

A unified genealogy of modern and ancient genomes*

Anthony Wilder Wohns^{1,2}, Yan Wong^{2†}, Ben Jeffery², Ali Akbari^{1,3,6}, Swapan Mallick^{1,4}, Ron Pinhasi⁵, Nick Patterson^{1,3,4,6}, David Reich^{1,3,4,6}, Jerome Kelleher^{2†}, Gil McVean^{2†‡}

¹Broad Institute of MIT and Harvard; Cambridge, MA 02142, USA.

²Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford; Oxford OX3 7LF, UK.

³Department of Human Evolutionary Biology, Harvard University; Cambridge, MA 02138, USA.

⁴Howard Hughes Medical Institute; Boston, MA 02115, USA.

⁵Department of Evolutionary Anthropology, University of Vienna; 1090 Vienna, Austria.

⁶Department of Genetics, Harvard Medical School; Boston, MA 02115, USA.

[†]These authors contributed equally to this work.

[‡]Corresponding author. Email: gil.mcvean@bdi.ox.ac.uk

Abstract: 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 non-parametric 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 non-parametric estimator of ancestor geographical location that recapitulates key events in human history.

One-Sentence Summary: The largest genealogy of modern and ancient genomes yet constructed delivers insights into human history and evolution.

* This manuscript has been accepted for publication in Science. This version has not undergone final editing. Please refer to the complete version of record at <http://www.sciencemag.org/>. The manuscript may not be reproduced or used in any manner that does not fall within the fair use provisions of the Copyright Act without the prior, written permission of AAAS.

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 data sets, 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 data sets 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).

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, necessary when considering genealogies within recombining organisms (22). Importantly, 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 non-parametric 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. We note that while 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 2,548 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). 154 individuals appear in more than one of these datasets (24). In addition, 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 coverage of 10.8x, 25.8x, 21.2x, and 25.3x) from the Afanasievo Culture, who lived ~4.6 thousand years ago (kya) in the Altai Mountains of Russia (Table S1). Finally, we used 3,589 published ancient samples from over 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 due to the lack of reliable phasing for the majority of 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- S5) (24, 31). Note that *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 (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 re-inferred the tree sequence. The Afanasievo family have high coverage and comparably reliable haplotype phasing and were included to demonstrate the ability of our approach to incorporate high quality ancient samples.

The integrated tree sequences of each autosome together 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 13,513,873 sites contain at least two mutations affecting more than one sample, implying that up to 17.5% of variable sites could result from more than one ancestral mutation. A high proportion of sites with over ~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), suggesting they are largely erroneous. Moreover, analysis of data simulated with an empirically-calibrated error profile and evaluation of enrichment of multiple mutations at sites with known elevated mutation rates, suggests that the majority of the multiple mutations we identify are likely explained by error, but a minority (c. 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 as it is 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 Interactive fig.S1). 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/Denisovan MRCA density from 2,000-5,000 generations ago

(Fig. 2B). Multiple populations show recent within-group TMRCA, 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 $\sim 1,000$ generations, which is caused by multiple MRCA 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 TMRCA (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 3,734 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 $\sim 86\%$ with a recall of $\sim 61\%$ (24). We find descendants among non-archaic 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 varies greatly among modern humans. In particular, we find that Papuans and Australians carry multiple fragments of Denisovan haplotypes that are largely unique to the individual (Fig. 3C). In contrast, other modern descendants of Denisovan haplotypes have fewer blocks which 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 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), supporting suggestions that the

proportions are similar between East Asians and West Eurasians (39) and inconsistent with other reports (26, 40).

Non-parametric inference of spatio-temporal 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, though 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' location, 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 find 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 kya the average location of ancestral haplotypes is in Northeast Africa and remains there until the oldest common ancestors are reached. Indeed, 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 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. In addition, migrations occurring within the last 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 kya, the estimated geographic center of human ancestors is still located in Africa, but many are also observed in the Middle East and Central Asia and a few are located in Papua New Guinea. At 140 kya, more ancestors are found in Papua New Guinea. This is almost 100 kya before the earliest documented human habitation of the region (48). However, our findings are potentially consistent with the proposed timescales of deeply diverged Denisovan lineages unique to Papuans (37) and possibly admixture with unsampled, "ghost" lineages. At 56 kya, some ancestral lineages are observed in the Americas, 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, since 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 focus on the temporal behavior of individual mutation frequencies in idealized populations (51, 52). More recently, modelling techniques have shifted to focus on the genealogical history of sampled genomes and the correlation structures arising through recombination (22, 53). Critically, a single (albeit extremely complex) set of ancestral relationships exist 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 new 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. While a large number of sites are inferred to carry multiple mutations, we find that the majority of these likely reflect genotype error and potentially errors arising from paralogy (particularly at sites requiring high numbers of mutations), though there remains a significant signal of recurrent mutation, as previously reported (62, 63). Similarly, we find some evidence for certain classes of error in ancient sequence leading to false “correction” of variant ages. We choose to retain all additional mutations in the analyses described in this paper, including those which are highly likely to reflect sequencing error, as 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 due to these errors in the sequencing datasets. Estimates from simulations show that genotype error may cause an upwards 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 impact 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 compression of cohorts against such a reference structure.

There exists room for improvement as well as new 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). In addition, 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 spatio-temporal 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.

References and Notes

1. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, The UK Biobank resource with deep phenotyping and genomic data. *Nature*. **562**, 203–209 (2018).
2. D. Taliun, D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, S. A. G. Taliun, A. Corvelo, S. M. Gogarten, H. M. Kang, A. N. Pitsillides, J. LeFaive, S. Lee, X. Tian, B. L. Browning, S. Das, A.-K. Emde, W. E. Clarke, D. P. Loesch, A. C. Shetty, T. W. Blackwell, A. V. Smith, Q. Wong, X. Liu, M. P. Conomos, D. M. Bobo, F. Aguet, C. Albert, A. Alonso, K. G. Ardlie, D. E. Arking, S. Aslibekyan, P. L. Auer, J. Barnard, R. G. Barr, L. Barwick, L. C. Becker, R. L. Beer, E. J. Benjamin, L. F. Bielak, J. Blangero, M. Boehnke, D. W. Bowden, J. A. Brody, E. G. Burchard, B. E. Cade, J. F. Casella, B. Chalazan, D. I. Chasman, Y.-D. I. Chen, M. H. Cho, S. H. Choi, M. K. Chung, C. B. Clish, A. Correa, J. E. Curran, B. Custer, D. Darbar, M. Daya, M. de Andrade, D. L. DeMeo, S. K. Dutcher, P. T. Ellinor, L. S. Emery, C. Eng, D. Fatkin, T. Fingerlin, L. Forer, M. Fornage, N. Franceschini, C. Fuchsberger, S. M. Fullerton, S. Germer, M. T. Gladwin, D. J. Gottlieb, X. Guo, M. E. Hall, J. He, N. L. Heard-Costa, S. R. Heckbert, M. R. Irvin, J. M. Johnsen, A. D. Johnson, R. Kaplan, S. L. R. Kardia, T. Kelly, S. Kelly, E. E. Kenny, D. P. Kiel, R. Klemmer, B. A. Konkle, C. Kooperberg, A. Köttgen, L. A. Lange, J. Lasky-Su, D. Levy, X. Lin, K.-H. Lin, C. Liu, R. J. F. Loos, L. Garman, R. Gerszten, S. A. Lubitz, K. L. Lunetta, A. C. Y. Mak, A. Manichaikul, A. K. Manning, R. A. Mathias, D. D. McManus, S. T. McGarvey, J. B. Meigs, D. A. Meyers, J. L. Mikulla, M. A. Minear, B. D. Mitchell, S. Mohanty, M. E. Montasser, C. Montgomery, A. C. Morrison, J. M. Murabito, A. Natale, P. Natarajan, S. C. Nelson, K. E. North, J. R. O'Connell, N. D. Palmer, N. Pankratz, G. M. Peloso, P. A. Peyser, J. Pleiness, W. S. Post, B. M. Psaty, D. C. Rao, S. Redline, A. P. Reiner, D. Roden, J. I. Rotter, I. Ruczinski, C. Sarnowski, S. Schoenherr, D. A. Schwartz, J.-S. Seo, S. Seshadri, V. A. Sheehan, W. H. Sheu, M. B. Shoemaker, N. L. Smith, J. A. Smith, N. Sotoodehnia, A. M.

Stilp, W. Tang, K. D. Taylor, M. Telen, T. A. Thornton, R. P. Tracy, D. J. Van Den Berg, R. S. Vasan, K. A. Viaud-Martinez, S. Vrieze, D. E. Weeks, B. S. Weir, S. T. Weiss, L.-C. Weng, C. J. Willer, Y. Zhang, X. Zhao, D. K. Arnett, A. E. Ashley-Koch, K. C. Barnes, E. Boerwinkle, S. Gabriel, R. Gibbs, K. M. Rice, S. S. Rich, E. K. Silverman, P. Qasba, W. Gan, N. Abe, L. Almasy, S. Ament, P. Anderson, P. Anugu, D. Applebaum-Bowden, T. Assimes, D. Avramopoulos, E. Barron-Casella, T. Beaty, G. Beck, D. Becker, A. Beitelshes, T. Benos, M. Bezerra, J. Bis, R. Bowler, U. Broeckel, J. Broome, K. Bunting, C. Bustamante, E. Buth, J. Cardwell, V. Carey, C. Carty, R. Casaburi, P. Castaldi, M. Chaffin, C. Chang, Y.-C. Chang, S. Chavan, B.-J. Chen, W.-M. Chen, L.-M. Chuang, R.-H. Chung, S. Comhair, E. Cornell, C. Crandall, J. Crapo, J. Curtis, C. Damcott, S. David, C. Davis, L. de las Fuentes, M. DeBaun, R. Deka, S. Devine, Q. Duan, R. Duggirala, J. P. Durda, C. Eaton, L. Ekunwe, A. El Boueiz, S. Erzurum, C. Farber, M. Flickinger, C. Frazar, M. Fu, L. Fulton, S. Gao, Y. Gao, M. Gass, B. Gelb, X. P. Geng, M. Geraci, A. Ghosh, C. Gignoux, D. Glahn, D.-W. Gong, H. Goring, S. Graw, D. Grine, C. C. Gu, Y. Guan, N. Gupta, J. Haessler, N. L. Hawley, B. Heavner, D. Herrington, C. Hersh, B. Hidalgo, J. Hixson, B. Hobbs, J. Hokanson, E. Hong, K. Hoth, C. A. Hsiung, Y.-J. Hung, H. Huston, C. M. Hwu, R. Jackson, D. Jain, M. A. Jhun, C. Johnson, R. Johnston, K. Jones, S. Kathiresan, A. Khan, W. Kim, G. Kinney, H. Kramer, C. Lange, E. Lange, L. Lange, C. Laurie, M. LeBoff, J. Lee, S. S. Lee, W.-J. Lee, D. Levine, J. Lewis, X. Li, Y. Li, H. Lin, H. Lin, K. H. Lin, S. Liu, Y. Liu, Y. Liu, J. Luo, M. Mahaney, N. T.-O. for P. M. (TOPMed) Consortium, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. **590**, 290–299 (2021).

3. D. Reich, *Who We Are and How We Got Here: Ancient DNA and the New Science of the Human Past* (Oxford University Press, Oxford, UK, 2018).
4. H. A. Lewin, G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. P. Gilbert, M. M. Goldstein, I. V. Grigoriev, K. J. Hackett, D. Haussler, E. D. Jarvis, W. E. Johnson, A. Patrinos, S. Richards, J. C. Castilla-Rubio, M.-A. van Sluys, P. S. Soltis, X. Xu, H. Yang, G. Zhang, Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci.* **115**, 4325–4333 (2018).
5. Lazaridis, N. Patterson, A. Mitnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, B. Berger, C. Economou, R. Bollongino, Q. Fu, K. I. Bos, S. Nordenfelt, H. Li, C. de Filippo, K. Prüfer, S. Sawyer, C. Posth, W. Haak, F. Hallgren, E. Fornander, N. Rohland, D. Delsate, M. Francken, J.-M. Guinet, J. Wahl, G. Ayodo, H. A. Babiker, G. Bailliet, E. Balanovska, O. Balanovsky, R. Barrantes, G. Bedoya, H. Ben-Ami, J. Bene, F. Berrada, C. M. Bravi, F. Brisighelli, G. B. J. Busby, F. Cali, M. Churnosov, D. E. C. Cole, D. Corach, L. Damba, G. van Driem, S. Dryomov, J.-M. Dugoujon, S. A. Fedorova, I. Gallego Romero, M. Gubina, M. Hammer, B. M. Henn, T. Hervig, U. Hodoglugil, A. R. Jha, S. Karachanak-Yankova, R. Khusainova, E. Khusnutdinova, R. Kittles, T. Kivisild, W. Klitz, V. Kučinskas, A. Kushniarevich, L. Laredj, S. Litvinov, T. Loukidis, R. W. Mahley, B. Melegh, E. Metspalu, J. Molina, J. Mountain, K. Näkkäläjärvi, D. Nesheva, T. Nyambo, L. Osipova, J. Parik, F. Platonov, O. Posukh, V. Romano, F. Rothhammer, I. Rudan, R. Ruizbakiev, H. Sahakyan, A. Sajantila, A. Salas, E. B. Starikovskaya, A. Tarekegn, D. Toncheva, S. Turdikulova, I. Uktveryte, O. Utevska, R. Vasquez, M. Villena, M. Voevoda, C. A. Winkler, L. Yepiskoposyan, P. Zalloua, T. Zemunik, A. Cooper, C. Capelli, M. G. Thomas, A. Ruiz-

- Linares, S. A. Tishkoff, L. Singh, K. Thangaraj, R. Vilems, D. Comas, R. Sukernik, M. Metspalu, M. Meyer, E. E. Eichler, J. Burger, M. Slatkin, S. Pääbo, J. Kelso, D. Reich, J. Krause, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. **513**, 409–413 (2014).
6. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).
 7. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Vilems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. **538**, 201–206 (2016).
 8. A. Bergström, S. A. McCarthy, R. Hui, M. A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, H. Blanché, J.-F. Deleuze, H. Cann, S. Mallick, D. Reich, M. S. Sandhu, P. Skoglund, A. Scally, Y. Xue, R. Durbin, C. Tyler-Smith, Insights into human genetic variation and population history from 929 diverse genomes. *Science*. **367**, eaay5012 (2020).
 9. S. Belsare, M. Levy-Sakin, Y. Mostovoy, S. Durinck, S. Chaudhuri, M. Xiao, A. S. Peterson, P.-Y. Kwok, S. Seshagiri, J. D. Wall, Evaluating the quality of the 1000 genomes project data. *BMC Genomics*. **20**, 620 (2019).
 10. L. Shi, Y. Guo, C. Dong, J. Huddleston, H. Yang, X. Han, A. Fu, Q. Li, N. Li, S. Gong, K. E. Lintner, Q. Ding, Z. Wang, J. Hu, D. Wang, F. Wang, L. Wang, G. J. Lyon, Y. Guan, Y. Shen, O. V. Evgrafov, J. A. Knowles, F. Thibaud-Nissen, V. Schneider, C.-Y. Yu, L. Zhou, E. E. Eichler, K.-F. So, K. Wang, Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* **7**, 12065 (2016).
 11. M. Wenger, P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Functammasan, A. Kolesnikov, N. D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C.-S. Chin, A. M. Phillippy, M. C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F. J. Sedlazeck, J. M. Zook, H. Li, S. Koren, A. Carroll, D. R. Rank, M. W. Hunkapiller, Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*. **37**, 1155–1162 (2019).
 12. S. Hwang, E. Kim, I. Lee, E. M. Marcotte, Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*. **5**, 17875 (2015).

13. J. Kelleher, Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, G. McVean, Inferring whole-genome histories in large population datasets. *Nature genetics*. **51**, 1330–1338 (2019).
14. L. L. Cavalli-Sforza, M. W. Feldman, The application of molecular genetic approaches to the study of human evolution. *Nature genetics*. **33**, 266–275 (2003).
15. N. Patterson, A. L. Price, D. Reich, Population Structure and Eigenanalysis. *PLOS Genetics*. **2**, 1–20 (2006).
16. J. K. Pritchard, M. Stephens, P. Donnelly, Inference of Population Structure Using Multilocus Genotype Data. *Genetics*. **155**, 945–959 (2000).
17. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of Population Structure using Dense Haplotype Data. *PLOS Genetics*. **8**, 1–16 (2012).
18. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history. *Genetics*. **192**, 1065–1093 (2012).
19. J. K. Pickrell, J. K. Pritchard, Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genetics*. **8**, 1–17 (2012).
20. L. Bonomi, Y. Huang, L. Ohno-Machado, Privacy challenges and research opportunities for genomic data sharing. *Nature Genetics*. **52**, 646–654 (2020).
21. J. Kelleher, A. M. Etheridge, G. McVean, Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*. **12**, 1–22 (2016).
22. R. R. Hudson, Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*. **23**, 183–201 (1983).
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).
24. Materials and methods are available as supplementary materials at the Science website.
25. K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. F. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*. **505**, 43–49 (2014).
26. K. Prüfer, C. de Filippo, S. Grote, F. Mafessoni, P. Korlević, M. Hajdinjak, B. Vernot, L. Skov, P. Hsieh, S. Peyrégne, D. Reher, C. Hopfe, S. Nagel, T. Maricic, Q. Fu, C. Theunert, R. Rogers, P. Skoglund, M. Chintalapati, M. Dannemann, B. J. Nelson, F. M. Key, P. Rudan, Ž. Kućan, I. Gušić, L. V. Golovanova, V. B. Doronichev, N. Patterson, D. Reich, E. E. Eichler, M. Slatkin, M. H. Schierup, A. M. Andrés, J. Kelso, M. Meyer, S. Pääbo, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*. **358**, 655–658 (2017).

