RESEARCH ARTICLE SUMMARY

HUMAN GENETICS

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§*

INTRODUCTION: Large-scale human genome-sequencing studies to date have been limited to large, metropolitan populations or to small numbers of genomes from each group. Much remains to be understood about the extent and structure of genetic variation in our species and how it was shaped by past population separations, admixture, adaptation, size changes, and gene flow from archaic human groups. Larger numbers of genome sequences from more diverse populations are needed to answer these questions.

RATIONALE: We sequenced 929 genomes from 54 geographically, linguistically, and culturally diverse human populations to an average of 35× coverage and analyzed the variation among them. We also physically resolved the haplotype phase of 26 of these genomes using linked-read sequencing.

RESULTS: We identified 67.3 million single-nucleotide polymorphisms, 8.8 million small insertions or deletions (indels), and 40,736 copy number variants. This includes hundreds of thousands of variants that had not been discovered by previous sequencing efforts, but which are common in one or more population. We demonstrate benefits to the study of population relationships of genome sequences over ascertained array genotypes, particularly when involving African populations.

Populations in central and southern Africa, the Americas, and Oceania each harbor tens to hundreds of thousands of private, common genetic variants. Most of these variants arose as new mutations rather than through archaic introgression, except in Oceanian populations, where many private variants derive from Denisovan admixture. Although some reach high frequencies, no variants are fixed between major geographical regions.

We estimate that the genetic separation between present-day human populations occurred mostly within the past 250,000 years. However, these early separations were gradual in nature and shaped by protracted gene flow. All populations thus still had some genetic contact more recently than this, but there is also evidence that a small fraction of present-day structure might be hundreds of thousands of years older. Most populations expanded in size over the past 10,000 years, but hunter-gatherer groups did not.

The low diversity among the Neanderthal haplotypes segregating in present-day populations indicates that, while more than one Neanderthal individual must have contributed genetic material to modern humans, there was likely only one major episode of admixture. By contrast, Denisovan haplotype diversity reflects a more complex history involving more than one episode of admixture.

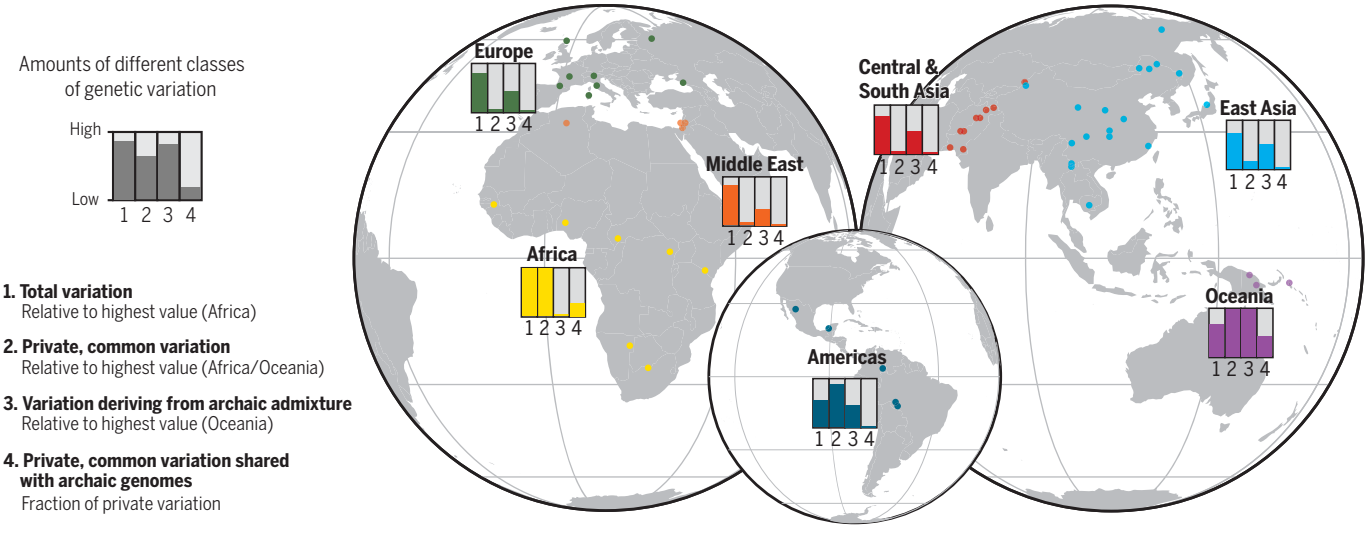
ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aay5012

We found small amounts of Neanderthal ancestry in West African genomes, most likely reflecting Eurasian admixture. Despite their very low levels or absence of archaic ancestry, African populations share many Neanderthal and Denisovan variants that are absent from Eurasia, reflecting how a larger proportion of the ancestral human variation has been maintained in Africa.

CONCLUSION: The discovery of substantial amounts of common genetic variation that was previously undocumented and is geographically restricted highlights the continued value of anthropologically informed study designs for understanding human diversity. The genome sequences presented here are a freely available resource with relevance to population history, medical genetics, anthropology, and linguistics. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: ab34@sanger.ac.uk (A.B.); cts@sanger.ac.uk (C.T.-S.)
†These authors contributed equally to this work.
‡Deceased.
§These authors contributed equally to this work.
Cite this article as A. Bergström *et al.*, *Science* **367**, eaay5012 (2020). DOI: 10.1126/science.aay5012



Structure of genetic variation across worldwide human populations. Shown is a schematic illustration of the approximate amounts of four different classes of genetic variation found in different geographical regions. The origins of the populations included in the study are indicated by dots.

RESEARCH ARTICLE

HUMAN GENETICS

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

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]§*

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 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 past 10,000 years, and a contrast between single Neanderthal but multiple Denisovan source populations contributing to present-day human populations.

