RESEARCH ARTICLE SUMMARY

HUMAN EVOLUTION

A unified genealogy of modern and ancient genomes

Anthony Wilder Wohns, Yan Wong†, Ben Jeffery, Ali Akbari, Swapan Mallick, Ron Pinhasi, Nick Patterson, David Reich, Jerome Kelleher†, Gil McVean*†

INTRODUCTION: The characterization of modern and ancient human genome sequences has revealed previously unknown features of our evolutionary past. As genome data generation continues to accelerate—through the sequencing of population-scale biobanks and ancient samples from around the world—so does the potential to generate an increasingly detailed understanding of how populations have evolved.

However, such genomic datasets are highly heterogeneous. Samples from diverse times, geographic locations, and populations are processed, sequenced, and analyzed using a variety of techniques. The resulting datasets contain genuine variation but also complex patterns of missingness and error. This makes combining data challenging and hinders efforts to generate the most complete picture of human genomic variation.

RATIONALE: To address these challenges, we use the foundational notion that the ancestral relationships of all humans who have ever lived can be described by a single genealogy or tree sequence, so named because it encodes the sequence of trees that link individuals to one another at every point in the genome. This tree sequence of humanity is immensely complex, but estimates of the structure are a powerful means of integrating diverse datasets and gaining greater insights into human genetic diversity. In this work, we introduce statistical and computational methods to infer such a unified genealogy of modern and ancient samples, validate the methods through a mixture of computer simulation and analysis of empirical data, and apply the methods to reveal features of human diversity and evolution.

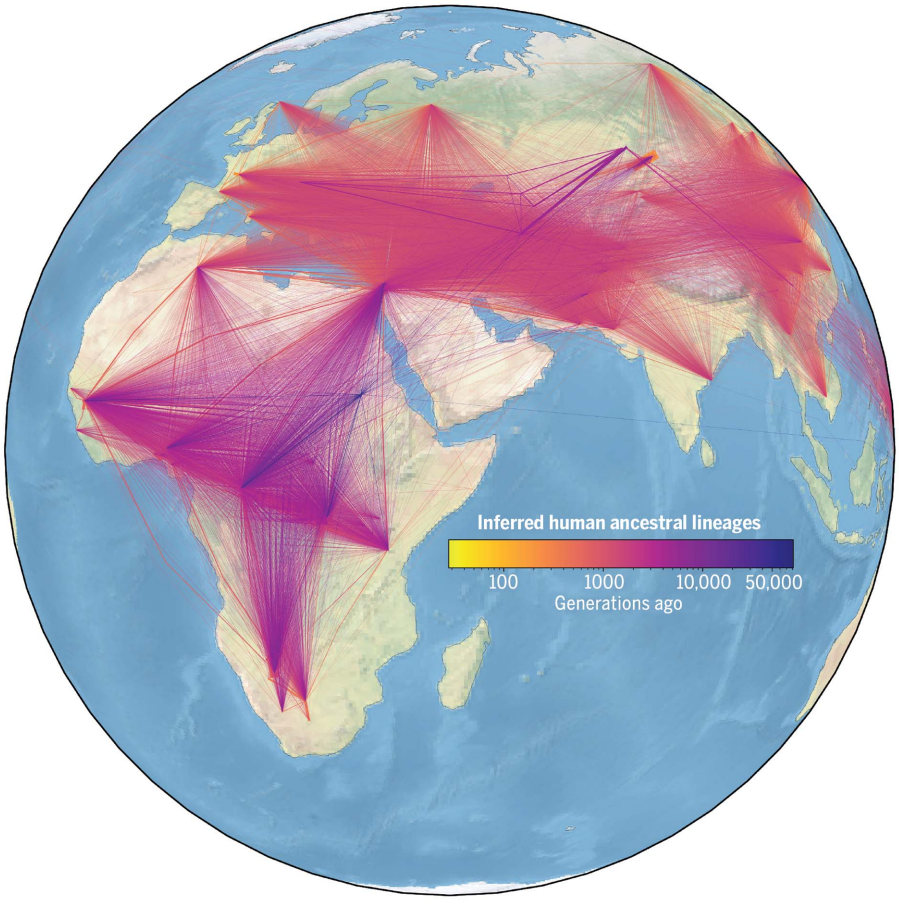
RESULTS: We present a unified tree sequence of 3601 modern and eight high-coverage ancient human genome sequences compiled from eight datasets. This structure is a lossless and compact representation of 27 million ancestral haplotype fragments and 231 million ancestral lineages linking genomes from these datasets back in time. The tree sequence also benefits from the use of an additional 3589 ancient samples compiled from more than 100 publications to constrain and date relationships.

Using simulations and empirical analyses, we demonstrate the ability to recover relationships between individuals and populations as well as to identify descendants of ancient samples. We calculate the distribution of the time to most recent common ancestry between the 215 populations of the constituent datasets, revealing patterns consistent with substantial variation in historical population size and evidence of archaic admixture in modern humans.

The tree sequence also offers insight into patterns of recurrent mutation and sequencing error in commonly used genetic datasets. We find pervasive signals of sequencing error as well as a small subset of variant sites that appear to be erroneous.

Finally, we introduce an estimator of ancestor geographic location that recapitulates key features of human history. We observe signals of very deep ancestral lineages in Africa, the out-of-Africa event, and archaic introgression in Oceania. The method motivates improved spatiotemporal inference methods that will better elucidate the paths and timings of historic migrations.

CONCLUSION: The profusion of genetic sequencing data creates challenges for integrating diverse data sources. Our results demonstrate that whole-genome genealogies provide a powerful platform for synthesizing genetic data and investigating human history and evolution. ■



Visualizing inferred human ancestral lineages over time and space. Each line represents an ancestor-descendant relationship in our inferred genealogy of modern and ancient genomes. The width of a line corresponds to how many times the relationship is observed, and lines are colored on the basis of the estimated age of the ancestor.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: gil.mcvean@bdi.ox.ac.uk
†These authors contributed equally to this work.
Cite this article as A. W. Wohns *et al.*, *Science* **375**, eabi8264 (2022). DOI: 10.1126/science.abi8264

RESEARCH ARTICLE

HUMAN EVOLUTION

A unified genealogy of modern and ancient genomes

Anthony Wilder Wohns[1,2], Yan Wong[2]†, Ben Jeffery[2], Ali Akbari[1,3,4], Swapan Mallick[1,5], Ron Pinhasi[6], Nick Patterson[1,3,4,5], David Reich[1,3,4,5], Jerome Kelleher[2]†, Gil McVean[2]*†

The sequencing of modern and ancient genomes from around the world has revolutionized our understanding of human history and evolution. However, the problem of how best to characterize ancestral relationships from the totality of human genomic variation remains unsolved. Here, we address this challenge with nonparametric methods that enable us to infer a unified genealogy of modern and ancient humans. This compact representation of multiple datasets explores the challenges of missing and erroneous data and uses ancient samples to constrain and date relationships. We demonstrate the power of the method to recover relationships between individuals and populations as well as to identify descendants of ancient samples. Finally, we introduce a simple nonparametric estimator of the geographical location of ancestors that recapitulates key events in human history.

Our ability to determine relationships among individuals, populations, and species is being transformed by population-scale biobanks of medical samples (*1*, *2*), collections of thousands of ancient genomes (*3*), and efforts to sequence millions of eukaryotic species for comparative genomic analyses (*4*). Such relationships, and the resulting distributions of genetic and phenotypic variation, reflect the complex set of selective, demographic, and molecular processes and events that have shaped species such as our own (*5*–*8*).

However, learning about evolutionary events and processes from the totality of genomic variation, in humans or other species, is challenging. Combining information from multiple datasets, even within a species, is technically demanding: Discrepancies between cohorts due to error (*9*), differing sequencing techniques (*10*, *11*), and variant processing (*12*) can lead to noise that can easily obscure genuine signal. Furthermore, few tools can cope with the vast datasets that arise from the combination of multiple sources (*13*). Also, statistical analysis typically relies on data-reduction techniques (*14*, *15*) or the fitting of parametric models (*16*–*19*), which may provide an incomplete picture of the complexities of evolutionary history. Finally, data access and governance restrictions often limit the ability to combine data sources (*20*).

[1]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [2]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7LF, UK. [3]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [4]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [5]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA. [6]Department of Evolutionary Anthropology, University of Vienna, 1090 Vienna, Austria.
*Corresponding author. Email: gil.mcvean@bdi.ox.ac.uk
†These authors contributed equally to this work.

The succinct tree sequence data structure provides a potential solution to many of these problems (*13*, *21*). Tree sequences extend the fundamental concept of a phylogenetic tree to multiple correlated trees along the genome, which is necessary when considering genealogies within recombining organisms (*22*). Notably, the tree sequence and the mapping of mutation events to it reflects the totality of what is knowable about genealogical relationships and the evolutionary history of individual variants. A tree sequence is defined as a graph with a set of nodes representing sampled chromosomes and ancestral haplotypes, edges connecting nodes representing lines of descent, and variable sites containing one or more mutations mapped onto the edges (Fig. 1A). Recombination events in the ancestral history of the sample create different edges and thus distinct but highly correlated trees along the genome. Tree sequences can not only be used to compress genetic data (*13*) but also lead to highly efficient algorithms for calculating population genetic statistics (*23*).

A unified genealogy of modern and ancient human genomes

Here, we introduce, validate, and apply nonparametric methods for inferring time-resolved tree sequences from multiple heterogeneous sources to efficiently infer a single, unified tree sequence of ancient and contemporary human genomes. Although humans are the focus of this study, the methods and approaches we introduce are valid for most recombining organisms.