27. F. Mafessoni, S. Grote, C. de Filippo, V. Slon, K. A. Kolobova, B. Viola, S. V. Markin, M. Chintalapati, S. Peyrégne, L. Skov, P. Skoglund, A. I. Krivoschapkin, A. P. Derevianko, M. Meyer, J. Kelso, B. Peter, K. Prüfer, S. Pääbo, A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl. Acad. Sci.* **117**, 15132–15136 (2020).
28. M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*. **338**, 222–226 (2012).
29. Q. Fu, H. Li, P. Moorjani, F. Jay, S. M. Slepchenko, A. A. Bondarev, P. L. F. Johnson, A. Aximu-Petri, K. Prüfer, C. de Filippo, M. Meyer, N. Zwyns, D. C. Salazar-García, Y. V. Kuzmin, S. G. Keates, P. A. Kosintsev, D. I. Razhev, M. P. Richards, N. V. Peristov, M. Lachmann, K. Douka, T. F. G. Higham, M. Slatkin, J.-J. Hublin, D. Reich, J. Kelso, T. B. Viola, S. Pääbo, Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*. **514**, 445–449 (2014).
30. J. Kelleher, W. Yan, B. Jeffery, A. W. Wohns, *tsinfer* (2021; <https://github.com/tskit-dev/tsinfer>). doi:10.5281/zenodo.5168051.
31. W. Wohns, Y. Wong, J. Ben, *tsdate* (2021; <https://github.com/tskit-dev/tsdate>). doi:10.5281/zenodo.5168040.
32. Bonné, The Samaritans: a demographic study. *Human biology*. **35**, 61–89 (1963).
33. P. K. Albers, G. McVean, Dating genomic variants and shared ancestry in population-scale sequencing data. *PLOS Biology*. **18**, 1–26 (2020).
34. L. Speidel, M. Forest, S. Shi, S. R. Myers, A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*. **51**, 1321–1329 (2019).
35. D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. F. Johnson, T. Maricic, J. M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. E. Eichler, M. Stoneking, M. Richards, S. Talamo, M. V. Shunkov, A. P. Derevianko, J.-J. Hublin, J. Kelso, M. Slatkin, S. Pääbo, Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*. **468**, 1053–1060 (2010).
36. D. Reich, N. Patterson, M. Kircher, F. Delfin, M. R. Nandineni, I. Pugach, A. M.-S. Ko, Y.-C. Ko, T. A. Jinam, M. E. Phipps, N. Saitou, A. Wollstein, M. Kayser, S. Pääbo, M. Stoneking, Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet*. **89**, 516–528 (2011).
37. G. S. Jacobs, G. Hudjashov, L. Saag, P. Kusuma, C. C. Darusallam, D. J. Lawson, M. Mondal, L. Pagani, F.-X. Ricaut, M. Stoneking, M. Metspalu, H. Sudoyo, J. S. Lansing, M. P. Cox, Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell*. **177**, 1010–1021 (2019).
38. V. M. Narasimhan, N. Patterson, P. Moorjani, N. Rohland, R. Bernardos, S. Mallick, I. Lazaridis, N. Nakatsuka, I. Olalde, M. Lipson, A. M. Kim, L. M. Olivieri, A. Coppa, M.

- Vidale, J. Mallory, V. Moiseyev, E. Kitov, J. Monge, N. Adamski, N. Alex, N. Broomandkhoshbacht, F. Candilio, K. Callan, O. Cheronet, B. J. Culleton, M. Ferry, D. Fernandes, S. Freilich, B. Gamarra, D. Gaudio, M. Hajdinjak, É. Harney, T. K. Harper, D. Keating, A. M. Lawson, M. Mah, K. Mandl, M. Michel, M. Novak, J. Oppenheimer, N. Rai, K. Sirak, V. Slon, K. Stewardson, F. Zalzal, Z. Zhang, G. Akhatov, A. N. Bagashev, A. Bagnera, B. Baitanayev, J. Bendezu-Sarmiento, A. A. Bissembaev, G. L. Bonora, T. T. Charginov, T. Chikisheva, P. K. Dashkovskiy, A. Derevianko, M. Dobeš, K. Douka, N. Dubova, M. N. Duisengali, D. Enshin, A. Epimakhov, A. V. Fribus, D. Fuller, A. Goryachev, A. Gromov, S. P. Grushin, B. Hanks, M. Judd, E. Kazizov, A. Khokhlov, A. P. Krygin, E. Kupriyanova, P. Kuznetsov, D. Luiselli, F. Maksudov, A. M. Mamedov, T. B. Mamirov, C. Meiklejohn, D. C. Merrett, R. Micheli, O. Mochalov, S. Mustafokulov, A. Nayak, D. Pettener, R. Potts, D. Razhev, M. Rykun, S. Sarno, T. M. Savenkova, K. Sikhymbaeva, S. M. Slepchenko, O. A. Soltobaev, N. Stepanova, S. Svyatko, K. Tabaldiev, M. Teschler-Nicola, A. A. Tishkin, V. V. Tkachev, S. Vasilyev, P. Velemínský, D. Voyakin, A. Yermolayeva, M. Zahir, V. S. Zubkov, A. Zubova, V. S. Shinde, C. Lalueza-Fox, M. Meyer, D. Anthony, N. Boivin, K. Thangaraj, D. J. Kennett, M. Frachetti, R. Pinhasi, D. Reich, The formation of human populations in South and Central Asia. *Science*. **365** (2019).
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).
 40. J. D. Wall, M. A. Yang, F. Jay, S. K. Kim, E. Y. Durand, L. S. Stevison, C. Gignoux, A. Woerner, M. F. Hammer, M. Slatkin, Higher levels of Neanderthal ancestry in East Asians than in Europeans. *Genetics*. **194**, 199–209 (2013).
 41. McDougall, F. H. Brown, J. G. Fleagle, Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*. **433**, 733–736 (2005).
 42. J.-J. Hublin, A. Ben-Ncer, S. E. Bailey, S. E. Freidline, S. Neubauer, M. M. Skinner, I. Bergmann, A. Le Cabec, S. Benazzi, K. Harvati, P. Gunz, New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*. **546**, 289–292 (2017).
 43. K. Wang, S. Goldstein, M. Bleasdale, B. Clist, K. Bostoen, P. Bakwa-Lufu, L. T. Buck, A. Crowther, A. Dème, R. J. McIntosh, J. Mercader, C. Ogola, R. C. Power, E. Sawchuk, P. Robertshaw, E. N. Wilmsen, M. Petraglia, E. Ndiema, F. K. Manthi, J. Krause, P. Roberts, N. Boivin, S. Schiffels, Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-Saharan Africa. *Science Advances*. **6** (2020).
 44. M. E. Prendergast, M. Lipson, E. A. Sawchuk, I. Olalde, C. A. Ogola, N. Rohland, K. A. Sirak, N. Adamski, R. Bernardos, N. Broomandkhoshbacht, K. Callan, B. J. Culleton, L. Eccles, T. K. Harper, A. M. Lawson, M. Mah, J. Oppenheimer, K. Stewardson, F. Zalzal, S. H. Ambrose, G. Ayodo, H. L. J. Gates, A. O. Gidna, M. Katongo, A. Kwekason, A. Z. P. Mabulla, G. S. Mudenda, E. K. Ndiema, C. Nelson, P. Robertshaw, D. J. Kennett, F. K. Manthi, D. Reich, Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. *Science*. **365** (2019).
 45. Kalkauskas, U. Perron, Y. Sun, N. Goldman, G. Baele, S. Guindon, N. De Maio, Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLOS Computational Biology*. **17**, 1–27 (2021).

46. L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes, A. Wilson, African populations and the evolution of human mitochondrial DNA. *Science*. **253**, 1503–1507 (1991).
47. P. A. Underhill, G. Passarino, A. A. Lin, P. Shen, M. Mirazón Lahr, R. A. Foley, P. J. Oefner, L. L. Cavalli-Sforza, The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Annals of Human Genetics*. **65**, 43–62 (2001).
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. *Journal of Archaeological Science*. **56**, 73–84 (2015).
49. B. Llamas, L. Fehren-Schmitz, G. Valverde, J. Soubrier, S. Mallick, N. Rohland, S. Nordenfelt, C. Valdiosera, S. M. Richards, A. Rohrlach, M. I. B. Romero, I. F. Espinoza, E. T. Cagigao, L. W. Jiménez, K. Makowski, I. S. L. Reyna, J. M. Lory, J. A. B. Torrez, M. A. Rivera, R. L. Burger, M. C. Ceruti, J. Reinhard, R. S. Wells, G. Politis, C. M. Santoro, V. G. Standen, C. Smith, D. Reich, S. Y. W. Ho, A. Cooper, W. Haak, Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Science Advances*. **2** (2016).
50. J. V. Moreno-Mayar, L. Vinner, P. de Barros Damgaard, C. de la Fuente, J. Chan, J. P. Spence, M. E. Allentoft, T. Vimala, F. Racimo, T. Pinotti, S. Rasmussen, A. Margaryan, M. Iraeta Orbegozo, D. Mylopotamitaki, M. Wooller, C. Bataille, L. Becerra-Valdivia, D. Chivall, D. Comeskey, T. Devière, D. K. Grayson, L. George, H. Harry, V. Alexandersen, C. Primeau, J. Erlandson, C. Rodrigues-Carvalho, S. Reis, M. Q. R. Bastos, J. Cybulski, C. Vullo, F. Morello, M. Vilar, S. Wells, K. Gregersen, K. L. Hansen, N. Lynnerup, M. Mirazón Lahr, K. Kjær, A. Strauss, M. Alfonso-Durruty, A. Salas, H. Schroeder, T. Higham, R. S. Malhi, J. T. Rasic, L. Souza, F. R. Santos, A.-S. Malaspinas, M. Sikora, R. Nielsen, Y. S. Song, D. J. Meltzer, E. Willerslev, Early human dispersals within the Americas. *Science*. **362** (2018).
51. R. A. Fisher, *The Genetical Theory of Natural Selection* (Clarendon, 1930).
52. S. Wright, Evolution in Mendelian populations. *Genetics*. **16**, 97–159 (1931).
53. J. F. C. Kingman, The coalescent. *Stochastic processes and their applications*. **13**, 235–248 (1982).
54. G. A. T. McVean, N. J. Cardin, Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*. **360**, 1387–1393 (2005).
55. M. D. Rasmussen, M. J. Hubisz, I. Gronau, A. Siepel, Genome-Wide Inference of Ancestral Recombination Graphs. *PLOS Genetics*. **10**, 1–27 (2014).
56. K. Harris, From a database of genomes to a forest of evolutionary trees. *Nature Genetics*. **51**, 1306–1307 (2019).
57. C. L. Scheib, R. Hui, E. D’Atanasio, A. W. Wohns, S. A. Inskip, A. Rose, C. Cessford, T. C. O’Connell, J. E. Robb, C. Evans, R. Patten, T. Kivisild, East Anglian early Neolithic monument burial linked to contemporary Megaliths. *Annals of Human Biology*. **46**, 145–149 (2019).

58. J. Stern, L. Speidel, N. A. Zaitlen, R. Nielsen, Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *The American Journal of Human Genetics*. **108**, 219–239 (2021).
59. L. Speidel, L. Cassidy, R. W. Davies, G. Hellenthal, P. Skoglund, S. R. Myers, Inferring population histories for ancient genomes using genome-wide genealogies. *bioRxiv* (2021).
60. N. K. Schaefer, B. Shapiro, R. E. Green, An ancestral recombination graph of human, Neanderthal, and Denisovan genomes. *Science Advances*. **7** (2021).
61. R. Nielsen, J. M. Akey, M. Jakobsson, J. K. Pritchard, S. Tishkoff, E. Willerslev, Tracing the peopling of the world through genomics. *Nature*. **541**, 302–310 (2017).
62. J. J. Michaelson, Y. Shi, M. Gujral, H. Zheng, D. Malhotra, X. Jin, M. Jian, G. Liu, D. Greer, A. Bhandari, W. Wu, R. Corominas, A. Peoples, A. Koren, A. Gore, S. Kang, G. N. Lin, J. Estabillio, T. Gadomski, B. Singh, K. Zhang, N. Akshoomoff, C. Corsello, S. McCarroll, L. M. Iakoucheva, Y. Li, J. Wang, J. Sebat, Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*. **151**, 1431–1442 (2012).
63. V. Nesta, D. Tafur, C. R. Beck, Hotspots of Human Mutation. *Trends in Genetics* (2020).
64. R. Hui, E. D’Atanasio, L. M. Cassidy, C. L. Scheib, T. Kivisild, Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Scientific Reports*. **10**, 18542 (2020).
65. K. P. Murphy, Y. Weiss, M. I. Jordan, Loopy belief propagation for approximate inference: An empirical study, in Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Stockholm, Sweden, 30 July - 1 August 1999.
66. tskit developers, *tskit* (2021; <https://github.com/tskit-dev/tskit>). doi:10.5281/zenodo.5465773.
67. Code to reproduce the analyses presented in this paper can be found at https://github.com/awohns/unified_genealogy_paper. doi:10.5281/zenodo.5172 d104.
68. Unified tree sequences of the HGDP, SGDP, and TGP autosomes are available from Zenodo at <https://zenodo.org/record/5495535>. doi:10.5281/zenodo.5495535.
69. Unified tree sequences of the HGDP, SGDP, TGP and high-coverage sequenced ancient autosomes are available from Zenodo at <https://zenodo.org/record/5512994>. doi: 10.5281/zenodo.5512994.
70. E. Lowy-Gallego, S. Fairley, X. Zheng-Bradley, M. Ruffier, L. Clarke, P. Flicek, The 1000 Genomes Project Consortium, Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Res*. **4**, 50 (2019).
71. M. E. Allentoft, M. Sikora, K.-G. Sjögren, S. Rasmussen, M. Rasmussen, J. Stenderup, P. B. Damgaard, H. Schroeder, T. Ahlström, L. Vinner, A.-S. Malaspinas, A. Margaryan, T. Higham, D. Chivall, N. Lynnerup, L. Harvig, J. Baron, P. D. Casa, P. Dąbrowski, P. R. Duffy, A. V. Ebel, A. Epimakhov, K. Frei, M. Furmanek, T. Gralak, A. Gromov, S. Gronkiewicz, G. Grupe, T. Hajdu, R. Jarysz, V. Khartanovich, A. Khokhlov, V. Kiss, J. Kolář, A. Kriiska, I. Lasak, C. Longhi, G. McGlynn, A. Merkevicius, I. Merkyte, M. Metspalu, R. Mkrtychyan, V. Moiseyev, L. Paja, G. Pálfi, D. Pokutta, \Lukasz Pospieszny,

- T. D. Price, L. Saag, M. Sablin, N. Shishlina, V. Smrčka, V. I. Soenov, V. Szeverényi, G. Tóth, S. V. Trifanova, L. Varul, M. Vicze, L. Yepiskoposyan, V. Zhitenev, L. Orlando, T. Sicheritz-Pontén, S. Brunak, R. Nielsen, K. Kristiansen, E. Willerslev, Population genomics of Bronze Age Eurasia. *Nature*. **522**, 167–172 (2015).
72. C. E. G. Amorim, S. Vai, C. Posth, A. Modi, I. Koncz, S. Hakenbeck, M. C. La Rocca, B. Mende, D. Bobo, W. Pohl, L. P. Baricco, E. Bedini, P. Francalacci, C. Giostra, T. Vida, D. Winger, U. von Freeden, S. Ghirotto, M. Lari, G. Barbujani, J. Krause, D. Caramelli, P. J. Geary, K. R. Veeramah, Understanding 6th-century barbarian social organization and migration through paleogenomics. *Nat. Commun.* **9**, 3547 (2018).
73. M. L. Antonio, Z. Gao, H. M. Moots, M. Lucci, F. Candilio, S. Sawyer, V. Oberreiter, D. Calderon, K. Devitofranceschi, R. C. Aikens, S. Aneli, F. Bartoli, A. Bedini, O. Cheronet, D. J. Cotter, D. M. Fernandes, G. Gasperetti, R. Grifoni, A. Guidi, F. La Pastina, E. Loreti, D. Manacorda, G. Matullo, S. Morretta, A. Nava, V. Fiocchi Nicolai, F. Nomi, C. Pavolini, M. Pentiricci, P. Pergola, M. Piranomonte, R. Schmidt, G. Spinola, A. Sperduti, M. Rubini, L. Bondioli, A. Coppa, R. Pinhasi, J. K. Pritchard, Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science*. **366**, 708–714 (2019).
74. S. Brace, Y. Diekmann, T. J. Booth, L. van Dorp, Z. Faltyskova, N. Rohland, S. Mallick, I. Olalde, M. Ferry, M. Michel, J. Oppenheimer, N. Broomandkhoshbacht, K. Stewardson, R. Martiniano, S. Walsh, M. Kayser, S. Charlton, G. Hellenthal, I. Armit, R. Schulting, O. E. Craig, A. Sheridan, M. Parker Pearson, C. Stringer, D. Reich, M. G. Thomas, I. Barnes, Ancient genomes indicate population replacement in Early Neolithic Britain. *Nature Ecology & Evolution*. **3**, 765–771 (2019).
75. F. Broushaki, M. G. Thomas, V. Link, S. López, L. van Dorp, K. Kirsanow, Z. Hofmanová, Y. Diekmann, L. M. Cassidy, D. Díez-del-Molino, A. Kousathanas, C. Sell, H. K. Robson, R. Martiniano, J. Blöcher, A. Scheu, S. Kreutzer, R. Bollongino, D. Bobo, H. Davoudi, O. Munoz, M. Currat, K. Abdi, F. Biglari, O. E. Craig, D. G. Bradley, S. Shennan, K. R. Veeramah, M. Mashkour, D. Wegmann, G. Hellenthal, J. Burger, Early Neolithic genomes from the eastern Fertile Crescent. *Science*. **353**, 499–503 (2016).
76. L. M. Cassidy, R. Martiniano, E. M. Murphy, M. D. Teasdale, J. Mallory, B. Hartwell, D. G. Bradley, 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).
77. P. de B. Damgaard, N. Marchi, S. Rasmussen, M. Peyrot, G. Renaud, T. Korneliussen, J. V. Moreno-Mayar, M. W. Pedersen, A. Goldberg, E. Usmanova, N. Baimukhanov, V. Loman, L. Hedeager, A. G. Pedersen, K. Nielsen, G. Afanasiev, K. Akmatov, A. Aldashev, A. Alpaslan, G. Baimbetov, V. I. Bazaliiskii, A. Beisenov, B. Boldbaatar, B. Boldgiv, C. Dorzhu, S. Ellingvag, D. Erdenebaatar, R. Dajani, E. Dmitriev, V. Evdokimov, K. M. Frei, A. Gromov, A. Goryachev, H. Hakonarson, T. Hegay, Z. Khachatryan, R. Khaskhanov, E. Kitov, A. Kolbina, T. Kubatbek, A. Kukushkin, I. Kukushkin, N. Lau, A. Margaryan, I. Merkyte, I. V. Mertz, V. K. Mertz, E. Mijiddorj, V. Moiyesev, G. Mukhtarova, B. Nurmukhanbetov, Z. Orozbekova, I. Panyushkina, K. Pieta, V. Smrčka, I. Shevnina, A. Logvin, K.-G. Sjögren, T. Štolcová, A. M. Taravella, K. Tashbaeva, A. Tkachev, T. Tulegenov, D. Voyakin, L. Yepiskoposyan, S. Undrakhbold, V. Varfolomeev, A. Weber, M. A. Wilson Sayres, N. Kradin, M. E.