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 to 700 thousand years ago (kya), the appearance of anatomical modernity in Africa in the past few hundred thousand years, an expansion out of Africa and the Near East 50 to 70 kya, 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 after multiple independent transitions from hunter-gatherer to food-producing lifestyles in the past 10,000 years (*1*). However, much 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 used low-coverage sequencing (*2*), whereas those sampling human groups more widely have mostly been limited to one to three genomes per population (*3*, *4*). The Human Genome Diversity Project (HGDP)–Centre d'Etude du Polymorphisme Humain (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 and 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 35× (minimum 25×) 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 analyzing 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 to 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 and 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 population sample (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 predefined variant sets used on genotyping arrays. We find that whereas analyses of the SNPs included on commonly used arrays accurately recapitulate relationships between non-African populations, they sometimes substantially 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 with 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 used 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 allele frequency differentiation (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, which are 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 did not find a general increase in the power to detect population relationships in the form of nonzero f_4-statistics when using all of the discovered

[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, 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, CA 94158, USA. [10]Centre d'Etude du Polymorphisme Humain, Fondation Jean Dausset, 75010 Paris, France. [11]GENMED Labex, ANR-10-LABX-0013 Paris, France. [12]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [13]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA. [14]Department of Medicine, University of Cambridge, Cambridge CB2 0QQ, UK. ***Corresponding author. Email: ab34@sanger.ac.uk (A.B.); cts@sanger.ac.uk (C.T.-S.)**
†These authors contributed equally to this work. ‡Deceased. §These authors contributed equally to this work.

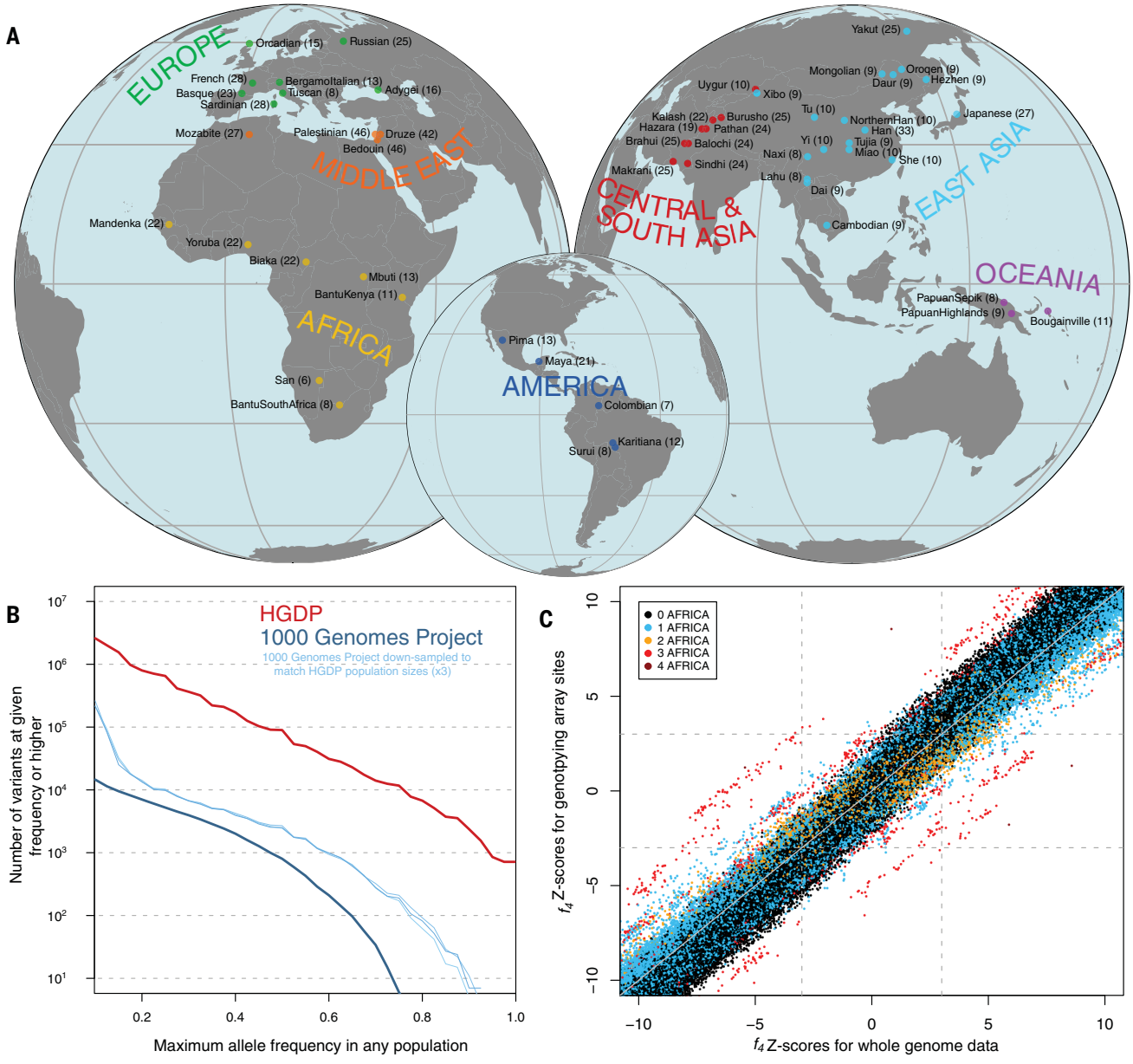


Fig. 1. Genome sequencing and variant discovery in 54 diverse human populations. (**A**) Geographical origins of the 54 populations from the HGDP-CEPH panel, with the number of sequenced individuals from each shown in parentheses. (**B**) Maximum allele frequencies of variants discovered in the HGDP dataset but not in the 1000 Genomes phase 3 dataset (red) and vice versa (dark blue). 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 caused by the 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 (light blue). 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 colored according to the number of African populations included in the statistic.

