find it plausible that at least some of the Denisovan ancestry in East Asian (maybe also other Eurasian and American) populations results from an admixture event separate from that in Oceanian populations. The structure in the East Asian patterns of Denisovan divergence (fig. S19B) might be due to partially overlapping distributions of divergence from the two components of admixture, or might possibly reflect an even more complicated history of admixture from several source populations at various locations and times.

Nucleotide diversity within archaic segments

Within one population, the expected number of nucleotide differences per site between two randomly drawn haplotypes is commonly known as nucleotide diversity (π). When comparing two populations, the same expected value between sequences randomly drawn from two populations (excluding all comparisons within the same population) is commonly known as absolute divergence (D_{XY}, also referred to as π_{XY}, π_B or d_{XY} in the literature) (*74*). Based on the "strict" set of result, three sets of D_{XY} between all pairs of populations were obtained: values calculated from only the Neanderthal segments in the genomes, from only the Denisovan segments in the genomes, and from only the unadmixed (also referred to as "modern" hereafter) segments of the genomes.

In this context, a "haplotype" refers to the collection of all Neanderthal (or Denisovan or modern) segments found in the same haploid genome. Since introgressed segments typically

span different genomic regions in different individuals, it is only meaningful to compare nucleotide differences in the overlapping regions of two haplotypes (fig. S20). To limit computational costs when calculating D_{XY} in Neanderthal segments, if the sample size of a population exceeds 10 individuals, only 20 haplotypes are randomly drawn for pairwise comparison with a maximum of 20 haplotypes from the other population (we found repeated draws produced very similar results); this cap was not applied to Denisovan segments, which are more scarce outside Oceania.

Neanderthal vs. unadmixed regions

The absolute divergence in Neanderthal regions ($D_{XY\text{-}N}$) and in unadmixed regions ($D_{XY\text{-}M}$) are colour-coded onto the upper-right and lower-left triangles, respectively, of a matrix in a heatmap (Fig. 6A). All $D_{XY\text{-}M}$ and $D_{XY\text{-}N}$ values were normalised such that variation within each is displayed using the same colour scale. Neighbour-joining trees were built using D_{XY} as distances (fig. S21). If there was a separate pulse of archaic admixture into some population(s), the recipient populations is expected to appear as an outgroup in the tree reconstructed from archaic segments. If this pulse was much more recent than the shared pulse elsewhere, the length of the tip branches would also appear shorter. The populations in Fig. 6A are ordered according to the neighbour-joining tree built with $D_{XY\text{-}M}$ values using San as outgroup (fig. S21A). The heatmap is generally symmetrical: the pattern in Neanderthal segments largely mirrors that in unadmixed segments, forming major clusters separating Oceanian, American, East Asian and European-Central/South Asian populations. This pattern is also reflected in the unrooted neighbour-joining tree built from $D_{XY\text{-}N}$ (fig. S21B).

The symmetry is broken on finer scales. $D_{XY\text{-}N}$ between the North African Mozabite and European/Middle Eastern populations appears lower than that between some Central/South Asian populations and the latter; in fact, it is almost as low as comparisons within Europe. But in terms of $D_{XY\text{-}M}$ all Central/South Asian populations are closer to European/Middle Eastern ones than to Mozabite. Most likely this is because the sub-Saharan African ancestry in Mozabite increases $D_{XY\text{-}M}$ to Europe/Middle East, but the Neanderthal ancestry in Mozabite remains what was received from the same source as Europe/Middle East. The values in the cluster in the lower right corner, including populations from Europe, the Middle East and Central/South Asia (excluding three with high genetic affinity to East Asia), were normalized again excluding other populations (fig. S21D). Now a European cluster can be distinguished, yet the relationships involving the Middle East and Central & South Asia are not well-defined. A history involving complex admixtures among the ancestors of these various groups, especially if including sources with no or very low levels of Neanderthal ancestry (including sub-Saharan Africans and the proposed basal Eurasian lineage (*38, 75*)), could have contributed to these patterns.

Denisovan vs. unadmixed regions

Fig. 6B shows the normalised absolute divergence in Denisovan segments ($D_{XY\text{-}D}$) and unadmixed regions ($D_{XY\text{-}M}$). The three Oceanian populations form a sister clade relative to all East Asian and American populations in unadmixed segments but exhibit very high divergence to all other populations in Denisovan segments. Their $D_{XY\text{-}D}$ to other non-African populations is largely homogenous, with only a faint affinity to East Asian populations. The Bougainville population displays lower $D_{XY\text{-}D}$ to East Asian populations than the other Oceanian populations, consistent with some Southeast Asian ancestry in Bougainville.

In Cambodians, the Denisovan segments show increased divergence to other East Asian populations, but decreased divergence to Oceanian populations in comparison to the

unadmixed segments. In the tree built from D_{XY-D} (fig. S21C), the Cambodian branch in also slightly longer in relation to other East Asian and American populations. Similarly to the much more prominent behavior of Oceanians, this evidence may suggest the presence of another component of Denisovan ancestry in Cambodians, with tentative connection to that in Oceania. One possibility is that this behavior might be driven by some fraction of South Asian related ancestry in Cambodians, which is likely not found in the other East Asian populations in the panel.

Correspondingly in the unrooted D_{XY-D} tree (fig. S21C), the branch leading to the Oceanian populations is so long that rooting by midpoint places the root there. Cambodians then lie basal to all other Eurasian and American groups. The relatively low levels of Denisovan ancestry in most parts of the world likely adds noise to these inter-population comparisons.

Neanderthal vs. Denisovan regions
Fig. S22 directly compares intra- and inter-population divergence in Neanderthal and Denisovan segments, again highlighting the distinct Denisovan ancestry in Oceania: in all comparisons excluding Oceanians, we again observe a strong correlation between D_{XY-D} and D_{XY-N}, with the former typically smaller than the latter; in contrast, almost all comparisons including Oceanian populations show higher D_{XY-D} than D_{XY-N}, and deviate from the otherwise largely linear relationship.

Archaic haplotype networks
We attempted to reconstruct the relationships between all the archaic segments identified in different individuals in a given region of the genome. Assuming two archaic sequences descend from the same ancestral sequence at the time of admixture and a mutation rate of 1.25×10^{-8} per site per generation, after 2,000 generations one would only expect to observe one difference per 20 kB. Long regions covered by as many archaic haplotypes as possible are therefore necessary to achieve a reasonable resolution. We searched for candidate regions in the genome by the following procedure:

- A multiple intersection of all Neanderthal/Denisovan segments (the "strict" set) on each chromosome was performed using multiIntersectBed from BEDTools (*61*) to obtain the total number of archaic haplotypes in a given genomic interval;
- The list of intervals was scanned to add new intervals by merging adjacent ones if a subset of individuals are present in both;
- A score is also assigned to each interval based on the length (L) and the number of samples (n):

$$s = n^w \cdot L$$

where w can be tuned to adjust the weight of including more haplotypes over extending the genomic region; here we fixed it at 1, hence the score equals the total length of archaic sequences in the interval;

- Intervals shorter than 50 kB or with fewer than 5 occurrences were removed;
- Non-overlapping intervals with the highest scores were collected following a greedy algorithm: within a minimal candidate set of intervals that do not overlap with any other, the interval with the highest score is selected and moved to the selected set, and any other intervals that overlap with it removed from the candidate set; next to be selected is the interval with the highest score among the remaining ones in the candidate set, and so on; the process repeats until no interval remains in the candidate set, and the algorithm moves on to the next set of intervals.

We constructed phylogenetic trees and haplotype networks using aligned archaic segments. Fewer than 4% of the genomic regions that we considered are longer than 0.2 Mb, so recombination between archaic segments is unlikely. When very few differences exist between haplotypes, or if recombination does occur between close haplotypes, haplotype networks have the advantage of allowing alternative links other than imposing a bifurcating tree with high uncertainty. The median joining network algorithm implemented in the pegas R package (*76*) was used to construct haplotype networks. Preliminary analysis also showed that the time to the most recent common ancestor (tMRCA) estimated from haplotype network analysis and maximum likelihood trees were highly correlated, although alternative links in the network tend to reduce tMRCA, especially when the sample size is large.

In each interval, polymorphic sites in all archaic haplotypes were retrieved to form a sequence alignment for haplotype network analysis. If a singleton allele among the archaic haplotypes was also present at a frequency above 1% in sub-Saharan African populations, we considered it was likely the result of phasing error and ignored it.

A total of 4,153 Neanderthal haplotype networks and 727 Denisovan ones were constructed using identified segments in all non-African genomes. A few examples are shown in fig. S23 and fig. S24. The sizes of the networks range from 3 to over 200 nodes. In some networks, haplotypes from the same geographical region cluster together (e.g. fig. S23A and fig. S24A), but in others identical haplotypes can be found across distant geographical regions (e.g. fig. S23B and fig. S24B). The Neanderthal haplotypes do not clearly tend to fall into separate clusters, as would be expected if there were multiple admixture events, whilst Denisovan haplotypes in Oceania are often separated from those in other geographical regions.

Age of archaic haplotype networks

To estimate the number of founding lineages contributing to extant haplotypes, we calculated the age (ρ) of each network (equivalent to tMRCA, or the height of a phylogenetic tree). The haplotype closest to the Vindija Neanderthal genome was assumed to be the root node. Following (*77*), ρ is measured as the average shortest distance from all nodes to the root:

$$\rho = \frac{1}{n}\sum_{i=1}^{m} n_i l_i$$

and the variance:

$$\sigma^2 = \frac{1}{n^2}\sum_{i=1}^{m} n_i^2 l_i$$

where n is the number of sequences, m the total number of edges, and n_i the number of samples whose shortest route to the root node passes through the ith edge. ρ can then be converted into time in years with the mutation rate and the number of comparable sites in the genomic region.

Fig. S25A shows the distribution of Neanderthal and Denisovan network ages in years. Filtering networks based on the length of the genomic region passing the mask, average B value of the genomic region, the number of missing sites in archaic genomes, the number of sites skipped (singletons also present in Africa), or the total number of polymorphic sites

does not alter the shape of the distribution visibly; nor do these values exhibit distinct distributions between the groups of largest and smallest networks.

The age distribution from the two archaic sources are similar, both reaching the highest density below 50k years. The median of the Neanderthal and Denisovan haplotype networks are 55,613 and 55,070 years, respectively. However, in the tail of the distributions we also observe networks that are hundreds of thousands of years old.

To explore how the number of introgressing lineages qualitatively changes the shape of the network age distribution and the accuracy of the inferred network age, we also constructed median joining networks on simulated haplotypes conditioned on the maximum number of introgressing haplotypes, using the simplified demographic history shown in fig. S25B. The sample sizes in Eurasia (including East Asia, Central/South Asia, the Middle East, and Europe), Oceania, and America populations were specified to match the geographical origin of actual Neanderthal haplotypes observed in each genomic region. The number of introgressing haplotypes was measured as the number of surviving lineages 2,000 generations ago, that is (backward in time) at the end of the bottleneck associated with admixture (which could also be an effect of negative selection). The duration and size of the bottleneck was arbitrarily selected to efficiently sample genealogies with few introgressing haplotypes, as the probability of getting a small number of surviving lineages at the time of introgression is too low without a bottleneck. For each genomic region used to construct a haplotype network, coalescent trees were repeatedly simulated with a matching sample size, until the number of introgressing haplotypes became equal or less than the desired maximum. Genetic sequences of a length matched to the genomic region after filtering were then generated from the tree. In principle, the choices of bottleneck severity and ancestral Neanderthal size will influence the distribution of the coalescent trees retained; yet we found that in practice the effects on the age distribution was minimal.

Three sets of 4,153 genealogies with at most 1, 2 and 4 founding haplotypes respectively were obtained, to match the 4,153 genomic regions used in Neanderthal haplotype network analysis. The same algorithm used on the empirical data was used to build haplotype networks for each simulated alignment dataset and estimate ρ. The distribution of ρ reasonably reflects the true tMRCA (fig. S25C). By allowing alternative links, the age estimated from haplotype networks can potentially underestimate the age of moderately old networks, but not the extremely old ones in the right tail of the distribution. The number of unique haplotypes in simulated and in empirical data align along the identity line with a strong correlation in all three sets of simulations, validating that the demographic model used in the simulations is a reasonable approximation of the true history.

Fig. 6C compares the distribution of network ages estimated from empirical data and three sets of simulated data. The empirical distribution clearly differed from that produced in simulations with only one founding haplotype, yet its overall shape appears shifted towards the left in comparison to the curves produced in simulations with a maximum of two and four haplotypes. The shift could result from negative selection against archaic haplotypes or sub-population structure not implemented in the simulations. The simulations with a maximum of two and four haplotypes generated very similar distributions. As few as two founding haplotypes (i.e one individual) appear sufficient to produce the number of very old haplotype networks observed. A much larger number of Neanderthal individuals could still have been involved, contributing a reduced number of distinct haplotypes depending on the genetic diversity of the Neanderthal population. The combined effects of negative selection, genetic

drift, dilution etc. could then have further reduced the diversity of the introgressed Neanderthal material to a very low level.

Number of founding lineages
The estimates of the ages of the haplotype networks reflect the genome-wide average number of founding archaic haplotypes. Here we estimate the number of founding archaic haplotypes in each genomic region as the number of surviving lineages in the tree at the time of admixture.

For each of the 4,135 genomic regions, a maximum likelihood tree was built from the sequence alignment and rooted by the haplotype closest to a San individual (HGDP00991). We chose to work with the tree structure rather than network here mainly because it more easily enables bootstrap analyses. The height at each node was determined by assigning height 0 to the tip farthest from the root, and positive heights to all other nodes, with the largest value at the root. Then the tree is truncated at the height corresponding to the expected number of differences per base pair in the time since admixture - assumed to be 2,000 generations - and the number lineages remaining connected to the root is counted. To gauge uncertainty, 1,000 non-parametric bootstrap replicates were performed for each genomic region.

Fig. S26 shows the number of founding haplotypes in 100 randomly sampled genomic regions, and Fig. 6B shows the distribution of the mean number from all genomic regions. The estimated number of haplotypes were mostly low: in over 70% of the trees, the value of two standard deviations below the mean is lower than 2. However, there are also cases where more than 10 or even 20 lineages existed at the time of introgression. A total of 17 genomic regions were estimated to have more than 20 founding Neanderthal haplotypes (table S9); their haplotype networks exhibit complicated structures radiating from one or two core haplotypes. It could mean that initially around a dozen (or more, depending on their relatedness) Neanderthal individuals contributed material, but except for very few regions of the genome, most Neanderthal lineages were subsequently lost through genetic drift and negative selection, so that only a handful of them remain among the diversity of Neanderthal segments in present-day modern humans.

Geographical separation
Deep splits in a haplotype network could reflect multiple sources of archaic gene flow. To measure the divergence between geographic regions, we define two regions as separate in a network if none of the nodes containing haplotypes from one region has a closest neighbour containing haplotypes from the other region, and vice versa. Table S10 displays the total number of haplotype networks analyzed in this way (which need to contain at least two haplotypes from each region), and the number of networks showing separation between pairs of geographical regions.

The diversification among modern human populations will have caused some divergence in the archaic segments even if they descended from the same source of admixture. The proportion of completely geographically separated Neanderthal haplotype networks is largely consistent with our understanding of non-African population history. Much fewer Denisovan haplotype networks are available for comparison, yet there is a deep split between Oceania and all other non-African regions: networks are completely separated in almost all cases, most strikingly between Oceania and East Asia, as well as between Oceania and Central/South Asia. Fisher's exact test confirms that the distributions of fully separated

Neanderthal and Denisovan networks are significantly different in these two comparisons (table S11). Another pair that show a near-significant difference is Central/South Asia and East Asia, but in this case they are better connected in Denisovan networks than in Neanderthal networks. Overall, if we take the Neanderthal networks as representative of a single-source scenario, the strong geographical separation between Oceania and other regions in the Denisovan haplotype network provides further evidence for different source populations of Denisovan haplotypes in Oceanian and Eurasian populations.