- Allentoft, L. Orlando, R. Nielsen, M. Sikora, E. Heyer, K. Kristiansen, E. Willerslev, 137 ancient human genomes from across the Eurasian steppes. *Nature*. **557**, 369–374 (2018).
78. P. de B. Damgaard, R. Martiniano, J. Kamm, J. V. Moreno-Mayar, G. Kroonen, M. Peyrot, G. Barjamovic, S. Rasmussen, C. Zacho, N. Baimukhanov, V. Zaibert, V. Merz, A. Biddanda, I. Merz, V. Loman, V. Evdokimov, E. Usmanova, B. Hemphill, A. Seguin-Orlando, F. E. Yediay, I. Ullah, K.-G. Sjögren, K. H. Iversen, J. Choin, C. de la Fuente, M. Ilardo, H. Schroeder, V. Moiseyev, A. Gromov, A. Polyakov, S. Omura, S. Y. Senyurt, H. Ahmad, C. McKenzie, A. Margaryan, A. Hameed, A. Samad, N. Gul, M. H. Khokhar, O. I. Goriunova, V. I. Bazaliiskii, J. Novembre, A. W. Weber, L. Orlando, M. E. Allentoft, R. Nielsen, K. Kristiansen, M. Sikora, A. K. Outram, R. Durbin, E. Willerslev, The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science*. **360**, eaar7711 (2018).
 79. S. S. Ebenesersdóttir, M. Sandoval-Velasco, E. D. Gunnarsdóttir, A. Jagadeesan, V. B. Guðmundsdóttir, E. L. Thordardóttir, M. S. Einarsson, K. H. S. Moore, Á. Sigurðsson, D. N. Magnúsdóttir, H. Jónsson, S. Snorraddóttir, E. Hovig, P. Møller, I. Kockum, T. Olsson, L. Alfredsson, T. F. Hansen, T. Werge, G. L. Cavalleri, E. Gilbert, C. Lalueza-Fox, J. W. Walser, S. Kristjánsson, S. Gopalakrishnan, L. Árnadóttir, Ó. Þ. Magnússon, M. T. P. Gilbert, K. Stefánsson, A. Helgason, Ancient genomes from Iceland reveal the making of a human population. *Science*. **360**, 1028–1032 (2018).
 80. M. Feldman, E. Fernández-Domínguez, L. Reynolds, D. Baird, J. Pearson, I. Hershkovitz, H. May, N. Goring-Morris, M. Benz, J. Gresky, R. A. Bianco, A. Fairbairn, G. Mustafaoğlu, P. W. Stockhammer, C. Posth, W. Haak, C. Jeong, J. Krause, Late Pleistocene human genome suggests a local origin for the first farmers of central Anatolia. *Nat. Commun.* **10**, 1218 (2019).
 81. M. Feldman, D. M. Master, R. A. Bianco, M. Burri, P. W. Stockhammer, A. Mitnik, A. J. Aja, C. Jeong, J. Krause, Ancient DNA sheds light on the genetic origins of early Iron Age Philistines. *Science Advances*. **5**, eaax0061 (2019).
 82. C. de la Fuente, M. C. Ávila-Arcos, J. Galimany, M. L. Carpenter, J. R. Homburger, A. Blanco, P. Contreras, D. Cruz Dávalos, O. Reyes, M. San Roman, A. Moreno-Estrada, P. F. Campos, C. Eng, S. Huntsman, E. G. Burchard, A.-S. Malaspinas, C. D. Bustamante, E. Willerslev, E. Llop, R. A. Verdugo, M. Moraga, Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia. *Proc. Natl. Acad. Sci.* **115**, E4006–E4012 (2018).
 83. D. M. Fernandes, D. Strapagiel, P. Borówka, B. Marciniak, E. Żądzińska, K. Sirak, V. Siska, R. Grygiel, J. Carlsson, A. Manica, W. Lorkiewicz, R. Pinhasi, A genomic Neolithic time transect of hunter-farmer admixture in central Poland. *Scientific Reports*. **8**, 14879 (2018).
 84. P. Flegontov, N. E. Altınışık, P. Changmai, N. Rohland, S. Mallick, N. Adamski, D. A. Bolnick, N. Broomandkhoshbacht, F. Candilio, B. J. Culleton, O. Flegontova, T. M. Friesen, C. Jeong, T. K. Harper, D. Keating, D. J. Kennett, A. M. Kim, T. C. Lamnidis, A. M. Lawson, I. Olalde, J. Oppenheimer, B. A. Potter, J. Raff, R. A. Sattler, P. Skoglund, K. Stewardson, E. J. Vajda, S. Vasilyev, E. Veselovskaya, M. G. Hayes, D. H. O'Rourke, J. Krause, R. Pinhasi, D. Reich, S. Schiffels, Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America. *Nature*. **570**, 236–240 (2019).

85. R. Fregel, F. L. Méndez, Y. Bokbot, D. Martín-Socas, M. D. Camalich-Massieu, J. Santana, J. Morales, M. C. Ávila-Arcos, P. A. Underhill, B. Shapiro, G. Wojcik, M. Rasmussen, A. E. R. Soares, J. Kapp, A. Sockell, F. J. Rodríguez-Santos, A. Mikdad, A. Trujillo-Mederos, C. D. Bustamante, Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. *Proc. Natl. Acad. Sci.* **115**, 6774–6779 (2018).
86. Q. Fu, A. Mittnik, P. L. F. Johnson, K. Bos, M. Lari, R. Bollongino, C. Sun, L. Giemsch, R. Schmitz, J. Burger, A. M. Ronchitelli, F. Martini, R. G. Cremonesi, J. Svoboda, P. Bauer, D. Caramelli, S. Castellano, D. Reich, S. Pääbo, J. Krause, A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. *Curr. Biol.* **23**, 553–559 (2013).
87. Q. Fu, M. Hajdinjak, O. T. Moldovan, S. Constantin, S. Mallick, P. Skoglund, N. Patterson, N. Rohland, I. Lazaridis, B. Nickel, B. Viola, K. Prüfer, M. Meyer, J. Kelso, D. Reich, S. Pääbo, An early modern human from Romania with a recent Neanderthal ancestor. *Nature*. **524**, 216–219 (2015).
88. Q. Fu, C. Posth, M. Hajdinjak, M. Petr, S. Mallick, D. Fernandes, A. Furtwängler, W. Haak, M. Meyer, A. Mittnik, B. Nickel, A. Peltzer, N. Rohland, V. Slon, S. Talamo, I. Lazaridis, M. Lipson, I. Mathieson, S. Schiffels, P. Skoglund, A. P. Derevianko, N. Drozdov, V. Slavinsky, A. Tsybankov, R. G. Cremonesi, F. Mallegni, B. Gély, E. Vacca, M. R. G. Morales, L. G. Straus, C. Neugebauer-Maresch, M. Teschler-Nicola, S. Constantin, O. T. Moldovan, S. Benazzi, M. Peresani, D. Coppola, M. Lari, S. Ricci, A. Ronchitelli, F. Valentin, C. Thevenet, K. Wehrberger, D. Grigorescu, H. Rougier, I. Crevecoeur, D. Flas, P. Semal, M. A. Mannino, C. Cupillard, H. Bocherens, N. J. Conard, K. Harvati, V. Moiseyev, D. G. Drucker, J. Svoboda, M. P. Richards, D. Caramelli, R. Pinhasi, J. Kelso, N. Patterson, J. Krause, S. Pääbo, D. Reich, The genetic history of Ice Age Europe. *Nature*. **534**, 200–205 (2016).
89. C. Gamba, E. R. Jones, M. D. Teasdale, R. L. McLaughlin, G. Gonzalez-Fortes, V. Mattiangeli, L. Domboróczki, I. Kővári, I. Pap, A. Anders, A. Whittle, J. Dani, P. Raczky, T. F. G. Higham, M. Hofreiter, D. G. Bradley, R. Pinhasi, Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5**, 5257 (2014).
90. G. González-Fortes, E. R. Jones, E. Lightfoot, C. Bonsall, C. Lazar, A. Grandal-d'Anglade, M. D. Garralda, L. Drak, V. Siska, A. Simalcsik, A. Boroneanț, J. R. V. Romaní, M. V. Rodríguez, P. Arias, R. Pinhasi, A. Manica, M. Hofreiter, 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).
91. G. González-Fortes, F. Tassi, E. Trucchi, K. Henneberger, J. L. A. Paijmans, D. Díez-Del-Molino, H. Schroeder, R. R. Susca, C. Barroso-Ruíz, F. J. Bermudez, C. Barroso-Medina, A. M. S. Bettencourt, H. A. Sampaio, A. Grandal-d'Anglade, A. Salas, A. de Lombera-Hermida, R. Fabregas Valcarce, M. Vaquero, S. Alonso, M. Lozano, X. P. Rodríguez-Alvarez, C. Fernández-Rodríguez, A. Manica, M. Hofreiter, G. Barbujani, A western route of prehistoric human migration from Africa into the Iberian Peninsula. *Proc Biol Sci.* **286**, 20182288 (2019).

92. T. Günther, H. Malmström, E. M. Svensson, A. Omrak, F. Sánchez-Quinto, G. M. Kılınç, M. Krzewińska, G. Eriksson, M. Fraser, H. Edlund, A. R. Munters, A. Coutinho, L. G. Simões, M. Vicente, A. Sjölander, B. Jansen Sellevold, R. Jørgensen, P. Claes, M. D. Shriver, C. Valdiosera, M. G. Netea, J. Apel, K. Lidén, B. Skar, J. Storå, A. Götherström, M. Jakobsson, Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLOS Biology*. **16**, 1–22 (2018).
93. T. Günther, C. Valdiosera, H. Malmström, I. Ureña, R. Rodriguez-Varela, Ó. O. Sverrisdóttir, E. A. Daskalaki, P. Skoglund, T. Naidoo, E. M. Svensson, J. M. Bermúdez de Castro, E. Carbonell, M. Dunn, J. Storå, E. Iriarte, J. L. Arsuaga, J.-M. Carretero, A. Götherström, M. Jakobsson, Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc. Natl. Acad. Sci.* **112**, 11917–11922 (2015).
94. M. Haber, C. Doumet-Serhal, C. Scheib, Y. Xue, P. Danecek, M. Mezzavilla, S. Youhanna, R. Martiniano, J. Prado-Martinez, M. Szpak, E. Matisoo-Smith, H. Schutkowski, R. Mikulski, P. Zalloua, T. Kivisild, C. Tyler-Smith, Continuity and Admixture in the Last Five Millennia of Levantine History from Ancient Canaanite and Present-Day Lebanese Genome Sequences. *The American Journal of Human Genetics*. **101**, 274–282 (2017).
95. M. Haber, C. Doumet-Serhal, C. L. Scheib, Y. Xue, R. Mikulski, R. Martiniano, B. Fischer-Genz, H. Schutkowski, T. Kivisild, C. Tyler-Smith, A Transient Pulse of Genetic Admixture from the Crusaders in the Near East Identified from Ancient Genome Sequences. *The American Journal of Human Genetics*. **104**, 977–984 (2019).
96. M. Hajdinjak, Q. Fu, A. Hübner, M. Petr, F. Mafessoni, S. Grote, P. Skoglund, V. Narasimham, H. Rougier, I. Crevecoeur, P. Semal, M. Soressi, S. Talamo, J.-J. Hublin, I. Gušić, Ž. Kućan, P. Rudan, L. V. Golovanova, V. B. Doronichev, C. Posth, J. Krause, P. Korlević, S. Nagel, B. Nickel, M. Slatkin, N. Patterson, D. Reich, K. Prüfer, M. Meyer, S. Pääbo, J. Kelso, Reconstructing the genetic history of late Neanderthals. *Nature*. **555**, 652–656 (2018).
97. É. Harney, H. May, D. Shalem, N. Rohland, S. Mallick, I. Lazaridis, R. Sarig, K. Stewardson, S. Nordenfelt, N. Patterson, I. HersHKovitz, D. Reich, Ancient DNA from Chalcolithic Israel reveals the role of population mixture in cultural transformation. *Nat. Commun.* **9**, 3336 (2018).
98. É. Harney, A. Nayak, N. Patterson, P. Joglekar, V. Mushrif-Tripathy, S. Mallick, N. Rohland, J. Sedig, N. Adamski, R. Bernardos, N. Broomandkhoshbacht, B. J. Culleton, M. Ferry, T. K. Harper, M. Michel, J. Oppenheimer, K. Stewardson, Z. Zhang, Harashawaradhana, M. S. Bartwal, S. Kumar, S. C. Diyundi, P. Roberts, N. Boivin, D. J. Kennett, K. Thangaraj, D. Reich, N. Rai, Ancient DNA from the skeletons of Roopkund Lake reveals Mediterranean migrants in India. *Nat. Commun.* **10**, 3670 (2019).
99. Z. Hofmanová, S. Kreutzer, G. Hellenthal, C. Sell, Y. Diekmann, D. Díez-del-Molino, L. van Dorp, S. López, A. Kousathanas, V. Link, K. Kirsanow, L. M. Cassidy, R. Martiniano, M. Strobel, A. Scheu, K. Kotsakis, P. Halstead, S. Triantaphyllou, N. Kyparissi-Apostolika, D. Urem-Kotsou, C. Ziota, F. Adaktylou, S. Gopalan, D. M. Bobo, L. Winkelbach, J. Blöcher, M. Unterländer, C. Leuenberger, Ç. Çilingiroğlu, B. Horejs, F. Gerritsen, S. J. Shennan, D. G. Bradley, M. Currat, K. R. Veeramah, D. Wegmann, M.

- G. Thomas, C. Papageorgopoulou, J. Burger, Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl. Acad. Sci.* **113**, 6886–6891 (2016).
100. M. Järve, L. Saag, C. L. Scheib, A. K. Pathak, F. Montinaro, L. Pagani, R. Flores, M. Guellil, L. Saag, K. Tambets, A. Kushniarevich, A. Solnik, L. Varul, S. Zadnikov, O. Petrauskas, M. Avramenko, B. Magomedov, S. Didenko, G. Toshev, I. Bruyako, D. Grechko, V. Okatenko, K. Gorbenko, O. Smyrnov, A. Heiko, R. Reid, S. Sapiehin, S. Sirotin, A. Tairov, A. Beisenov, M. Starodubtsev, V. Vasilev, A. Nechvaloda, B. Atabiev, S. Litvinov, N. Ekomasova, M. Dzhaubermezov, S. Voroniatov, O. Utevska, I. Shramko, E. Khusnutdinova, M. Metspalu, N. Savelev, A. Kriiska, T. Kivisild, R. Villems, 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).
 101. C. Jeong, O. Balanovsky, E. Lukianova, N. Kahbatkyzy, P. Flegontov, V. Zaporozhchenko, A. Immel, C.-C. Wang, O. Ixan, E. Khussainova, B. Bekmanov, V. Zaibert, M. Lavryashina, E. Pocheshkhova, Y. Yusupov, A. Agdzhoyan, S. Koshel, A. Bukin, P. Nymadawa, S. Turdikulova, D. Dalimova, M. Churnosov, R. Skhalyakho, D. Daragan, Y. Bogunov, A. Bogunova, A. Shtrunov, N. Dubova, M. Zhabagin, L. Yepiskoposyan, V. Churakov, N. Pislegin, L. Damba, L. Saroyants, K. Dibirova, L. Atramentova, O. Utevska, E. Idrisov, E. Kamenshchikova, I. Evseeva, M. Metspalu, A. K. Outram, M. Robbeets, L. Djansugurova, E. Balanovska, S. Schiffels, W. Haak, D. Reich, J. Krause, The genetic history of admixture across inner Eurasia. *Nature Ecology & Evolution.* **3**, 966–976 (2019).
 102. C. Jeong, A. T. Ozga, D. B. Witonsky, H. Malmström, H. Edlund, C. A. Hofman, R. W. Hagan, M. Jakobsson, C. M. Lewis, M. S. Aldenderfer, A. Di Rienzo, C. Warinner, 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.* **113**, 7485–7490 (2016).
 103. C. Jeong, S. Wilkin, T. Amgalantugs, A. S. Bouwman, W. T. T. Taylor, R. W. Hagan, S. Bromage, S. Tsolmon, C. Trachsel, J. Grossmann, J. Littleton, C. A. Makarewicz, J. Krigbaum, M. Burri, A. Scott, G. Davaasambuu, J. Wright, F. Irmer, E. Myagmar, N. Boivin, M. Robbeets, F. J. Rühli, J. Krause, B. Frohlich, J. Hendy, C. Warinner, Bronze Age population dynamics and the rise of dairy pastoralism on the eastern Eurasian steppe. *Proc. Natl. Acad. Sci.* **115**, E11248–E11255 (2018).
 104. E. R. Jones, G. Gonzalez-Fortes, S. Connell, V. Siska, A. Eriksson, R. Martiniano, R. L. McLaughlin, M. Gallego Llorente, L. M. Cassidy, C. Gamba, T. Meshveliani, O. Bar-Yosef, W. Müller, A. Belfer-Cohen, Z. Matskevich, N. Jakeli, T. F. G. Higham, M. Currat, D. Lordkipanidze, M. Hofreiter, A. Manica, R. Pinhasi, D. G. Bradley, Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* **6**, 8912 (2015).
 105. E. R. Jones, G. Zarina, V. Moiseyev, E. Lightfoot, P. R. Nigst, A. Manica, R. Pinhasi, D. G. Bradley, The Neolithic Transition in the Baltic Was Not Driven by Admixture with Early European Farmers. *Curr. Biol.* **27**, 576–582 (2017).
 106. H. Kanzawa-Kiriyama, K. Kryukov, T. A. Jinam, K. Hosomichi, A. Saso, G. Suwa, S. Ueda, M. Yoneda, A. Tajima, K. Shinoda, I. Inoue, N. Saitou, A partial nuclear genome

of the Jomons who lived 3000 years ago in Fukushima, Japan. *Journal of Human Genetics*. **62**, 213–221 (2017).