SNPs, most of which are rare, compared with 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 found 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 the Mbuti and Yoruba since the divergence of non-Africans. An excess sharing of the San with the Mandenka relative to the Mbuti at low allele frequencies may similarly reflect low amounts of West African–related admixture into the San (Fig. 2C) (*17*). The known Denisovan admixture in Oceanian populations manifests itself, without

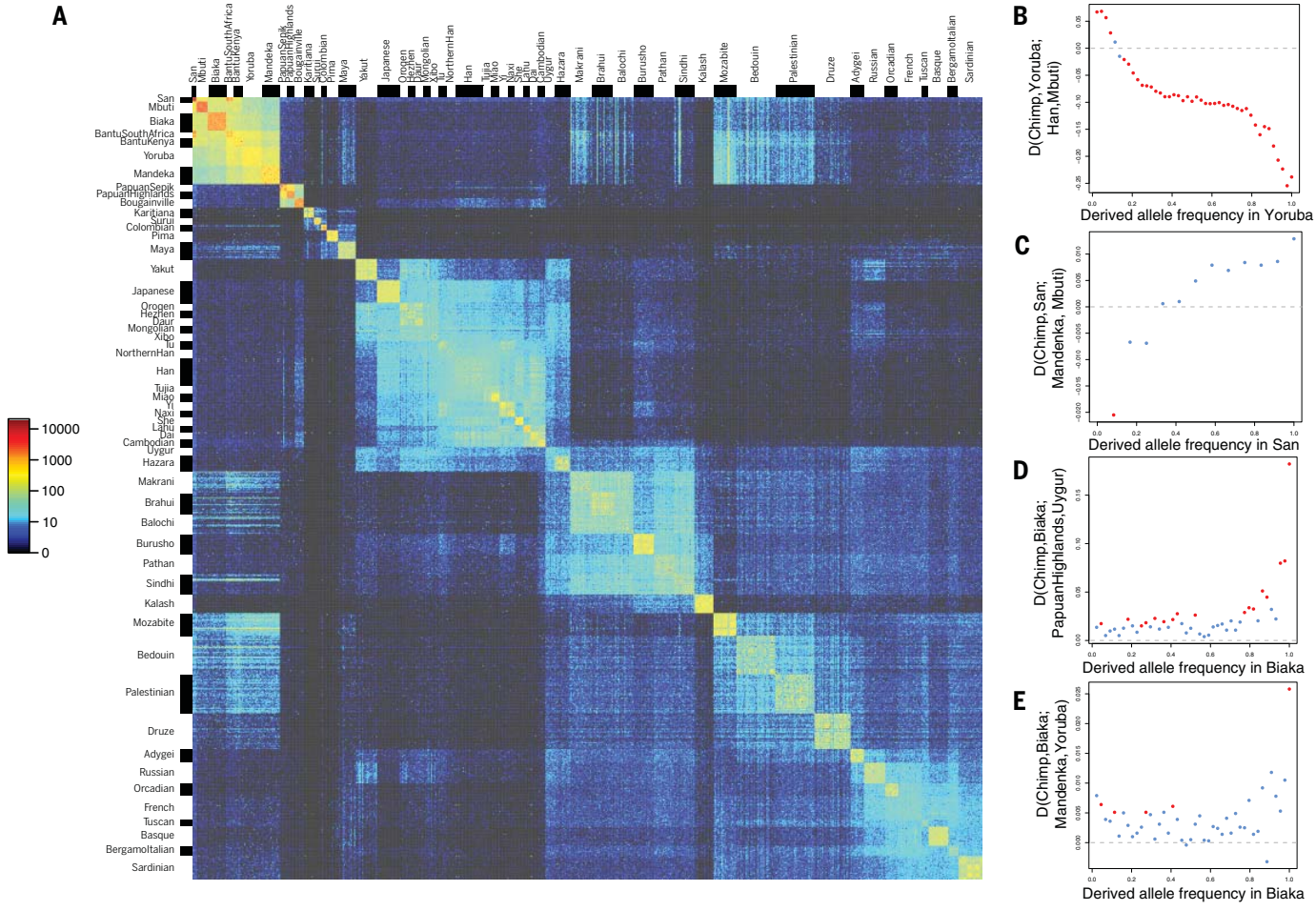


Fig. 2. Insights into population relationships from low-frequency variants. (**A**) Heatmap of pairwise counts of doubleton alleles (alleles observed exactly twice across the dataset) among all 929 individuals, grouped by population. (**B** to **E**) *D*-statistics of the form *D(Chimp,X;A,B)*, stratified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

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 the Yoruba than to the Mandenka, another West African population (Fig. 2E), which would be consistent with the 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 and S13). An F* lineage representing the deepest known split in the FT branch that is carried by most non-African men was found only once across the 1205 males of the 1000 Genomes Project (*19*). Here, we found it in five out of seven sampled males in the Lahu from the 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.

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 found no such private variants that are fixed in a given continent or major region (Fig. 3, A to C). The highest frequencies are reached by a few tens of variants present at >70% (and a few thousands at >50%) in Africa, the Americas, and Oceania. By contrast, the highest frequency variants private to Europe, East Asia, the Middle East, or Central and South Asia reach just 10 to 30%. This likely reflects greater genetic connectivity within Eurasia owing to culturally driven migrations and admixture in the past 10,000 years, events that 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 ~20 that are fixed in our small sample of six individuals.