It is notable that in the only two networks where Denisovan haplotypes in Oceania are not fully separated from those in Central/South Asia, Oceania and East Asia are also connected. Out of the five networks where Oceania and East Asia are not separated, three involve Denisovan haplotypes from Cambodia as a bridge: in two cases the Cambodian haplotype is the sole connection, in another one a haplotype from Lahu is also involved. One possibility is that Cambodian ancestry contains a component that is somewhat intermediate between or similarly related to Oceanian and East Asian ancestries (e.g. South Asian related). But since this connection to Oceania is not observed in unadmixed regions of the genome (Fig. 6B), it could also possibly suggest another independent component of Denisovan ancestry in Cambodians, whose source population is closer to the source in Oceania.

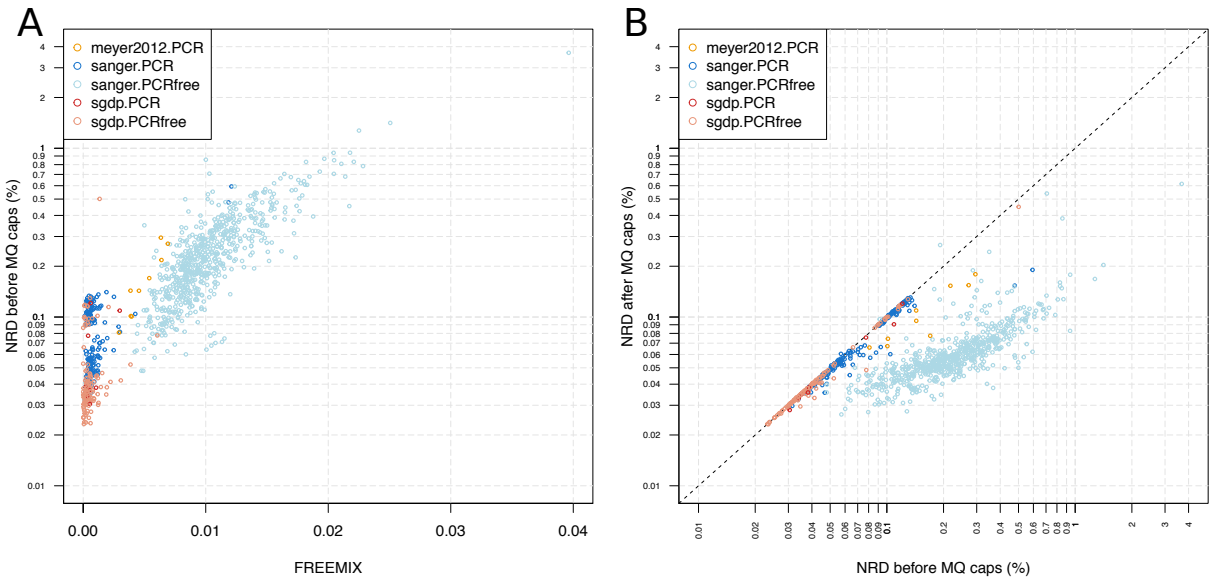


Fig. S1. Sample specific mapping quality caps offsets the effects of index hopping on genotype accuracy. **(A)** The FREEMIX contamination estimate strongly correlates with array genotype discordance (non-reference discordance, NRD) in the multiplexed Sanger PCR-free libraries. **(B)** The application of sample specific mapping quality caps as a function of the FREEMIX estimate decreases the genotype discordance.

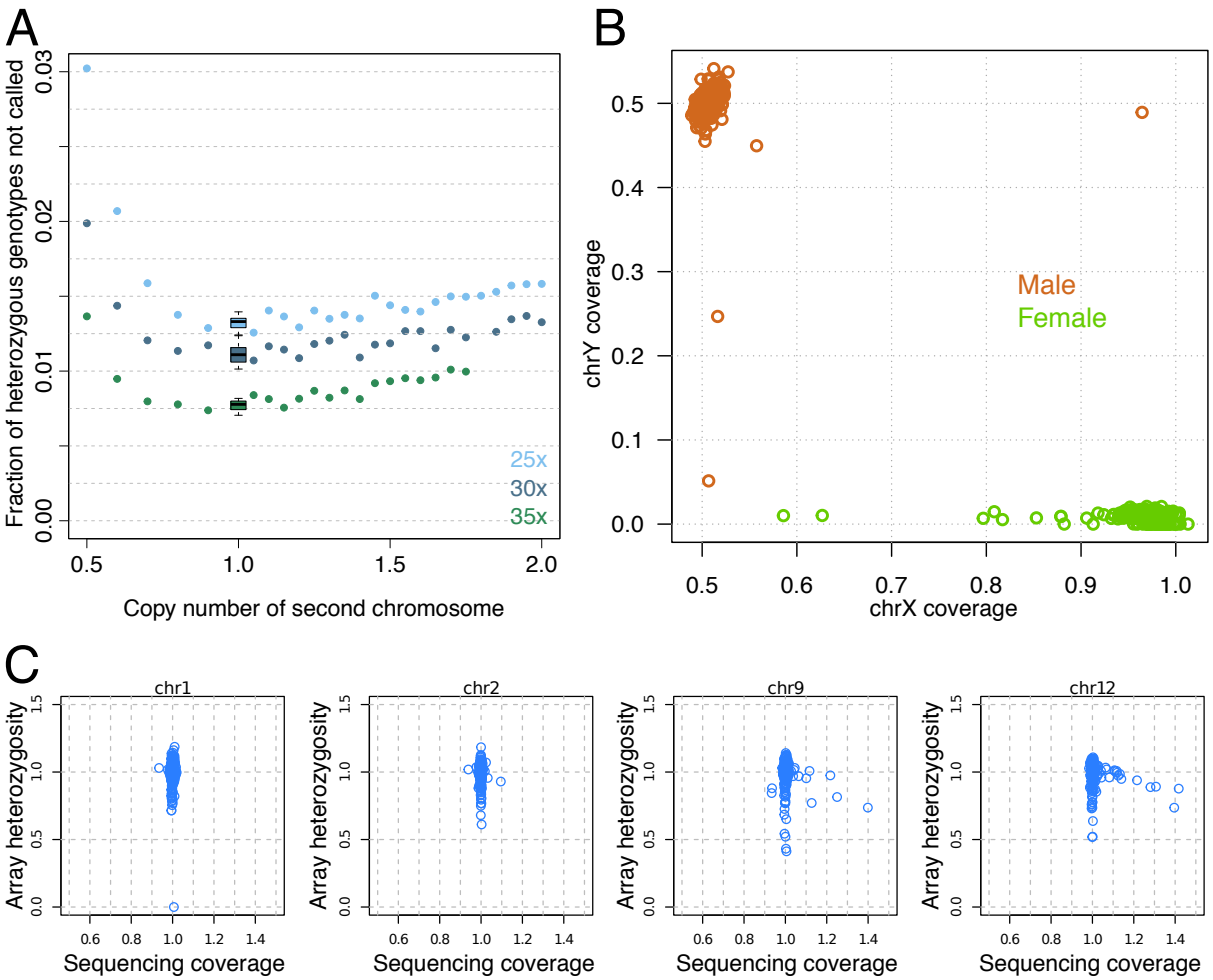


Fig. S2: The effects of cell-line chromosomal copy number alterations. **(A)** The fraction of heterozygous genotypes not correctly called in pseudo-diploid datasets constructed from two male X chromosomes down-sampled to varying degrees corresponding to copy number alterations of varying levels of severity, relative to a case with perfectly balanced copy number. Ten replicate experiments were performed for the balanced copy number case, the results of which are represented by box plots. The experiment was performed at three different levels of overall coverage; 25x, 30x and 35x. **(B)** Sequencing coverage of chromosomes X and Y for all sequenced samples, relative to the genome-wide coverage, coloured by the self-identified gender of the sample donor. **(C)** Sequencing coverage in our generated sequencing data of a chromosome in a sample against the fraction of heterozygous genotypes on that chromosome in that sample (normalized by chromosome and population) in a previously published array genotype dataset (8), displayed for two typical chromosomes as well as chromosomes 9 and 12 which display the largest number of whole-chromosome copy number alterations.

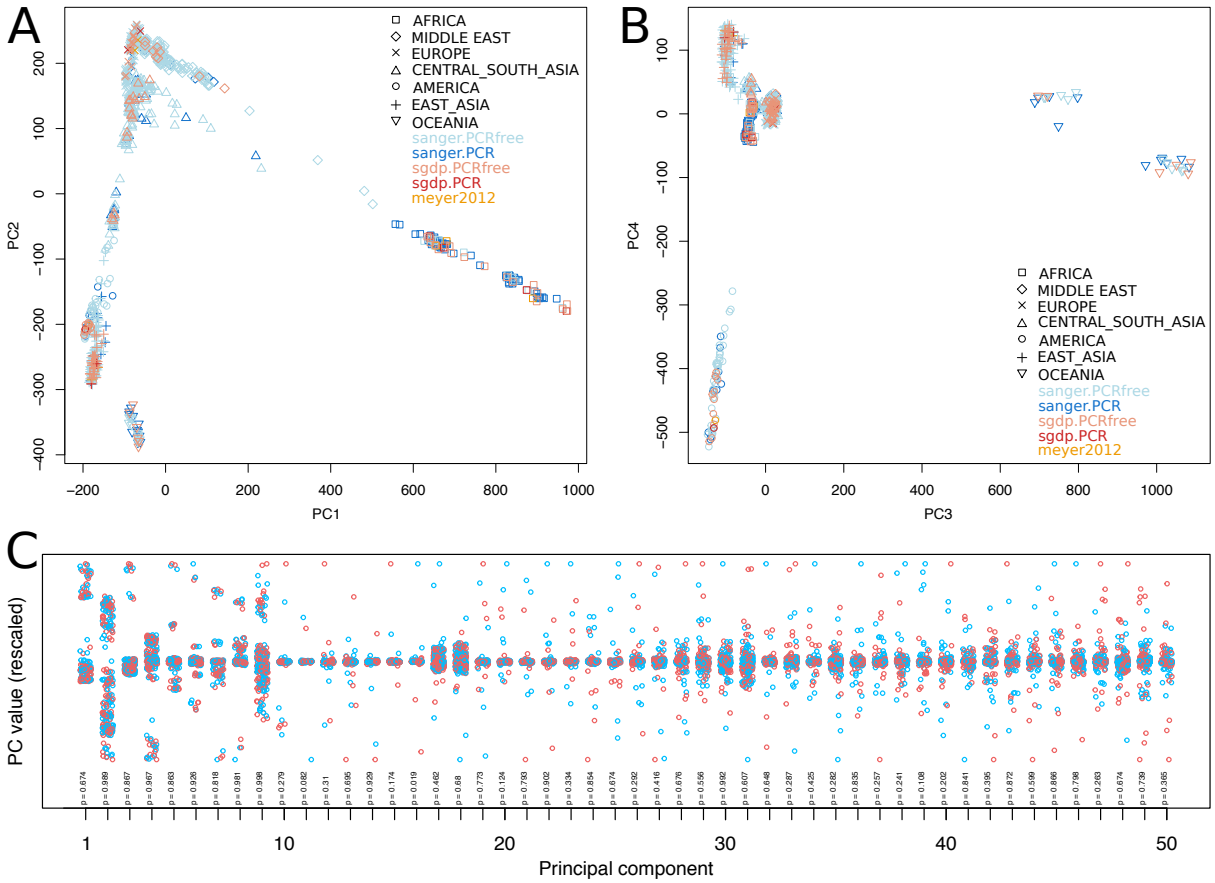


Fig. S3: Testing for genotype batch effects between libraries of different sources. (A) The first two principal components (from analyses using genotypes on sites ascertained as polymorphic in archaic genomes, calculated using AKT: https://github.com/Illumina/akt) with sample colours indicating the combination of library source and library type, and symbol type indicating geographical region. Library source and type does not visibly seem to affect the clustering of samples. **(B)** The third and fourth principal components. **(C)** For each of the first 50 principal components, sample colour indicates whether their source is Sanger (blue) or SGDP (red). A subset of Sanger libraries was sampled at random to match the population composition of the SGDP libraries, and a Wilcoxon rank sum test was performed to test for differences in the placement along each principal component between the SGDP and the Sanger libraries. The resulting p-values are displayed under the data for each component. Only one component, PC16, displays a difference with $p < 0.05$, though this is driven by a few outlier samples and is not statistically significant considering the 50 tests performed.

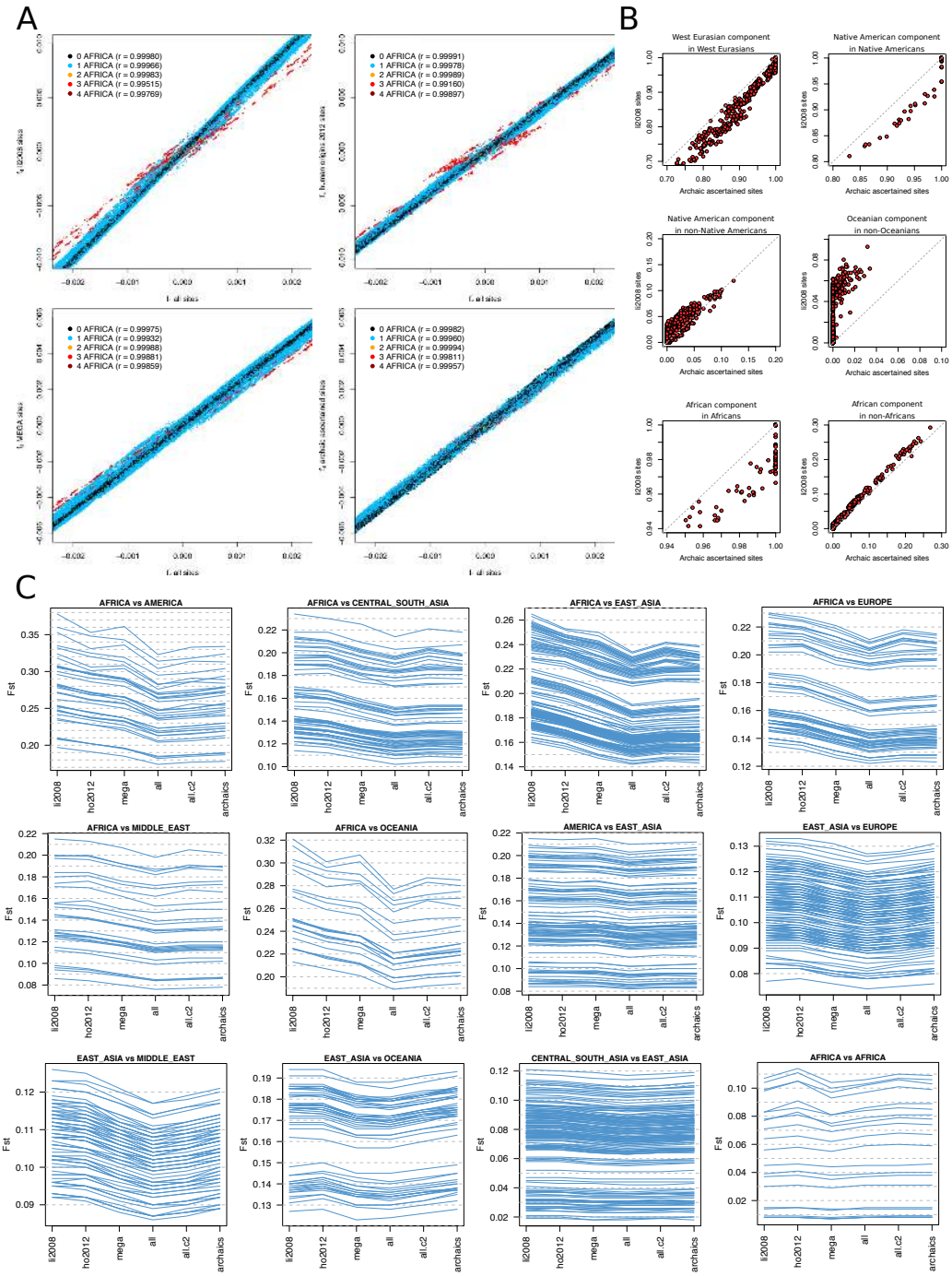


Fig. S4. The effects of variant ascertainment on population genetic analyses. **(A)** Comparisons of all possible f_4-statistics involving the 54 populations calculated using different sets of ascertained sites, against the values calculated using all discovered variants. Points are coloured according to the number of African populations included in the statistic. **(B)** Estimates of individual ancestry components using ADMIXTURE at k=5, at which the five components correspond well to five major regional ancestries. Compared to using the Li 2008 array sites, the sites ascertained in archaic genomes leads to cleaner ancestry estimates with less "leaking" of a given regional ancestry component into individuals outside that region. **(C)** Estimates of F_{ST} for a selection of population pairs, grouped by region, using the Li 2008 array sites ("li2008"), the Human Origin 2012 sites ("ho2012"), the MEGA array sites ("mega"), all variants discovered in the sequencing data ("all"), all variants excluding singletons ("all.c2") and variants ascertained in archaic genomes ("archaics"). Each line connects the F_{ST} values for one pair of populations across the different set of sites.