107. Keller, A. Graefen, M. Ball, M. Matzas, V. Boisguerin, F. Maixner, P. Leidinger, C. Backes, R. Khairat, M. Forster, B. Stade, A. Franke, J. Mayer, J. Spangler, S. McLaughlin, M. Shah, C. Lee, T. T. Harkins, A. Sartori, A. Moreno-Estrada, B. Henn, M. Sikora, O. Semino, J. Chiaroni, S. Rootsi, N. M. Myres, V. M. Cabrera, P. A. Underhill, C. D. Bustamante, E. E. Vigl, M. Samadelli, G. Cipollini, J. Haas, H. Katus, B. D. O'Connor, M. R. J. Carlson, B. Meder, N. Blin, E. Meese, C. M. Pusch, A. Zink, New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.* **3**, 698 (2012).
108. D. J. Kennett, S. Plog, R. J. George, B. J. Culleton, A. S. Watson, P. Skoglund, N. Rohland, S. Mallick, K. Stewardson, L. Kistler, S. A. LeBlanc, P. M. Whiteley, D. Reich, G. H. Perry, Archaeogenomic evidence reveals prehistoric matrilineal dynasty. *Nat. Commun.* **8**, 14115 (2017).
109. G. M. Kılınç, A. Omrak, F. Özer, T. Günther, A. M. Büyükkarakaya, E. Bıçakçı, D. Baird, H. M. Dönertaş, A. Ghalichi, R. Yaka, D. Koptekin, S. C. Açıkan, P. Parvizi, M. Krzewińska, E. A. Daskalaki, E. Yüncü, N. D. Dağtaş, A. Fairbairn, J. Pearson, G. Mustafaoğlu, Y. S. Erdal, Y. G. Çakan, İ. Togan, M. Somel, J. Storå, M. Jakobsson, A. Götherström, The Demographic Development of the First Farmers in Anatolia. *Curr. Biol.* **26**, 2659–2666 (2016).
110. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*. **15**, 356 (2014).
111. M. Krzewińska, A. Kjellström, T. Günther, C. Hedenstierna-Jonson, T. Zachrisson, A. Omrak, R. Yaka, G. M. Kılınç, M. Somel, V. Sobrado, J. Evans, C. Knipper, M. Jakobsson, J. Storå, A. Götherström, Genomic and Strontium Isotope Variation Reveal Immigration Patterns in a Viking Age Town. *Curr. Biol.* **28**, 2730–2738.e10 (2018).
112. M. Krzewińska, G. M. Kılınç, A. Juras, D. Koptekin, M. Chyleński, A. G. Nikitin, N. Shcherbakov, I. Shuteleva, T. Leonova, L. Kraeva, F. A. Sungatov, A. N. Sultanova, I. Potekhina, S. Łukasik, M. Krenz-Niedbala, L. Dalén, V. Sinika, M. Jakobsson, J. Storå, A. Götherström, Ancient genomes suggest the eastern Pontic-Caspian steppe as the source of western Iron Age nomads. *Science Advances*. **4**, eaat4457 (2018).
113. T. C. Lamnidis, K. Majander, C. Jeong, E. Salmela, A. Wessman, V. Moiseyev, V. Khartanovich, O. Balanovsky, M. Ongyerth, A. Weihmann, A. Sajantila, J. Kelso, S. Pääbo, P. Onkamo, W. Haak, J. Krause, S. Schiffels, Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nat. Commun.* **9**, 5018 (2018).
114. Lazaridis, D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland, S. Mallick, D. Fernandes, M. Novak, B. Gamarra, K. Sirak, S. Connell, K. Stewardson, E. Harney, Q. Fu, G. Gonzalez-Forbes, E. R. Jones, S. A. Roodenberg, G. Lengyel, F. Bocquentin, B. Gasparian, J. M. Monge, M. Gregg, V. Eshed, A. -S. Mizrahi, C. Meiklejohn, F. Gerritsen, L. Bejenaru, M. Blüher, A. Campbell, G. Cavalleri, D. Comas, P. Froguel, E. Gilbert, S. M. Kerr, P. Kovacs, J. Krause, D. McGettigan, M. Merrigan, D. A. Merriwether, S. O'Reilly, M. B. Richards, O. Semino, M. Shamoon-Pour, G. Stefanescu, M. Stumvoll, A. Tönjes, A. Torroni, J. F. Wilson, L. Yengo, N. A. Hovhannisyan, N.

- Patterson, R. Pinhasi, D. Reich, Genomic insights into the origin of farming in the ancient Near East. *Nature*. **536**, 419–424 (2016).
115. Lazaridis, A. Mittnik, N. Patterson, S. Mallick, N. Rohland, S. Pfrengle, A. Furtwängler, A. Peltzer, C. Posth, A. Vasilakis, P. J. P. McGeorge, E. Konsolaki-Yannopoulou, G. Korres, H. Martlew, M. Michalodimitrakakis, M. Özşait, N. Özşait, A. Papathanasiou, M. Richards, S. A. Roodenberg, Y. Tzedakis, R. Arnott, D. M. Fernandes, J. R. Hughey, D. M. Lotakis, P. A. Navas, Y. Maniatis, J. A. Stamatoyannopoulos, K. Stewardson, P. Stockhammer, R. Pinhasi, D. Reich, J. Krause, G. Stamatoyannopoulos, Genetic origins of the Minoans and Mycenaeans. *Nature*. **548**, 214–218 (2017).
 116. J. Lindo, A. Achilli, U. A. Perego, D. Archer, C. Valdiosera, B. Petzelt, J. Mitchell, R. Worl, E. J. Dixon, T. E. Fifield, M. Rasmussen, E. Willerslev, J. S. Cybulski, B. M. Kemp, M. DeGiorgio, R. S. Malhi, Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity. *Proc. Natl. Acad. Sci.* **114**, 4093–4098 (2017).
 117. J. Lindo, R. Haas, C. Hofman, M. Apata, M. Moraga, R. A. Verdugo, J. T. Watson, C. V. Llave, D. Witonsky, C. Beall, C. Warinner, J. Novembre, M. Aldenderfer, A. D. Rienzo, The genetic prehistory of the Andean highlands 7000 years BP through European contact. *Science Advances*. **4**, eaau4921 (2018).
 118. M. Lipson, P. Skoglund, M. Spriggs, F. Valentin, S. Bedford, R. Shing, H. Buckley, I. Phillip, G. K. Ward, S. Mallick, N. Rohland, N. Broomandkhoshbacht, O. Cheronet, M. Ferry, T. K. Harper, M. Michel, J. Oppenheimer, K. Sirak, K. Stewardson, K. Auckland, A. V. S. Hill, K. Maitland, S. J. Oppenheimer, T. Parks, K. Robson, T. N. Williams, D. J. Kennett, A. J. Mentzer, R. Pinhasi, D. Reich, Population Turnover in Remote Oceania Shortly after Initial Settlement. *Curr. Biol.* **28**, 1157-1165.e7 (2018).
 119. M. Lipson, A. Szécsényi-Nagy, S. Mallick, A. Pósa, B. Stégmár, V. Keerl, N. Rohland, K. Stewardson, M. Ferry, M. Michel, J. Oppenheimer, N. Broomandkhoshbacht, E. Harney, S. Nordenfelt, B. Llamas, B. Gusztáv Mende, K. Köhler, K. Oross, M. Bondár, T. Marton, A. Osztás, J. Jakucs, T. Paluch, F. Horváth, P. Csengeri, J. Koós, K. Sebők, A. Anders, P. Raczky, J. Regenye, J. P. Barna, S. Fábrián, G. Serlegi, Z. Toldi, E. Gyöngyvér Nagy, J. Dani, E. Molnár, G. Pálfi, L. Márk, B. Melegh, Z. Bánfai, L. Domboróczki, J. Fernández-Eraso, J. Antonio Mujika-Alustiza, C. Alonso Fernández, J. Jiménez Echevarría, R. Bollongino, J. Orschiedt, K. Schierhold, H. Meller, A. Cooper, J. Burger, E. Bánffy, K. W. Alt, C. Lalueza-Fox, W. Haak, D. Reich, Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature*. **551**, 368–372 (2017).
 120. M. Lipson, I. Ribot, S. Mallick, N. Rohland, I. Olalde, N. Adamski, N. Broomandkhoshbacht, A. M. Lawson, S. López, J. Oppenheimer, K. Stewardson, R. N. Asombang, H. Bocherens, N. Bradman, B. J. Culleton, E. Cornelissen, I. Crevecoeur, P. de Maret, F. L. M. Fomine, P. Lavachery, C. M. Mindzie, R. Orban, E. Sawchuk, P. Semal, M. G. Thomas, W. Van Neer, K. R. Veeramah, D. J. Kennett, N. Patterson, G. Hellenthal, C. Lalueza-Fox, S. MacEachern, M. E. Prendergast, D. Reich, Ancient West African foragers in the context of African population history. *Nature*. **577**, 665–670 (2020).

121. M. Lipson, O. Cheronet, S. Mallick, N. Rohland, M. Oxenham, M. Pietruszewsky, T. O. Pryce, A. Willis, H. Matsumura, H. Buckley, K. Domett, G. H. Nguyen, H. H. Trinh, A. A. Kyaw, T. T. Win, B. Pradier, N. Broomandkhoshbacht, F. Candilio, P. Changmai, D. Fernandes, M. Ferry, B. Gamarra, E. Harney, J. Kampuansai, W. Kutanan, M. Michel, M. Novak, J. Oppenheimer, K. Sirak, K. Stewardson, Z. Zhang, P. Flegontov, R. Pinhasi, D. Reich, Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science*. **361**, 92–95 (2018).
122. M. Gallego Llorente, E. R. Jones, A. Eriksson, V. Siska, K. W. Arthur, J. W. Arthur, M. C. Curtis, J. T. Stock, M. Coltorti, P. Pieruccini, S. Stretton, F. Brock, T. Higham, Y. Park, M. Hofreiter, D. G. Bradley, J. Bhak, R. Pinhasi, A. Manica, Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science*. **350**, 820–822 (2015).
123. A.-S. Malaspinas, O. Lao, H. Schroeder, M. Rasmussen, M. Raghavan, I. Moltke, P. F. Campos, F. S. Sagredo, S. Rasmussen, V. F. Gonçalves, A. Albrechtsen, M. E. Allentoft, P. L. F. Johnson, M. Li, S. Reis, D. V. Bernardo, M. DeGiorgio, A. T. Duggan, M. Bastos, Y. Wang, J. Stenderup, J. V. Moreno-Mayar, S. Brunak, T. Sicheritz-Ponten, E. Hodges, G. J. Hannon, L. Orlando, T. D. Price, J. D. Jensen, R. Nielsen, J. Heinemeier, J. Olsen, C. Rodrigues-Carvalho, M. M. Lahr, W. A. Neves, M. Kayser, T. Higham, M. Stoneking, S. D. J. Pena, E. Willerslev, Two ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil. *Curr. Biol.* **24**, R1035–R1037 (2014).
124. H. Malmström, T. Günther, E. M. Svensson, A. Juras, M. Fraser, A. R. Munters, Ł. Pospieszny, M. Törv, J. Lindström, A. Götherström, J. Storå, M. Jakobsson, The genomic ancestry of the Scandinavian Battle Axe Culture people and their relation to the broader Corded Ware horizon. *Proceedings of the Royal Society B: Biological Sciences*. **286**, 20191528 (2019).
125. R. Martiniano, A. Caffell, M. Holst, K. Hunter-Mann, J. Montgomery, G. Müldner, R. L. McLaughlin, M. D. Teasdale, W. van Rheeën, J. H. Veldink, L. H. van den Berg, O. Hardiman, M. Carroll, S. Roskams, J. Oxley, C. Morgan, M. G. Thomas, I. Barnes, C. McDonnell, M. J. Collins, D. G. Bradley, Genomic signals of migration and continuity in Britain before the Anglo-Saxons. *Nat. Commun.* **7**, 10326 (2016).
126. R. Martiniano, L. M. Cassidy, R. Ó'Maoldúin, R. McLaughlin, N. M. Silva, L. Manco, D. Fidalgo, T. Pereira, M. J. Coelho, M. Serra, J. Burger, R. Parreira, E. Moran, A. C. Valera, E. Porfirio, R. Boaventura, A. M. Silva, D. G. Bradley, The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLOS Genetics*. **13**, 1–24 (2017).
127. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, S. A. Roodenberg, E. Harney, K. Stewardson, D. Fernandes, M. Novak, K. Sirak, C. Gamba, E. R. Jones, B. Llamas, S. Dryomov, J. Pickrell, J. L. Arsuaga, J. B. de Castro, E. Carbonell, F. Gerritsen, A. Khokhlov, P. Kuznetsov, M. Lozano, H. Meller, O. Mochalov, V. Moiseyev, M. A. R. Guerra, J. Roodenberg, J. M. Vergès, J. Krause, A. Cooper, K. W. Alt, D. Brown, D. Anthony, C. Lalueza-Fox, W. Haak, R. Pinhasi, D. Reich, Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*. **528**, 499–503 (2015).

128. Mathieson, S. Alpaslan-Roodenberg, C. Posth, A. Szécsényi-Nagy, N. Rohland, S. Mallick, I. Olalde, N. Broomandkoshbacht, F. Candilio, O. Cheronet, D. Fernandes, M. Ferry, B. Gamarra, G. G. Fortes, W. Haak, E. Harney, E. Jones, D. Keating, B. Krause-Kyora, I. Kucukkalipci, M. Michel, A. Mittnik, K. Nägele, M. Novak, J. Oppenheimer, N. Patterson, S. Pfrengle, K. Sirak, K. Stewardson, S. Vai, S. Alexandrov, K. W. Alt, R. Andreescu, D. Antonović, A. Ash, N. Atanassova, K. Bacvarov, M. B. Gusztáv, H. Bocherens, M. Bolus, A. Boroneanț, Y. Boyadzhiev, A. Budnik, J. Burmaz, S. Chohadzhiev, N. J. Conard, R. Cottiaux, M. Čuka, C. Cupillard, D. G. Drucker, N. Elenski, M. Francken, B. Galabova, G. Ganetsovski, B. Gély, T. Hajdu, V. Handzhyska, K. Harvati, T. Higham, S. Iliev, I. Janković, I. Karavanić, D. J. Kennett, D. Komšo, A. Kozak, D. Labuda, M. Lari, C. Lazar, M. Leppek, K. Leshtakov, D. L. Vetro, D. Los, I. Lozanov, M. Malina, F. Martini, K. McSweeney, H. Meller, M. Mendušić, P. Mirea, V. Moiseyev, V. Petrova, T. D. Price, A. Simalcsik, L. Sineo, M. Šlaus, V. Slavchev, P. Stanev, A. Starović, T. Szeniczey, S. Talamo, M. Teschler-Nicola, C. Thevenet, I. Valchev, F. Valentin, S. Vasilyev, F. Veljanovska, S. Venelinova, E. Veselovskaya, B. Viola, C. Virag, J. Zaninović, S. Zäuner, P. W. Stockhammer, G. Catalano, R. Krauß, D. Caramelli, G. Zariņa, B. Gaydarska, M. Lillie, A. G. Nikitin, I. Potekhina, A. Papathanasiou, D. Borić, C. Bonsall, J. Krause, R. Pinhasi, D. Reich, The genomic history of southeastern Europe. *Nature*. **555**, 197–203 (2018).
129. H. McColl, F. Racimo, L. Vinner, F. Demeter, T. Gakuhari, J. Víctor Moreno-Mayar, G. Van Driem, U. G. Wilken, A. Seguin-Orlando, C. De la Fuente Castro, S. Wasef, R. Shoocongdej, V. Souksavatdy, T. Sayavongkhamdy, M. M. Saidin, M. E. Allentoft, T. Sato, A. S. Malaspinas, F. A. Aghakhanian, T. Korneliussen, A. Prohaska, A. Margaryan, P. De Barros Damgaard, S. Kaewsutthi, P. Lertrit, T. M. H. Nguyen, H. chun Hung, T. M. Tran, H. N. Truong, G. H. Nguyen, S. Shahidan, K. Wiradnyana, H. Matsumae, N. Shigehara, M. Yoneda, H. Ishida, T. Masuyama, Y. Yamada, A. Tajima, H. Shibata, A. Toyoda, T. Hanihara, S. Nakagome, T. Deviese, A. M. Bacon, P. Durringer, J. L. Ponche, L. Shackelford, E. Patole-Edoumba, A. T. Nguyen, B. Bellina-Pryce, J. C. Galipaud, R. Kinaston, H. Buckley, C. Pottier, S. Rasmussen, T. Higham, R. A. Foley, M. M. Lahr, L. Orlando, M. Sikora, M. E. Phipps, H. Oota, C. Higham, D. M. Lambert, E. Willerslev, The prehistoric peopling of Southeast Asia. *Science*. **361**, 88–92 (2018).
130. Mittnik, C.-C. Wang, S. Pfrengle, M. Daubaras, G. Zariņa, F. Hallgren, R. Allmäe, V. Khartanovich, V. Moiseyev, M. Törv, A. Furtwängler, A. Andrades Valtueña, M. Feldman, C. Economou, M. Oinonen, A. Vasks, E. Balanovska, D. Reich, R. Jankauskas, W. Haak, S. Schiffels, J. Krause, The genetic prehistory of the Baltic Sea region. *Nat Commun*. **9**, 442 (2018).
131. Mittnik, K. Massy, C. Knipper, F. Wittenborn, R. Friedrich, S. Pfrengle, M. Burri, N. Carlich-Witjes, H. Deeg, A. Furtwängler, M. Harbeck, K. von Heyking, C. Kociumaka, I. Kucukkalipci, S. Lindauer, S. Metz, A. Staskiewicz, A. Thiel, J. Wahl, W. Haak, E. Pernicka, S. Schiffels, P. W. Stockhammer, J. Krause, Kinship-based social inequality in Bronze Age Europe. *Science*. **366**, 731–734 (2019).
132. M. Mondal, F. Casals, T. Xu, G. M. Dall’Olio, M. Pybus, M. G. Netea, D. Comas, H. Laayouni, Q. Li, P. P. Majumder, J. Bertranpetit, Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nature Genetics*. **48**, 1066–1070 (2016).