To generate a unified genealogy of modern and ancient human genomes, we integrated data from three modern datasets: the 1000 Genomes Project (TGP), which contains 2548 sequenced individuals from 26 populations (*6*); the Human Genome Diversity Project (HGDP), which consists of 929 sequenced individuals from 54 populations (*8*); and the Simons Genome Diversity Project (SGDP), with 278 sequenced individuals from 142 populations (*7*). In total, 154 individuals appear in more than one of these datasets (*24*). Additionally, we included data from three high-coverage sequenced Neanderthal genomes (*25*–*27*), a single Denisovan genome (*28*), and high-coverage whole-genome data from a nuclear family of four (a mother, a father, and their two sons, with average coverages of 10.8×, 25.8×, 21.2×, and 25.3×, respectively) from the Afanasievo culture, who lived ~4.6 thousand years ago (ka) in the Altai Mountains of Russia (table S1). Finally, we used 3589 published ancient samples from >100 publications compiled by the Reich Laboratory (*24*) and three sequenced ancient samples—Loschbour, LBK-Stuttgart, and Ust'-Ishim (*5*, *29*)—to constrain allele age estimates. These ancient genomes were not included in the final tree sequence because of the lack of reliable phasing for most of the samples.

We built a unified genealogy from these datasets using an iterative approach (Fig. 1B). We first merged the modern datasets and inferred a tree sequence for each autosome using *tsinfer*, version 0.2 (*24*, *30*). We then estimated the age of ancestral haplotypes with *tsdate*, a Bayesian approach that infers the age of ancestral haplotypes with good accuracy and scaling properties (Fig. 1C and figs. S1 to S5) (*24*, *31*). Notably, *tsdate* can be used to date any valid tree sequence, not only those inferred by *tsinfer*. *tsdate* can also use ancient samples to improve date estimates (Fig. 1D). We identified 6,412,717 variants present in both ancient and modern samples. A lower bound on variant age is provided by the estimated archaeological date of the oldest ancient sample in which the derived allele is found. Where this was inconsistent with the initial inferred value (for 559,431 or 8.7% of variants), we used the archaeological date as the variant age.

Finally, we integrated the Afanasievo family and four archaic sequences with the modern samples and reinferred the tree sequence. The Afanasievo family has high coverage and comparably reliable haplotype phasing and was included to demonstrate the ability of our approach to incorporate high-quality ancient samples.

The integrated tree sequences of each autosome combined contain 26,958,720 inferred ancestral haplotype fragments, 231,073,278 edges, 91,172,114 variable sites, and 245,631,834 mutations. We infer that 38.7% of variant sites require more than one change in allelic state in the tree sequence to explain the data. This may indicate either recurrent mutations or errors, all of which are represented by additional mutations in the tree sequence. If we discount mutations that are likely indicative of sequencing errors (*24*), we find that

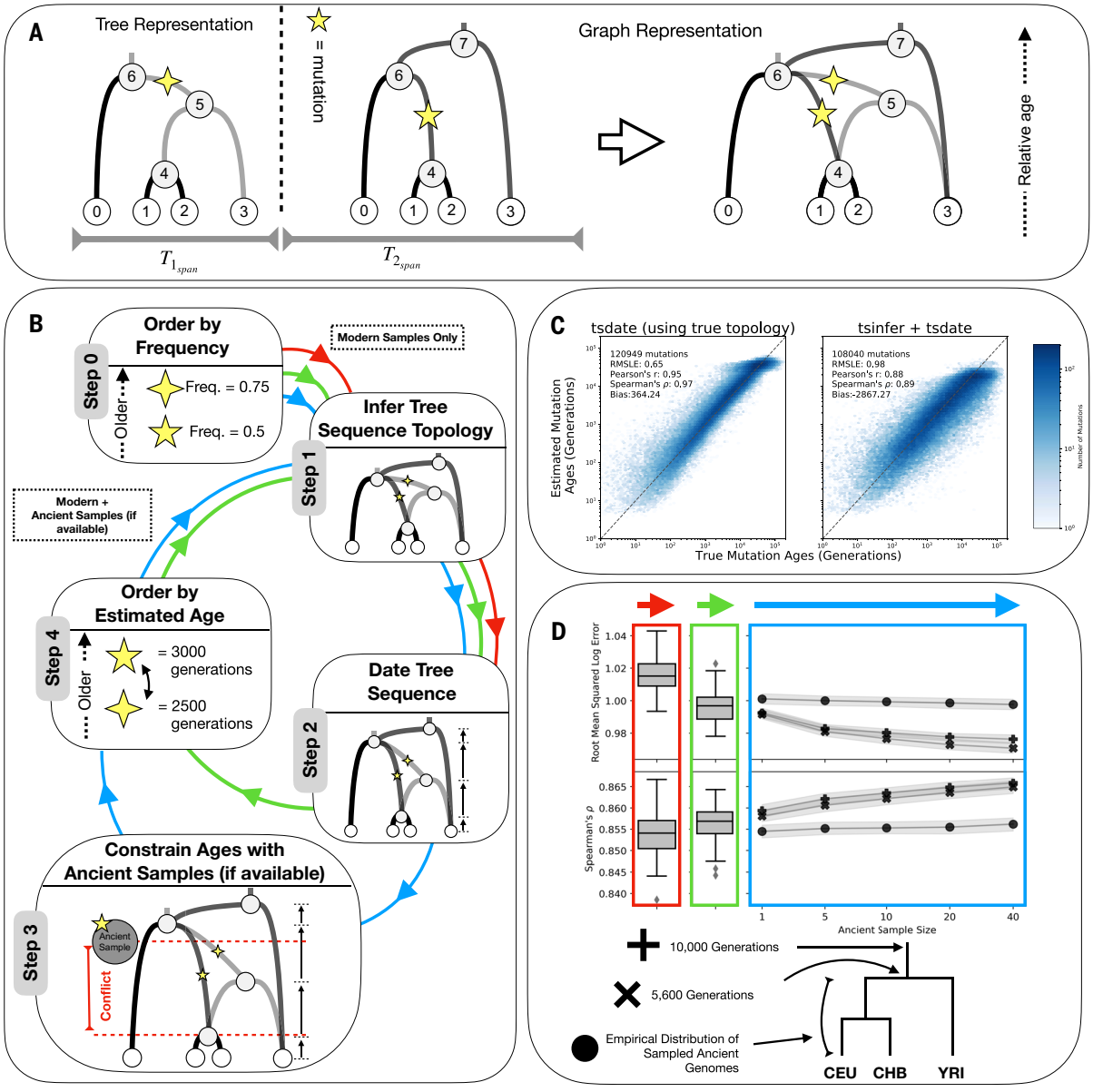


Fig. 1. Schematic overview and validation of the inference methodology. (**A**) An example tree sequence topology with four samples (nodes 0 to 3), two marginal trees, four ancestral haplotypes (nodes 4 to 7), and two mutations. T_{span} measures the genomic span of each marginal tree topology, with the dotted line indicating the location of a recombination event. The graph representation is equivalent to the tree representation. (**B**) Schematic representation of the inference methodology. Step 0: Alleles are ordered by frequency (freq.); the mutation represented by the four-point star is considered to be older. Step 1: The tree sequence topology is inferred with *tsinfer* using modern samples. Step 2: The tree sequence is dated with *tsdate*. Step 3: Node date estimates are constrained with the known age of ancient samples. Step 4: Ancestral haplotypes are reordered by the estimated age of their focal mutation; the five-pointed star mutation is now inferred to be older. The algorithm returns to step 1 to reinfer the tree sequence topology with ancient samples. Arrows refer to modes of operation: steps 0, 1, and 2 only (red); steps 0, 1, 2, 4, 1, and 2 (green); or steps 0, 1, 2, 3, 4, 1, and 2 (blue) (*24*). (**C**) Scatter plots and accuracy metrics comparing simulated (*x* axis) and inferred (*y* axis) mutation ages from *msprime* neutral coalescent simulations, using *tsdate* with the simulated topology (left) and inferred topology from *tsinfer* (right). RMSLE, root mean squared log error. (**D**) Accuracy metrics, RMSLE (top), and Spearman rank correlation coefficient (*r*) (bottom), with modern samples only (first panel), after one round of iteration (second panel), and with increasing numbers of ancient samples (third panel) [colored arrows as in (B)]. Ancient samples from three eras of human history are considered, as in the schematic (*24*). CEU, Utah residents with Northern and Western European Ancestry; CHB, Han Chinese; YRI, Yorubans.

13,513,873 sites contain at least two mutations affecting more than one sample, which implies that up to 17.5% of variable sites could result from more than one ancestral mutation. A high proportion of sites with more than ~100 mutations on chromosome 20 have sequencing or alignment quality issues as defined by the TGP accessibility mask (*6*) or are in minimal linkage disequilibrium to their surrounding sites (fig. S6), which suggests that they are largely erroneous. Moreover, analysis of data simulated with an empirically calibrated error profile and evaluation of the enrichment of multiple mutations at sites with known elevated mutation rates suggests that most of

the multiple mutations we identify are likely explained by error, but a minority (~20%) are the result of genuine recurrence or back mutation (*24*). We chose to retain such sites so that our inferred tree sequences are lossless representations of the original data sources; however, future iterative approaches to the removal of such probable errors are likely to improve use cases, such as imputation.

To characterize fine-scale patterns of relatedness between the 215 populations of the constituent datasets, we estimated the time to the most recent common ancestor (TMRCA) between pairs of haplotypes from these populations at the 122,637 distinct trees in the tree sequence of chromosome 20 (~300 billion pairwise TMRCAs). In this and other analyses, we present data from this chromosome because they are representative of genome-wide patterns. After performing hierarchical clustering on the average pairwise TMRCA values, we find that samples do not cluster by data source (which would indicate artifacts) but reflect patterns of global relatedness (Fig. 2 and the external interactive figure). We conclude that our method of integrating datasets is therefore robust to biases introduced by different datasets.