Most of these geographically restricted variants reflect new 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

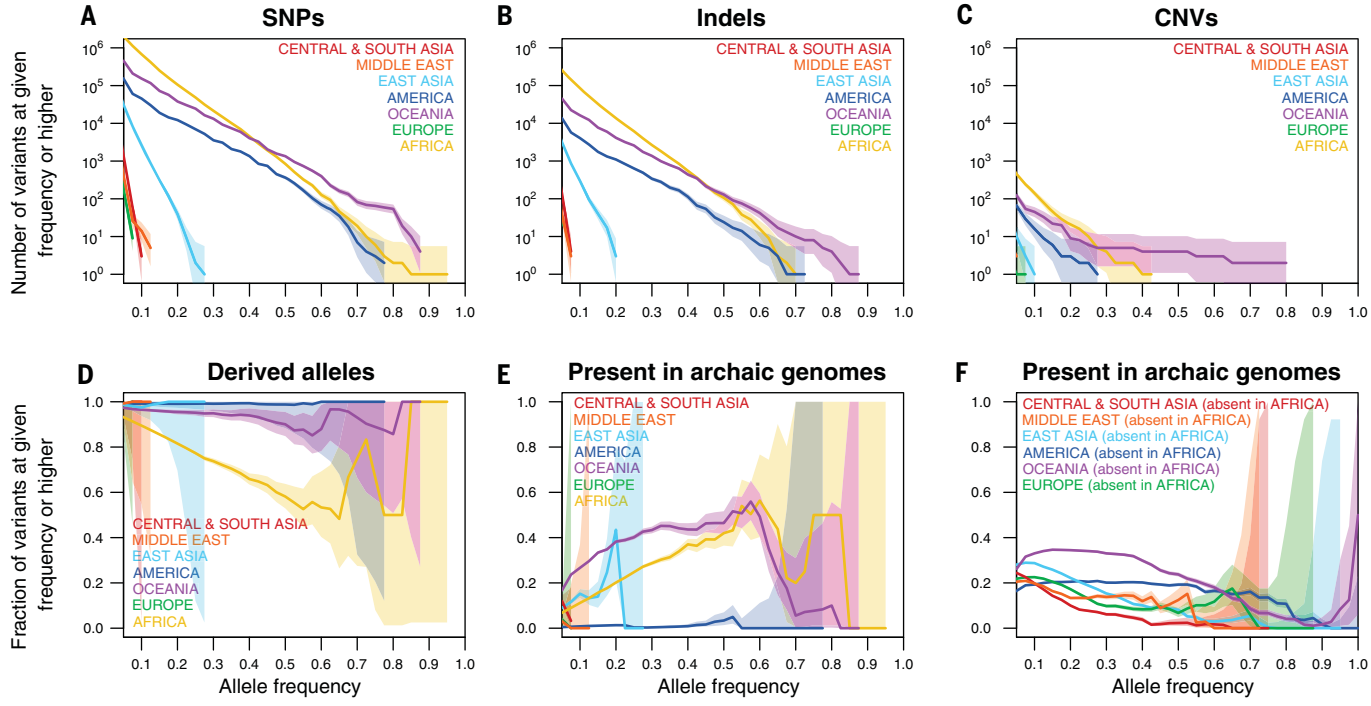


Fig. 3. Counts and properties of geographically private variants. (**A** to **C**) Counts of region-specific variants. The vertical axis displays the number of variants private to a given geographical region that have an allele frequency in that region equal to or higher than the corresponding value on the horizontal axis. Shaded areas denote 95% Poisson confidence intervals. (A) SNPs. (B) Indels. (C) CNVs. (**D**) 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**) 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 in (E) but now counting variants that are present in the given region and absent in Africa regardless of their frequency elsewhere.

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 ≥35% of private variants present at ≥20% frequency are shared with the Denisovan genome. Generally, ≥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 ≤80% of such common variants are more likely to have derived from new mutations, which thus have been a stronger force than archaic admixture in introducing new 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 ~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 and 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 CNVs 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 analyzing single human genomes, and haplotype phasing errors can cause artifacts 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 past 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 past 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 that 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 about 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 artifact and will require further validation, the inferred growth rate exceeds even those of large European and East Asian populations in the past 10,000 years, suggesting