133. J. V. Moreno-Mayar, B. A. Potter, L. Vinner, M. Steinrücken, S. Rasmussen, J. Terhorst, J. A. Kamm, A. Albrechtsen, A.-S. Malaspinas, M. Sikora, J. D. Reuther, J. D. Irish, R. S. Malhi, L. Orlando, Y. S. Song, R. Nielsen, D. J. Meltzer, E. Willerslev, Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*. **553**, 203–207 (2018).
134. J. V. Moreno-Mayar, L. Vinner, P. de Barros Damgaard, C. de la Fuente, J. Chan, J. P. Spence, M. E. Allentoft, T. Vimala, F. Racimo, T. Pinotti, S. Rasmussen, A. Margaryan, M. Iraeta Orbegozo, D. Mylopotamitaki, M. Wooller, C. Bataille, L. Becerra-Valdivia, D. Chivall, D. Comeskey, T. Devière, D. K. Grayson, L. George, H. Harry, V. Alexandersen, C. Primeau, J. Erlandson, C. Rodrigues-Carvalho, S. Reis, M. Q. R. Bastos, J. Cybulski, C. Vullo, F. Morello, M. Vilar, S. Wells, K. Gregersen, K. L. Hansen, N. Lynnerup, M. Mirazón Lahr, K. Kjær, A. Strauss, M. Alfonso-Durruty, A. Salas, H. Schroeder, T. Higham, R. S. Malhi, J. T. Rasic, L. Souza, F. R. Santos, A.-S. Malaspinas, M. Sikora, R. Nielsen, Y. S. Song, D. J. Meltzer, E. Willerslev, Early human dispersals within the Americas. *Science*. **362** (2018), doi:10.1126/science.aav2621.
135. G. Nikitin, P. Stadler, N. Kotova, M. Teschler-Nicola, T. D. Price, J. Hoover, D. J. Kennett, I. Lazaridis, N. Rohland, M. Lipson, D. Reich, Interactions between earliest Linearbandkeramik farmers and central European hunter gatherers at the dawn of European Neolithization. *Sci Rep*. **9**, 19544 (2019).
136. Ning, C.-C. Wang, S. Gao, Y. Yang, X. Zhang, X. Wu, F. Zhang, Z. Nie, Y. Tang, M. Robbeets, J. Ma, J. Krause, Y. Cui, Ancient Genomes Reveal Yamnaya-Related Ancestry and a Potential Source of Indo-European Speakers in Iron Age Tianshan. *Curr. Biol*. **29**, 2526–2532 (2019).
137. Olalde, H. Schroeder, M. Sandoval-Velasco, L. Vinner, I. Lobón, O. Ramirez, S. Civit, P. García Borja, D. C. Salazar-García, S. Talamo, J. María Fullola, F. Xavier Oms, M. Pedro, P. Martínez, M. Sanz, J. Daura, J. Zilhão, T. Marquès-Bonet, M. T. P. Gilbert, C. Lalueza-Fox, A Common Genetic Origin for Early Farmers from Mediterranean Cardial and Central European LBK Cultures. *Mol Biol Evol*. **32**, 3132–3142 (2015).
138. Olalde, M. E. Allentoft, F. Sánchez-Quinto, G. Santpere, C. W. K. Chiang, M. DeGiorgio, J. Prado-Martinez, J. A. Rodríguez, S. Rasmussen, J. Quilez, O. Ramírez, U. M. Marigorta, M. Fernández-Callejo, M. E. Prada, J. M. V. Encinas, R. Nielsen, M. G. Netea, J. Novembre, R. A. Sturm, P. Sabeti, T. Marquès-Bonet, A. Navarro, E. Willerslev, C. Lalueza-Fox, Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature*. **507**, 225–228 (2014).
139. Olalde, S. Brace, M. E. Allentoft, I. Armit, K. Kristiansen, T. Booth, N. Rohland, S. Mallick, A. Szécsényi-Nagy, A. Mittnik, E. Altena, M. Lipson, I. Lazaridis, T. K. Harper, N. Patterson, N. Broomandkhoshbacht, Y. Diekmann, Z. Faltyskova, D. Fernandes, M. Ferry, E. Harney, P. de Knijff, M. Michel, J. Oppenheimer, K. Stewardson, A. Barclay, K. W. Alt, C. Liesau, P. Ríos, C. Blasco, J. V. Miguel, R. M. García, A. A. Fernández, E. Bánffy, M. Bernabò-Brea, D. Billoin, C. Bonsall, L. Bonsall, T. Allen, L. Büster, S. Carver, L. C. Navarro, O. E. Craig, G. T. Cook, B. Cunliffe, A. Denaire, K. E. Dinwiddy, N. Dodwell, M. Ernée, C. Evans, M. Kuchařík, J. F. Farré, C. Fowler, M. Gazenbeek, R. G. Pena, M. Haber-Uriarte, E. Haduch, G. Hey, N. Jowett, T. Knowles, K. Massy, S. Pfengle, P. Lefranc, O. Lemercier, A. Lefebvre, C. H. Martínez, V. G. Olmo, A. B. Ramírez, J. L. Maurandi, T. Majó, J. I. McKinley, K. McSweeney, B. G. Mende, A.

- Modi, G. Kulcsár, V. Kiss, A. Czene, R. Patay, A. Endrődi, K. Köhler, T. Hajdu, T. Szeniczey, J. Dani, Z. Bernert, M. Hoole, O. Cheronet, D. Keating, P. Velemínský, M. Dobeš, F. Candilio, F. Brown, R. F. Fernández, A.-M. Herrero-Corral, S. Tusa, E. Carnieri, L. Lentini, A. Valenti, A. Zanini, C. Waddington, G. Delibes, E. Guerra-Doce, B. Neil, M. Brittain, M. Luke, R. Mortimer, J. Desideri, M. Besse, G. Brücken, M. Furmanek, A. Hałuszko, M. Mackiewicz, A. Rapiński, S. Leach, I. Soriano, K. T. Lillios, J. L. Cardoso, M. P. Pearson, P. Włodarczyk, T. D. Price, P. Prieto, P.-J. Rey, R. Risch, M. A. Rojo Guerra, A. Schmitt, J. Serrallongue, A. M. Silva, V. Smrčka, L. Vergnaud, J. Zilhão, D. Caramelli, T. Higham, M. G. Thomas, D. J. Kennett, H. Fokkens, V. Heyd, A. Sheridan, K.-G. Sjögren, P. W. Stockhammer, J. Krause, R. Pinhasi, W. Haak, I. Barnes, C. Lalueza-Fox, D. Reich, The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature*. **555**, 190–196 (2018).
140. Olalde, S. Mallick, N. Patterson, N. Rohland, V. Villalba-Mouco, M. Silva, K. Dulias, C. J. Edwards, F. Gandini, M. Pala, P. Soares, M. Ferrando-Bernal, N. Adamski, N. Broomandkhoshbacht, O. Cheronet, B. J. Culleton, D. Fernandes, A. M. Lawson, M. Mah, J. Oppenheimer, K. Stewardson, Z. Zhang, J. M. Jiménez Arenas, I. J. Toro Moyano, D. C. Salazar-García, P. Castanyer, M. Santos, J. Tremoleda, M. Lozano, P. García Borja, J. Fernández-Eraso, J. Mujika-Alustiza, C. Barroso, F. J. Bermúdez, E. Viguera Mínguez, J. Burch, N. Coromina, D. Vivó, A. Cebrià, J. M. Fullola, O. García-Puchol, J. I. Morales, F. X. Oms, T. Majó, J. M. Vergès, A. Díaz-Carvajal, I. Ollich-Castanyer, F. J. López-Cachero, A. M. Silva, C. Alonso-Fernández, G. Delibes de Castro, J. Jiménez Echevarría, A. Moreno-Márquez, G. Pascual Berlanga, P. Ramos-García, J. Ramos-Muñoz, E. Vijande Vila, G. Aguilera Arzo, Á. Esparza Arroyo, K. T. Lillios, J. Mack, J. Velasco-Vázquez, A. Waterman, L. Benítez de Lugo Enrich, M. Benito Sánchez, B. Agustí, F. Codina, G. de Prado, A. Estalrich, A. Fernández Flores, C. Finlayson, G. Finlayson, S. Finlayson, F. Giles-Guzmán, A. Rosas, V. Barciela González, G. García Atiénzar, M. S. Hernández Pérez, A. Llanos, Y. Carrión Marco, I. Collado Beneyto, D. López-Serrano, M. Sanz Tormo, A. C. Valera, C. Blasco, C. Liesau, P. Ríos, J. Daura, M. J. de Pedro Michó, A. A. Diez-Castillo, R. Flores Fernández, J. Francès Farré, R. Garrido-Pena, V. S. Gonçalves, E. Guerra-Doce, A. M. Herrero-Corral, J. Juan-Cabanilles, D. López-Reyes, S. B. McClure, M. Merino Pérez, A. Oliver Foix, M. Sanz Borràs, A. C. Sousa, J. M. Vidal Encinas, D. J. Kennett, M. B. Richards, K. Werner Alt, W. Haak, R. Pinhasi, C. Lalueza-Fox, D. Reich, The genomic history of the Iberian Peninsula over the past 8000 years. *Science*. **363**, 1230–1234 (2019).
141. Omrak, T. Günther, C. Valdiosera, E. M. Svensson, H. Malmström, H. Kiesewetter, W. Aylward, J. Storå, M. Jakobsson, A. Götherström, Genomic Evidence Establishes Anatolia as the Source of the European Neolithic Gene Pool. *Curr. Biol.* **26**, 270–275 (2016).
142. J. K. Pickrell, N. Patterson, C. Barbieri, F. Berthold, L. Gerlach, T. Güldemann, B. Kure, S. W. Mpoloka, H. Nakagawa, C. Naumann, M. Lipson, P.-R. Loh, J. Lachance, J. Mountain, C. D. Bustamante, B. Berger, S. A. Tishkoff, B. M. Henn, M. Stoneking, D. Reich, B. Pakendorf, The genetic prehistory of southern Africa. *Nat Commun.* **3**, 1143 (2012).
143. Posth, N. Nakatsuka, I. Lazaridis, P. Skoglund, S. Mallick, T. C. Lamnidis, N. Rohland, K. Nägele, N. Adamski, E. Bertolini, N. Broomandkhoshbacht, A. Cooper, B. J. Culleton, T. Ferraz, M. Ferry, A. Furtwängler, W. Haak, K. Harkins, T. K. Harper, T. Hünemeier,

- A. M. Lawson, B. Llamas, M. Michel, E. Nelson, J. Oppenheimer, N. Patterson, S. Schiffels, J. Sedig, K. Stewardson, S. Talamo, C.-C. Wang, J.-J. Hublin, M. Hubbe, K. Harvati, A. Nuevo Delaunay, J. Beier, M. Francken, P. Kaulicke, H. Reyes-Centeno, K. Rademaker, W. R. Trask, M. Robinson, S. M. Gutierrez, K. M. Prufer, D. C. Salazar-García, E. N. Chim, L. Müller Plumm Gomes, M. L. Alves, A. Liryo, M. Inglez, R. E. Oliveira, D. V. Bernardo, A. Barioni, V. Wesolowski, N. A. Scheifler, M. A. Rivera, C. R. Plens, P. G. Messineo, L. Figuti, D. Corach, C. Scabuzzo, S. Eggers, P. DeBlasis, M. Reindel, C. Méndez, G. Politis, E. Tomasto-Cagigao, D. J. Kennett, A. Strauss, L. Fehren-Schmitz, J. Krause, D. Reich, Reconstructing the Deep Population History of Central and South America. *Cell*. **175**, 1185–1197 (2018).
144. Posth, K. Nägele, H. Colleran, F. Valentin, S. Bedford, K. W. Kami, R. Shing, H. Buckley, R. Kinaston, M. Walworth, G. R. Clark, C. Reepmeyer, J. Flexner, T. Maric, J. Moser, J. Gresky, L. Kiko, K. J. Robson, K. Auckland, S. J. Oppenheimer, A. V. S. Hill, A. J. Mentzer, J. Zech, F. Petchey, P. Roberts, C. Jeong, R. D. Gray, J. Krause, A. Powell, Language continuity despite population replacement in Remote Oceania. *Nat Ecol Evol*. **2**, 731–740 (2018).
145. M. E. Prendergast, M. Lipson, E. A. Sawchuk, I. Olalde, C. A. Ogola, N. Rohland, K. A. Sirak, N. Adamski, R. Bernardos, N. Broomandkhoshbacht, K. Callan, B. J. Culleton, L. Eccles, T. K. Harper, A. M. Lawson, M. Mah, J. Oppenheimer, K. Stewardson, F. Zalzal, S. H. Ambrose, G. Ayodo, H. L. J. Gates, A. O. Gidna, M. Katongo, A. Kwekason, A. Z. P. Mabulla, G. S. Mudenda, E. K. Ndiema, C. Nelson, P. Robertshaw, D. J. Kennett, F. K. Manthi, D. Reich, Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. *Science*. **365** (2019).
146. M. Raghavan, P. Skoglund, K. E. Graf, M. Metspalu, A. Albrechtsen, I. Moltke, S. Rasmussen, T. W. J. Stafford, L. Orlando, E. Metspalu, M. Karmin, K. Tambets, S. Rootsi, R. Mägi, P. F. Campos, E. Balanovska, O. Balanovsky, E. Khusnutdinova, S. Litvinov, L. P. Osipova, S. A. Fedorova, M. I. Voevoda, M. DeGiorgio, T. Sicheritz-Ponten, S. Brunak, S. Demeshchenko, T. Kivisild, R. Villems, R. Nielsen, M. Jakobsson, E. Willerslev, Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*. **505**, 87–91 (2014).
147. M. Raghavan, M. DeGiorgio, A. Albrechtsen, I. Moltke, P. Skoglund, T. S. Korneliussen, B. Grønnow, M. Appelt, H. C. Gulløv, T. M. Friesen, W. Fitzhugh, H. Malmström, S. Rasmussen, J. Olsen, L. Melchior, B. T. Fuller, S. M. Fahrni, T. J. Stafford, V. Grimes, M. A. P. Renouf, J. Cybulski, N. Lynnerup, M. M. Lahr, K. Britton, R. Knecht, J. Arneborg, M. Metspalu, O. E. Cornejo, A.-S. Malaspinas, Y. Wang, M. Rasmussen, V. Raghavan, T. V. O. Hansen, E. Khusnutdinova, T. Pierre, K. Dneprovsky, C. Andreasen, H. Lange, M. G. Hayes, J. Coltrain, V. A. Spitsyn, A. Götherström, L. Orlando, T. Kivisild, R. Villems, M. H. Crawford, F. C. Nielsen, J. Dissing, J. Heinemeier, M. Meldgaard, C. Bustamante, D. H. O'Rourke, M. Jakobsson, M. T. P. Gilbert, R. Nielsen, E. Willerslev, The genetic prehistory of the New World Arctic. *Science*. **345**, 1255832 (2014).
148. M. Raghavan, M. Steinrücken, K. Harris, S. Schiffels, S. Rasmussen, M. DeGiorgio, A. Albrechtsen, C. Valdiosera, M. C. Ávila-Arcos, A.-S. Malaspinas, A. Eriksson, I. Moltke, M. Metspalu, J. R. Homburger, J. Wall, O. E. Cornejo, J. V. Moreno-Mayar, T. S. Korneliussen, T. Pierre, M. Rasmussen, P. F. Campos, P. de Barros Damgaard, M. E.