In this genealogy, numerous features of human history are immediately apparent, such as the deep divergence of archaic and modern humans, the effects of the out-of-Africa event (Fig. 2A), and a subtle increase in Oceanian and

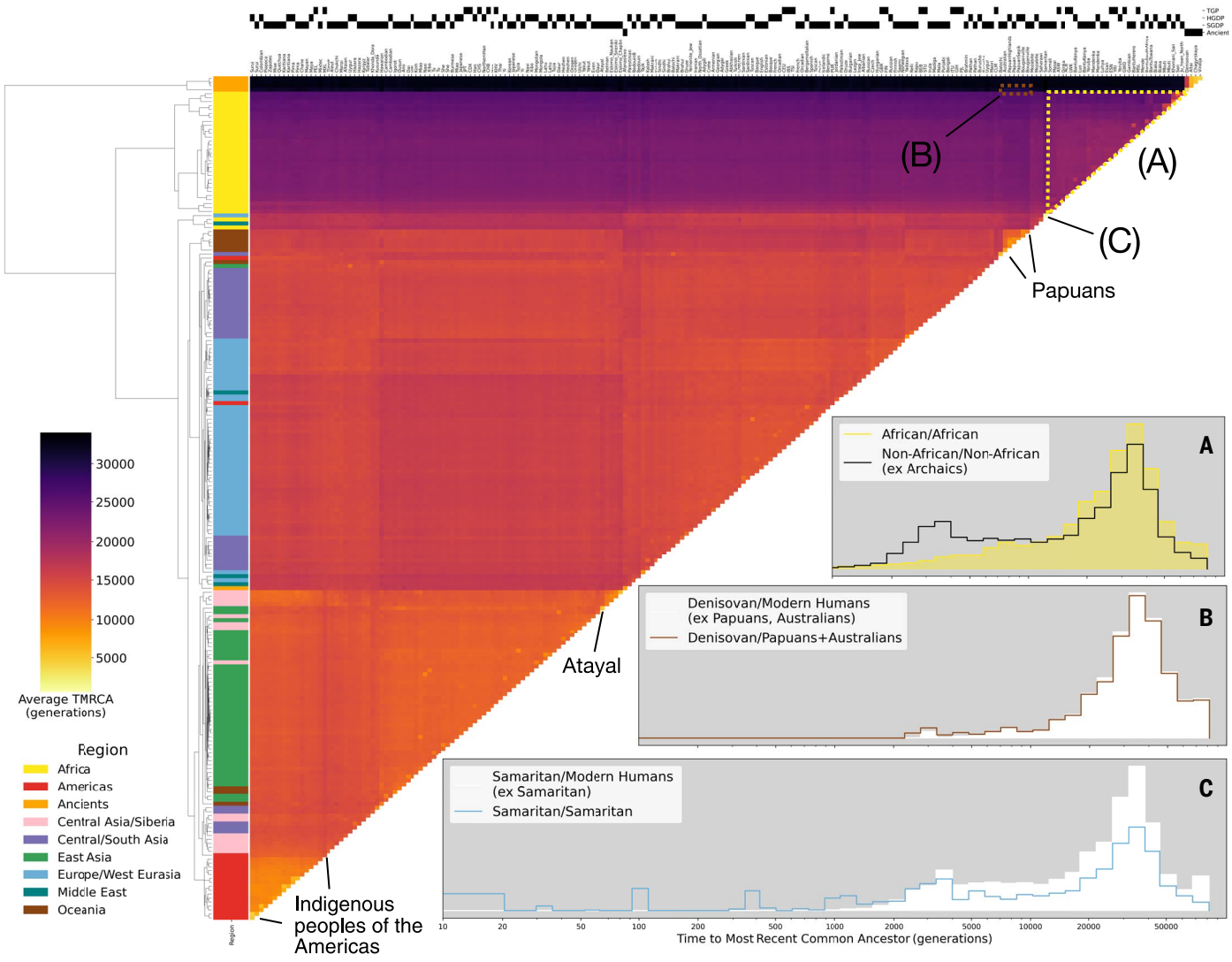


Fig. 2. Clustered heatmap showing the average TMRCA on chromosome 20 for haplotypes within pairs of the 215 populations in the HGDP, TGP, SGDP, and ancient samples. Each cell in the heatmap is colored by the logarithmic mean TMRCA of samples from the two populations. Hierarchical clustering of rows and columns has been performed using the unweighted pair group method with arithmetic mean (UPGMA) algorithm on the value of the pairwise average TMRCAs. Row colors are given by the region of origin for each population, as shown in the legend. The source of genomic samples for each population is indicated in the shaded boxes above the column labels. Three population relationships are highlighted using span-weighted histograms of the TMRCA distributions: (**A**) Average distribution of TMRCAs between all non-African populations (black line) compared with African/African TMRCAs (solid yellow). (**B**) Denisovan and Papuan/Australian TMRCAs (solid line) compared with the Denisovan against all nonarchaic populations (solid white). This subtle but specific signal of elevated recent ancestry between the Denisovan and Papuans/Australians is particularly evident in the external interactive figure. (**C**) TMRCAs between the two Samaritan chromosomes (solid line) compared with the Samaritans/all other modern humans (solid white). Selected populations with particularly recent within-group TMRCAs are indicated. Duplicate samples appearing in more than one modern dataset are included in this analysis. The external interactive figure is an interactive version of this figure that is available at https://awohns.github.io/unified_genealogy/interactive_figure.html.

Denisovan most recent common ancestor (MRCA) density from 2000 to 5000 generations ago (Fig. 2B). Multiple populations show recent within-group TMRCAs, which is suggestive of recent bottlenecks or consanguinity. The most extreme cases occur when a population consists of a single individual in our dataset, such as the Samaritan individual from the SGDP, where we see a logarithmic average within-group TMRCA of ~1000 generations, which is caused by multiple MRCAs at very recent times (Fig. 2C) and is consistent with a severe bottleneck and consanguinity in recent centuries (*32*). Indigenous populations in the Americas, an Atayal individual from Taiwan, and Papuans also exhibit particularly recent within-group TMRCAs (Fig. 2).

Tree sequence–based analysis of descent from ancient sequences

To validate the dating methodology, we used simulations to show that the integration of ancient samples improves derived allele age estimates under a range of demographic histories (Fig. 1D). To provide empirical validation of the method, we tested how best to infer allele ages that are consistent with observations from ancient samples. Thus, we inferred and dated a tree sequence of TGP chromosome 20 (without using ancient samples) and compared the resulting point estimates and upper and lower bounds on allele age with results from *GEVA* (*33*) and *Relate* (*34*). This resulted in a set of 659,804 variant sites where all three methods provide an allele age estimate. Of these, 76,889 derived alleles are observed within the combined set of 3734 ancient genome samples, thus putting a lower bound on allele age. The estimated allele ages from *tsdate* and *Relate* showed the greatest compatibility with ancient lower bounds, despite the fact that the mean age estimate from *tsdate* is more recent than that of *Relate* (Fig. 3A) (*24*).

Next, to assess the ability of the unified tree sequence to recover known relationships between ancient and modern populations, we considered the patterns of descent to modern samples from Archaic sequences on chromosome 20. Simulations indicate that this approach detects introgressed genetic material from Denisovans at a precision of ~86% with a recall of ~61% (*24*). We find descendants among nonarchaic individuals, including both modern individuals and the Afanasievo, for 13% of the span of the Denisovan haplotypes on chromosome 20. The highest degree of descent among modern humans is in Oceanian populations, as previously reported (*28*, *35*–*37*) (Fig. 3B). However, the tree sequence also reveals how both the extent and nature of descent from Denisovan haplotypes vary greatly among modern humans. In particular, we find that Papuans and Australians carry multiple fragments of Denisovan haplotypes that are largely specific to the individual (Fig. 3C). By contrast, other modern descendants of Denisovan haplotypes have fewer blocks that are more widely shared, often between geographically distant individuals.

Examining the other ancient samples in the unified genealogy, we find the greatest amount of descent from the haplotypes of the Afanasievo