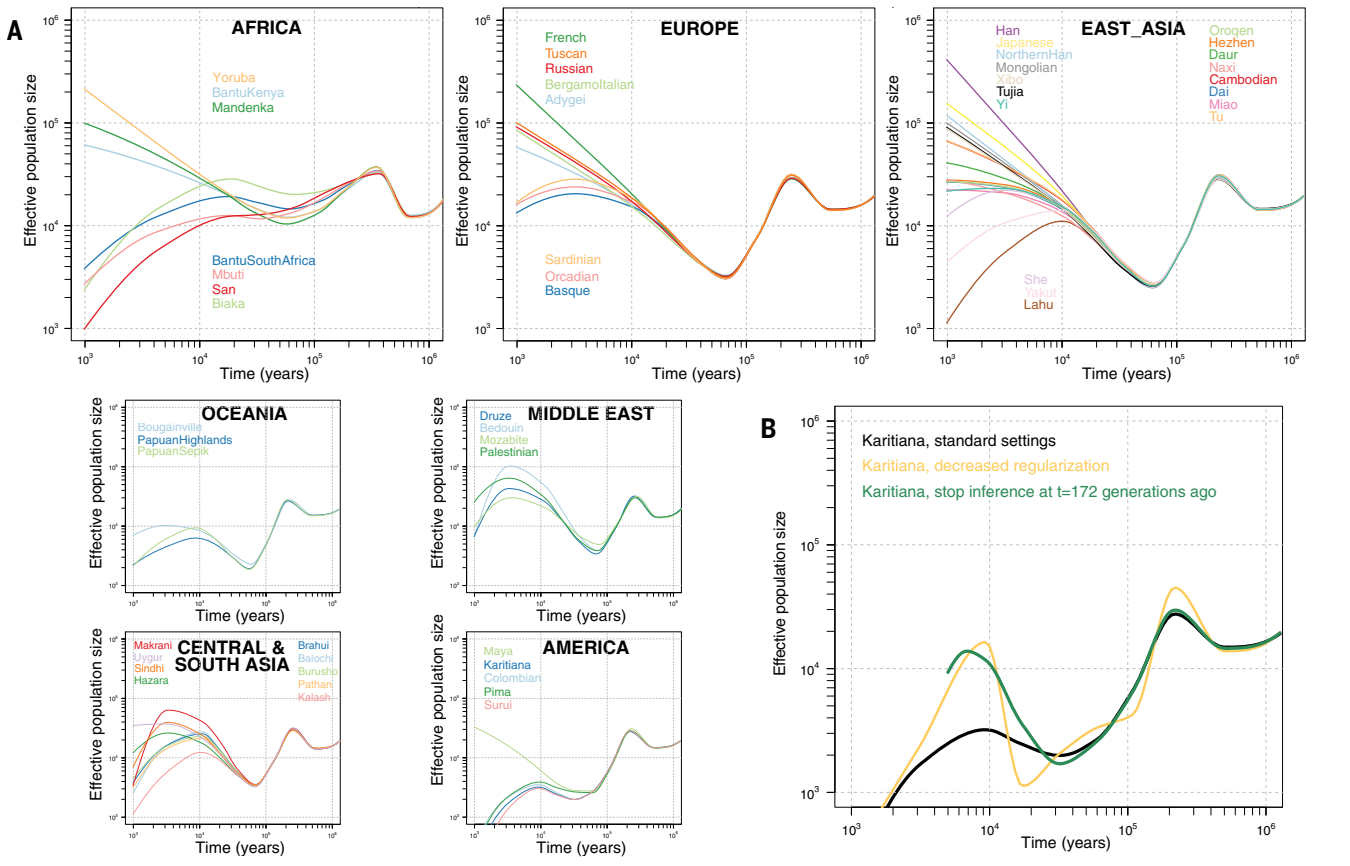


Fig. 4. Effective population size histories of 54 diverse populations. (**A**) Effective population sizes for all populations inferred using SMC++, computed using composite likelihoods across six different distinguished individuals per population. Our ability to infer recent size histories in some South 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 past few thousand years from the time period of inference leads to curves displaying massive growth approximately in the period 10 to 20 kya.

that this could be one of the most substantial growth episodes in modern human population history.

Although 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 past 10,000 years. This might require yet larger sample sizes, new analytical methods that exploit other features of genetic variation (*28*), or both.

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 of the split time between two populations, we use 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 with 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 rainforest hunter-gatherer groups, the Mbuti and the Biaka, at ~62 kya; between the Mbuti and the West African Yoruba at ~69 kya; between the Yoruba and the southern African San at ~126 kya; and between the San and both the Biaka and the Mbuti at ~110 kya. Non-Africans have separation midpoints from the Yoruba at ~76 kya, from the Biaka at ~96 kya, from the Mbuti at ~123 kya, and, representing the deepest split in the dataset, from the San at ~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 the San and the Biaka until at least 50 kya, and between the Mbuti, the Biaka, and the 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 is that there lived populations already at this time that 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, whereas most of the ancestry was connected until much more recently. In light of such composite ancestries