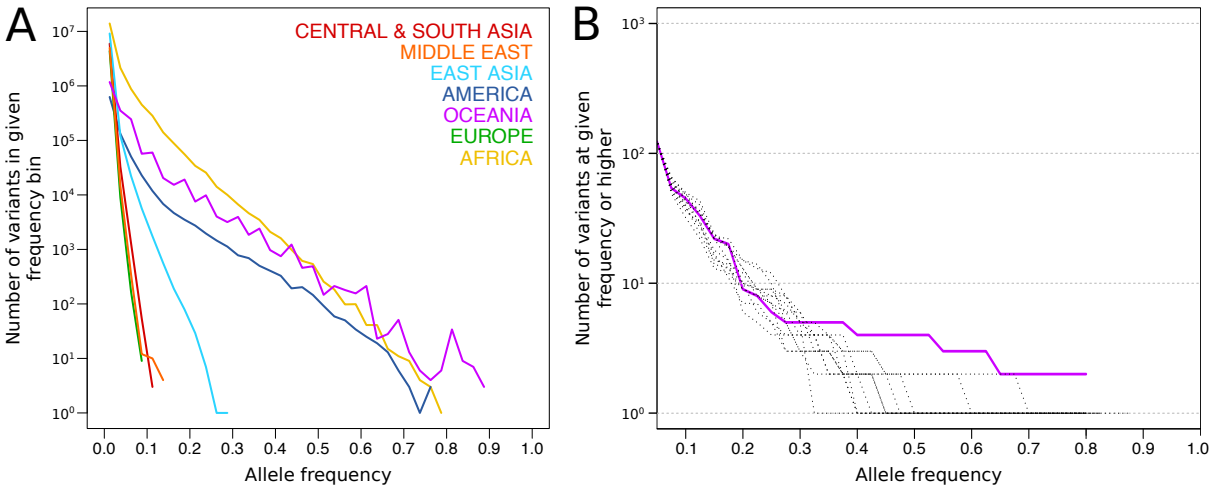


Fig. S5. Further properties of geographically restricted variants. (A) Non-cumulative counts of region-specific SNPs. As Fig. 3A, but non-cumulative, displaying the number of variants in each 2.5% frequency bin. The larger degree of fluctuation between bins for Oceania reflects the lower sample size for this region. **(B)** Assessing the statistical significance of the high number of high-frequency private Oceanian CNVs**.** The solid line displays the number of CNVs private to Oceanian populations that have an allele frequency in those populations equal to or higher than the corresponding value at the horizontal axis. The dashed lines are the results of 1000 random samples of equal numbers of private Oceanian SNPs, displaying only the most extreme 12 samples out of these that contained a variant reaching as high a frequency as the highest-frequency observed CNV.

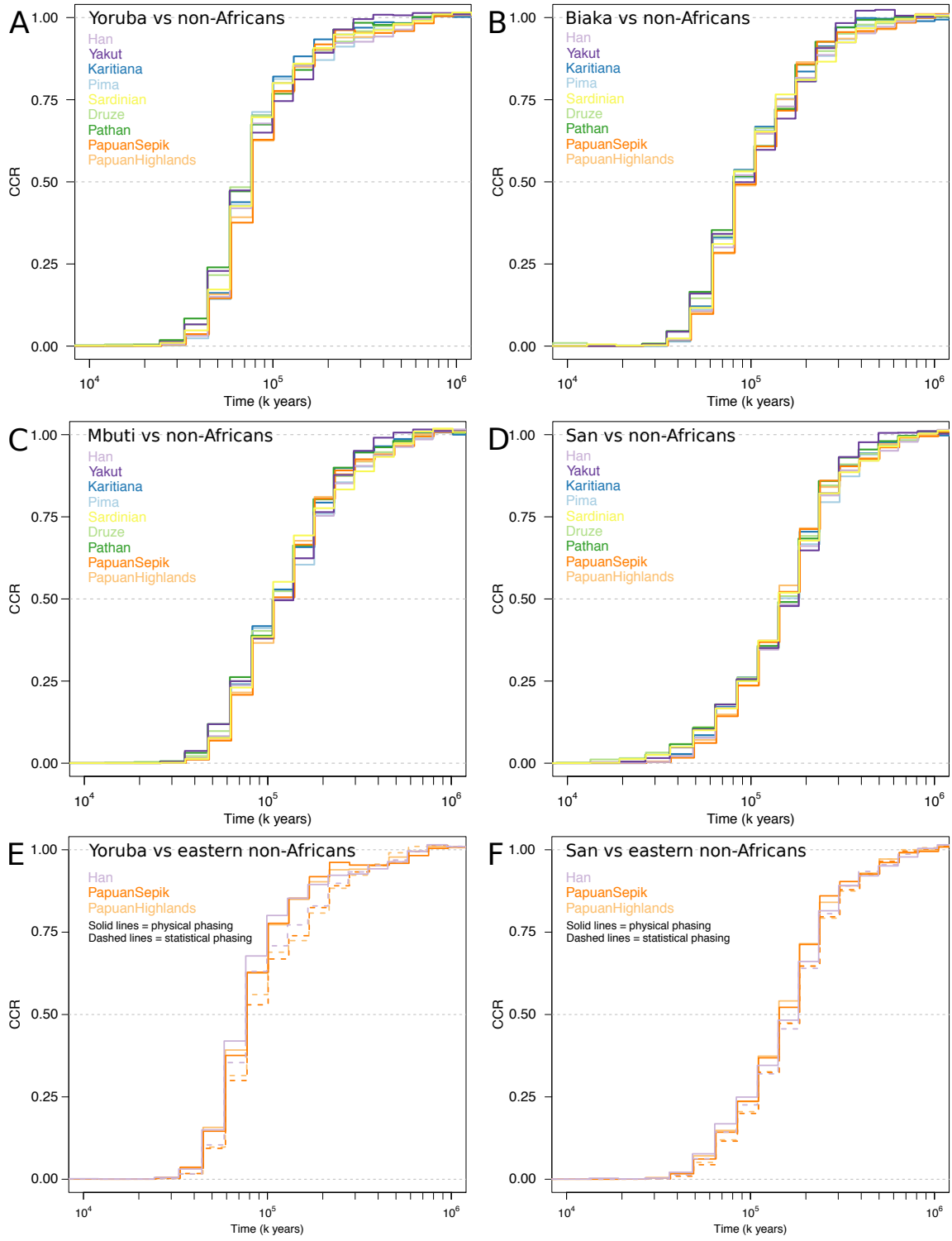


Fig. S6. MSMC2 analyses of divergences between African and non-African populations. Curves were computed using 4 physically phased haplotypes per population. **(A)** Yoruba versus non-African populations. **(B)** Biaka versus non-African populations. **(C)** Mbuti versus versus non-African populations. **(D)** San versus non-African populations. **(E)** Yoruba versus eastern non-African populations, comparing the results obtained with physical phasing to those obtained with statistical phasing. **(F)** San versus eastern non-African populations, comparing the results obtained with physical phasing to those obtained with statistical phasing.

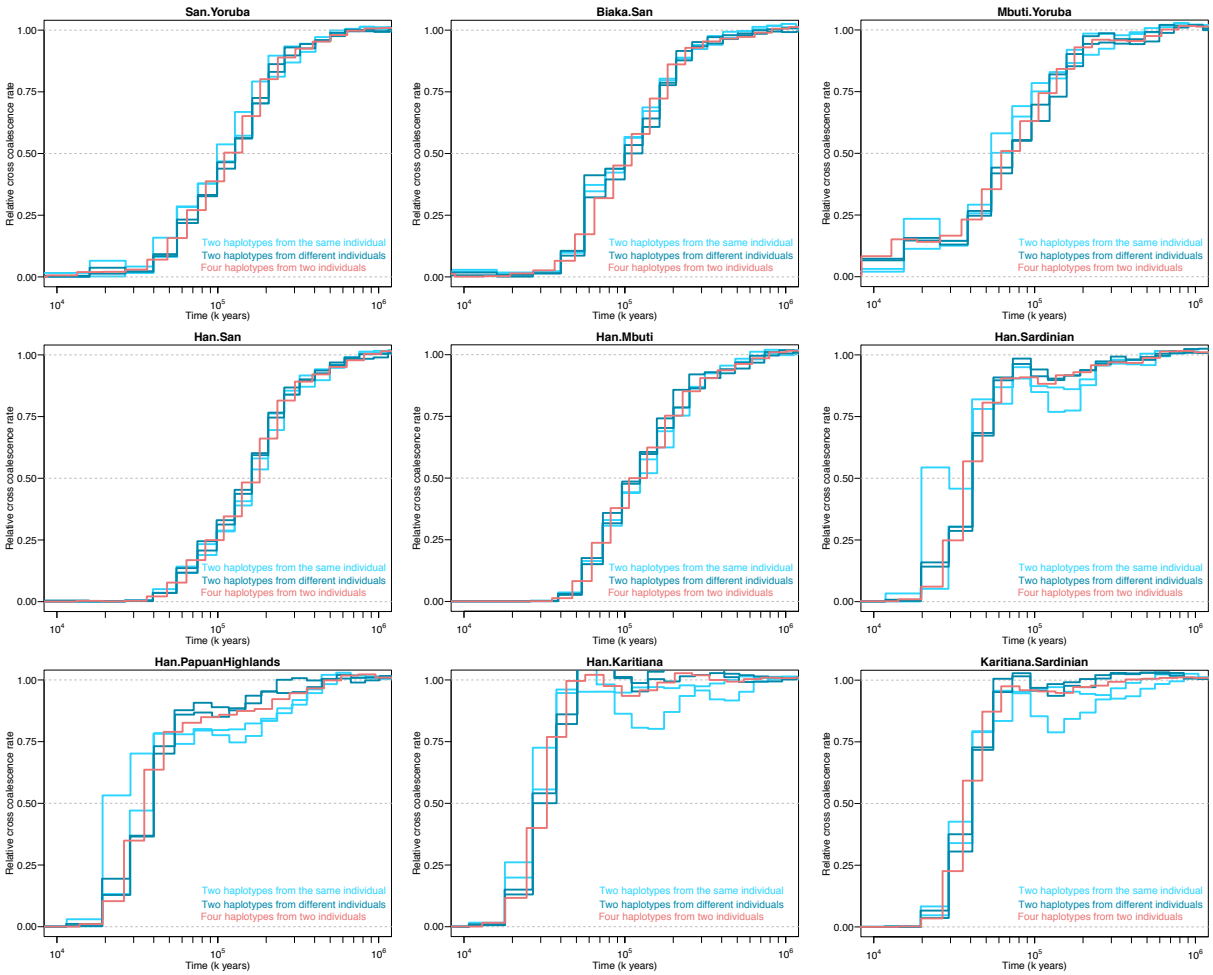


Fig. S7. Testing diploid batch effects in MSMC2. Curves are displayed that use different numbers and configurations of haplotypes from each of the two populations. The "four haplotypes from two individuals" curves are from the standard runs that are also displayed in Fig. 5 (a total of eight haplotypes in the run across the two populations). The "Two haplotypes from the same individual" curves use two haplotypes from one individual from each population (a total of four haplotypes in the run). The "Two haplotypes from different individuals" curves also use two haplotypes from each population (a total of four haplotypes in the run), but from two different individuals. For these latter two runs with four haplotypes, there are two independent replicate curves using different pairs of individuals.