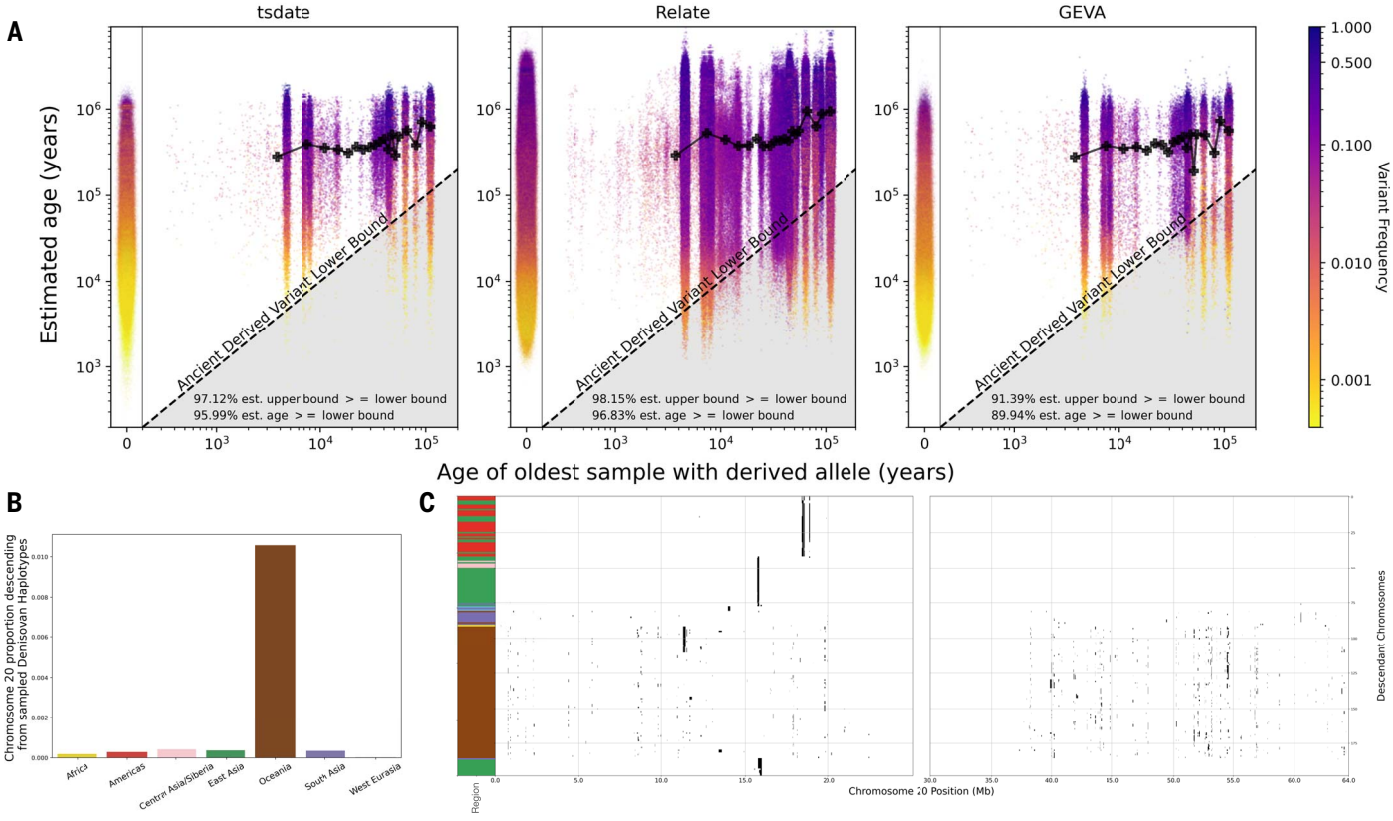


Fig. 3. Validation of inference methods using ancient samples. (**A**) Comparison of mutation age estimates from *tsdate*, *Relate*, and *GEVA* with 3734 ancient samples at 76,889 variants on chromosome 20 (note that *Relate* estimates ages separately for each population in which a variant is found). The radiocarbon-dated age of the oldest ancient sample carrying a derived allele at each variant site in the TGP is used as the lower bound on the age of the mutation (diagonal lines). Mutations below this line have an estimated age that is inconsistent with the age of the ancient sample. Black lines on each plot show the moving average of allele age estimates from each method as a function of oldest ancient sample age. Plots to the left show the distribution of allele age estimates for modern-only variants from each respective method. Additional metrics are reported in each plot. (**B**) Percentage of chromosome 20 for modern samples in each region that is inferred to descend from Denisovan haplotypes, calculated with the genomic descent statistic (*57*). (**C**) Tracts of descent along chromosome 20 descending from Denisovan haplotypes in modern samples with at least 100 kilobases (kb) of total descent (colors as in Fig. 2).

family among individuals in Western Eurasia and South Asia (fig. S7A), consistent with findings from the genetically similar Yamnaya peoples and supporting a contemporaneous diffusion of Afanasievo-like genetic material via multiple routes (*38*). For the Neanderthals, where there are three samples of different ages, our simulations indicate that interpretation of the descent statistics is complicated by varying levels of precision and recall among lineages. Nevertheless, recall is highest at regions where introgressing and sampled archaic lineages share more recent common ancestry, and precision is higher for the Vindija sample, which is more closely related to introgressing lineages. Examining patterns of descent from Vindija haplotypes across autosomes indicates that modern non-African groups carry similar levels of Vindija-like material (fig. S8), which supports suggestions that the proportions are similar between East Asians and West Eurasians (*39*) and is inconsistent with other studies (*26*, *40*).

Nonparametric inference of spatiotemporal dynamics in human history

Tree sequence–based analysis of ancient samples demonstrates an ability to characterize patterns of recent descent. We developed a simple estimator of ancestor spatial location that uses the coordinates of descendants of a node combined with the structure of the tree sequence to provide an estimate of ancestors' geographic position (*24*). Briefly, this is accomplished by determining the coordinates of a parent node in the tree sequence as the midpoint of its immediate children (*24*), an approach that performs well in simple simulations (fig. S9). The approach can use information on the location of ancient samples, although it does not attempt to capture the geographical plausibility of different locations and routes. The inferred locations are thus a model-free estimate of ancestors' locations, informed by the tree sequence topology and geographic distribution of samples.

We applied our method to the unified tree sequence of chromosome 20, excluding TGP individuals (which lack precise location information). We found that the inferred ancestor location recovers multiple key events in human history (Fig. 4 and movie S1). Despite the fact that the geographic center of sampled individuals is in Central Asia, by 72 ka, the average location of ancestral haplotypes is in Northeast Africa and remains there until the oldest common ancestors are reached. In fact, the inferred geographic center of gravity of the 100 oldest ancestral haplotypes (which have an average age of ~2 million years) is located in Sudan at 19.4°N, 33.7°E. These findings reflect the depth of African lineages in the inferred tree sequence and are compatible with well-dated early modern human fossils from eastern and northern Africa (*41*, *42*). We caution that if we analyzed data from a grid sampling of populations in Africa, the geographic center of gravity of independent lineages at different time depths would shift. Additionally, migrations occurring within the past few thousand years (*43*, *44*) mean that present-day distributions of groups in Africa and elsewhere may not represent those of their ancestors, and thus we may have a distorted picture of ancient geographic distributions (*45*). Nevertheless, the deep tree structure is geographically centered in Africa in autosomal data, just as it is for mitochondrial DNA and Y chromosomes (*46*, *47*).

By 280 ka, the estimated geographic center of human ancestors is still located in Africa, but many ancestors are also observed in the Middle East and Central Asia, and a few are located in Papua New Guinea. At 140 ka, more ancestors

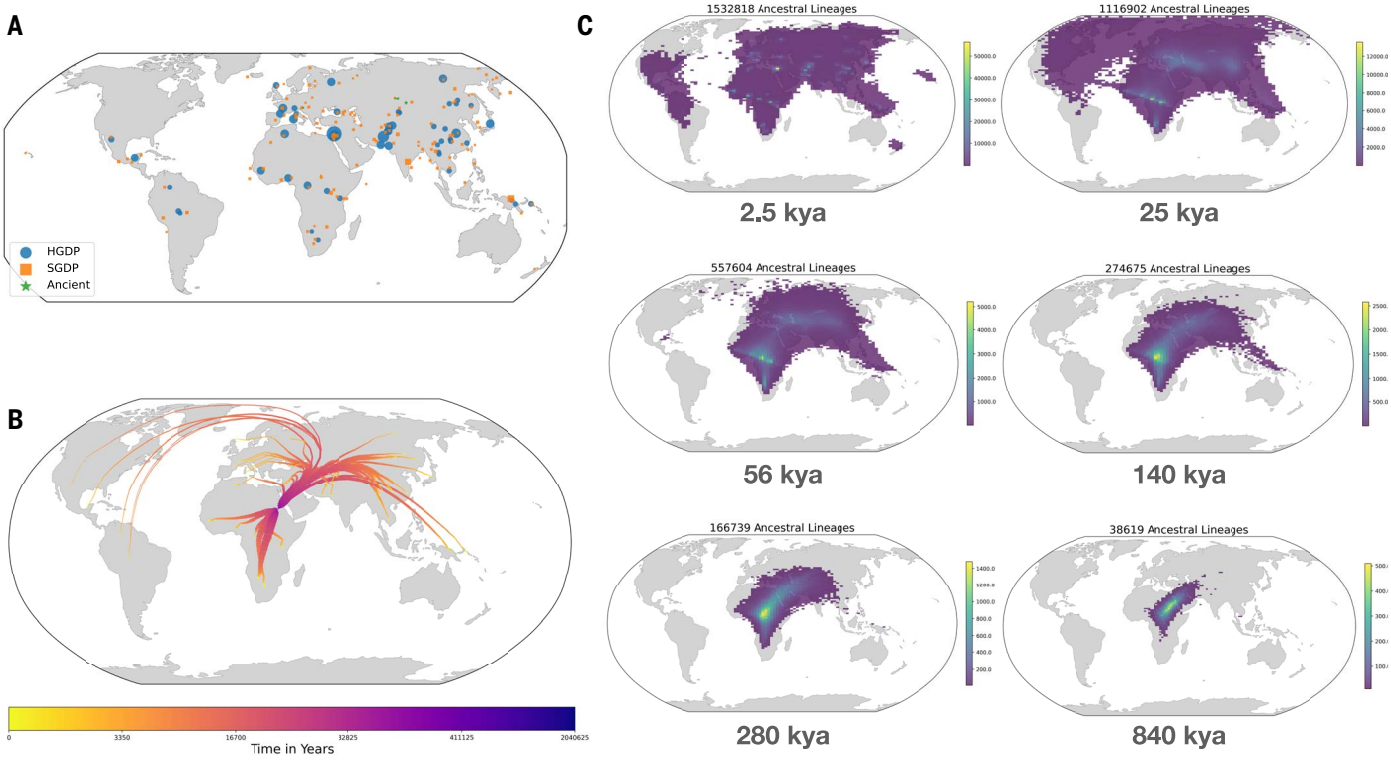


Fig. 4. Visualization of the nonparametric estimator of ancestor geographic location for HGDP, SGDP, Neanderthal, Denisovan, and Afanasievo samples on chromosome 20. (**A**) Geographic location of samples used to infer ancestral geography. The size of each symbol is proportional to the number of samples in that population. (**B**) The average location of the ancestors of each HGDP population from time t = 0 to ~2 million years ago. The width of lines is proportional to the number of ancestors of each population over time. The ancestor of a population is defined as an inferred ancestral haplotype with at least one descendant in that population. (**C**) Two-dimensional histograms showing the inferred geographical location of HGDP ancestral lineages at six time points. Histogram bins with <10 ancestors are not shown. The geographic concentration of ancestors at more recent times is an artifact of uneven sampling and our geographic inference method.

are found in Papua New Guinea. This is almost 100 thousand years before the earliest documented human habitation of the region (*48*). However, our findings are potentially consistent with the proposed time scales of deeply diverged Denisovan lineages specific to Papuans (*37*) and possibly with admixture with unsampled ghost lineages. At 56 ka, some ancestral lineages are observed in the Americas, which is earlier than the estimated migration times to the Americas (*49*). This effect is possibly attributable to the presence of ancestors that predate the migration and did not live in the Americas but whose descendants now exist solely in this region (*50*); the same effect may also explain observations from Papua New Guinea. Additional ancient samples and more-sophisticated inference approaches are required to distinguish between these hypotheses because there remains considerable uncertainty about the true age of any single ancestor (*24*). Nevertheless, these results demonstrate the ability of inference methods applied to tree sequences to capture key features of human history in a manner that does not require complex parametric modeling.