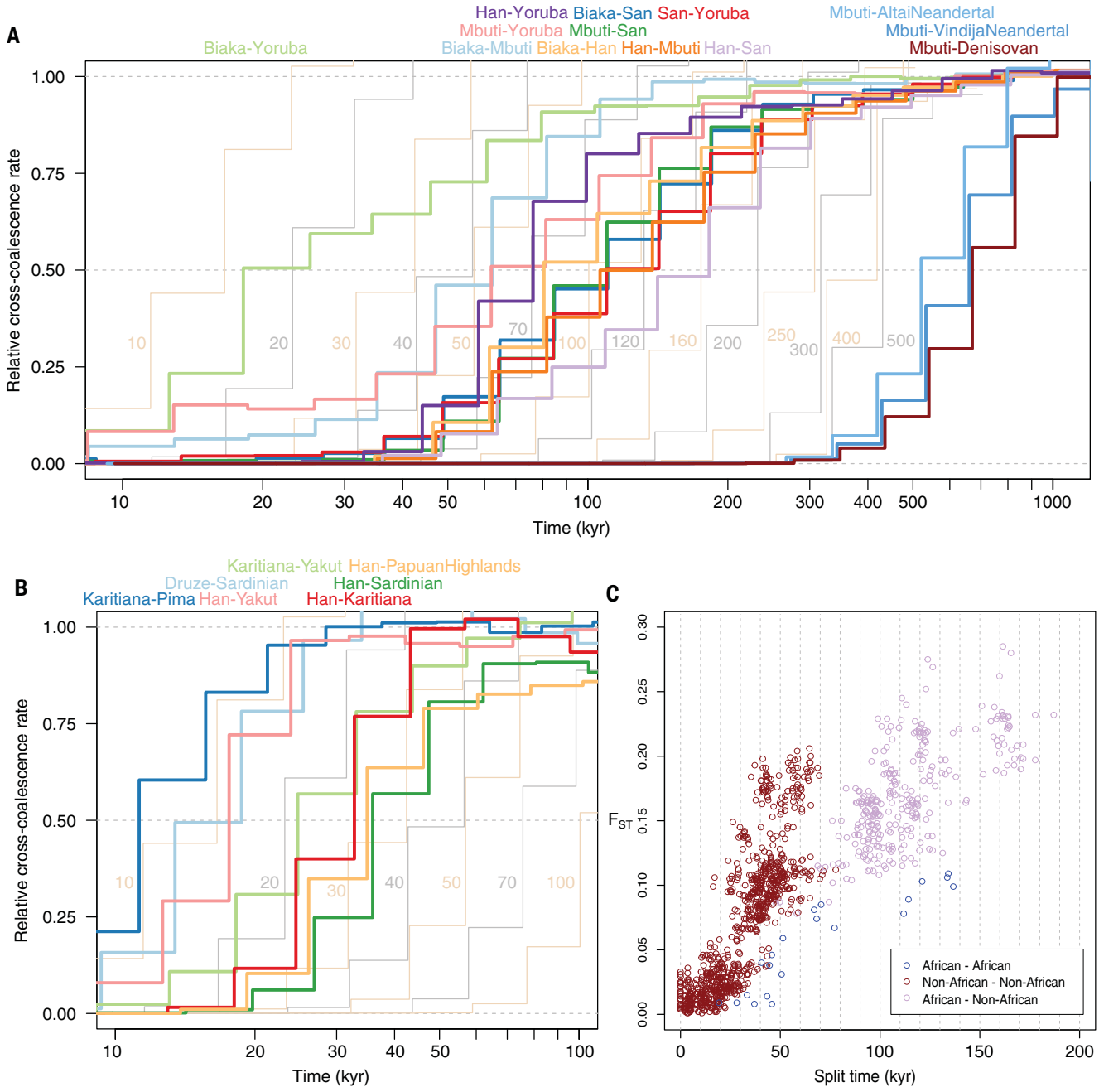


Fig. 5. Time depth and mode of population separations. (**A**) MSMC2 cross-population results for pairs of African populations, including the Han Chinese as a representative of non-Africans, as well as between archaic populations and the Mbuti as a representative of modern humans. Curves between modern human groups were computed using four physically phased haplotypes per population, and 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 gray curves. (**B**) MSMC2 cross-population results as in (A) but 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 because we could not obtain reliable momi2 fits for these.

in present-day human populations and the clear deviation of our MSMC2 results from instant split behaviors, we argue 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 past 250,000 years, 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. Although 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

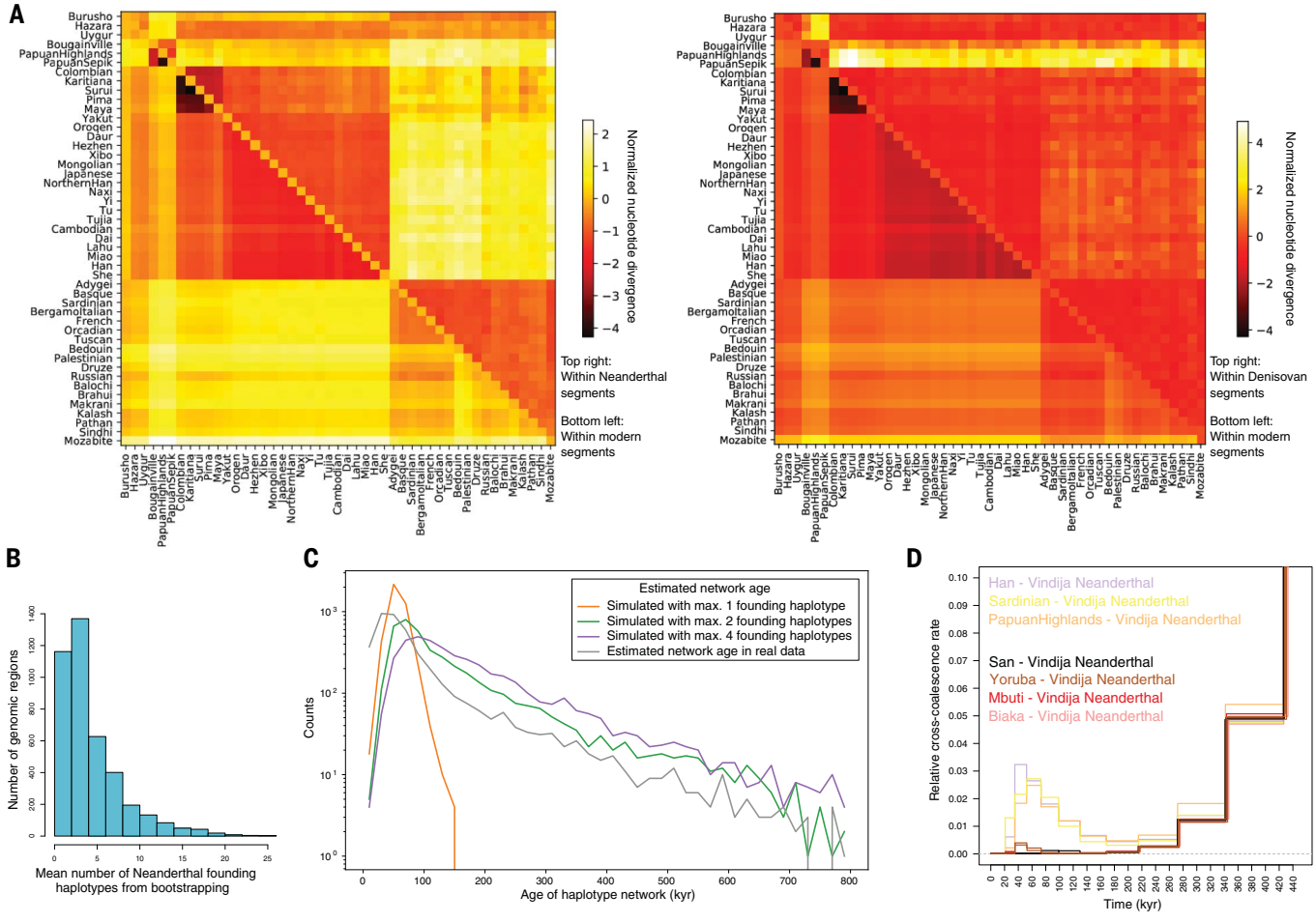


Fig. 6. Archaic haplotypes in modern human populations. (**A**) Nucleotide divergence D_{XY} within segments deriving from archaic admixture and within other segments in non-African populations. (**B**) Mean number of archaic founding haplotypes estimated by constructing maximum likelihood trees for 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**) Distribution of estimated ages of archaic haplotype networks in the present-day human population. The distribution is compared with 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) populations 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).