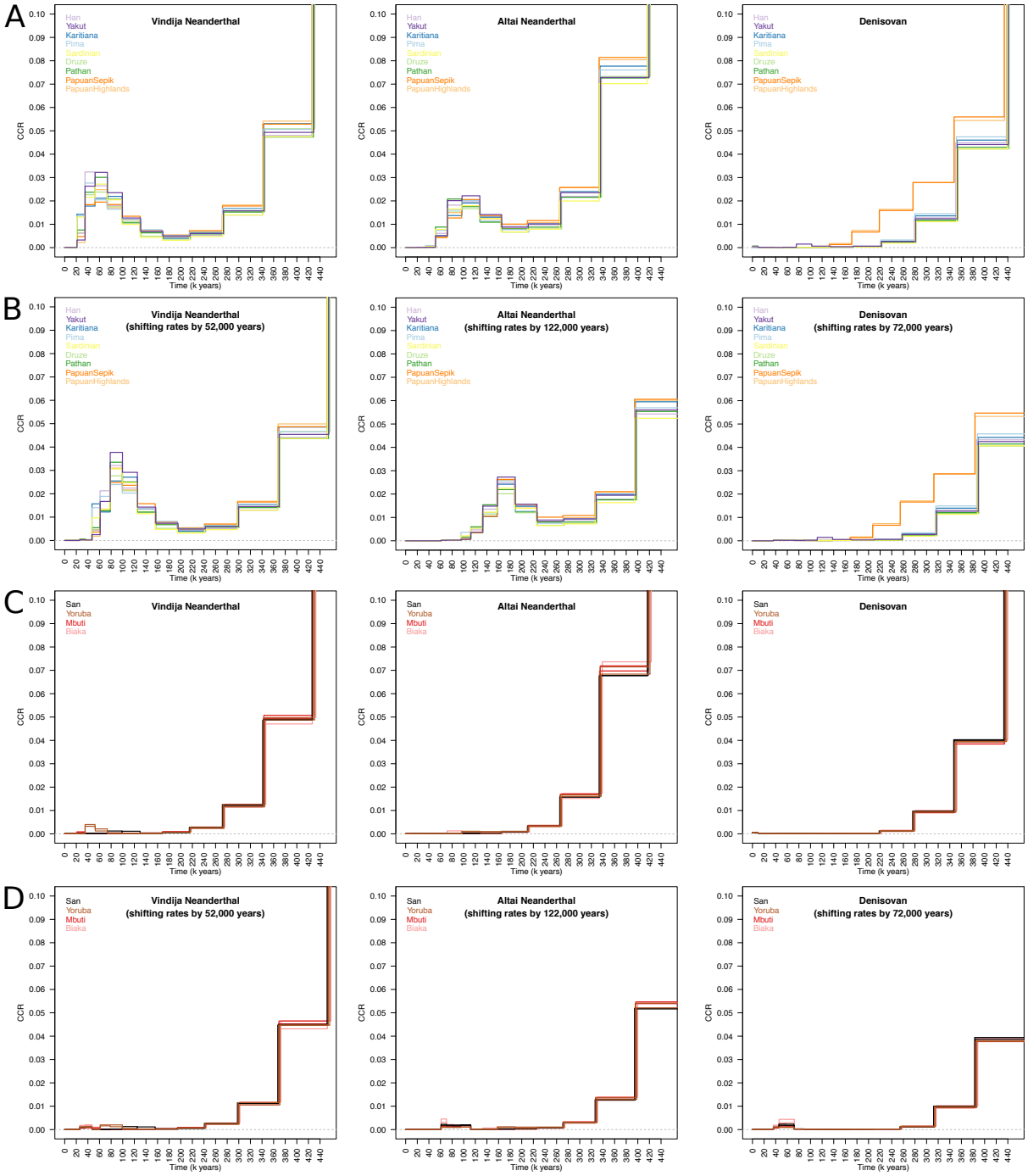


Fig. S8. Signals of archaic gene flow in MSMC2 curves. Results of cross-population MSMC2 runs, zooming in on the signal of Neanderthal genome flow in modern human genomes (note the highly reduced range of the vertical axis). **(A)** Archaic genomes against non-African genomes. **(B)** Archaic genomes against non-African genomes, attempting to adjust for the age of the archaic specimens. **(C)** Archaic genomes against African genomes. For each African population, curves for two different individuals are displayed in the same colour. **(D)** Archaic genomes against African genomes, attempting to adjust for the age of the archaic specimens.

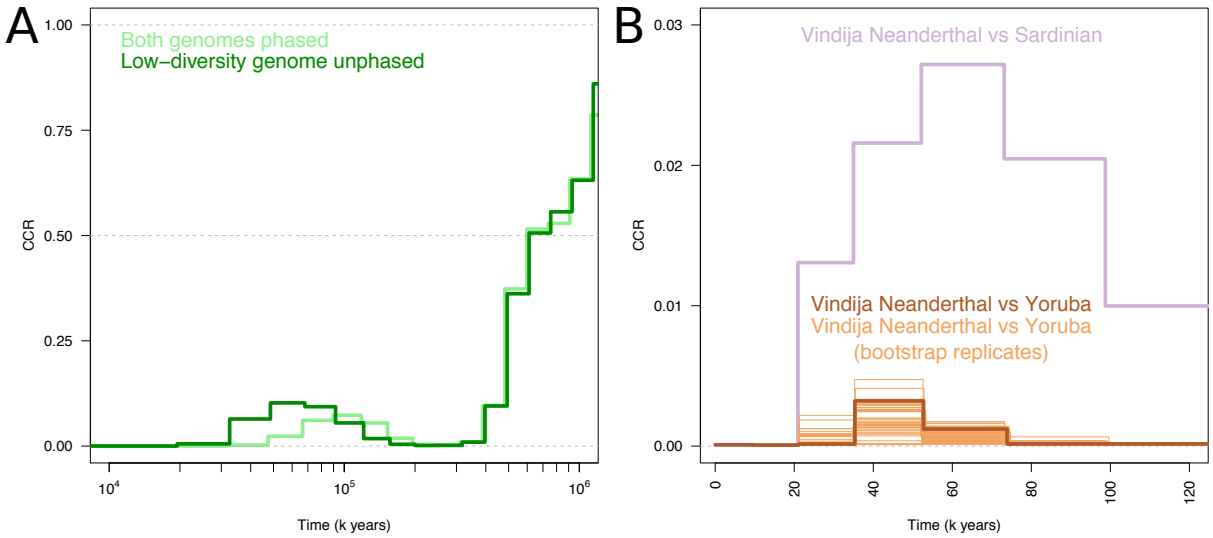


Fig. S9. Evaluating the robustness of running MSMC2 on unphased, archaic genomes. **(A)** Coalescent simulations were performed to approximately mirror the history of divergence and admixture between modern humans and Neanderthals, with a divergence at 500 kya, tenfold reduction in effective size in the second population from 400 kya and admixture from the second into the first population at 50 kya. The inference was rerun after stripping the haplotype phase from the genome sampled from the lower-diversity population, but this does not substantially affect the inferred relative cross-coalescence curve. **(B)** Block bootstrapping across the genome was performed to evaluate the robustness of the Neanderthal gene flow signal in a Yoruban genome. The curves for all 50 bootstrap replicates are plotted together. The curve for a Sardinian genome is included for comparison purposes.

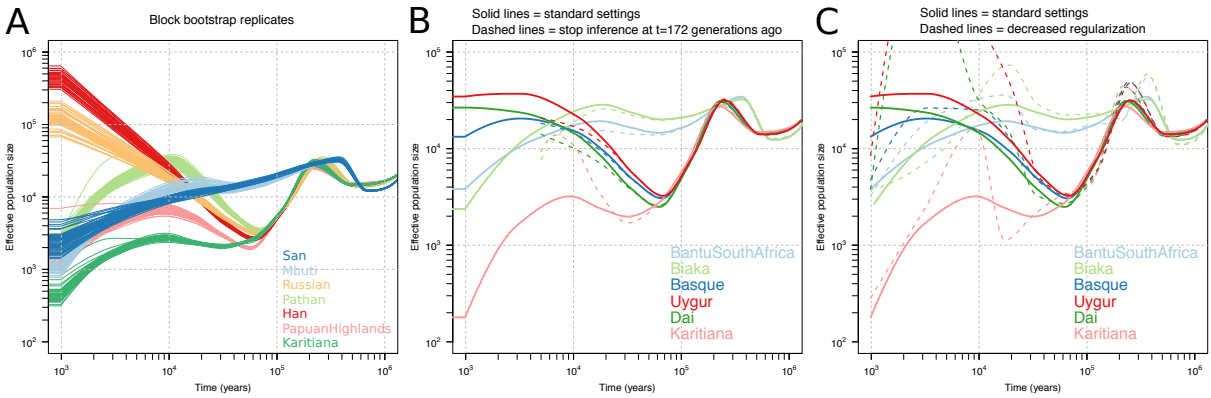


Fig. S10. Technical assessments of SMC++ effective population size inferences. **(A)** For selected populations, each individual curve displayed is an independent SMC++ run performed on a bootstrap dataset constructed from blocks across the genome. 50 replicates per population were analysed. **(B)** The effects of stopping the inference earlier on selected populations. Most populations are not affected much by this, but the Native American (Karitiana) population displays a pronounced period of growth between 10 and 20 kya. **(C)** The effects of decreasing the regularization parameter on selected populations. Many population display clearly artefactual behaviours, but they do not follow the same trajectory as the Native American population.

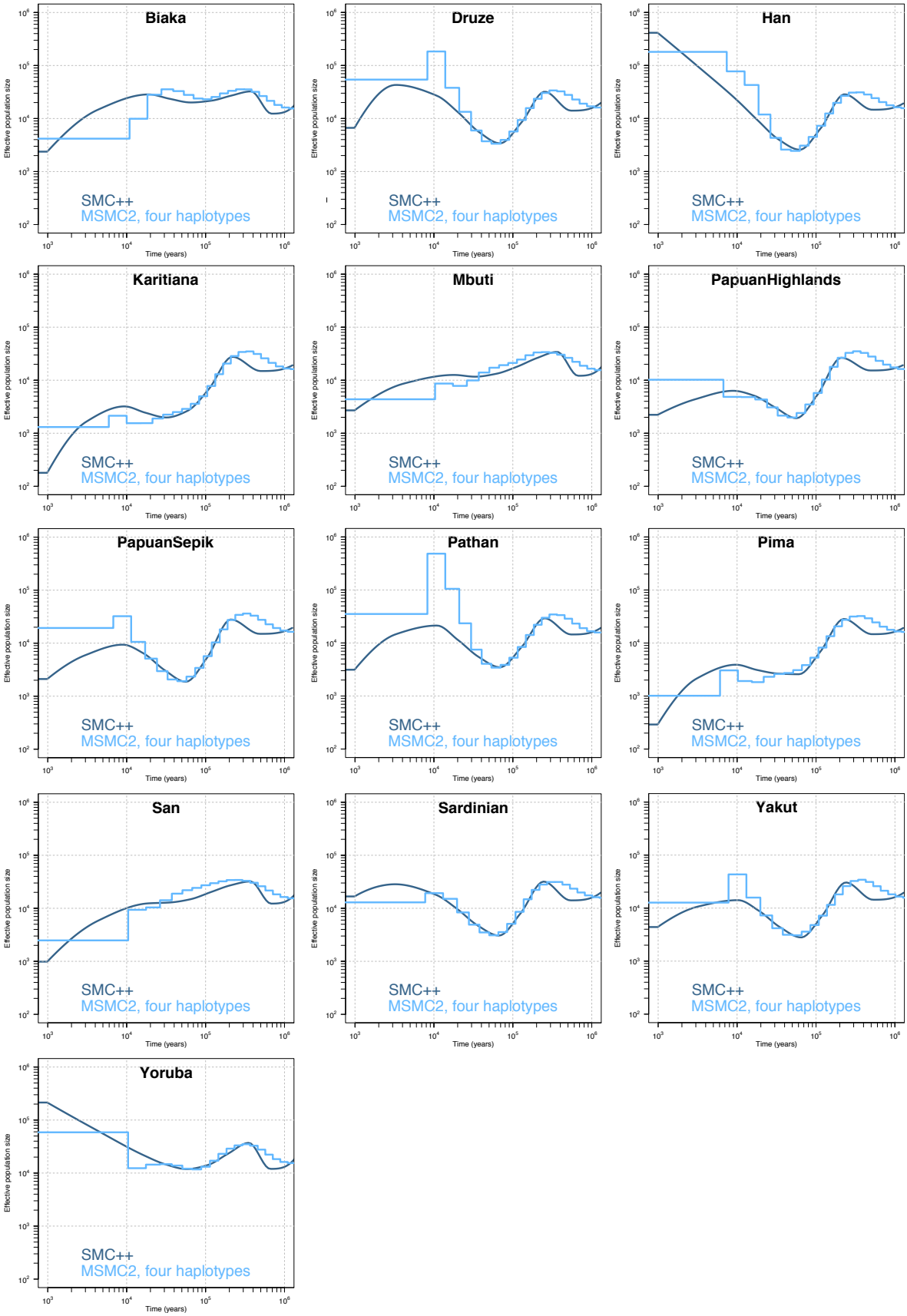


Fig. S11. Comparison of effective population size histories inferred using SMC++ and MSMC2. The SMC++ runs used all individuals from the given population, across six alternative choices of the distinguished individual so as to produce composite likelihoods. The MSMC2 runs used two physically phased genomes (four haplotypes) per population.

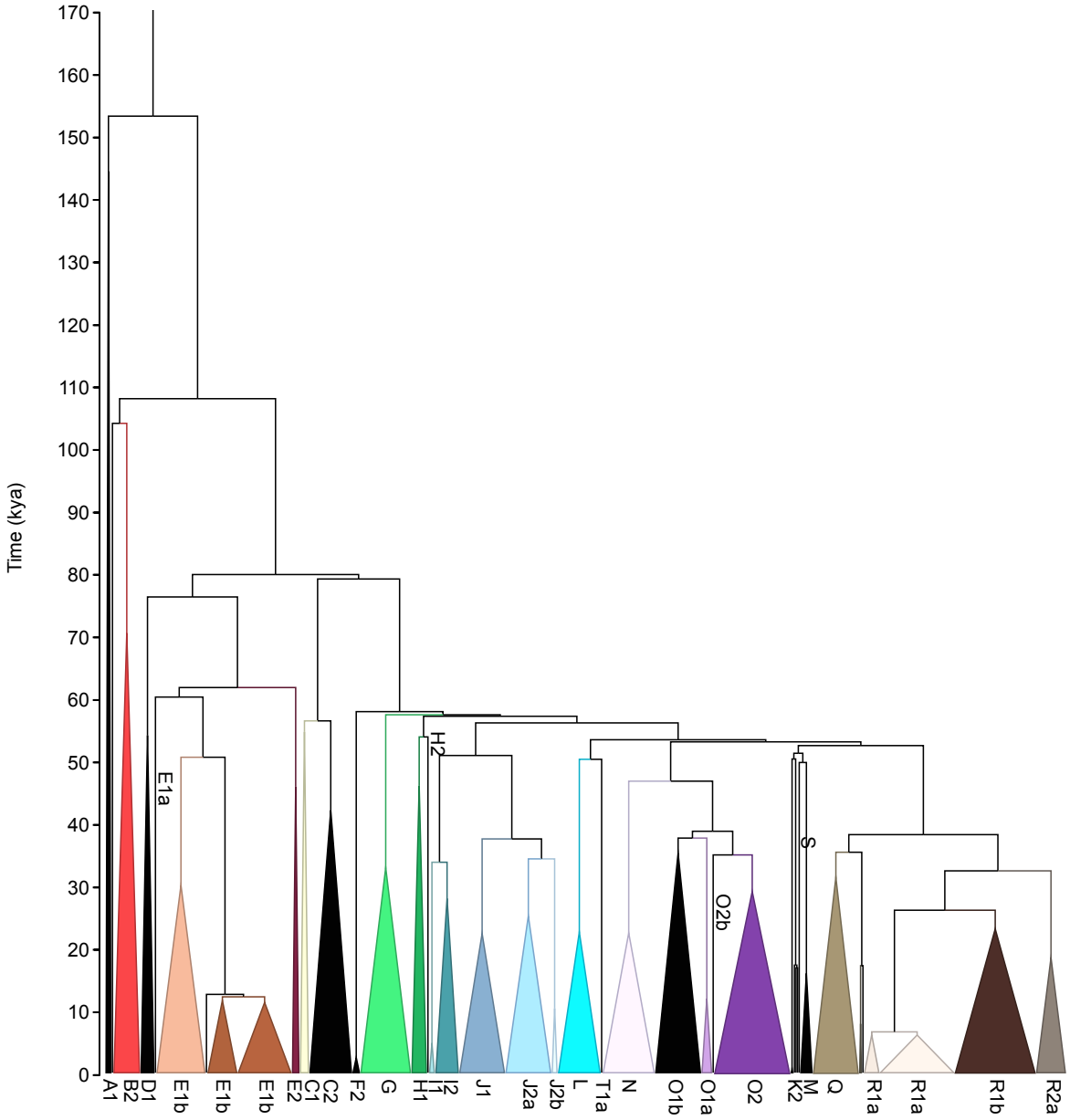


Fig. S12. High-level Y-chromosomal phylogeny. Branch lengths are proportional to the estimated times between splits. Coloured triangles represent collapsed major clades, with the width of the triangles proportional to the number of samples in each clade.