Discussion

A central theme in evolutionary biology is how best to represent and analyze genomic diversity to learn about the processes, forces, and events that have shaped organismal history. Historically, many modeling approaches have focused on the temporal behavior of individual mutation frequencies in idealized populations (*51*, *52*). More recently, modeling techniques have shifted to focus on the genealogical history of sampled genomes and the correlation structures that arise through recombination (*22*, *53*). Notably, a single (albeit extremely complex) set of ancestral relationships exists that, coupled with how mutation events have altered genetic material through descent, describes what we observe today.

However, developing efficient methods for inferring the underlying genealogy has proved challenging (*54*, *55*). The methods described here produce high-quality dated genealogies that include thousands of modern and ancient samples. These genealogies cannot be entirely accurate; nevertheless, they enable a wealth of analyses that reveal features of human evolution (*23*, *56*–*60*). That our highly simplistic geographic estimator captures key events suggests that more-sophisticated approaches, coupled with the ongoing program of sequencing ancient samples, will continue to generate insights into our history. Specifically, the methods developed here provide a framework for testing different models of human migration and demographic history, such as Neanderthal absorption models (*61*), using a parametric and explicitly spatial simulation framework. However, the accuracy of any ancestral geographic inference method will be limited when the distribution of sampled individuals does not reflect the location of the samples' ancestors.

Our study also highlights the importance of accommodating genotype error and recurrent mutation in the analysis of genomic variation. Although a large number of sites are inferred to carry multiple mutations, we find that most of these likely reflect genotype error and potentially errors arising from paralogy (particularly at sites requiring high numbers of mutations), although there remains a substantial signal of recurrent mutation, as previously reported (*62*, *63*). Similarly, we find some evidence for certain classes of error in ancient sequences leading to false correction of variant ages. We choose to retain all additional mutations in the analyses described in this paper, including those that are highly likely to reflect sequencing error, because this reflects the input data used to build the tree sequence, and any effort to remove mutations corresponding to errors will itself introduce bias. We caution that the absolute ages we report have some degree of error, in part as a result of these errors in the sequencing datasets. Estimates from simulations show that genotype error may cause an upward bias of up to 16% in age estimates derived from modern samples (fig. S3), but we also find that removing sites that are highly likely to be erroneous has a marginal effect on age estimates (fig. S10). Improving methods to detect and correct or mitigate against the effect of genotype errors is an important direction for future research.

Because the tree sequence approach aims to capture the structure of human relationships and genomic diversity, it provides a principled basis for combining data from multiple different sources, not just correcting errors but also enabling tasks such as imputing missing data. Although additional work is required to integrate other types of mutation, a reference tree sequence for human variation—along with the tools to use it appropriately (*13*, *23*)—potentially represents a basis for harmonizing much larger and wider sets of genomic data sources and enabling cross–data source analyses. We note that reference tree sequences could also enable data sharing and preserve privacy in genomic analysis (*20*) through the compression of cohorts against such a reference structure.

There exists room for improvement as well as opportunities for genomic analyses that use the dated tree sequence structure. Our approach requires phased genomes, a particular challenge for ancient samples. However, it should be possible to use a diploid version of the matching algorithm in *tsinfer* to jointly solve phasing and imputation. This also has the potential to alleviate biases introduced by using modern and genetically distant reference panels for ancient samples (*64*). Additionally, our approach to age inference within *tsdate* only provides an approximate solution to the cycles that are inherent in genealogical histories (*65*) and could be extended to model heterogeneity in mutation rates. There are also many possible approaches for improving the sophistication of spatiotemporal ancestor inference.

The unified genealogy presented in this work represents a foundation for building a comprehensive understanding of human genomic diversity, including both modern and ancient samples, which enables applications ranging from improving genome interpretation to deciphering our earliest roots. Although much work is still required to build the genealogy of everyone, the methods presented here provide a solution to this fundamental task.

Materials and methods summary

Dated tree sequences were constructed from the TGP (*6*), the SGDP (*7*), the HGDP (*8*), three Neanderthal genomes (*25*–*27*), and the Denisovan genome (*28*), and we added a dataset from a nuclear family of four from the Afanasievo culture who lived ~4.6 ka, sequenced to a depth of between 10.8× and 25.8×. First, tree sequence topologies were estimated using *tsinfer* (*13*), updated (version 0.2.0) to handle missing data and detect potential genotype errors and recurrent mutations. Subsequently, the dates of ancestral haplotypes were obtained with a new algorithm, *tsdate*—an approximate Bayesian method that estimates a joint posterior distribution for the nodes in a tree sequence using mutations inferred from the input sequences, inferred ancestors, and tree sequence topology. Ages in the unified genealogy were constrained by radiocarbon-dated ancient samples from the Allen Ancient DNA resource (*5*–*7*, *18*, *24*, *25*, *26*, *28*, *29*, *38*, *44*, *50*, *66*–*172*) as well as from the Loschbour, LBK-Stuttgart, and Ust'-Ishim (*5*, *29*) sequenced ancient samples. The geographic location of ancestral haplotypes was estimated from the inferred tree sequence topology using a weighted sum of the daughter node geographic locations converted to Cartesian coordinates. Coalescent simulations for method evaluation were performed using *msprime* (*21*) and *stdpopsim* (*173*). Full details of algorithms, data sources, data processing steps, and simulations are provided in the supplementary materials.

REFERENCES AND NOTES

1. C. Bycroft *et al.*, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018). doi: 10.1038/s41586-018-0579-z; pmid: 30305743
2. D. Taliun *et al.*, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021). doi: 10.1038/s41586-021-03205-y; pmid: 33568819
3. D. Reich, *Who We Are and How We Got Here: Ancient DNA and the New Science of the Human Past* (Oxford Univ. Press, 2018).
4. H. A. Lewin *et al.*, Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4325–4333 (2018). doi: 10.1073/pnas.1720115115; pmid: 29686065