- Allentoft, J. Lindo, E. Metspalu, R. Rodríguez-Varela, J. Mansilla, C. Henrickson, A. Seguin-Orlando, H. Malmström, T. J. Stafford, S. S. Shringarpure, A. Moreno-Estrada, M. Karmin, K. Tambets, A. Bergström, Y. Xue, V. Warmuth, A. D. Friend, J. Singarayer, P. Valdes, F. Balloux, I. Lebreiro, J. L. Vera, H. Rangel-Villalobos, D. Pettener, D. Luiselli, L. G. Davis, E. Heyer, C. P. E. Zollikofer, M. S. Ponce de León, C. I. Smith, V. Grimes, K.-A. Pike, M. Deal, B. T. Fuller, B. Arriaza, V. Standen, M. F. Luz, F. Ricaut, N. Guidon, L. Osipova, M. I. Voevoda, O. L. Posukh, O. Balanovsky, M. Lavryashina, Y. Bogunov, E. Khusnutdinova, M. Gubina, E. Balanovska, S. Fedorova, S. Litvinov, B. Malyarchuk, M. Derenko, M. J. Mosher, D. Archer, J. Cybulski, B. Petzelt, J. Mitchell, R. Worl, P. J. Norman, P. Parham, B. M. Kemp, T. Kivisild, C. Tyler-Smith, M. S. Sandhu, M. Crawford, R. Villems, D. G. Smith, M. R. Waters, T. Goebel, J. R. Johnson, R. S. Malhi, M. Jakobsson, D. J. Meltzer, A. Manica, R. Durbin, C. D. Bustamante, Y. S. Song, R. Nielsen, E. Willerslev, POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*. **349**, aab3884 (2015).
149. M. Rasmussen, Y. Li, S. Lindgreen, J. S. Pedersen, A. Albrechtsen, I. Moltke, M. Metspalu, E. Metspalu, T. Kivisild, R. Gupta, M. Bertalan, K. Nielsen, M. T. P. Gilbert, Y. Wang, M. Raghavan, P. F. Campos, H. M. Kamp, A. S. Wilson, A. Gledhill, S. Tridico, M. Bunce, E. D. Lorenzen, J. Binladen, X. Guo, J. Zhao, X. Zhang, H. Zhang, Z. Li, M. Chen, L. Orlando, K. Kristiansen, M. Bak, N. Tommerup, C. Bendixen, T. L. Pierre, B. Grønnow, M. Meldgaard, C. Andreassen, S. A. Fedorova, L. P. Osipova, T. F. G. Higham, C. B. Ramsey, T. V. O. Hansen, F. C. Nielsen, M. H. Crawford, S. Brunak, T. Sicheritz-Pontén, R. Villems, R. Nielsen, A. Krogh, J. Wang, E. Willerslev, Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*. **463**, 757–762 (2010).
150. M. Rasmussen, S. L. Anzick, M. R. Waters, P. Skoglund, M. DeGiorgio, T. W. J. Stafford, S. Rasmussen, I. Moltke, A. Albrechtsen, S. M. Doyle, G. D. Poznik, V. Gudmundsdottir, R. Yadav, A.-S. Malaspinas, S. S. 5th White, M. E. Allentoft, O. E. Cornejo, K. Tambets, A. Eriksson, P. D. Heintzman, M. Karmin, T. S. Korneliussen, D. J. Meltzer, T. L. Pierre, J. Stenderup, L. Saag, V. M. Warmuth, M. C. Lopes, R. S. Malhi, S. Brunak, T. Sicheritz-Pontén, I. Barnes, M. Collins, L. Orlando, F. Balloux, A. Manica, R. Gupta, M. Metspalu, C. D. Bustamante, M. Jakobsson, R. Nielsen, E. Willerslev, The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*. **506**, 225–229 (2014).
151. M. Rasmussen, M. Sikora, A. Albrechtsen, T. S. Korneliussen, J. V. Moreno-Mayar, G. D. Poznik, C. P. E. Zollikofer, M. P. de León, M. E. Allentoft, I. Moltke, H. Jónsson, C. Valdiosera, R. S. Malhi, L. Orlando, C. D. Bustamante, T. W. J. Stafford, D. J. Meltzer, R. Nielsen, E. Willerslev, The ancestry and affiliations of Kennewick Man. *Nature*. **523**, 455–458 (2015).
152. R. Rodríguez-Varela, T. Günther, M. Krzewińska, J. Storå, T. H. Gillingwater, M. MacCallum, J. L. Arsuaga, K. Dobney, C. Valdiosera, M. Jakobsson, A. Götherström, L. Girdland-Flink, Genomic Analyses of Pre-European Conquest Human Remains from the Canary Islands Reveal Close Affinity to Modern North Africans. *Curr. Biol*. **27**, 3396–3402 (2017).
153. L. Saag, L. Varul, C. L. Scheib, J. Stenderup, M. E. Allentoft, L. Saag, L. Pagani, M. Reidla, K. Tambets, E. Metspalu, A. Kriiska, E. Willerslev, T. Kivisild, M. Metspalu,

Extensive Farming in Estonia Started through a Sex-Biased Migration from the Steppe. *Curr. Biol.* **27**, 2185–2193 (2017).

154. L. Saag, M. Laneman, L. Varul, M. Malve, H. Valk, M. A. Razzak, I. G. Shirobokov, V. I. Khartanovich, E. R. Mikhaylova, A. Kushniarevich, C. L. Scheib, A. Solnik, T. Reisberg, J. Parik, L. Saag, E. Metspalu, S. Rootsi, F. Montinaro, M. Remm, R. Mägi, E. D’Atanasio, E. R. Crema, D. Díez-del-Molino, M. G. Thomas, A. Kriiska, T. Kivisild, R. Villems, V. Lang, M. Metspalu, K. Tambets, The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East. *Curr. Biol.* **29**, 1701–1711.e16 (2019).
155. L. Scheib, H. Li, T. Desai, V. Link, C. Kendall, G. Dewar, P. W. Griffith, A. Mörseburg, J. R. Johnson, A. Potter, S. L. Kerr, P. Endicott, J. Lindo, M. Haber, Y. Xue, C. Tyler-Smith, M. S. Sandhu, J. G. Lorenz, T. D. Randall, Z. Faltyskova, L. Pagani, P. Danecek, T. C. O’Connell, P. Martz, A. S. Boraas, B. F. Byrd, A. Leventhal, R. Cambra, R. Williamson, L. Lesage, B. Holguin, E. Y.-D. Soto, J. Rosas, M. Metspalu, J. T. Stock, A. Manica, A. Scally, D. Wegmann, R. S. Malhi, T. Kivisild, Ancient human parallel lineages within North America contributed to a coastal expansion. *Science*. **360**, 1024–1027 (2018).
156. Sánchez-Quinto, H. Malmström, M. Fraser, L. Girdland-Flink, E. M. Svensson, L. G. Simões, R. George, N. Hollfelder, G. Burenhult, G. Noble, K. Britton, S. Talamo, N. Curtis, H. Brzobohata, R. Sumberova, A. Götherström, J. Storå, M. Jakobsson, Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society. *Proc. Natl. Acad. Sci.* **116**, 9469–9474 (2019).
157. V. J. Schuenemann, A. Peltzer, B. Welte, W. P. van Pelt, M. Molak, C.-C. Wang, A. Furtwängler, C. Urban, E. Reiter, K. Nieselt, B. Teßmann, M. Francken, K. Harvati, W. Haak, S. Schiffels, J. Krause, Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. *Nat Commun.* **8**, 15694 (2017).
158. S. Schiffels, W. Haak, P. Pääjänen, B. Llamas, E. Popescu, L. Loe, R. Clarke, A. Lyons, R. Mortimer, D. Sayer, C. Tyler-Smith, A. Cooper, R. Durbin, Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nat. Commun.* **7**, 10408 (2016).
159. M. Schlebusch, H. Malmström, T. Günther, P. Sjödén, A. Coutinho, H. Edlund, A. R. Munters, M. Vicente, M. Steyn, H. Soodyall, M. Lombard, M. Jakobsson, Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*. **358**, 652–655 (2017).
160. H. Schroeder, M. Sikora, S. Gopalakrishnan, L. M. Cassidy, P. Maisano Delser, M. Sandoval Velasco, J. G. Schraiber, S. Rasmussen, J. R. Homburger, M. C. Ávila-Arcos, M. E. Allentoft, J. V. Moreno-Mayar, G. Renaud, A. Gómez-Carballa, J. E. Laffoon, R. J. A. Hopkins, T. F. G. Higham, R. S. Carr, W. C. Schaffer, J. S. Day, M. Hoogland, A. Salas, C. D. Bustamante, R. Nielsen, D. G. Bradley, C. L. Hofman, E. Willerslev, Origins and genetic legacies of the Caribbean Taino. *Proc Natl Acad Sci U S A.* **115**, 2341–2346 (2018).
161. H. Schroeder, A. Margaryan, M. Szmyt, B. Theulot, P. Włodarczak, S. Rasmussen, S. Gopalakrishnan, A. Szczepanek, T. Konopka, T. Z. T. Jensen, B. Witkowska, S. Wilk, M. M. Przybył, Łukasz Pospieszny, K.-G. Sjögren, Z. Belka, J. Olsen, K. Kristiansen, E.

- Willerslev, K. M. Frei, M. Sikora, N. N. Johannsen, M. E. Allentoft, Unraveling ancestry, kinship, and violence in a Late Neolithic mass grave. *Proc. Natl. Acad. Sci.* **116**, 10705–10710 (2019).
162. Seguin-Orlando, T. S. Korneliussen, M. Sikora, A.-S. Malaspinas, A. Manica, I. Moltke, A. Albrechtsen, A. Ko, A. Margaryan, V. Moiseyev, T. Goebel, M. Westaway, D. Lambert, V. Khartanovich, J. D. Wall, P. R. Nigst, R. A. Foley, M. M. Lahr, R. Nielsen, L. Orlando, E. Willerslev, Genomic structure in Europeans dating back at least 36,200 years. *Science*. **346**, 1113–1118 (2014).
 163. V. Shinde, V. M. Narasimhan, N. Rohland, S. Mallick, M. Mah, M. Lipson, N. Nakatsuka, N. Adamski, N. Broomandkhoshbacht, M. Ferry, A. M. Lawson, M. Michel, J. Oppenheimer, K. Stewardson, N. Jadhav, Y. J. Kim, M. Chatterjee, A. Munshi, A. Panyam, P. Waghmare, Y. Yadav, H. Patel, A. Kaushik, K. Thangaraj, M. Meyer, N. Patterson, N. Rai, D. Reich, An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers. *Cell*. **179**, 729-735.e10 (2019).
 164. M. Sikora, V. V. Pitulko, V. C. Sousa, M. E. Allentoft, L. Vinner, S. Rasmussen, A. Margaryan, P. de Barros Damgaard, C. de la Fuente, G. Renaud, M. A. Yang, Q. Fu, I. Dupanloup, K. Giampoudakis, D. Nogués-Bravo, C. Rahbek, G. Kroonen, M. Peyrot, H. McColl, S. V. Vasilyev, E. Veselovskaya, M. Gerasimova, E. Y. Pavlova, V. G. Chasnyk, P. A. Nikolskiy, A. V. Gromov, V. I. Khartanovich, V. Moiseyev, P. S. Grebenyuk, A. Y. Fedorchenko, A. I. Lebedintsev, S. B. Slobodin, B. A. Malyarchuk, R. Martiniano, M. Meldgaard, L. Arppe, J. U. Palo, T. Sundell, K. Mannermaa, M. Putkonen, V. Alexandersen, C. Primeau, N. Baimukhanov, R. S. Malhi, K.-G. Sjögren, K. Kristiansen, A. Wessman, A. Sajantila, M. M. Lahr, R. Durbin, R. Nielsen, D. J. Meltzer, L. Excoffier, E. Willerslev, The population history of northeastern Siberia since the Pleistocene. *Nature*. **570**, 182–188 (2019).
 165. M. Sikora, A. Seguin-Orlando, V. C. Sousa, A. Albrechtsen, T. Korneliussen, A. Ko, S. Rasmussen, I. Dupanloup, P. R. Nigst, M. D. Bosch, G. Renaud, M. E. Allentoft, A. Margaryan, S. V. Vasilyev, E. V. Veselovskaya, S. B. Borutskaya, T. Deviese, D. Comeskey, T. Higham, A. Manica, R. Foley, D. J. Meltzer, R. Nielsen, L. Excoffier, M. Mirazon Lahr, L. Orlando, E. Willerslev, Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science*. **358**, 659–662 (2017).
 166. P. Skoglund, J. C. Thompson, M. E. Prendergast, A. Mittnik, K. Sirak, M. Hajdinjak, T. Salie, N. Rohland, S. Mallick, A. Peltzer, A. Heinze, I. Olalde, M. Ferry, E. Harney, M. Michel, K. Stewardson, J. I. Cerezo-Román, C. Chiumia, A. Crowther, E. Gomanichindebvu, A. O. Gidna, K. M. Grillo, I. T. Helenius, G. Hellenthal, R. Helm, M. Horton, S. López, A. Z. P. Mabulla, J. Parkington, C. Shipton, M. G. Thomas, R. Tibesasa, M. Welling, V. M. Hayes, D. J. Kennett, R. Ramesar, M. Meyer, S. Pääbo, N. Patterson, A. G. Morris, N. Boivin, R. Pinhasi, J. Krause, D. Reich, Reconstructing Prehistoric African Population Structure. *Cell*. **171**, 59–71 (2017).
 167. P. Skoglund, S. Mallick, M. C. Bortolini, N. Chennagiri, T. Hünemeier, M. L. Petzl-Erler, F. M. Salzano, N. Patterson, D. Reich, Genetic evidence for two founding populations of the Americas. *Nature*. **525**, 104–108 (2015).
 168. P. Skoglund, C. Posth, K. Sirak, M. Spriggs, F. Valentin, S. Bedford, G. R. Clark, C. Reepmeyer, F. Petchey, D. Fernandes, Q. Fu, E. Harney, M. Lipson, S. Mallick, M.

- Novak, N. Rohland, K. Stewardson, S. Abdullah, M. P. Cox, F. R. Friedlaender, J. S. Friedlaender, T. Kivisild, G. Koki, P. Kusuma, D. A. Merriwether, F.-X. Ricaut, J. T. S. Wee, N. Patterson, J. Krause, R. Pinhasi, D. Reich, Genomic insights into the peopling of the Southwest Pacific. *Nature*. **538**, 510–513 (2016).
169. P. Skoglund, H. Malmström, A. Omrak, M. Raghavan, C. Valdiosera, T. Günther, P. Hall, K. Tambets, J. Parik, K.-G. Sjögren, J. Apel, E. Willerslev, J. Storå, A. Götherström, M. Jakobsson, Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science*. **344**, 747–750 (2014).
170. V. Slon, F. Mafessoni, B. Vernot, C. de Filippo, S. Grote, B. Viola, M. Hajdinjak, S. Peyrégne, S. Nagel, S. Brown, K. Douka, T. Higham, M. B. Kozlikin, M. V. Shunkov, A. P. Derevianko, J. Kelso, M. Meyer, K. Prüfer, S. Pääbo, The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature*. **561**, 113–116 (2018).
171. M. Unterländer, F. Palstra, I. Lazaridis, A. Pilipenko, Z. Hofmanová, M. Groß, C. Sell, J. Blöcher, K. Kirsanow, N. Rohland, B. Rieger, E. Kaiser, W. Schier, D. Pozdniakov, A. Khokhlov, M. Georges, S. Wilde, A. Powell, E. Heyer, M. Currat, D. Reich, Z. Samashev, H. Parzinger, V. I. Molodin, J. Burger, Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. *Nat. Commun.* **8**, 14615 (2017).
172. Valdiosera, T. Günther, J. C. Vera-Rodríguez, I. Ureña, E. Iriarte, R. Rodríguez-Varela, L. G. Simões, R. M. Martínez-Sánchez, E. M. Svensson, H. Malmström, L. Rodríguez, J.-M. Bermúdez de Castro, E. Carbonell, A. Alday, J. Hernández Vera, A. Götherström, J.-M. Carretero, J. L. Arsuaga, C. I. Smith, M. Jakobsson, 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).
173. M. van de Loosdrecht, A. Bouzouggar, L. Humphrey, C. Posth, N. Barton, A. Aximu-Petri, B. Nickel, S. Nagel, E. H. Talbi, M. A. El Hajraoui, S. Amzazi, J.-J. Hublin, S. Pääbo, S. Schiffels, M. Meyer, W. Haak, C. Jeong, J. Krause, Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. *Science*. **360**, 548–552 (2018).
174. C. M. van den Brink, R. Beeri, D. Kirzner, E. Bron, A. Cohen-Weinberger, E. Kamaisky, T. Gonen, L. Gershuny, Y. Nagar, D. Ben-Tor, N. Sukenik, O. Shamir, E. F. Maher, D. Reich, 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).
175. K. R. Veeramah, A. Rott, M. Groß, L. van Dorp, S. López, K. Kirsanow, C. Sell, J. Blöcher, D. Wegmann, V. Link, Z. Hofmanová, J. Peters, B. Trautmann, A. Gairhos, J. Haberstroh, B. Pääffgen, G. Hellenthal, B. Haas-Gebhard, M. Harbeck, J. Burger, 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).
176. N. Vyas, A. Al-Meer, 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).
177. C.-C. Wang, S. Reinhold, A. Kalmykov, A. Wissgott, G. Brandt, C. Jeong, O. Cheronet, M. Ferry, E. Harney, D. Keating, S. Mallick, N. Rohland, K. Stewardson, A. R. Kantorovich, V. E. Maslov, V. G. Petrenko, V. R. Erlikh, B. Ch. Atabiev, R. G.

- Magomedov, P. L. Kohl, K. W. Alt, S. L. Pichler, C. Gerling, H. Meller, B. Vardanyan, L. Yeganyan, A. D. Rezepkin, D. Mariaschk, N. Berezina, J. Gresky, K. Fuchs, C. Knipper, S. Schiffels, E. Balanovska, O. Balanovsky, I. Mathieson, T. Higham, Y. B. Berezin, A. Buzhilova, V. Trifonov, R. Pinhasi, A. B. Belinskij, D. Reich, S. Hansen, J. Krause, W. Haak, Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. *Nat. Commun.* **10**, 590 (2019).
178. M. A. Yang, X. Gao, C. Theunert, H. Tong, A. Aximu-Petri, B. Nickel, M. Slatkin, M. Meyer, S. Pääbo, J. Kelso, Q. Fu, 40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia. *Curr. Biol.* **27**, 3202–3208 (2017).
179. P. Zalloua, C. J. Collins, A. Gosling, S. A. Biagini, B. Costa, O. Kardailsky, L. Nigro, W. Khalil, F. Calafell, E. Matisoo-Smith, Ancient DNA of Phoenician remains indicates discontinuity in the settlement history of Ibiza. *Sci Rep.* **8**, 17567 (2018).
180. Picard toolkit. *Broad Institute, GitHub repository* (2019), (available at <http://broadinstitute.github.io/picard/>).
181. R. M. Kuhn, D. Haussler, W. J. Kent, The UCSC genome browser and associated tools. *Briefings in Bioinformatics.* **14**, 144–161 (2012).
182. S. R. Browning, B. L. Browning, Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics.* **81**, 1084–1097 (2007).
183. S. E. Hunt, W. McLaren, L. Gil, A. Thormann, H. Schuilenburg, D. Sheppard, A. Parton, I. M. Armean, S. J. Trevanion, P. Flicek, F. Cunningham, Ensembl variation resources. *Database.* **2018** (2018), doi:10.1093/database/bay119.
184. N. Li, M. Stephens, Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics.* **165**, 2213–2233 (2003).
185. K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Y. Wayne, S. K. W. Tsui, H. Xue, J. T.-F. Wong, L. M. Galver, J.-B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J.-F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P.-Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L.-C. Tsui, W. Mak, Y. Qiang Song, P. K. H. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. Vernon Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M.