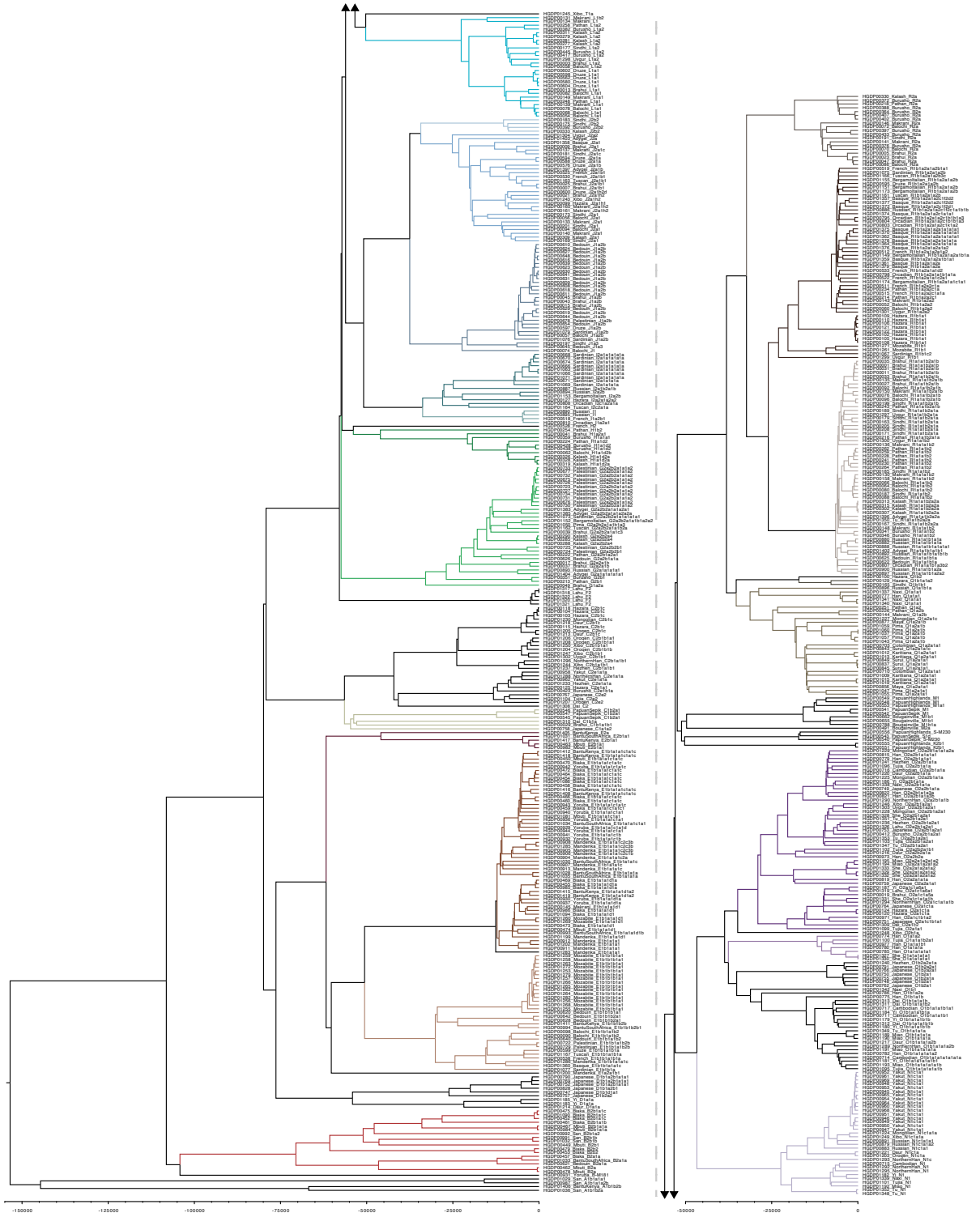


Fig. S13. Detailed Y-chromosomal phylogeny. Branch lengths are proportional to the estimated times between splits. The phylogeny is divided into two parts for display purposes: two arrows connect branches in the lower half of the phylogeny, on the left side of the dashed line, to their continuations in the upper half of the phylogeny, on the right side of the dashed line. Branch colours correspond to those in fig. S12. The haplogroup call for each individual is displayed alongside the sample name and population.

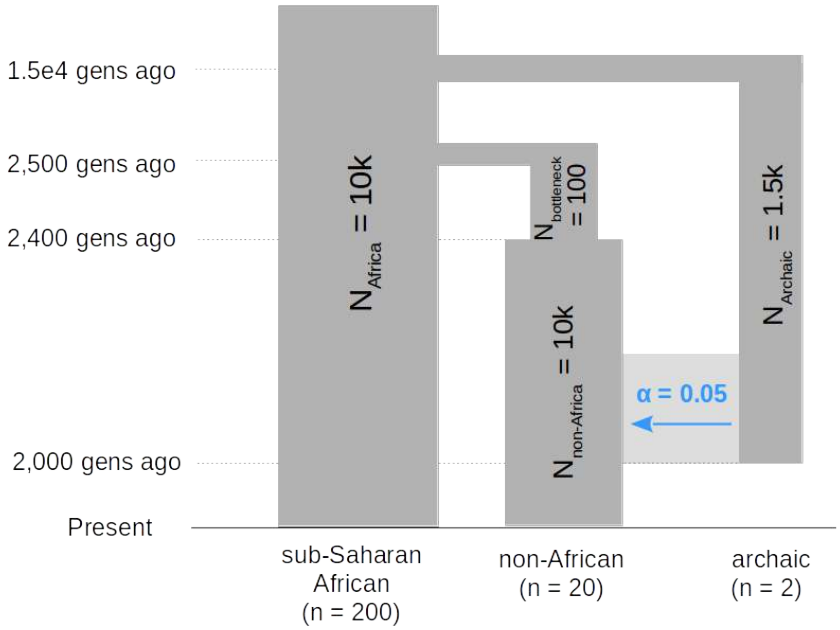


Fig. S14. Demographic model underlying simulations with only one source of archaic gene flow.

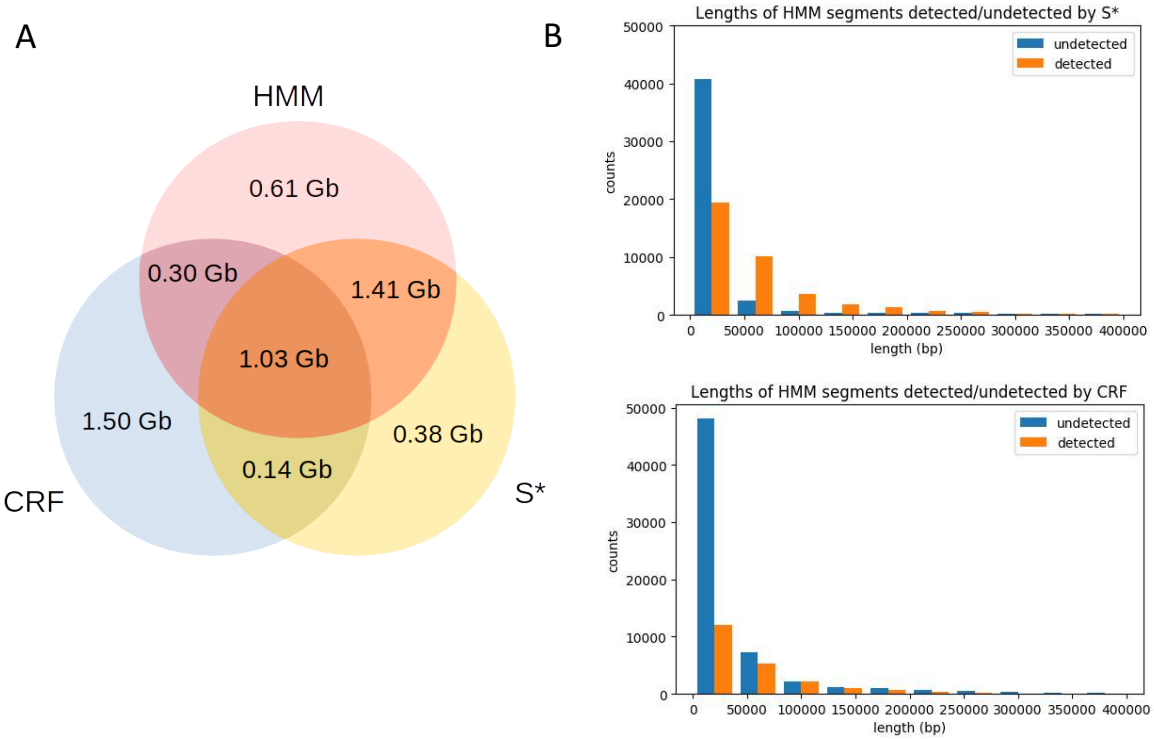


Fig. S15. Comparison of HMM for archaic haplotype detection to other methods. (A) Amount and overlaps of Neanderthal segments identified on chromosome 1 of 544 individuals from the 1000 Genomes Project. **(B)** Lengths of Neanderthal segments detected by the HMM and detected/undetected by other methods.

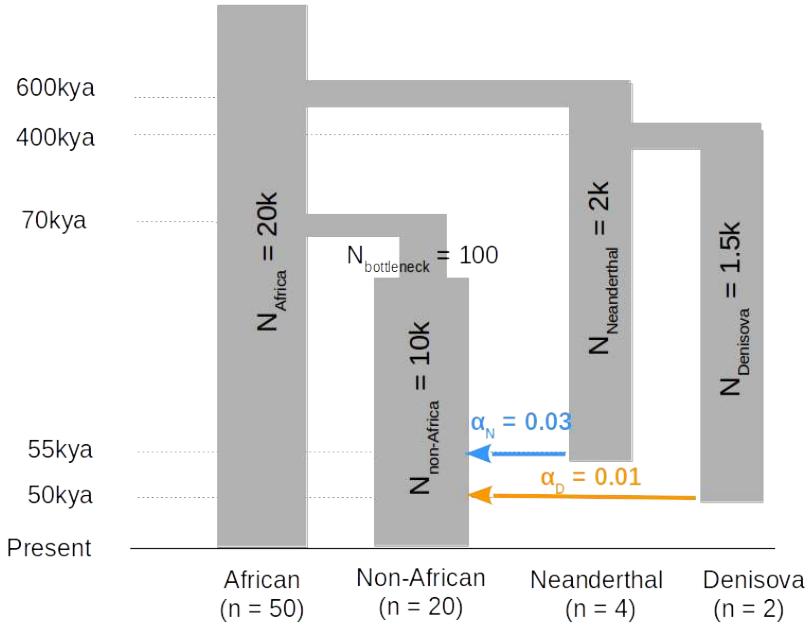


Fig. S16. Demographic model underlying simulations with both Neanderthal and Denisovan gene flow.

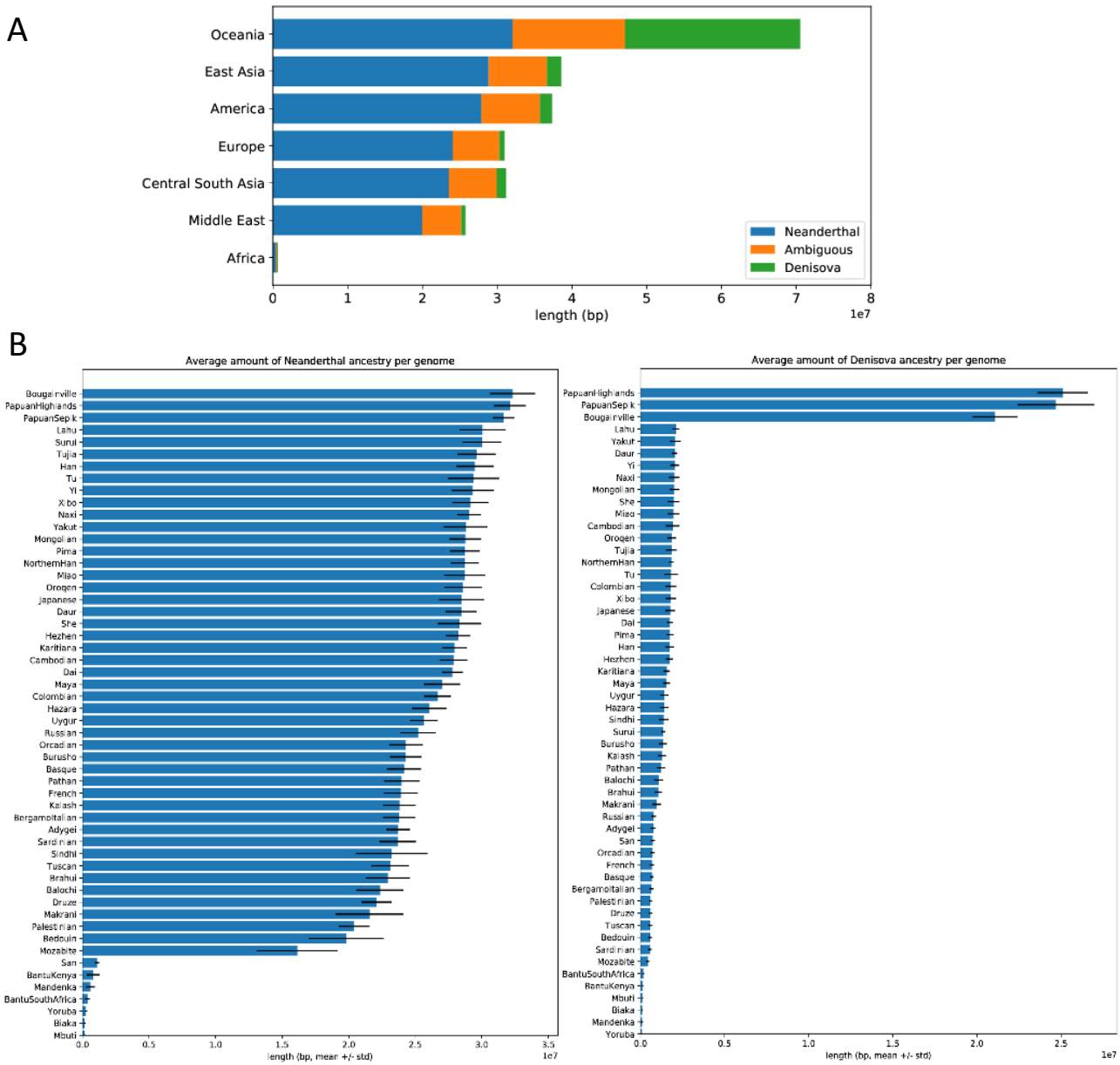
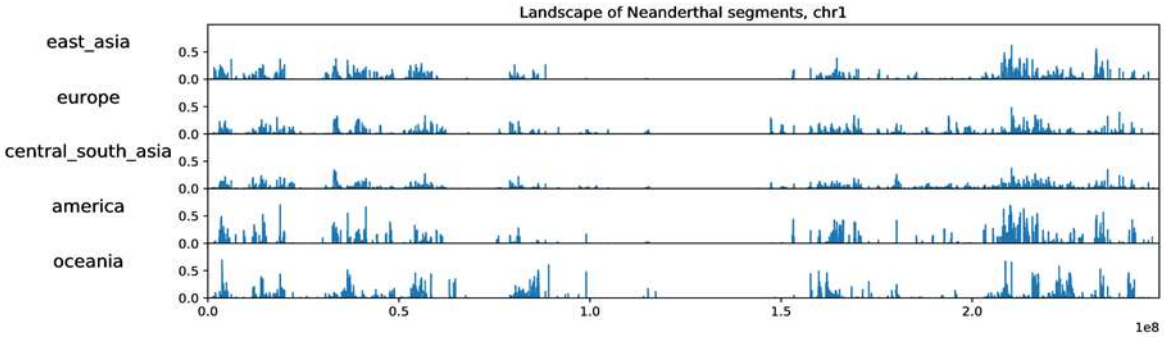


Fig. S17. Amounts of archaic segments identified. (A) Average amount of archaic segments identified per genome by geographical region. **(B)** Average amount of Neanderthal and Denisovan segments identified per genome, by population.