5. I. Lazaridis *et al*., Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014). doi: 10.1038/nature13673; pmid: 25230663
6. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi: 10.1038/nature15393; pmid: 26432245
7. 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
8. A. Bergström *et al*., Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020). doi: 10.1126/science.aay5012; pmid: 32193295
9. S. Belsare *et al*., Evaluating the quality of the 1000 genomes project data. *BMC Genomics* **20**, 620 (2019). doi: 10.1186/s12864-019-5957-x; pmid: 31416423
10. L. Shi *et al*., Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* **7**, 12065 (2016). doi: 10.1038/ncomms12065; pmid: 27356984
11. A. M. Wenger *et al*., Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019). doi: 10.1038/s41587-019-0217-9; pmid: 31406327
12. S. Hwang, E. Kim, I. Lee, E. M. Marcotte, Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* **5**, 17875 (2015). doi: 10.1038/srep17875; pmid: 26639839
13. J. Kelleher *et al*., Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330–1338 (2019). doi: 10.1038/s41588-019-0483-y; pmid: 31477934
14. L. L. Cavalli-Sforza, M. W. Feldman, The application of molecular genetic approaches to the study of human evolution. *Nat. Genet.* **33**, 266–275 (2003). doi: 10.1038/ng1113; pmid: 12610536
15. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLOS Genet.* **2**, e190 (2006). doi: 10.1371/journal.pgen.0020190; pmid: 17194218
16. J. K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000). doi: 10.1093/genetics/155.2.945; pmid: 10835412
17. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLOS Genet.* **8**, e1002453 (2012). doi: 10.1371/journal.pgen.1002453; pmid: 22291602
18. N. Patterson *et al*., Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012). doi: 10.1534/genetics.112.145037; pmid: 22960212
19. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genet.* **8**, e1002967 (2012). doi: 10.1371/journal.pgen.1002967; pmid: 23166502
20. L. Bonomi, Y. Huang, L. Ohno-Machado, Privacy challenges and research opportunities for genomic data sharing. *Nat. Genet.* **52**, 646–654 (2020). doi: 10.1038/s41588-020-0651-0; pmid: 32601475
21. J. Kelleher, A. M. Etheridge, G. McVean, Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Comput. Biol.* **12**, e1004842 (2016). doi: 10.1371/journal.pcbi.1004842; pmid: 27145223
22. R. R. Hudson, Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983). doi: 10.1016/0040-5809(83)90013-8; pmid: 6612631
23. P. Ralph, K. Thornton, J. Kelleher, Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes. *Genetics* **215**, 779–797 (2020). doi: 10.1534/genetics.120.303253; pmid: 32357960
24. Materials and methods are available as supplementary materials online.
25. 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
26. 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
27. F. Mafessoni *et al*., A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 15132–15136 (2020). doi: 10.1073/pnas.2004944117; pmid: 32546518
28. 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
29. Q. Fu *et al*., Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014). doi: 10.1038/nature13810; pmid: 25341783
30. J. Kelleher, Y. Wong, B. Jeffery, A. W. Wohns, I. Rebollo, D. Rodriguez Fernandez, tskit-dev/tsinfer: 0.2.1, version 0.2.1, Zenodo (2021); https://doi.org/10.5281/zenodo.5168051.
31. A. W. Wohns, Y. Wong, B. Jeffery, tskit-dev/tsdate: Minor feature and bugfix release, version 0.1.4, Zenodo (2021); https://doi.org/10.5281/zenodo.5168040.
32. B. Bonne, The Samaritans: A demographic study. *Hum. Biol.* **35**, 61–89 (1963). pmid: 13968819
33. P. K. Albers, G. McVean, Dating genomic variants and shared ancestry in population-scale sequencing data. *PLOS Biol.* **18**, e3000586 (2020). doi: 10.1371/journal.pbio.3000586 pmid: 31951611
34. L. Speidel, M. Forest, S. Shi, S. R. Myers, A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019). doi: 10.1038/s41588-019-0484-x; pmid: 31477933
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. 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). doi: 10.1016/j.ajhg.2011.09.005; pmid: 21944045
37. 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
38. V. M. Narasimhan *et al*., The formation of human populations in South and Central Asia. *Science* **365**, eaat7487 (2019). doi: 10.1126/science.aat7487; pmid: 31488661
39. L. Chen, A. B. Wolf, W. Fu, L. Li, J. M. Akey, Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell* **180**, 677–687.e16 (2020). doi: 10.1016/j.cell.2020.01.012; pmid: 32004458
40. J. D. Wall *et al*., Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics* **194**, 199–209 (2013). doi: 10.1534/genetics.112.148213; pmid: 23410836
41. I. McDougall, F. H. Brown, J. G. Fleagle, Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **433**, 733–736 (2005). doi: 10.1038/nature03258; pmid: 15716951
42. J.-J. Hublin *et al*., New fossils from Jebel Irhoud, Morocco and the pan-African origin of *Homo sapiens*. *Nature* **546**, 289–292 (2017). doi: 10.1038/nature22336; pmid: 28593953
43. K. Wang *et al*., Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-Saharan Africa. *Sci. Adv.* **6**, eaaz0183 (2020). doi: 10.1126/sciadv.aaz0183; pmid: 32582847
44. M. E. Prendergast *et al*., Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. *Science* **365**, eaaw6275 (2019). doi: 10.1126/science.aaw6275; pmid: 31147405
45. A. Kalkauskas *et al*., Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLOS Comput. Biol.* **17**, e1008561 (2021). doi: 10.1371/journal.pcbi.1008561
46. L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes, A. C. Wilson, African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507 (1991). doi: 10.1126/science.1840702; pmid: 1840702
47. P. A. Underhill *et al*., The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65**, 43–62 (2001). doi: 10.1046/j.1469-1809.2001.6510043.x; pmid: 11415522
48. J. F. O'Connell, J. Allen, The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. *J. Archaeol. Sci.* **56**, 73–84 (2015). doi: 10.1016/j.jas.2015.02.020
49. 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
50. J. V. Moreno-Mayar *et al*., Early human dispersals within the Americas. *Science* **362**, eaav2621 (2018). doi: 10.1126/science.aav2621; pmid: 30409807
51. R. A. Fisher, *The Genetical Theory of Natural Selection* (Clarendon, 1930).
52. S. Wright, Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931). doi: 10.1093/genetics/16.2.97; pmid: 17246615
53. J. F. C. Kingman, The coalescent. *Stochastic Process. Appl.* **13**, 235–248 (1982). doi: 10.1016/0304-4149(82)90011-4
54. G. A. T. McVean, N. J. Cardin, Approximating the coalescent with recombination. *Phil. Trans. R. Soc. B* **360**, 1387–1393 (2005). doi: 10.1098/rstb.2005.1673; pmid: 16048782
55. M. D. Rasmussen, M. J. Hubisz, I. Gronau, A. Siepel, Genome-wide inference of ancestral recombination graphs. *PLOS Genet.* **10**, e1004342 (2014). doi: 10.1371/journal.pgen.1004342; pmid: 24831947
56. K. Harris, From a database of genomes to a forest of evolutionary trees. *Nat. Genet.* **51**, 1306–1307 (2019). doi: 10.1038/s41588-019-0492-x; pmid: 31477932
57. C. L. Scheib *et al*., East Anglian early Neolithic monument burial linked to contemporary Megaliths. *Ann. Hum. Biol.* **46**, 145–149 (2019). doi: 10.1080/03014460.2019.1623912; pmid: 31184205
58. A. J. Stern, L. Speidel, N. A. Zaitlen, R. Nielsen, Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *Am. J. Hum. Genet.* **108**, 219–239 (2021). doi: 10.1016/j.ajhg.2020.12.005; pmid: 33440170
59. L. Speidel *et al*., Inferring population histories for ancient genomes using genome-wide genealogies. *Mol. Biol. Evol.* **38**, 3497–3511 (2021). doi: 10.1093/molbev/msab174
60. N. K. Schaefer, B. Shapiro, R. E. Green, An ancestral recombination graph of human, Neanderthal, and Denisovan genomes. *Sci. Adv.* **7**, eabc0776 (2021). doi: 10.1126/sciadv.abc0776; pmid: 34272242
61. R. Nielsen *et al*., Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017). doi: 10.1038/nature21347; pmid: 28102248
62. J. J. Michaelson *et al*., Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431–1442 (2012). doi: 10.1016/j.cell.2012.11.019; pmid: 23260136
63. A. V. Nesta, D. Tafur, C. R. Beck, Hotspots of Human Mutation. *Trends Genet.* **37**, 717–729 (2021). doi: 10.1016/j.tig.2020.10.003; pmid: 33199048
64. R. Hui, E. D'Atanasio, L. M. Cassidy, C. L. Scheib, T. Kivisild, Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Sci. Rep.* **10**, 18542 (2020). doi: 10.1038/s41598-020-75387-w; pmid: 33122697
65. K. P. Murphy, Y. Weiss, M. I. Jordan, in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, K. B. Laskey, H. Prade, Eds. (Morgan Kaufmann Publishers Inc., 1999), pp. 467–475.
66. M. E. Allentoft *et al*., Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015). doi: 10.1038/nature14507; pmid: 26062507
67. C. E. G. Amorim *et al*., Understanding 6th-century barbarian social organization and migration through paleogenomics. *Nat. Commun.* **9**, 3547 (2018). doi: 10.1038/s41467-018-06024-4; pmid: 30206220
68. M. L. Antonio *et al*., Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science* **366**, 708–714 (2019). doi: 10.1126/science.aay6826; pmid: 31699931
69. S. Brace *et al*., Ancient genomes indicate population replacement in Early Neolithic Britain. *Nat. Ecol. Evol.* **3**, 765–771 (2019). doi: 10.1038/s41559-019-0871-9; pmid: 30988490
70. F. Broushaki *et al*., Early Neolithic genomes from the eastern Fertile Crescent. *Science* **353**, 499–503 (2016). doi: 10.1126/science.aaf7943; pmid: 27417496
71. L. M. Cassidy *et al*., Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 368–373 (2016). doi: 10.1073/pnas.1518445113; pmid: 26712024
72. P. B. Damgaard *et al*., 137 ancient human genomes from across the Eurasian steppes. *Nature* **557**, 369–374 (2018). doi: 10.1038/s41586-018-0094-2; pmid: 29743675
73. P. de Barros Damgaard *et al*., The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* **360**, eaar7711 (2018). doi: 10.1126/science.aar7711; pmid: 29743352
74. S. S. Ebenesersdóttir *et al*., Ancient genomes from Iceland reveal the making of a human population. *Science* **360**, 1028–1032 (2018). doi: 10.1126/science.aar2625; pmid: 29853688
75. M. Feldman *et al*., Late Pleistocene human genome suggests a local origin for the first farmers of central Anatolia. *Nat. Commun.* **10**, 1218 (2019). doi: 10.1038/s41467-019-09209-7; pmid: 30890703
76. M. Feldman *et al*., Ancient DNA sheds light on the genetic origins of early Iron Age Philistines. *Sci. Adv.* **5**, eaax0061 (2019). doi: 10.1126/sciadv.aax0061; pmid: 31281897