- Munro, Z. Steve Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, T. Johnson, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. Wright Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. Ota Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, A second generation human haplotype map of over 3.1 million SNPs. *Nature*. **449**, 851–861 (2007).
186. J. Haldane, The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet*. **8**, 299–309 (1919).
 187. P. Donnelly, S. Leslie, The coalescent and its descendants. *arXiv* (2010) (available at <https://arxiv.org/abs/1006.1514>).
 188. Y. M. Rosen, B. J. Paten, An average-case sublinear forward algorithm for the haploid Li and Stephens model. *Algorithms for Molecular Biology*. **14**, 11 (2019).
 189. R. R. Hudson, Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, 203–217 (1983).
 190. C. Wiuf, P. Donnelly, Conditional genealogies and the age of a neutral mutant. *Theoretical Population Biology*. **56**, 183–201 (1999).
 191. S. Wright, Isolation by distance. *Genetics*. **28**, 114–138 (1943).
 192. G. Malécot, *Mathématiques de l'hérédité* (1948).
 193. J. Felsenstein, A pain in the torus: some difficulties with models of isolation by distance. *The american naturalist*. **109**, 359–368 (1975).
 194. N. H. Barton, J. Kelleher, A. M. Etheridge, A new model for extinction and recolonization in two dimensions: quantifying phylogeography. *Evolution: International journal of organic evolution*. **64**, 2701–2715 (2010).
 195. C. J. Battey, P. L. Ralph, A. D. Kern, Space is the Place: Effects of Continuous Spatial Structure on Analysis of Population Genetic Data. *Genetics*. **215**, 193–214 (2020).
 196. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian Phylogeography Finds Its Roots. *PLOS Computational Biology*. **5**, 1–16 (2009).
 197. M. M. Osmond, G. Coop, Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies. *bioRxiv* (2021).

198. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*. **5**, 1–11 (2009).
199. J. R. Adrion, C. B. Cole, N. Dukler, J. G. Galloway, A. L. Gladstein, G. Gower, C. C. Kyriazis, A. P. Ragsdale, G. Tsambos, F. Baumdicker, J. Carlson, R. A. Cartwright, A. Durvasula, I. Gronau, B. Y. Kim, P. McKenzie, P. W. Messer, E. Noskova, D. Ortega-Del Vecchyo, F. Racimo, T. J. Struck, S. Gravel, R. N. Gutenkunst, K. E. Lohmueller, P. L. Ralph, D. R. Schrider, A. Siepel, J. Kelleher, A. D. Kern, A community-maintained standard library of population genetic models. *eLife*. **9**, e54967 (2020).
200. M. Kendall, C. Colijn, Mapping phylogenetic trees to reveal distinct patterns of evolution. *Molecular biology and evolution*. **33**, 2735–2743 (2016).
201. D. F. Robinson, L. R. Foulds, Comparison of phylogenetic trees. *Mathematical biosciences*. **53**, 131–147 (1981).
202. M. K. Kuhner, J. Yamato, Practical Performance of Tree Comparison Metrics. *Systematic Biology*. **64**, 205–214 (2014).
203. K. Douka, V. Slon, Z. Jacobs, C. B. Ramsey, M. V. Shunkov, A. P. Derevianko, F. Mafessoni, M. B. Kozlikin, B. Li, R. Grün, D. Comeskey, T. Deviese, S. Brown, B. Viola, L. Kinsley, M. Buckley, M. Meyer, R. G. Roberts, S. Pääbo, J. Kelso, T. Higham, Age estimates for hominin fossils and the onset of the Upper Palaeolithic at Denisova Cave. *Nature*. **565**, 640–644 (2019).
204. S. Sankararaman, S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, S. Pääbo, N. Patterson, D. Reich, The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*. **507**, 354–357 (2014).
205. B. Vernot, J. M. Akey, Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. *Science*. **343**, 1017–1021 (2014).
206. B. Vernot, S. Tucci, J. Kelso, J. G. Schraiber, A. B. Wolf, R. M. Gitterman, M. Dannemann, S. Grote, R. C. McCoy, H. Norton, L. B. Scheinfeldt, D. A. Merriwether, G. Koki, J. S. Friedlaender, J. Wakefield, S. Pääbo, J. M. Akey, Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science*. **352**, 235–239 (2016).
207. L. Skov, M. Coll Macià, G. Sveinbjörnsson, F. Mafessoni, E. A. Lucotte, M. S. Einarisdóttir, H. Jonsson, B. Halldorsson, D. F. Gudbjartsson, A. Helgason, M. H. Schierup, K. Stefansson, The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes. *Nature*. **582**, 78–83 (2020).
208. M. J. Hubisz, A. L. Williams, A. Siepel, Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLOS Genetics*. **16**, 1–24 (2020).
209. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy. *Nature*. **585**, 357–362 (2020).

210. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*. **17**, 261–272 (2020).

Acknowledgments: We thank the Oxford Big Data research computing team, specifically A. Huffman and R. Esnouf, and 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., G.M. Methodology: A.W.W., Y.W., A.K., S.M., N.P., J.K., G.M. Software: A.W.W., Y.W., J.K. Investigation: A.W.W., Y.W., B.J., R.P., D.R. Formal analysis: A.W.W., Y.W., B.J., A.K., S.M. Validation: A.W.W., Y.W., B.J. Visualization: A.W.W., Y.W., B.J. Data curation: A.W.W., Y.W., B.J., A.K., S.M. Resources: A.W.W., Y.W., A.K., S.M., R.P., D.R., J.K. Funding acquisition: D.R., J.K., G.M. Project administration: A.W.W., D.R. Supervision: Y.W., D.R., J.K., G.M. Writing—original draft: A.W.W., Y.W., G.M. Writing—review & editing: A.W.W., Y.W., D.R., J.K., G.M.

Competing interests: G.M. is a director of and shareholder in Genomics plc and a partner in Peptide Groove LLP.

Data and materials availability: Newly reported sequencing data from the Afanasievo family is available from the European Nucleotide Archive, accession number PRJEB43093; phased variant data for the family is 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 at doi:10.5281/zenodo.5168051 and available at <https://tsinfer.readthedocs.io/> under the GNU General Public License v3.0 (30), *tsdate* is deposited to Zenodo at doi:10.5281/zenodo.5168040 and available at <https://tsdate.readthedocs.io/> under the MIT License (31), and *tskit* is deposited to Zenodo at doi:10.5281/zenodo.5465773 and available at <https://tskit.readthedocs.io/> under the MIT License (66). All code used to perform analyses in this paper is deposited to Zenodo at doi:10.5281/zenodo.5172104 and can be found at https://github.com/awohns/unified_genealogy_paper (67). Unified tree sequences of the HGDP, SGDP, and TGP autosomes are available from Zenodo at <https://doi.org/10.5281/zenodo.5495535> (68). Unified tree sequences of the HGDP, SGDP, TGP, and high coverage ancient autosomes are available at <https://doi.org/10.5281/zenodo.5512994> (69). Tree sequences were compressed using the *tszip* utility; see the documentation at <https://tszip.readthedocs.io/> for further details.

Supplementary Materials

Materials and Methods

Supplementary Text

Figs. S1 to S21

Tables S1 to S3

References (68–210)

Movie S1

Interactive Figure S1

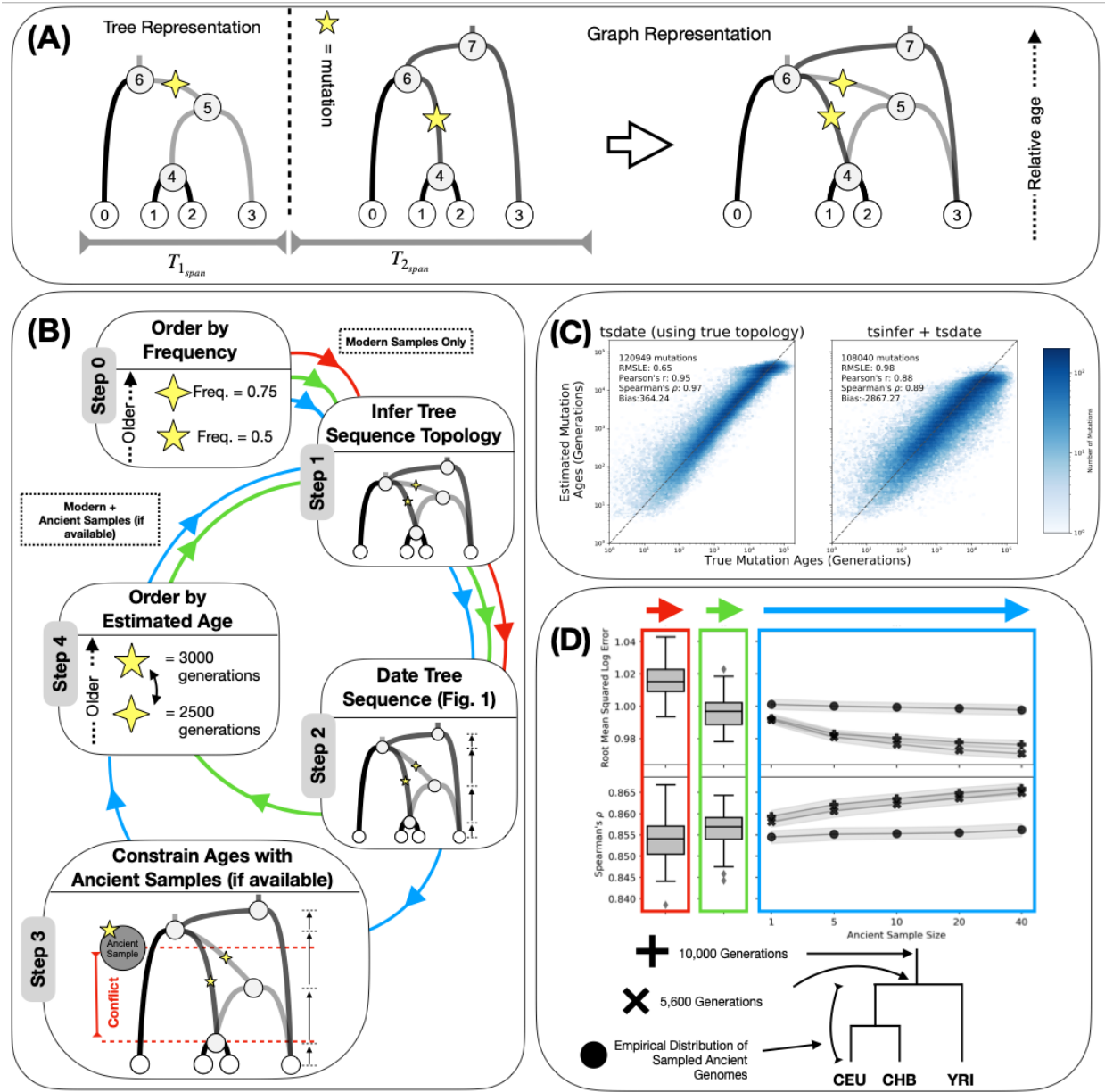


Fig. 1. Schematic overview and validation of the inference methodology. (A) An example tree sequence topology with four samples (nodes 0-3), two marginal trees, four ancestral haplotypes (nodes 4-7), and two mutations. T_{span} measures the genomic span of each marginal tree topology, with the dotted 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; 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 older. The algorithm returns to Step 1 to re-infer 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) and Steps 0, 1, 2, 3, 4, 1, 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). **(D)** Accuracy metrics, root-mean squared log error (top) and Spearman rank correlation coefficient (bottom), with modern samples only (first panel), after one round of iteration (second panel) and with increasing numbers of ancient samples (colored arrows as in panel B). Ancient samples from three eras of human history are considered as in the schematic (24).

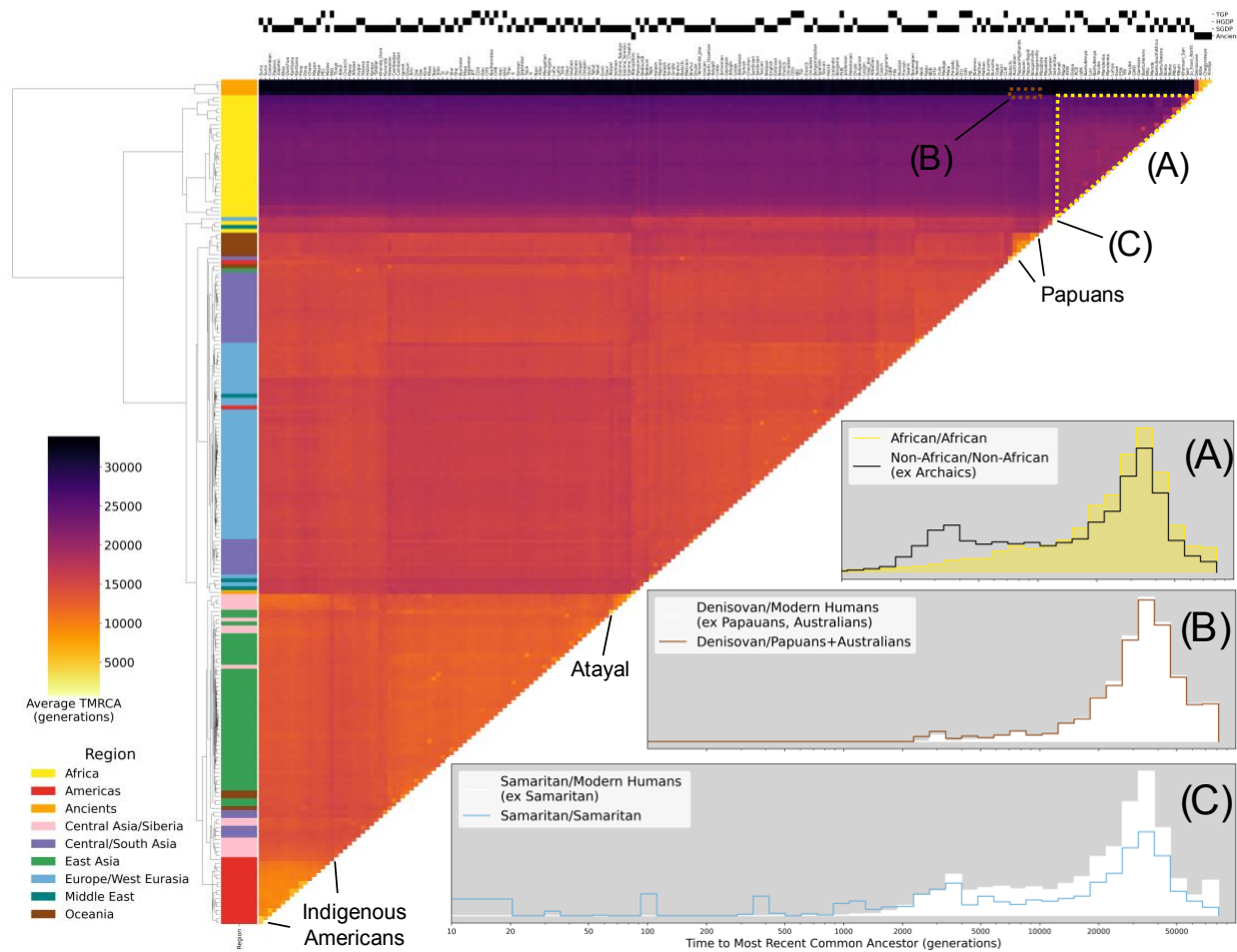


Fig. 2. Clustered heatmap showing the average time to the most recent common ancestor (TMRCAs) on chromosome 20 for haplotypes within pairs of the 215 populations in the HGDP, TGP, SGDP, and ancient samples. Each cell in the heatmap is colored by the logarithmic mean TMRCA of samples from the two populations. Hierarchical clustering of rows and columns has been performed using the UPGMA algorithm on the value of the pairwise average TMRCA. 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 TMRCA between all non-African populations (black line) compared to African/African TMRCA (solid yellow). **(B)** Denisovan and Papuan/Australian TMRCA (solid line), compared to the Denisovan against all non-Archaic populations (solid white). This subtle but unique signal of elevated recent ancestry between the Denisovan and Papuans/Australians is particularly evident in Interactive fig. S1 at https://awohns.github.io/unified_genealogy/interactive_figure.html. **(C)** TMRCA between the two Samaritan chromosomes (solid line), compared to the Samaritans/all other modern humans (solid white). Selected populations with particularly recent within-group TMRCA are indicated. Duplicate samples appearing in more than one modern dataset are included in this analysis. Interactive Figure S1 is an interactive version of this figure and is available at: https://awohns.github.io/unified_genealogy/interactive_figure.html.