A Distribution of Neanderthal segments along chromosome



B Distribution of Denisova segments along chromosome 1

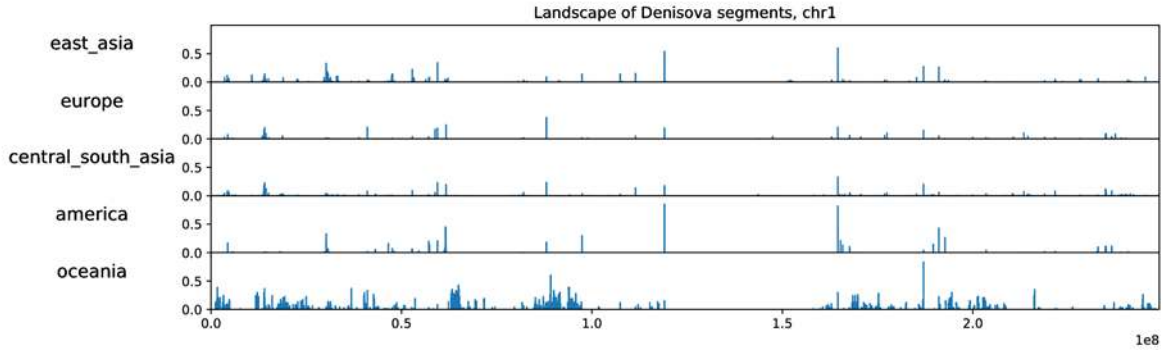


Fig. S18. Distribution of archaic segments (”strict” criteria) along chromosome 1 by geographical region.

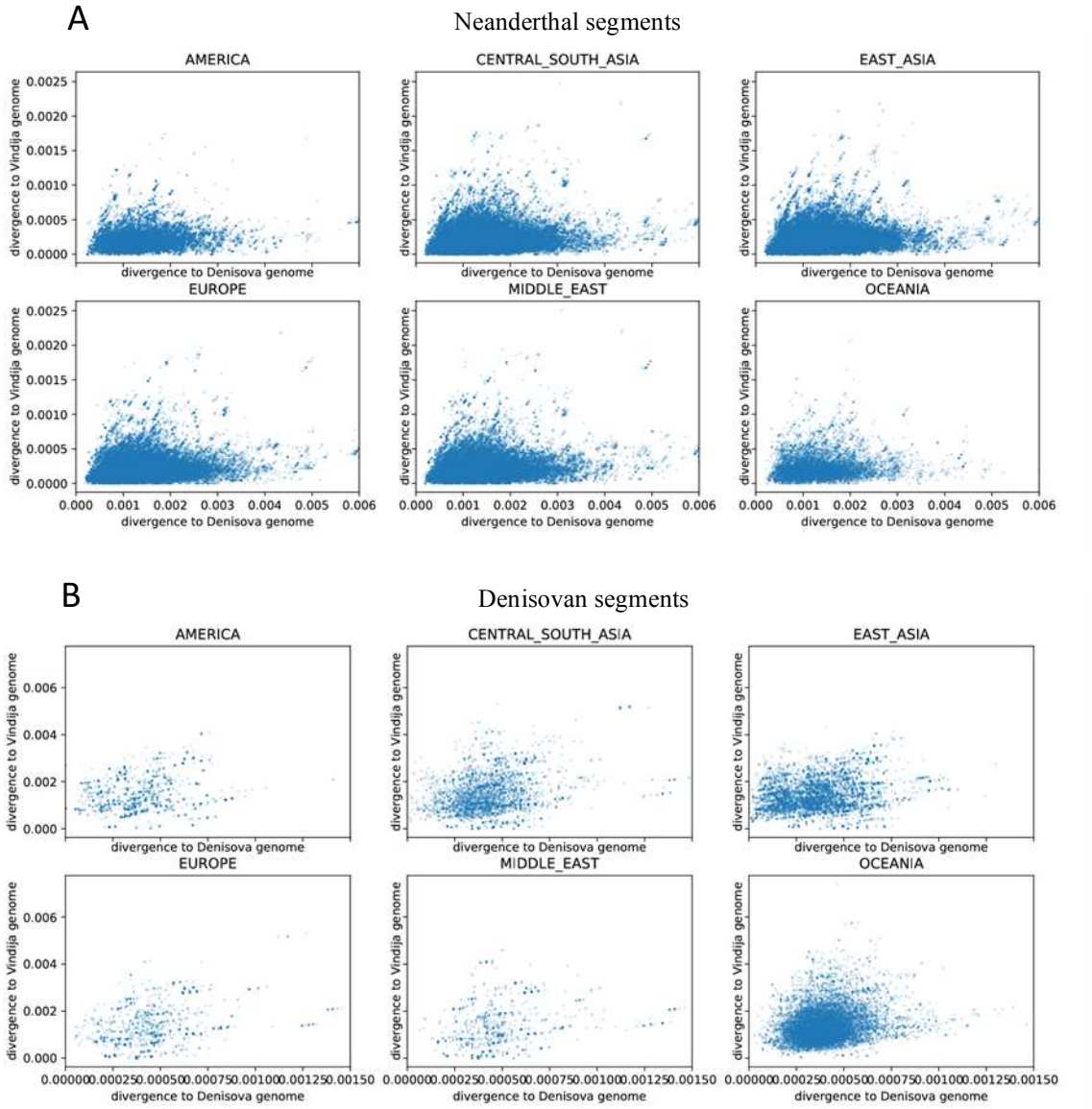


Fig. S19. Divergence of each identified archaic segment to the Vindija Neanderthal and Altai Denisovan genomes across geographical regions. (A) Identified Neanderthal segments. **(B)** Identified Denisovan segments.

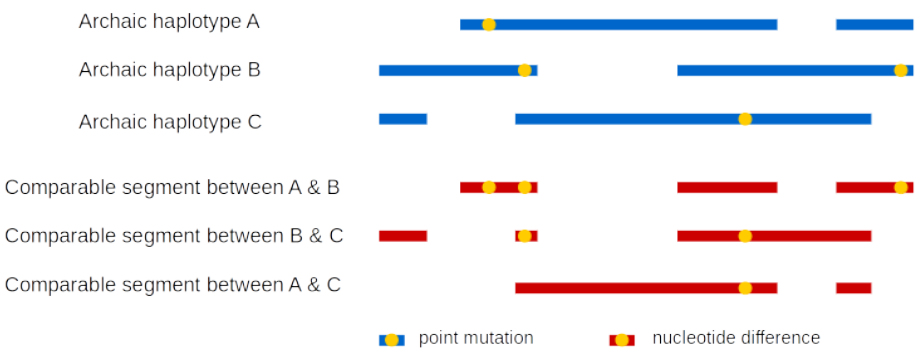


Fig. S20. Schematic showing comparable regions between three archaic haplotypes and nucleotide differences.

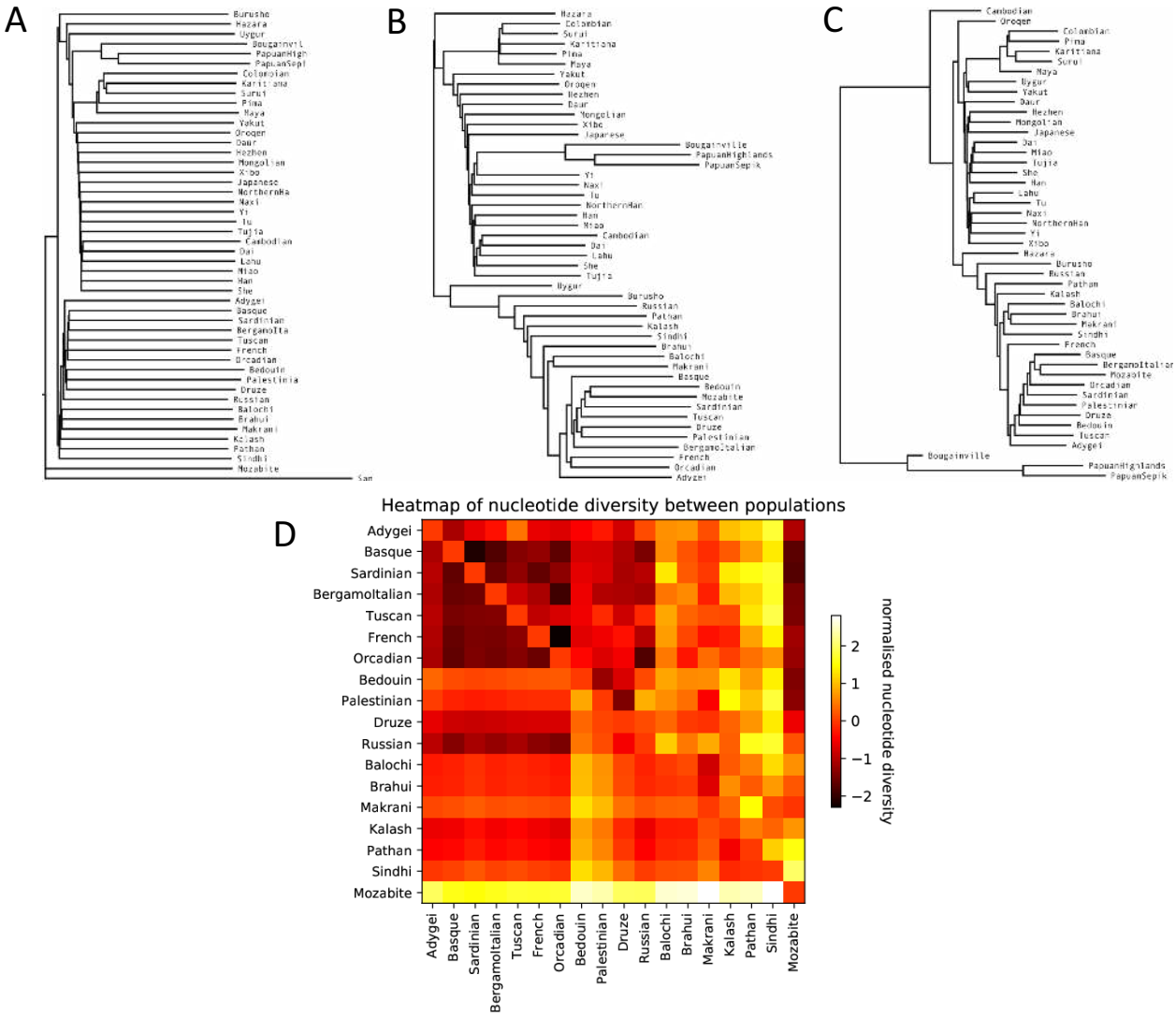


Fig. S21. Population relationships in different classes of genome segments. (A) Neighbour-joining tree built from D_{XY} measured in unadmixed segments of the genome, rooted by San as outgroup. **(B)** Neighbour-joining tree built from D_{XY} measured in Neanderthal segments in the genome, rooted by midpoint. **(C)** Neighbour-joining tree built from D_{XY} measured in Denisovan segments in the genome, rooted by midpoint. **(D)** Heatmap comparing normalised D_{XY} measured in Neanderthal (top right) vs. unadmixed (bottom left) regions of the genome, for west Eurasian populations only.

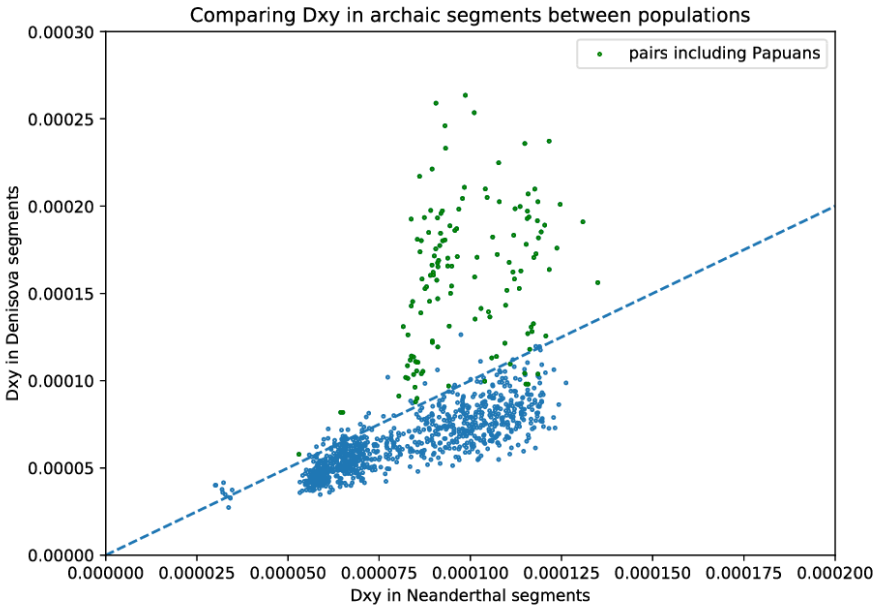


Fig. S22. Absolute divergence (D_{XY}) between all pairs of non-African populations measured in Denisovan and Neanderthal segments of the genome.

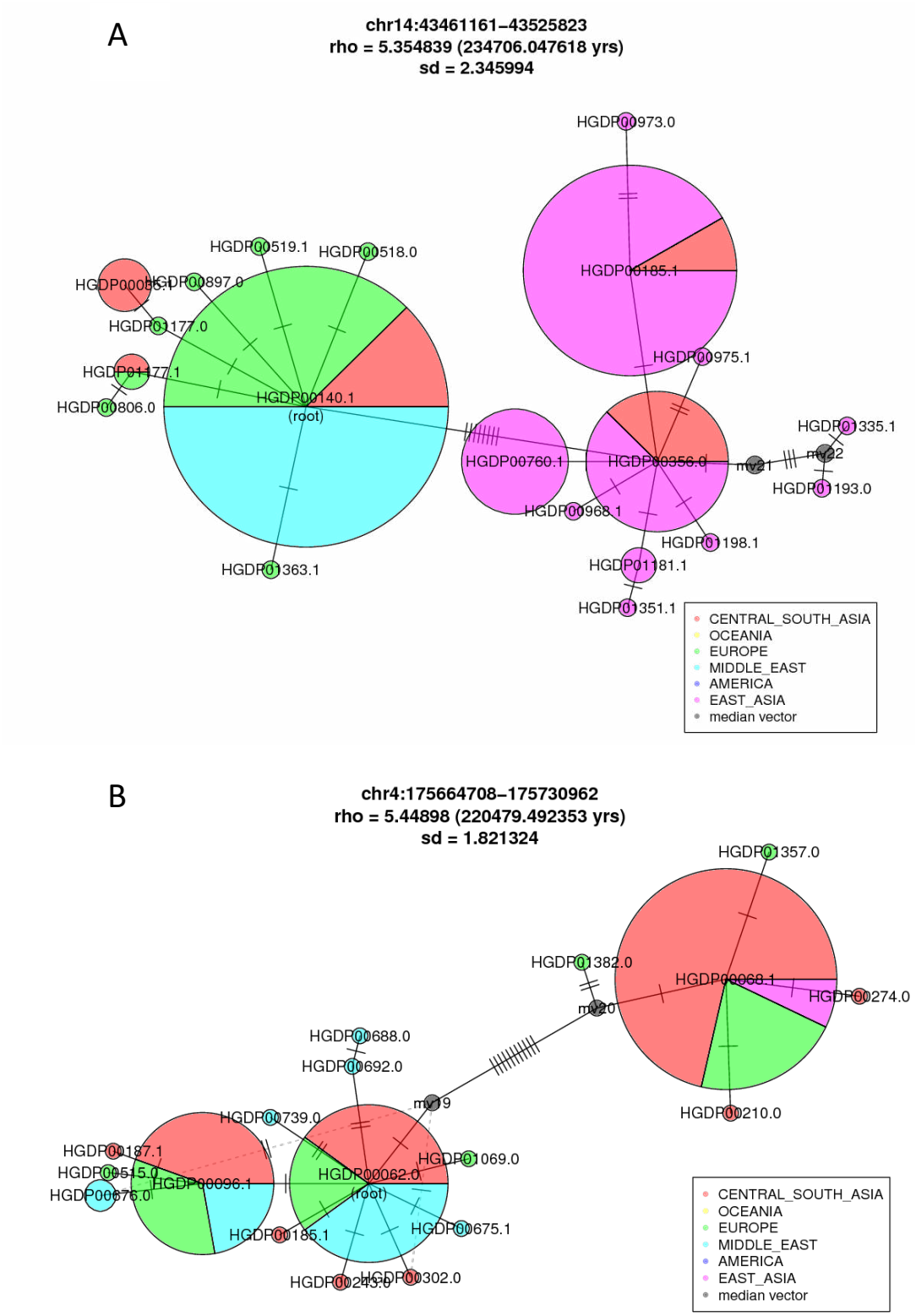


Fig. S23. Examples of Neanderthal haplotype networks. Each circle represents a distinct haplotype, labelled by one sample name and coloured by the geographical origins of the samples, and the radius is proportional to the number of samples carrying that haplotype. The number of bars on the edges equals the number of mutations between haplotypes. Small grey circles labelled "mv" represents median vectors reconstructed in the median joining algorithm. Dashed lines: alternative links.