77. C. de la Fuente *et al*., Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4006–E4012 (2018). doi: 10.1073/pnas.1715688115; pmid: 29632188
78. D. M. Fernandes *et al*., A genomic Neolithic time transect of hunter-farmer admixture in central Poland. *Sci. Rep.* **8**, 14879 (2018). doi: 10.1038/s41598-018-33067-w; pmid: 30291256
79. P. Flegontov *et al*., Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America. *Nature* **570**, 236–240 (2019). doi: 10.1038/s41586-019-1251-y; pmid: 31168094
80. R. Fregel *et al*., Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 6774–6779 (2018). doi: 10.1073/pnas.1800851115; pmid: 29895688
81. Q. Fu *et al*., A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23**, 553–559 (2013). doi: 10.1016/j.cub.2013.02.044; pmid: 23523248
82. Q. Fu *et al*., An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015). doi: 10.1038/nature14558; pmid: 26098372
83. Q. Fu *et al*., The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016). doi: 10.1038/nature17993; pmid: 27135931
84. C. Gamba *et al*., Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5**, 5257 (2014). doi: 10.1038/ncomms6257; pmid: 25334030
85. G. González-Fortes *et al*., Paleogenomic Evidence for Multi-generational Mixing between Neolithic Farmers and Mesolithic Hunter-Gatherers in the Lower Danube Basin. *Curr. Biol.* **27**, 1801–1810.e10 (2017). doi: 10.1016/j.cub.2017.05.023; pmid: 28552360
86. G. González-Fortes *et al*., A western route of prehistoric human migration from Africa into the Iberian Peninsula. *Proc. Biol. Sci.* **286**, 20182288 (2019). doi: 10.1098/rspb.2018.2288; pmid: 30963949
87. T. Günther *et al*., Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLOS Biol.* **16**, e2003703 (2018). doi: 10.1371/journal.pbio.2003703; pmid: 29315301
88. T. Günther *et al*., Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11917–11922 (2015). doi: 10.1073/pnas.1509851112; pmid: 26351665
89. M. Haber *et al*., Continuity and Admixture in the Last Five Millennia of Levantine History from Ancient Canaanite and Present-Day Lebanese Genome Sequences. *Am. J. Hum. Genet.* **101**, 274–282 (2017). doi: 10.1016/j.ajhg.2017.06.013; pmid: 28757201
90. M. Haber *et al*., A Transient Pulse of Genetic Admixture from the Crusaders in the Near East Identified from Ancient Genome Sequences. *Am. J. Hum. Genet.* **104**, 977–984 (2019). doi: 10.1016/j.ajhg.2019.03.015; pmid: 31006515
91. M. Hajdinjak *et al*., Reconstructing the genetic history of late Neanderthals. *Nature* **555**, 652–656 (2018). doi: 10.1038/nature26151; pmid: 29562232
92. É. Harney *et al*., Ancient DNA from Chalcolithic Israel reveals the role of population mixture in cultural transformation. *Nat. Commun.* **9**, 3336 (2018). doi: 10.1038/s41467-018-05649-9; pmid: 30127404
93. É. Harney *et al*., Ancient DNA from the skeletons of Roopkund Lake reveals Mediterranean migrants in India. *Nat. Commun.* **10**, 3670 (2019). doi: 10.1038/s41467-019-11357-9; pmid: 31431628
94. Z. Hofmanová *et al*., Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 6886–6891 (2016). doi: 10.1073/pnas.1523951113; pmid: 27274049
95. M. Järve *et al*., Shifts in the Genetic Landscape of the Western Eurasian Steppe Associated with the Beginning and End of the Scythian Dominance. *Curr. Biol.* **29**, 2430–2441.e10 (2019). doi: 10.1016/j.cub.2019.06.019; pmid: 31303491
96. C. Jeong *et al*., The genetic history of admixture across inner Eurasia. *Nat. Ecol. Evol.* **3**, 966–976 (2019). doi: 10.1038/s41559-019-0878-2; pmid: 31036896
97. C. Jeong *et al*., Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7485–7490 (2016). doi: 10.1073/pnas.1520844113; pmid: 27325755
98. C. Jeong *et al*., Bronze Age population dynamics and the rise of dairy pastoralism on the eastern Eurasian steppe. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E11248–E11255 (2018). doi: 10.1073/pnas.1813608115; pmid: 30397125
99. E. R. Jones *et al*., Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* **6**, 8912 (2015). doi: 10.1038/ncomms9912; pmid: 26567969
100. E. R. Jones *et al*., The Neolithic Transition in the Baltic Was Not Driven by Admixture with Early European Farmers. *Curr. Biol.* **27**, 576–582 (2017). doi: 10.1016/j.cub.2016.12.060; pmid: 28162894
101. H. Kanzawa-Kiriyama *et al*., A partial nuclear genome of the Jomons who lived 3000 years ago in Fukushima, Japan. *J. Hum. Genet.* **62**, 213–221 (2017). doi: 10.1038/jhg.2016.110; pmid: 27581845
102. A. Keller *et al*., New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.* **3**, 698 (2012). doi: 10.1038/ncomms1701; pmid: 22426219
103. D. J. Kennett *et al*., Archaeogenomic evidence reveals prehistoric matrilineal dynasty. *Nat. Commun.* **8**, 14115 (2017). doi: 10.1038/ncomms14115; pmid: 28221340
104. G. M. Kılınç *et al*., The Demographic Development of the First Farmers in Anatolia. *Curr. Biol.* **26**, 2659–2666 (2016). doi: 10.1016/j.cub.2016.07.057; pmid: 27498567
105. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014). doi: 10.1186/s12859-014-0356-4; pmid: 25420514
106. M. Krzewińska *et al*., Genomic and Strontium Isotope Variation Reveal Immigration Patterns in a Viking Age Town. *Curr. Biol.* **28**, 2730–2738.e10 (2018). doi: 10.1016/j.cub.2018.06.053; pmid: 30146150
107. M. Krzewińska *et al*., Ancient genomes suggest the eastern Pontic-Caspian steppe as the source of western Iron Age nomads. *Sci. Adv.* **4**, eaat4457 (2018). doi: 10.1126/sciadv.aat4457; pmid: 30417088
108. T. C. Lamnidis *et al*., Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nat. Commun.* **9**, 5018 (2018). doi: 10.1038/s41467-018-07483-5; pmid: 30479341
109. 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
110. I. Lazaridis *et al*., Genetic origins of the Minoans and Mycenaeans. *Nature* **548**, 214–218 (2017). doi: 10.1038/nature23310; pmid: 28783727
111. J. Lindo *et al*., Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4093–4098 (2017). doi: 10.1073/pnas.1620410114; pmid: 28377518
112. J. Lindo *et al*., The genetic prehistory of the Andean highlands 7000 years BP though European contact. *Sci. Adv.* **4**, eaau4921 (2018). doi: 10.1126/sciadv.aau4921; pmid: 30417096
113. M. Lipson *et al*., Population Turnover in Remote Oceania Shortly after Initial Settlement. *Curr. Biol.* **28**, 1157–1165.e7 (2018). doi: 10.1016/j.cub.2018.02.051; pmid: 29501328
114. M. Lipson *et al*., Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature* **551**, 368–372 (2017). doi: 10.1038/nature24476; pmid: 29144465
115. M. Lipson *et al*., Ancient West African foragers in the context of African population history. *Nature* **577**, 665–670 (2020). doi: 10.1038/s41586-020-1929-1; pmid: 31969706
116. M. Lipson *et al*., Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* **361**, 92–95 (2018). doi: 10.1126/science.aat3188; pmid: 29773666
117. M. Gallego Llorente *et al*., Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science* **350**, 820–822 (2015). doi: 10.1126/science.aad2879; pmid: 26449472
118. A.-S. Malaspinas *et al*., Two ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil. *Curr. Biol.* **24**, R1035–R1037 (2014). doi: 10.1016/j.cub.2014.09.078; pmid: 25455029
119. H. Malmström *et al*., The genomic ancestry of the Scandinavian Battle Axe Culture people and their relation to the broader Corded Ware horizon. *Proc. Biol. Sci.* **286**, 20191528 (2019). doi: 10.1098/rspb.2019.1528; pmid: 31594508
120. R. Martiniano *et al*., Genomic signals of migration and continuity in Britain before the Anglo-Saxons. *Nat. Commun.* **7**, 10326 (2016). doi: 10.1038/ncomms10326; pmid: 26783717
121. R. Martiniano *et al*., The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLOS Genet.* **13**, e1006852 (2017). doi: 10.1371/journal.pgen.1006852; pmid: 28749934
122. I. Mathieson *et al*., Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015). doi: 10.1038/nature16152; pmid: 26595274
123. I. Mathieson *et al*., The genomic history of southeastern Europe. *Nature* **555**, 197–203 (2018). doi: 10.1038/nature25778; pmid: 29466330
124. H. McColl *et al*., The prehistoric peopling of Southeast Asia. *Science* **361**, 88–92 (2018). doi: 10.1126/science.aat3628; pmid: 29976827
125. A. Mittnik *et al*., The genetic prehistory of the Baltic Sea region. *Nat. Commun.* **9**, 442 (2018). doi: 10.1038/s41467-018-02825-9; pmid: 29382937
126. A. Mittnik *et al*., Kinship-based social inequality in Bronze Age Europe. *Science* **366**, 731–734 (2019). doi: 10.1126/science.aax6219; pmid: 31601705
127. M. Mondal *et al*., Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nat. Genet.* **48**, 1066–1070 (2016). doi: 10.1038/ng.3621; pmid: 27455350
128. J. V. Moreno-Mayar *et al*., Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* **553**, 203–207 (2018). doi: 10.1038/nature25173; pmid: 29323294
129. A. G. Nikitin *et al*., Interactions between earliest Linearbandkeramik farmers and central European hunter gatherers at the dawn of European Neolithization. *Sci. Rep.* **9**, 19544 (2019). doi: 10.1038/s41598-019-56029-2; pmid: 31863024
130. C. Ning *et al*., Ancient Genomes Reveal Yamnaya-Related Ancestry and a Potential Source of Indo-European Speakers in Iron Age Tianshan. *Curr. Biol.* **29**, 2526–2532.e4 (2019). doi: 10.1016/j.cub.2019.06.044; pmid: 31353181
131. I. Olalde *et al*., A Common Genetic Origin for Early Farmers from Mediterranean Cardial and Central European LBK Cultures. *Mol. Biol. Evol.* **32**, 3132–3142 (2015). doi: 10.1093/molbev/msv181; pmid: 26337550
132. I. Olalde *et al*., Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507**, 225–228 (2014). doi: 10.1038/nature12960; pmid: 24463515
133. I. Olalde *et al*., The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* **555**, 190–196 (2018). doi: 10.1038/nature25738; pmid: 29466337
134. I. Olalde *et al*., The genomic history of the Iberian Peninsula over the past 8000 years. *Science* **363**, 1230–1234 (2019). doi: 10.1126/science.aav4040; pmid: 30872528
135. A. Omrak *et al*., Genomic Evidence Establishes Anatolia as the Source of the European Neolithic Gene Pool. *Curr. Biol.* **26**, 270–275 (2016). doi: 10.1016/j.cub.2015.12.019; pmid: 26748850
136. J. K. Pickrell *et al*., The genetic prehistory of southern Africa. *Nat. Commun.* **3**, 1143 (2012). doi: 10.1038/ncomms2140; pmid: 23072811
137. C. Posth *et al*., Reconstructing the Deep Population History of Central and South America. *Cell* **175**, 1185–1197.e22 (2018). doi: 10.1016/j.cell.2018.10.027; pmid: 30415837
138. C. Posth *et al*., Language continuity despite population replacement in Remote Oceania. *Nat. Ecol. Evol.* **2**, 731–740 (2018). doi: 10.1038/s41559-018-0498-2; pmid: 29487365
139. M. Raghavan *et al*., Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014). doi: 10.1038/nature12736; pmid: 24256729
140. M. Raghavan *et al*., The genetic prehistory of the New World Arctic. *Science* **345**, 1255832 (2014). doi: 10.1126/science.1255832; pmid: 25170159
141. 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
142. M. Rasmussen *et al*., Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010). doi: 10.1038/nature08835; pmid: 20148029
143. M. Rasmussen *et al*., The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014). doi: 10.1038/nature13025; pmid: 24522598