sites in these genomes. The midpoint estimates suggest that modern and archaic populations separated 550 to 700 kya (Fig. 5A), which is consistent 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 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 consistent with previous estimates (*3*, *4*, *22*), with all populations sharing most of their ancestry within the past 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 that those results were caused by some artifact 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 the 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 most of their ancestry within the past 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 after their shared bottleneck event (*34*).

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 to 3.6%) Denisovan ancestry in Papuan highlanders (*15*), substantially lower than the first estimate of 4 to 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 that 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 and 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 because of 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 and 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. 6, B and C). Coalescent simulations suggest that, genome-wide, as few as two to four founding haplotypes are sufficient to produce the observed distribution of haplotype network sizes.

By 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 and table S10), corresponding to a deep divergence between the Denisovan source populations. East Asian populations also harbor 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 found that non-Africans display clear modes of nonzero 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 and fig. S8). The Denisovan gene flow into Oceanians is also visible in these analyses but is less pronounced and substantially shifted backward 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 that is similar in shape but much less pronounced than that in non-Africans (Fig. 6D and fig. S9). Other African populations do not clearly display the same behavior. These results provide evidence for low amounts of Neanderthal ancestry in West Africa, consistent with previous results that were based on other approaches (*16*, *20*), and we estimate this at 0.18 ± 0.06% in the Yoruba using an f_4-ratio (assuming that the Mbuti have 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 the 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 owing 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 versus 26.4%). However, more Denisovan variants survive in Africans (18.9 versus 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 that were segregating in late archaic populations.

Discussion

Although 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 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 catalog 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 major population expansion after 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 harbor 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 the 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 characterize variation in a population compared with genotyping using arrays, sequencing to low coverage, or sequencing just small numbers of genomes. 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, and especially useful in diverse populations. However, short-read sequencing still imposes limitations on the ability to identify more complex structural variation. We expect that 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, will unveil these additional layers of human genetic diversity.

Although 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 10 individuals from each of the ~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 toward 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 ≤20 or a DP (depth) value ≥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 that is largely based on the 1000 Genomes Project (*2*) "strict mask" and restricted analyses to these regions. Batch effects between library types observed when analyzing 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 that we used in our analyses, along with the seven regional and/or continental labels previously used (*6*). We constructed 10x Genomics linked-read libraries (*14*) and sequenced these on Illumina HiSeq X machines for 26 individuals from 13 globally representative populations to physically resolve their haplotype phase and to 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 between the 13 populations for which we had physically phased genomes for two individuals each. By means of site-frequency spectrum modeling 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 analyzed 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.

A more detailed description of the materials and methods is available in the supplementary materials.

REFERENCES AND NOTES

1. R. Nielsen *et al.*, Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017). doi: 10.1038/nature21347; pmid: 28102248
2. The 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi: 10.1038/nature15393; pmid: 26432245
3. S. Mallick *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016). doi: 10.1038/nature18964; pmid: 27654912
4. L. Pagani *et al.*, Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538**, 238–242 (2016). doi: 10.1038/nature19792; pmid: 27654910
5. H. M. Cann *et al.*, A human genome diversity cell line panel. *Science* **296**, 261–262 (2002). doi: 10.1126/science.296.5566.261b; pmid: 11954565
6. N. A. Rosenberg *et al.*, Genetic structure of human populations. *Science* **298**, 2381–2385 (2002). doi: 10.1126/science.1078311; pmid: 12493913
7. M. Jakobsson *et al.*, Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998–1003 (2008). doi: 10.1038/nature06742; pmid: 18288195
8. J. Z. Li *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008). doi: 10.1126/science.1153717; pmid: 18292342
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). doi: 10.1093/molbev/msp243; pmid: 19822636
10. N. Patterson *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012). doi: 10.1534/genetics.112.145037; pmid: 22960212
11. M. Meyer *et al.*, A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012). doi: 10.1126/science.1224344; pmid: 22936568
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). doi: 10.1186/2041-2223-5-13; pmid: 25254093
13. M. Raghavan *et al.*, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015). doi: 10.1126/science.aab3884; pmid: 26198033
14. G. X. Zheng *et al.*, Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016). doi: 10.1038/nbt.3432; pmid: 26829319
15. Materials and methods are available as supplementary materials.
16. K. Prüfer *et al.*, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014). doi: 10.1038/nature12886; pmid: 24352235