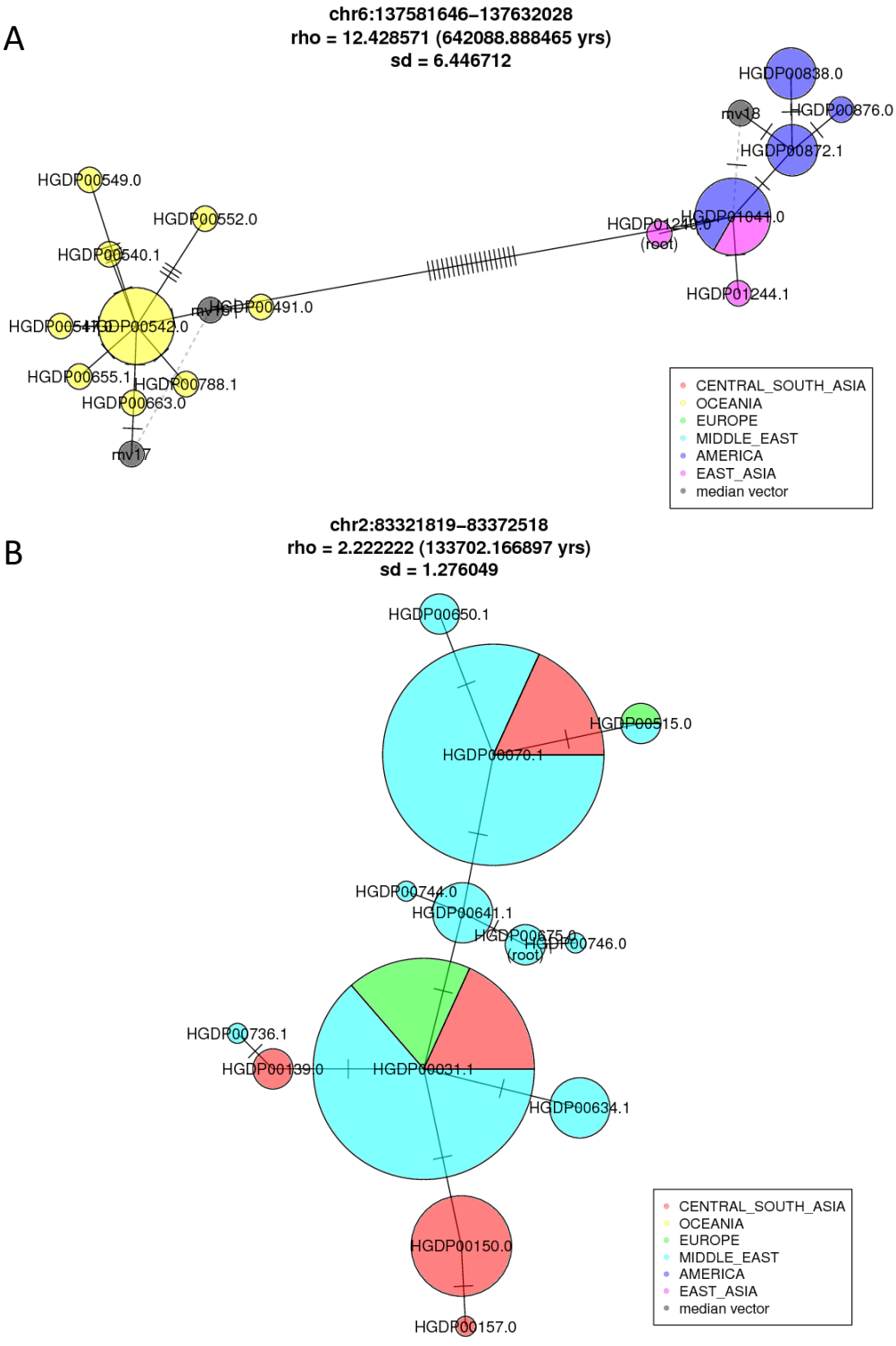


Fig. S24: Examples of Denisovan haplotype networks. Each circle represents a distinct haplotype, labelled by one sample name and coloured by the geographical origins of the samples, and the radius is proportional to the number of samples carrying that haplotype. The number of bars on the edges equals the number of mutations between haplotypes. Small grey circles labelled "mv" represents median vectors reconstructed in the median joining algorithm. Dashed lines: alternative links.

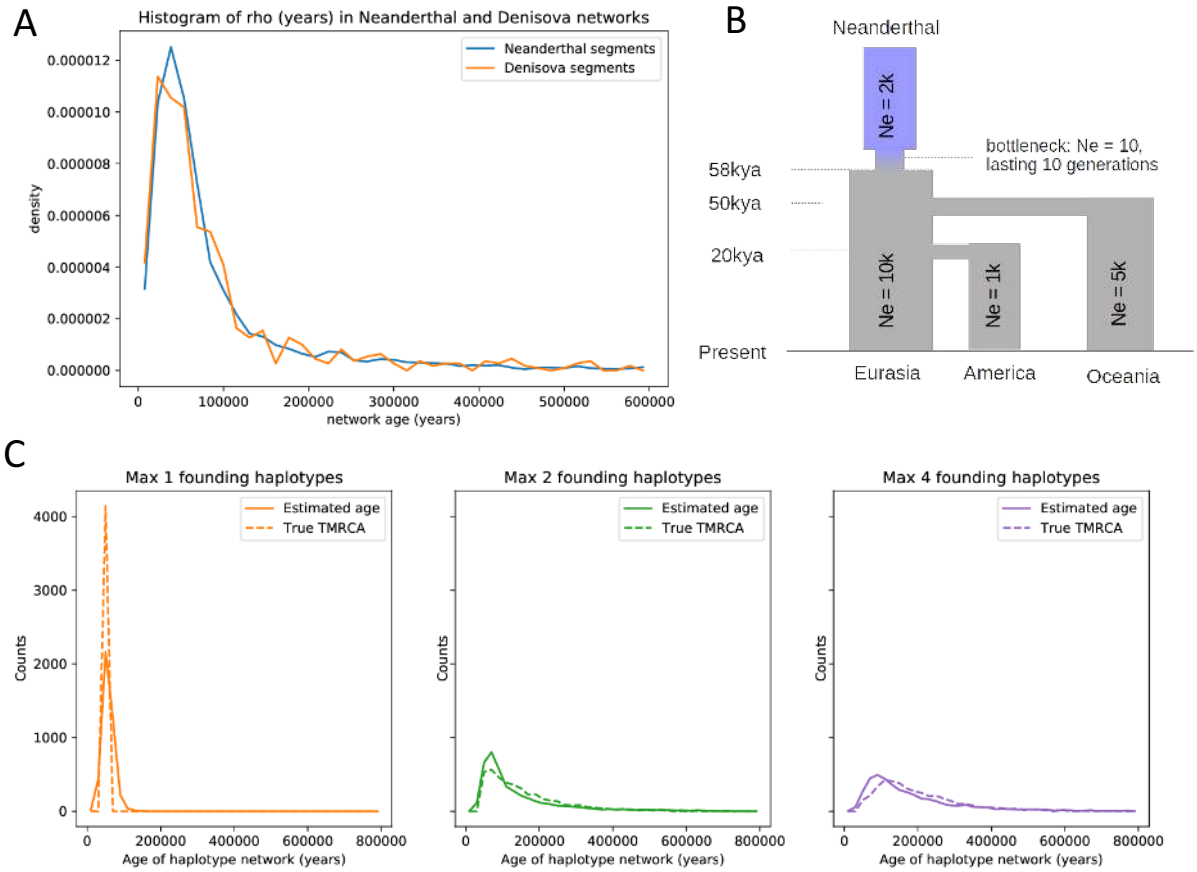


Fig. S25. Archaic haplotype network ages. (A) The distribution of Neanderthal and Denisovan haplotype network ages. **(B)** Demographic model used in simulations exploring different numbers of founding Neanderthal haplotypes. **(C)** Distribution of haplotype network ages estimated from simulations, and true tMRCA.

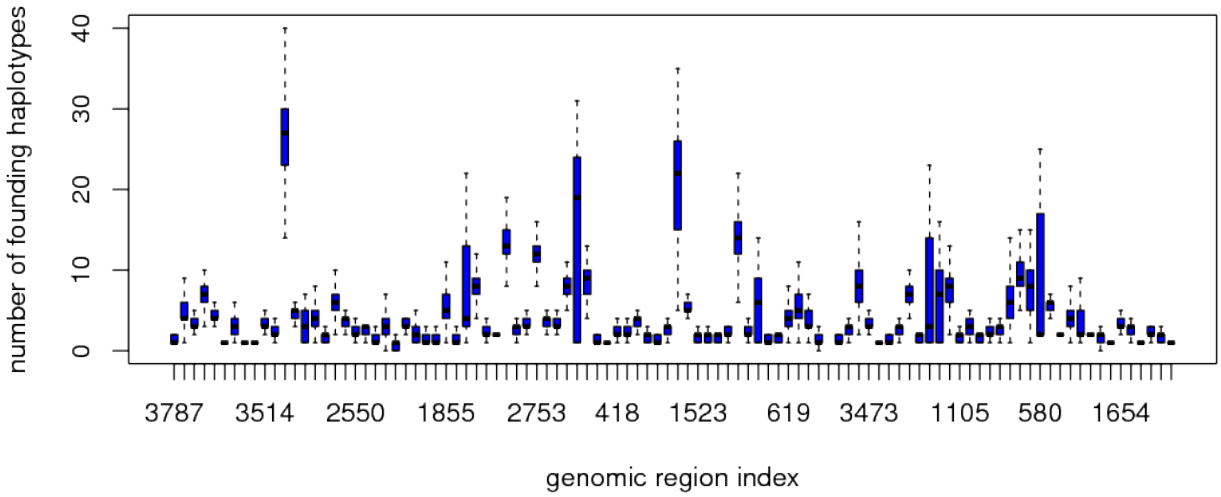


Fig. S26. The number of founding Neanderthal lineages across the genome. A boxplot showing the number of founding Neanderthal lineages from 1,000 bootstraps in 100 genomic regions randomly selected from the total set of 4,135.

Table S1. Overview of populations included in the dataset. The "Lat." and "Lon." columns indicate the approximate latitude and longitude, respectively, of the geographical place of origin of the population. The "Total" column indicates the total number of genome sequences from the given population, and the subsequent columns break that number down by sequencing source and library type.

Population	Region	Lat.	Lon.	Total	Sanger PCR-free	Sanger PCR	SGDP PCR-free	SGDP PCR	Meyer PCR
Adygei	Europe	44	39	16	11	3	2	0	0
Balochi	Central & South Asia	30.5	66.5	24	19	3	2	0	0
BantuKenya	Africa	-3	37	11	0	9	2	0	0
BantuSouthAfrica	Africa	-25.6	24.25	8	1	3	4	0	0
Basque	Europe	43	0	23	21	0	2	0	0
Bedouin	Middle East	31	35	46	41	3	2	0	0
BergamoItalian	Europe	46	10	12	10	0	2	0	0
Biaka	Africa	4	17	22	3	17	2	0	0
Bougainville	Oceania	-6	155	11	5	4	2	0	0
Brahui	Central & South Asia	30.5	66.5	25	20	3	2	0	0
Burusho	Central & South Asia	36.5	74	24	19	3	2	0	0
Cambodian	East Asia	12	105	9	4	3	2	0	0
Colombian	America	3	-68	7	3	2	2	0	0
Dai	East Asia	21	100	9	4	0	3	1	1
Daur	East Asia	48.5	124	9	5	3	1	0	0
Druze	Middle East	32	35	42	37	3	2	0	0
French	Europe	46	2	28	24	0	2	1	1
Han	East Asia	32.3	114	33	29	0	2	1	1
Hazara	Central & South Asia	33.5	70	19	13	4	2	0	0
Hezhen	East Asia	47.5	133.5	9	7	0	2	0	0
Japanese	East Asia	37.5	139	27	25	0	2	0	0
Kalash	Central & South Asia	36	71.5	22	17	3	2	0	0
Karitiana	America	-10	-63	12	6	2	2	1	1
Lahu	East Asia	22	100	8	6	0	2	0	0
Makrani	Central & South Asia	26	64	25	20	3	2	0	0
Mandenka	Africa	12	-12	22	5	13	2	1	1
Maya	America	19	-91	21	17	2	2	0	0
Mbuti	Africa	1	29	13	1	7	3	1	1
Miao	East Asia	28	109	10	5	3	2	0	0
Mongolian	East Asia	48.5	119	9	5	2	2	0	0
Mozabite	Middle East	32	3	27	22	3	2	0	0
Naxi	East Asia	26	100	8	6	0	2	0	0
NorthernHan	East Asia	34.7	107.8	10	10	0	0	0	0
Orcadian	Europe	59	-3	15	13	0	2	0	0
Oroqen	East Asia	50.4	126.5	9	7	0	2	0	0
Palestinian	Middle East	32	35	46	40	3	3	0	0
PapuanHighlands	Oceania	-6.1	145.4	9	4	2	3	0	0
PapuanSepik	Oceania	-4	143	8	2	5	1	0	0
Pathan	Central & South Asia	33.5	70.5	24	19	3	2	0	0
Pima	America	29	-108	13	8	3	2	0	0
Russian	Europe	61	40	25	23	0	2	0	0
San	Africa	-21	20	6	1	1	3	1	0
Sardinian	Europe	40	9	28	24	0	2	1	1
She	East Asia	27	119	10	8	0	2	0	0
Sindhi	Central & South Asia	25.5	69	24	19	3	2	0	0
Surui	America	-11	-62	8	3	3	2	0	0
Tu	East Asia	36	101	10	8	0	2	0	0
Tujia	East Asia	29	109	9	5	2	2	0	0
Tuscan	Europe	43	11	8	6	0	2	0	0
Uygur	Central & South Asia	44	81	10	5	3	2	0	0
Xibo	East Asia	43.5	81.5	9	4	3	2	0	0
Yakut	East Asia	63	129.5	25	20	3	2	0	0
Yi	East Asia	28	103	10	8	0	2	0	0
Yoruba	Africa	8	5	22	1	17	2	1	1

Table S2. Details on the results of the 10x Genomics Chromium sequencing experiments. Metrics were calculated by the 10x Genomics Loupe software.