144. M. Rasmussen *et al*., The ancestry and affiliations of Kennewick Man. *Nature* **523**, 455–458 (2015). doi: 10.1038/nature14625; pmid: 26087396
145. R. Rodríguez-Varela *et al*., Genomic Analyses of Pre-European Conquest Human Remains from the Canary Islands Reveal Close Affinity to Modern North Africans. *Curr. Biol.* **27**, 3396–3402.e5 (2017). doi: 10.1016/j.cub.2017.09.059; pmid: 29107554
146. L. Saag *et al*., Extensive Farming in Estonia Started through a Sex-Biased Migration from the Steppe. *Curr. Biol.* **27**, 2185–2193.e6 (2017). doi: 10.1016/j.cub.2017.06.022; pmid: 28712569
147. L. Saag *et al*., The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East. *Curr. Biol.* **29**, 1701–1711.e16 (2019). doi: 10.1016/j.cub.2019.04.026; pmid: 31080083
148. C. L. Scheib *et al*., Ancient human parallel lineages within North America contributed to a coastal expansion. *Science* **360**, 1024–1027 (2018). doi: 10.1126/science.aar6851; pmid: 29853687
149. F. Sánchez-Quinto *et al*., Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 9469–9474 (2019). doi: 10.1073/pnas.1818037116; pmid: 30988179
150. V. J. Schuenemann *et al*., Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. *Nat. Commun.* **8**, 15694 (2017). doi: 10.1038/ncomms15694; pmid: 28556824
151. S. Schiffels *et al*., Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nat. Commun.* **7**, 10408 (2016). doi: 10.1038/ncomms10408; pmid: 26783965
152. C. M. Schlebusch *et al*., Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**, 652–655 (2017). doi: 10.1126/science.aao6266; pmid: 28971970
153. H. Schroeder *et al*., Origins and genetic legacies of the Caribbean Taino. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2341–2346 (2018). doi: 10.1073/pnas.1716839115; pmid: 29463742
154. H. Schroeder *et al*., Unraveling ancestry, kinship, and violence in a Late Neolithic mass grave. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 10705–10710 (2019). doi: 10.1073/pnas.1820210116; pmid: 31061125
155. A. Seguin-Orlando *et al*., Genomic structure in Europeans dating back at least 36,200 years. *Science* **346**, 1113–1118 (2014). doi: 10.1126/science.aaa0114; pmid: 25378462
156. V. Shinde *et al*., An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers. *Cell* **179**, 729–735.e10 (2019). doi: 10.1016/j.cell.2019.08.048; pmid: 31495572
157. M. Sikora *et al*., The population history of northeastern Siberia since the Pleistocene. *Nature* **570**, 182–188 (2019). doi: 10.1038/s41586-019-1279-z; pmid: 31168093
158. M. Sikora *et al*., Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science* **358**, 659–662 (2017). doi: 10.1126/science.aao1807; pmid: 28982795
159. P. Skoglund *et al*., Reconstructing Prehistoric African Population Structure. *Cell* **171**, 59–71.e21 (2017). doi: 10.1016/j.cell.2017.08.049; pmid: 28938123
160. P. Skoglund *et al*., Genetic evidence for two founding populations of the Americas. *Nature* **525**, 104–108 (2015). doi: 10.1038/nature14895; pmid: 26196601
161. P. Skoglund *et al*., Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**, 510–513 (2016). doi: 10.1038/nature19844; pmid: 27698418
162. P. Skoglund *et al*., Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* **344**, 747–750 (2014). doi: 10.1126/science.1253448; pmid: 24762536
163. V. Slon *et al*., The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* **561**, 113–116 (2018). doi: 10.1038/s41586-018-0455-x; pmid: 30135579
164. M. Unterländer *et al*., Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. *Nat. Commun.* **8**, 14615 (2017). doi: 10.1038/ncomms14615; pmid: 28256537
165. C. Valdiosera *et al*., Four millennia of Iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of Eurasia. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3428–3433 (2018). doi: 10.1073/pnas.1717762115; pmid: 29531053
166. M. van de Loosdrecht *et al*., Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. *Science* **360**, 548–552 (2018). doi: 10.1126/science.aar8380; pmid: 29545507
167. C. M. van den Brink *et al*., A Late Bronze Age II clay coffin from Tel Shaddud in the Central Jezreel Valley, Israel: Context and historical implications. *Levant* **49**, 105–135 (2017). doi: 10.1080/00758914.2017.1368204
168. K. R. Veeramah *et al*., Population genomic analysis of elongated skulls reveals extensive female-biased immigration in Early Medieval Bavaria. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3494–3499 (2018). doi: 10.1073/pnas.1719880115; pmid: 29531040
169. D. N. Vyas, A. Al-Meeri, C. J. Mulligan, Testing support for the northern and southern dispersal routes out of Africa: An analysis of Levantine and southern Arabian populations. *Am. J. Phys. Anthropol.* **164**, 736–749 (2017). doi: 10.1002/ajpa.23312; pmid: 28913852
170. C.-C. Wang *et al*., Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. *Nat. Commun.* **10**, 590 (2019). doi: 10.1038/s41467-018-08220-8; pmid: 30713341
171. M. A. Yang *et al*., 40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia. *Curr. Biol.* **27**, 3202–3208.e9 (2017). doi: 10.1016/j.cub.2017.09.030; pmid: 29033327
172. P. Zalloua *et al*., Ancient DNA of Phoenician remains indicates discontinuity in the settlement history of Ibiza. *Sci. Rep.* **8**, 17567 (2018). doi: 10.1038/s41598-018-35667-y; pmid: 30514893
173. J. R. Adrion *et al*., A community-maintained standard library of population genetic models. *eLife* **9**, e54967 (2020). doi: 10.7554/eLife.54967; pmid: 32573438
174. J. Kelleher *et al*., tskit-dev/tskit: C API 0.99.14, version C_0.99.14, Zenodo (2021); https://doi.org/10.5281/zenodo.5465773.
175. A. W. Wohns, Y. Wong, B. Jeffery, awohns/unified_genealogy_paper: Revised manuscript, version 0.1.1, Zenodo (2021); https://doi.org/10.5281/zenodo.5784137.
176. A. W. Wohns *et al*., A unified genealogy of modern and ancient genomes: Unified, inferred tree sequences of 1000 Genomes, Human Genome Diversity, and Simons Genome Diversity Projects, version 1.0.0, Zenodo (2021); https://doi.org/10.5281/zenodo.5495535.
177. A. W. Wohns *et al*., A unified genealogy of modern and ancient genomes: Unified, inferred tree sequences of 1000 Genomes, Human Genome Diversity, and Simons Genome Diversity Projects with ancient samples, version 1.0.0, Zenodo (2021); https://doi.org/10.5281/zenodo.5512994.

ACKNOWLEDGMENTS

We thank the Oxford Big Data research computing team, specifically A. Huffman and R. Esnouf, as well as D. Lieberman, K. Lohse, and E. Castedo Ellerman for comments. **Funding:** This work was supported by Wellcome Trust grant 100956/Z/13/Z (to G.M.), the Li Ka Shing Foundation (to G.M.), the Robertson Foundation (to J.K.), the Rhodes Trust (to A.W.W.), NIH (NIGMS) grant GM100233 (to D.R.), the Paul Allen Foundation (to D.R.), the John Templeton Foundation grant 61220 (to D.R.), and the Howard Hughes Medical Institute (to D.R.). The computational aspects of this research were supported by the Wellcome Trust (core award 203141/Z/16/Z) and the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. **Author contributions:** Conceptualization: A.W.W., Y.W., J.K., and G.M. Methodology: A.W.W., Y.W., A.A., S.M., N.P., J.K., and G.M. Software: A.W.W., Y.W., and J.K. Investigation: A.W.W., Y.W., B.J., R.P., and D.R. Formal analysis: A.W.W., Y.W., B.J., A.A., and S.M. Validation: A.W.W., Y.W., and B.J. Visualization: A.W.W., Y.W., and B.J. Data curation: A.W.W., Y.W., B.J., A.A., and S.M. Resources: A.W.W., Y.W., A.A., S.M., R.P., D.R., and J.K. Funding acquisition: D.R., J.K., and G.M. Project administration: A.W.W. and D.R. Supervision: Y.W., D.R., J.K., and G.M. Writing–original draft: A.W.W., Y.W., and G.M. Writing–review and editing: A.W.W., Y.W., D.R., J.K., and G.M. **Competing interests:** G.M. is a director of and shareholder in Genomics plc and a partner in Peptide Groove LLP. The authors declare no other competing interests. **Data and materials availability:** Newly reported sequencing data from the Afanasievo family are available from the European Nucleotide Archive, accession number PRJEB43093, and phased variant data for the family are available from the European Variation Archive, accession number PRJEB46983. All publicly available datasets used in this paper are available from their original publications. *tsinfer* is deposited to Zenodo (*30*) and is available at https://tsinfer.readthedocs.io/ under the GNU General Public License v3.0, *tsdate* is deposited to Zenodo (*31*) and is available at https://tsdate.readthedocs.io/ under the MIT License, and *tskit* is deposited to Zenodo (*174*) and is available at https://tskit.readthedocs.io/ under the MIT License. All code used to perform analyses in this paper is deposited to Zenodo (*175*) and can be found at https://github.com/awohns/unified_genealogy_paper. Unified tree sequences of the HGDP, SGDP, and TGP autosomes are available from Zenodo (*176*). Unified tree sequences of the HGDP, SGDP, TGP, and high-coverage ancient autosomes are available from Zenodo (*177*). Tree sequences were compressed using the *tszip* utility; see the documentation at https://tszip.readthedocs.io/ for further details.

SUPPLEMENTARY MATERIALS

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

9 April 2021; accepted 23 December 2021
10.1126/science.abi8264