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). doi: 10.1073/pnas.1313787111; pmid: 24550290
18. P. Skoglund *et al*., Reconstructing prehistoric African population structure. *Cell* **171**, 59–71 (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). doi: 10.1038/ng.3559; pmid: 27111036
20. K. Prüfer *et al*., A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017). doi: 10.1126/science.aao1887; pmid: 28982794
21. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011). doi: 10.1038/nature10231; pmid: 21753753
22. S. Schiffels, R. Durbin, Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014). doi: 10.1038/ng.3015; pmid: 24952747
23. S. Song, E. Sliwerska, S. Emery, J. M. Kidd, Modeling human population separation history using physically phased genomes. *Genetics* **205**, 385–395 (2017). doi: 10.1534/genetics.116.192963; pmid: 28049708
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). doi: 10.1038/ng.3748; pmid: 28024154
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). doi: 10.1073/pnas.96.19.10597; pmid: 10485871
26. B. Llamas *et al*., Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci. Adv.* **2**, e1501385 (2016). doi: 10.1126/sciadv.1501385; pmid: 27051878
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.e3 (2019). doi: 10.1016/j.cub.2018.11.029; pmid: 30581024
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). doi: 10.1016/j.ajhg.2015.07.012; pmid: 26299365
29. A. S. Malaspinas *et al*., A genomic history of Aboriginal Australia. *Nature* **538**, 207–214 (2016). doi: 10.1038/nature18299; pmid: 27654914
30. A. Scally, The mutation rate in human evolution and demographic inference. *Curr. Opin. Genet. Dev.* **41**, 36–43 (2016). doi: 10.1016/j.gde.2016.07.008; pmid: 27589081
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). doi: 10.1002/ajpa.20188; pmid: 15795887
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). doi: 10.1038/ncomms16046; pmid: 28675384
33. J. Kamm, J. Terhorst, R. Durbin, Y. S. Song, Efficiently inferring the demographic history of many populations with allele count data. *J. Am. Stat. Assoc.* **0**, 1–16 (2019). doi: 10.1080/01621459.2019.1635482
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). doi: 10.1534/genetics.112.144758; pmid: 23172852
35. D. Reich *et al*., Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010). doi: 10.1038/nature09710; pmid: 21179161
36. P. Qin, M. Stoneking, Denisovan ancestry in East Eurasian and Native American populations. *Mol. Biol. Evol.* **32**, 2665–2674 (2015). doi: 10.1093/molbev/msv141; pmid: 26104010
37. B. Vernot *et al*., Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239 (2016). doi: 10.1126/science.aad9416; pmid: 26989198
38. I. Lazaridis *et al*., Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016). doi: 10.1038/nature19310; pmid: 27459054
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.e9 (2018). doi: 10.1016/j.cell.2018.02.031; pmid: 29551270
40. G. S. Jacobs *et al*., Multiple deeply divergent Denisovan ancestries in Papuans. *Cell* **177**, 1010–1021.e32 (2019). doi: 10.1016/j.cell.2019.02.035; pmid: 30981557
41. I. Lazaridis *et al*., Paleolithic DNA from the Caucasus reveals core of West Eurasian ancestry. bioRxiv 423079 [Preprint]. 21 September 2018. https://doi.org/10.1101/423079.
42. E. Garrison *et al*., Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018). doi: 10.1038/nbt.4227; pmid: 30125266
43. G. F. Simons, C. D. Fennig, *Ethnologue: Languages of the World* (SIL International, ed. 21, 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). doi: 10.1101/gr.107524.110; pmid: 20644199
45. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLOS Genet.* **2**, e190 (2006). doi: 10.1371/journal.pgen.0020190; pmid: 17194218
46. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009). doi: 10.1101/gr.094052.109; pmid: 19648217

ACKNOWLEDGMENTS

We thank the sample donors who made this research possible, as well as the CEPH Biobank (BIORESOURCES, Paris) 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 S. Fairley and colleagues at the International Genome Sample Resource for incorporating and hosting data. We also 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. by Wellcome grant 207492. A.B. and P.S. were supported by the Francis Crick Institute (grant 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 852558) and the Wellcome Trust (grant 217223/Z/19/Z). R.H. was supported by a Gates Cambridge scholarship. P.H. was supported by the Estonian Research Council (grant PUT1036). D.R. is an investigator of the Howard Hughes Medical Institute. **Author contributions:** Y.X., R.D., and C.T.-S. conceived the study. A.B., S.A.M., Q.A., H.B., J.-F.D., H.W., S.M., D.R., M.S.S., Y.X., R.D., and C.T.-S. coordinated data generation. A.B., S.A.M., M.A.A., P.D., Y.C., and S.M. processed genome-sequencing data. A.B., R.H., M.A.A., S.F., P.H., J.K., and P.S. performed population genomic analyses. P.H., P.S., A.S., Y.X., R.D., and C.T.-S. supervised analyses. A.B. and C.T.-S. wrote the manuscript with input from R.H., M.A.A., D.R., P.S., A.S., Y.X., R.D., and all other authors. A.B., R.H., M.A.A., S.F., and P.H. wrote the supplementary materials. **Competing interests:** R.D. is a founder and consultant for Congenica Ltd. and holds stock in Illumina from previous consulting. **Data and materials availability:** Raw read alignments are available from the European Nucleotide Archive under study accession no. 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 no. 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 (http://www.cephb.fr/en/hgdp_panel.php).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/367/6484/eaay5012/suppl/DC1
Materials and Methods
Figs. S1 to S26
Tables S1 to S11
References (*47*–*77*)

View/request a protocol for this paper from *Bio-protocol*.

26 June 2019; accepted 4 February 2020
10.1126/science.aay5012

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 and Chris Tyler-Smith

Science **367** (6484), eaay5012.
DOI: 10.1126/science.aay5012

Genomes from around the globe

Genomic sequencing of diverse human populations to understand overall genetic diversity has lagged behind in-depth examination of specific populations. To add to our understanding of human genetic diversity, Bergström *et al.* generated whole-genome sequences surveying individuals in the Human Genome Diversity Project, which is a panel of global populations that has been instrumental in understanding the history of human populations. The authors' study adds data about African, Oceanian, and Amerindian populations and indicates that diversity tends to result from differences at the single-nucleotide level rather than copy number variation. An analysis of archaic sequences in modern populations identifies ancestral genetic variation in African populations that likely predates modern humans and has been lost in most non-African populations.

Science, this issue p. eaay5012

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.