Sample	Population	Average Molecule Length	Linked Reads per Molecule	Coverage	DNA fragments >20kb (%)	DNA fragments > 100kb (%)	Phase block N50	SNPs phased (%)
HGDP00460	Biaka	16299	12	33.0	37.8	2.07	1,524,737	98.9
HGDP00472	Biaka	8258	6	30.2	17.5	2.79	191,984	97.5
HGDP00562	Druze	8294	6	24.0	12.8	2.12	119,503	97.0
HGDP00580	Druze	15756	20	31.1	33.3	3.47	417,052	98.7
HGDP00774	Han	12518	7	31.2	25.8	2.02	222,937	97.7
HGDP00819	Han	22113	15	35.2	55.8	2.00	601,565	98.7
HGDP01013	Karitiana	30225	22	32.9	69.6	3.14	596,564	98.3
HGDP01019	Karitiana	7713	4	23.2	11.3	1.91	81,409	95.7
HGDP00450	Mbuti	14468	10	35.0	33.5	2.03	990,820	98.7
HGDP01081	Mbuti	12242	9	32.7	24.4	2.70	560,811	98.6
HGDP00549	PapuanHighlands	8724	5	25.8	14.3	1.78	126,844	96.8
HGDP00551	PapuanHighlands	14200	15	31.2	26.0	3.21	206,857	97.7
HGDP00542	PapuanSepik	11604	6	29.1	24.4	1.81	176,726	97.4
HGDP00547	PapuanSepik	11157	11	29.0	16.9	3.48	139,899	97.0
HGDP00224	Pathan	7214	5	27.6	10.7	2.41	81,721	96.9
HGDP00228	Pathan	7589	5	24.6	10.9	1.89	101,084	97.0
HGDP01043	Pima	27111	28	33.2	64.7	2.72	591,200	98.6
HGDP01056	Pima	31234	13	26.7	70.8	3.28	636,630	98.9
HGDP01029	San	11063	11	29.7	17.7	2.52	406,739	98.5
HGDP01032	San	22654	20	33.7	56.2	2.78	2,941,155	98.9
HGDP00670	Sardinian	12126	14	32.9	24.4	2.64	292920	98.4
HGDP01067	Sardinian	33980	37	35.9	74.2	4.73	1,463,166	98.8
HGDP00946	Yakut	7376	6	30.4	10.2	2.34	80,637	96.3
HGDP00954	Yakut	5820	4	23.6	8.94	2.21	40,003	94.9
HGDP00930	Yoruba	7913	6	30.9	12.0	2.54	129,564	97.7
HGDP00931	Yoruba	9279	7	32.5	15.3	2.86	256,356	98.3

Table S3. Evaluation of statistical phasing accuracy. Switch error rates were measured for 26 individuals against experimentally phased haplotypes obtained using 10x Genomics linked reads for the same individuals.

Individual	Population	Region	Switch error rate with singletons	Switch error rate without singletons
HGDP00460	Biaka	Africa	0.0140	0.0044
HGDP00472	Biaka	Africa	0.0157	0.0064
HGDP00450	Mbuti	Africa	0.0144	0.0051
HGDP01081	Mbuti	Africa	0.0155	0.0052
HGDP01029	San	Africa	0.0378	0.0131
HGDP01032	San	Africa	0.0374	0.0122
HGDP00930	Yoruba	Africa	0.0092	0.0033
HGDP00931	Yoruba	Africa	0.0097	0.0027
HGDP01013	Karitiana	America	0.0086	0.0047
HGDP01019	Karitiana	America	0.0066	0.0052
HGDP01043	Pima	America	0.0076	0.0039
HGDP01056	Pima	America	0.0081	0.0052
HGDP00224	Pathan	Central & South Asia	0.0130	0.0043
HGDP00228	Pathan	Central & South Asia	0.0124	0.0040
HGDP00774	Han	East Asia	0.0156	0.0049
HGDP00819	Han	East Asia	0.0167	0.0054
HGDP00946	Yakut	East Asia	0.0061	0.0033
HGDP00954	Yakut	East Asia	0.0088	0.0046
HGDP00670	Sardinian	Europe	0.0107	0.0033
HGDP01067	Sardinian	Europe	0.0115	0.0038
HGDP00562	Druze	Middle East	0.0067	0.0031
HGDP00580	Druze	Middle East	0.0125	0.0038
HGDP00549	PapuanHighlands	Oceania	0.0237	0.0117
HGDP00551	PapuanHighlands	Oceania	0.0281	0.0120
HGDP00542	PapuanSepik	Oceania	0.0281	0.0119
HGDP00547	PapuanSepik	Oceania	0.0278	0.0115

Table S4. Defining regions for analyses of region-specific variants. These were used to count variants that were present in the "Ingroup" and absent from the "Outgroup". Labels in all capital letters denote continental-level region labels, while those in lower case letters denote populations. The notation "anc(X)" denotes estimates of individual fractions for an ancestry component X, as estimated using ADMIXTURE at k=5 on SNPs ascertained as polymorphic in archaic genomes. Numbers in parentheses denote the number of individuals falling into each set.

Region label	Ingroup	Outgroup
AFRICA	AFRICA (n=104)	!AFRICA & anc(AFRICA) <= 0.01 (n=627)
CENTRAL_AFRICA	Biaka or Mbuti (n=35)	(Biaka or Mbuti) (n=894)
San	San (n=6)	!(San or BantuSouthAfrica) (n=915)
OCEANIA	OCEANIA (n=28)	!OCEANIA (n=901)
AMERICA	AMERICA & anc(AMERICA) => 0.95 (n=40)	!AMERICA (n=868)
CENTRAL_AMERICA	(Pima or Maya) & anc(AMERICA) => 0.95 (n=27)	!(Pima or Maya) (n=895)
SOUTH_AMERICA	(Colombian or Surui or Karitiana) & anc(AMERICA) => 0.95 (n=27)	!(Colombian or Surui or Karitiana) (n=902) (n=902)
EUROPE	EUROPE (n=155)	!EUROPE & !(AMERICA & anc(WestEurasian) >= 0.01) (n=749)
EAST_ASIA	EAST_ASIA (n=223)	!EAST_ASIA & !Hazara & !Uygur & !(OCEANIA & anc(EastEurasian) >= 0.01) (n=665)
MIDDLE_EAST	MIDDLE_EAST (n=161)	!MIDDLE_EAST (n=768)
CENTRAL_SOUTH_ASIA	CENTRAL_SOUTH_ASIA & anc(AFRICA) <= 0.05 (n=179)	!CENTRAL_SOUTH_ASIA _ASIA (n=732)
NON_AFRICA_X	region X as defined elsewhere	AFRICA & anc(AFRICA) >= 0.91 (n=101)

Table S5: Encoding of HMM emission types. A value of "0" in the genotype column means ancestral allele in all genomes in the panel, while "1" means derived allele in at least one genome.

Genotype			**Emission type**
Sub-Saharan African panel	Sample of interest	Archaic panel	
1	0	1	1
1	1	0	2
0	1	1	3

Table S6. HMM parameters estimated from simulations.

Initial state distribution (π)	$[0.9655 \quad 0.0345]$
Emission matrix (E)	$\begin{bmatrix} 0.1777 & 0.8208 & 1.336\times10^{-3} \\ 1.691\times10^{-3} & 2.015\times10^{-4} & 0.9981 \end{bmatrix}$

Table S7: Proportion of archaic segments that are still detected after parameter changes.

	Neanderthal segments						Denisovan segments					
Parameter change	**America**	**Central & South Asia**	**East Asia**	**Europe**	**Middle East**	**Oceania**	**America**	**Central & South Asia**	**East Asia**	**Europe**	**Middle East**	**Oceania**
$\pi_1 = 0.017$	0.9981	0.9954	0.9956	0.9967	0.9984	0.9949	0.9834	0.9671	0.9896	0.9465	0.9476	0.9914
$\pi_1 = 0.067$	0.9931	0.9951	0.9932	0.9975	0.9967	0.9939	0.9907	0.9939	0.9971	0.9988	0.9917	0.9922
t = 1000	0.9931	0.9900	0.9930	0.9912	0.9908	0.9881	0.9741	0.9219	0.9849	0.9412	0.9027	0.9851
t = 4000	0.9823	0.9884	0.9832	0.9900	0.9868	0.9857	0.9882	0.9829	0.9441	0.9916	0.9769	0.9742
$E_{0,0} = 0.1954$	0.9979	0.9990	0.9981	0.9987	0.9963	0.9980	0.9992	0.9987	0.9995	0.9989	1.0000	0.9991
$E_{0,1} = 0.9031$	0.9912	0.9937	0.9911	0.9908	0.9936	0.9945	0.9495	0.9554	0.9842	0.9667	0.8818	0.9849
$E_{0,2} = 1.469e-3$	0.9992	0.9986	0.9980	0.9995	0.9992	0.9977	0.9841	0.9940	0.9942	0.9813	0.9970	0.9943
$E_{0,0} = 0.1599$	0.9999	0.9991	0.9983	0.9998	0.9998	0.9972	0.9841	0.9928	0.9937	0.9813	0.9970	0.9935
$E_{0,1} = 0.7388$	0.9902	0.9914	0.9893	0.9919	0.9899	0.9884	0.9767	0.9669	0.9899	0.9477	0.9602	0.9760
$E_{0,2} = 1.202e-3$	0.9996	0.9995	0.9995	0.9986	0.9971	0.9993	0.9967	0.9994	0.9982	1.0000	1.0000	0.9992
$E_{1,0} = 1.860e-4$	0.9996	0.9999	0.9994	0.9991	1.0000	0.9998	0.9985	1.0000	0.9981	1.0000	0.9979	0.9981
$E_{1,1} = 2.216e-4$	1.0000	0.9999	1.0000	0.9995	0.9972	0.9998	1.0000	1.0000	0.9995	1.0000	1.0000	0.9991
$E_{1,0} = 1.860e-4$	0.9977	0.9990	0.9979	0.9991	0.9989	0.9980	0.9992	0.9987	0.9995	0.9989	1.0000	0.9974
$E_{1,1} = 2.216e-4$	1.0000	1.0000	0.9998	0.9999	1.0000	1.0000	1.0000	1.0000	0.9986	1.0000	1.0000	0.9972
$E_{1,2} = 0.8983$	0.9358	0.9394	0.9531	0.9428	0.9524	0.9328	0.7981	0.5987	0.8425	0.5340	0.3825	0.9116

Table S8. Intersection of genomic regions covered by at least two archaic segments between non-African populations. Each value represents the probability of finding a genomic region in the column label conditioned on finding it in the row label.

Neanderthal							
Geographic region	**Total length (bp)**	**Conditional probability to also be found in**					
		America	CS Asia	E Asia	Europe	Middle East	Oceania
America	204886844	-	0.9117	0.9105	0.7235	0.6155	0.3579
C&S Asia	671828352	0.2780	-	0.6099	0.6458	0.6056	0.2409
E Asia	525811986	0.3548	0.7793	-	0.5513	0.4900	0.3000
Europe	482248609	0.3074	0.8997	0.6011	-	0.7981	0.2399
Middle East	453085016	0.2783	0.8979	0.5687	0.8495	-	0.2261
Oceania	218296694	0.3359	0.7413	0.7226	0.5300	0.4693	-

Denisovan							
Geographic region	**Total length (bp)**	**Conditional probability to also be found in**					
		America	CS Asia	E Asia	Europe	Middle East	Oceania
America	13333796	-	0.5761	0.8382	0.3202	0.2470	0.2309
C&S Asia	55284712	0.1389	-	0.3665	0.2106	0.1773	0.1980
E Asia	56280344	0.1986	0.3600	-	0.1330	0.0852	0.1878
Europe	14067884	0.3035	0.8276	0.5321	-	0.6230	0.2146
Middle East	13893642	0.2370	0.7054	0.3451	0.6308	-	0.2091
Oceania	190235182	0.0162	0.0575	0.0556	0.0159	0.0153	-

Table S9. Genomic regions with more than 20 founding Neanderthal haplotypes.

Chr	Start	End	#Haplotypes	#Unique haplotypes	Rho	sd	Rho (years)	sd (years)
9	110937947	111009555	220	75	6.059	3.814	223447.6	140658.6
6	66848428	66915134	163	68	2.178	0.349	89369.29	14324.72
12	113841761	113933611	150	85	2.94	0.316	80828.57	8698.694
5	58495484	58599651	198	99	2.581	0.404	78918.59	12363.74
1	217246438	217327394	200	89	2.325	0.593	74454.43	18987.48
19	56087294	56138132	228	67	1.110	0.104	70633.69	6593.898
10	62941526	62993549	209	52	1.187	0.071	70207.82	4177.36
12	114080296	114130307	204	69	1.304	0.089	66588.11	4520.688
2	13827358	13919411	125	73	2.136	0.165	60264.87	4664.095
1	216557045	216647205	218	85	2.147	0.155	58608.5	4234.904
1	32911081	32992108	211	68	1.422	0.268	57411.51	10832.91
4	28482612	28545486	191	64	1.277	0.080	54515.13	3432.053
12	20849814	20933980	151	63	1.709	0.217	52310.35	6651.919
9	126565708	126646783	190	66	1.389	0.162	46613.15	5425.191
1	212466385	212548707	196	78	1.301	0.044	43196.05	1458.018
9	94515802	94565893	148	42	0.534	0.011	38530.75	781.0288
1	33403459	33454985	263	63	0.669	0.025	34713.19	1273.397

Table S10. The number of completely geographically separated haplotype network between pairs of regions out of the total number of comparable networks.

Neanderthal haplotype networks, separated / total

	Central/South Asia	East Asia	Europe	Middle East	Oceania
America	254 / 862	184 / 878	221 / 618	245 / 506	159 / 263
Central/South Asia	-	228 / 2064	133 / 2190	139 / 2032	261 / 690
East Asia	-	-	338 / 1356	429 / 1162	187 / 714
Europe	-	-	-	81 / 1932	249 / 448
Middle East	-	-	-	-	230 / 393

Denisovan haplotype networks, separated / total

	Central/South Asia	East Asia	Europe	Middle East	Oceania
America	6 / 36	8 / 59	4 / 15	3 / 9	10 / 12
Central/South Asia	-	4 / 88	3 / 42	3 / 32	37 / 39
East Asia	-	-	5 / 27	5 / 14	35 / 40
Europe	-	-	-	2 / 31	7 / 9
Middle East	-	-	-	-	6 / 8

Table S11. p-values from Fisher's exact test on different distributions of separated/connected networks in Neanderthal vs. Denisovan haplotypes between pairs of regions.

	Central&South Asia	East Asia	Europe	Middle East	Oceania
America	0.1320	0.2419	0.5908	0.5067	0.1375
Central&South Asia	-	0.0534	0.7398	0.4802	3.075×10^{-13}
East Asia	-	-	0.6523	1	5.207×10^{-15}
Europe	-	-	-	0.3801	0.3100
Middle East	-	-	-	-	0.4793