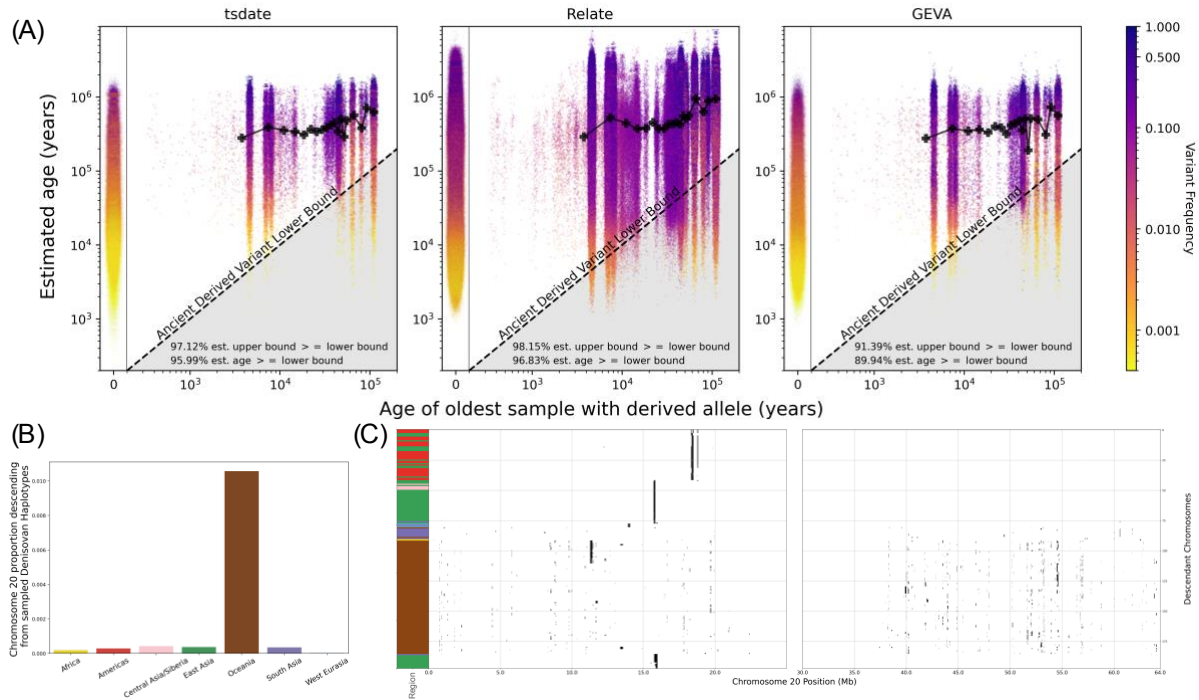


Fig.3. Validation of inference methods using ancient samples. (A) Comparison of mutation age estimates from *tsdate*, *Relate* and *GEVA* with 3,734 ancient samples at 76,889 variants on chromosome 20 (note that *Relate* estimates ages separately for each population in which a variant is found). The radiocarbon- dated age of the oldest ancient sample carrying a derived allele at each variant site in the 1000 Genomes Project 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).

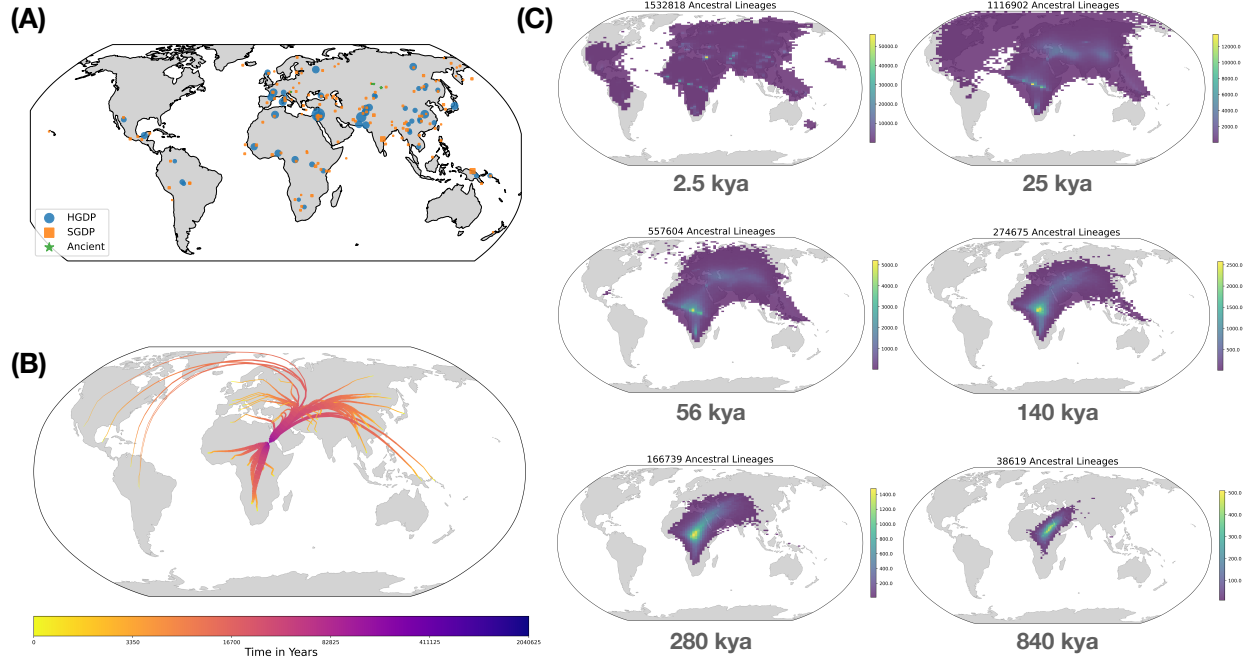


Fig. 4. Visualization of the non-parametric estimator of ancestor geographic location for HGDP, SGDP, Neanderthal, Denisovan, and Afanasievo samples on chromosome 20. (A)

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) 2d-histograms showing the inferred geographical location of HGDP ancestral lineages at six time-points. Histogram bins with fewer than 10 ancestors are not shown. Note that the geographic concentration of ancestors at more recent times is an artifact of uneven sampling and our geographic inference method.

Materials and Methods

Details of dataset preparation, construction of the unified genealogy, and novel algorithms are described in this section. Simulation-based and empirical analyses can be found in the Supplementary Text.

1 Dataset Preparation

Downloading and preparing publicly available data

The TGP Phase 3 GRCh37 VCF was accessed from TGP release 20130502 (6), while the TGP GRCh38 variant calls were accessed from the European Variation Archive under accession number ERZ822766 (70). SGDP data was downloaded from https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data/PS2_multisample_public (7). The HGDP statistically phased dataset was downloaded from <ftp://ngs.sanger.ac.uk/production/hgdp/hgdp-wgs.20190516/statphase/> (8). Note that HGDP does not provide phased data from the X chromosome. The Reich laboratory's Allen Ancient DNA resource, version 42.4, was downloaded from <https://reichdata.hms.harvard.edu/pub/datasets/amh-repo/curated-releases/index-v42.4.html>. This resource provides genotypes for 3589 genotyped ancient individuals at up to 1.23 million variant sites in the genome, compiled from over 100 publications (5–7, 18, 24, 25, 27, 28, 38, 71–179). The sequenced Denisovan, Vindija, Altai, Ust'-Ishim, Loschbour, and LBK-Stuttgart samples were downloaded from <http://cdna.eva.mpg.de/neandertal/Vindija/VCF/> and the Chagyrskaya Neanderthal from <http://ftp.eva.mpg.de/neandertal/Chagyrskaya/>.

The GRCh38 reference assembly was used in all analyses with the exception of the TGP variant age estimation analysis where GRCh37 was used. Only a GRCh37 version of the SGDP

and ancient datasets were available, so these datasets were lifted over to GRCh38 using the Picard tool (180) and the hg19 to hg38 UCSC LiftOver chain (181).

The four archaic individuals are phased using Beagle 5.1 without a reference panel and without imputation (182). The lack of a suitable reference panel likely results in extensive phasing error, but the high levels of homozygosity in archaic genomes mean that errors will only occur between heterozygous sites. Furthermore, errors will only effect downstream analyses when a descendant haplotype crosses a switch error. Indeed, these errors seem to have a limited effect, as running the inference pipeline with unphased archaic genomes does not change any of the conclusions we draw.

Afanasievo Family

We generated between 5 to 8 lanes of shotgun sequencing data using an Illumina HiSeqX10 instrument (2 x 101 cycles and reading out the indices with 2 x 7 cycles) from four individuals from the Afanasievo culture, using sequencing libraries for which in-solution enrichment data was previously published in Narasimhan *et al.* (2019) (38). We merged the paired sequences and mapped to the human genome reference sequence using the samse command in BWA v.0.7.15-r114053 with the parameters -n 0.01, -o 2, and -l 16500, and removed duplicated molecules as described in the original publication that generated in-solution enrichment data on these libraries. The average coverage measured on the autosomal SNP targets used for the in-solution enrichment experiment was 10.8x for I3388 (the mother in the family), 25.8x for I3950 (the father in the family), 21.2x for I6714 (one of the sons in the family), and 25.3x for I3949 (the other son of the family).

The Afanasievo “quartet” allows for reliable phasing of ancient genomes. We used bcftools (v1.10.2) to calculate genotype likelihoods of biallelic SNPs in the Afanasievo samples. We used Beagle 4.0 (182) by providing a genotype likelihood file (gl argument) and pedigree file

(ped argument) to phase this family without any reference panel and imputation. We excluded sites with $\max(\text{GP}) < 0.99$.

Table S1 contains further details of the provenance and sequencing of these samples.

2 Inferring a unified genealogy

Once VCF files from each dataset were prepared, we created `.samples` files, a file format used with `tsinfer` (30), for each constituent dataset using all biallelic SNPs with high-confidence ancestral alleles from Ensembl release 100 (183). The `.samples` files were then combined using the `merge` functionality in `tsinfer`, where the genotypes of variants missing from at least one sample were marked as “missing data” to be imputed by `tsinfer`. This approach uses a reference panel of inferred ancestral haplotypes to impute missing data at the 95.6% of sites in our unified genealogy that have at least one missing genotype. Variants which are biallelic in both datasets being merged, but which have different derived alleles, are included in the merged dataset as multiallelic sites. A tree sequence was then inferred and dated separately for the short and long arms of each autosome using modern samples from the HGDP, SGDP, and TGP datasets (note that results from chromosome 20 are derived from running this pipeline on the whole chromosome, unless otherwise noted). Empirical lower bounds on allele age were gathered from all ancient samples (including unphased and genotyped samples). Variants present in an ancient individual but missing in all modern datasets were not used. Ancient lower bounds at each variant site were compared with `tsdate` age estimates from the inferred tree sequence of modern samples (31). The estimated age of a site in the inferred tree sequence is defined as the oldest most recent common ancestor which possesses a derived variant at each site. At sites where the empirical lower bound from ancient samples is older than our estimated age, this bound was used as the variant age. With these constrained age estimates, the tree sequence was re-inferred using the four archaic individuals and the Afanasievo quartet.

Note that we do not redate the tree sequence containing ancient samples since the prior of `tsdate` does not currently incorporate ancient samples. Details of this process can be found at <https://github.com/awohns/unified-genealogy-paper/tree/master/all-data> (67).

3 Tree sequence inference algorithm: `tsinfer` version 0.2

`Tsinfer` is a scalable method for inferring tree sequence topologies using genetic variation data which can accurately infer genetic genealogies even in the presence of selection and local population structure (13), both important effects in datasets that include both modern and ancient human samples. Here we update `tsinfer` to version 0.2 by incorporating two new features: provision for inexact matching in the copying process and support for missing data (30).

Mismatch, error, and recurrent mutation

The `tsinfer` algorithm is a two-step process (13). First, partial ancestral haplotypes for the sampled DNA sequences are constructed on the basis of shared, derived alleles at a set of sites. Second, a matching algorithm is used to determine patterns of descent between ancestors, and from samples to ancestors. It is this second, matching phase, based on the Li and Stephens (184) (LS) copying process, which is updated here. Matching is based on a Hidden Markov Model (HMM) in which a hidden state is inferred at an array of positions (“inference sites”) along a haplotype. The hidden state is the identity of the most recent ancestor of the current haplotype, from among the n possible older haplotypes at that position. In other words, the hidden state specifies the ancestor of which the current haplotype is an immediate copy. In previous versions of `tsinfer`, we supported only exact haplotype matching – haplotypes had to match perfectly against their ancestors, and a non-matching site required a recombinational switch to a different ancestor. We have now implemented the full LS model by incorporating a mismatch term which

allows for inexact matching. More specifically, the underlying HMM is now parameterised by both *transition* and *emission* probabilities, corresponding to the effects of recombination and mismatch (arising from error or mutation) respectively.

Transition probabilities are relevant when moving from one position to the next along the genome. With a recombination rate (i.e. genetic distance, or mean number of crossovers) of r between the current and previous position, as calculated from a standard genetic map (185), the probability, p_r , of a detectable recombination is given by the standard map function (186)

$$p_r = (1 - e^{-2r})/2. \quad (1)$$

As in the Li and Stephens formulation (184, 187), if a recombination occurs, the hidden state may switch to any of the n ancestors with equal probability, giving transition probabilities of p_r/n to any alternative state, and a probability of $1 - p_r + p_r/n$ of the state remaining unchanged (188).

Emission probabilities allow for mismatch between the hidden state and the observed haplotype: an emission probability of 0 means that the observed allele at a site is always that indicated by the hidden state, an emission probability of 1 means the observed allele never matches that indicated by the hidden state. For biallelic sites with $a = 2$ alleles, an emission probability of $\frac{1}{2}$ thus indicates that the identity of the ancestor at this site has no influence on the observed allele. As it is only the *relative* values of emission probabilities, compared to the transition probabilities, that are important, we parameterize emission via a *mismatch ratio*, ϕ . This is used to calculate a single emission probability used at all sites, m , as a function of the median recombination rate between adjacent inference sites, \tilde{r} . More specifically, we set

$$m = (1 - e^{-a\phi\tilde{r}})/a. \quad (2)$$

For biallelic sites ($a = 2$) and small \tilde{r} , Equation 2 returns an emission probability approximately

ϕ times the median recombination probability (\tilde{p}_r). If a mismatch is inferred from the Viterbi decoding of the forward algorithm, it is incorporated into the resulting tree sequence by placing an additional mutation in the haplotype being matched. Note, however, that such mutations may either reflect true recurrent or back mutations, or represent errors of various sorts, such as in sequencing or in the approximations made in the `tsinfer` algorithm.

Choosing appropriate mismatch ratios

Probabilities of recombination between adjacent sites can be provided in the form of a genetic map. However, optimal mismatch probabilities will depend on factors such as sequencing error and variation in mutation rates along the genome. To establish suitable mismatch ratios, we therefore evaluated inference performance on both simulated and real data. [fig. S11](#) shows the effect of different mismatch probabilities, using a simulated 10 megabase (Mb) region of 1,500 human genomes, both with and without added error in sequencing and ancestral state polarisation. Different metrics disagree slightly on the optimal values used to minimise difference between the simulated and inferred trees. However, error metrics are consistently low when mismatch ratios in both the ancestor and sample matching phases are set to between 0.001 and 10. This range is also suggested as optimal by two different proxy measures of tree sequence complexity based on file size. [fig. S12](#) shows roughly the same pattern when inferring tree sequences from real data, although in these cases only file size measures are available — no ground truth exists for comparison. In all cases, good results are obtained by mismatch ratios close to unity, where the probability of a mismatch is set equal to the median probability of recombination between adjacent inference sites (marked as dashed and dotted lines on the plots). A mismatch ratio of 1 is therefore used in all further analyses; the breadth of the plateau in parameter space indicates similarly accurate results are obtained using mismatch ratios within an order of magnitude either side of this value. More details are given in [Section S1](#).

Missing data

Missing data is accommodated in `tsinfer` by using older ancestors as a “reference panel” for imputation. In the core `tsinfer` HMM, samples and ancestors copy from older ancestors. If the sample or ancestor contains missing data at a site, the missing genotypes are imputed from the most recent ancestor without missing data. The approach provides a principled approach to imputing missing data for both contemporary and ancient genomes, as only older ancestral haplotypes are used.

4 Age inference algorithm: `tsdate`

The `tsdate` algorithm is an approximate Bayesian method for inferring the age of ancestral nodes in a tree sequence topology (31). The tree sequence may be either simulated or inferred, for example by `tsinfer` (13) or `Relate` (34).

There are two main components to the algorithm: the construction of a prior on node age and propagation of likelihoods up and down the tree sequence using the inside-outside algorithm.

Conditional Coalescent Prior

The first step in the algorithm is to assign a prior distribution to the age of ancestral nodes in the tree sequence. A coalescent prior is an obvious choice (22, 53, 189). However, since the number of lineages remaining at any given time is unknown, rather than use a fully tree sequence-aware prior we instead use the theoretical results of Wiuf & Donnelly (1999) (190) to find a marginal prior for each node. This approach uses the number of modern samples descending from each ancestral node to find the mean and variance of node age under the “conditional coalescent.”

We begin by considering a coalescent tree with n sequences, partitioned into i and $n - i$. We condition on the event E that all samples in i coalesce before sharing a common ancestor with a sample outside of i . Conditioning on E affects the order of coalescence events, but not

the exponentially distributed waiting times. We seek to find the mean and variance of the time τ when all i samples have coalesced. At this time there will be only one ancestor of i remaining, but depending on how many lineages outside of i have coalesced, the number of total ancestors, α , could range from 2 to $n - i + 1$.

Equation (10) in Wiuf & Donnelly (1999) (190) shows that the expectation of τ is

$$E(\tau) = \frac{i-1}{n}. \quad (3)$$

The variance can be found using Equation 3 and the expectation of τ^2 , which is calculated as

$$E(\tau^2) = \sum_{m=2}^{n-i+1} P(\alpha = m) E(\tau^2 | \alpha = m),$$

where $E(\tau^2 | \alpha)$ is given by Equation 10 in Wiuf & Donnelly (1999) (190) as

$$E(\tau^2 | \alpha) = 8 \sum_{j=\alpha}^n \frac{1}{j^2} + \frac{8}{n} - \frac{8}{\alpha} - \frac{8}{n\alpha},$$

and $P(\alpha = m)$ is Corollary 2 of Wiuf & Donnelly (1999) (190) when only one ancestor in i remains:

$$P(\alpha = m) = \frac{\binom{n-m-1}{i-2} \binom{m}{2}}{\binom{n}{i+1}}.$$

Coalescent simulations using `msprime` show that the distribution of the logarithm of the age of a node with a fixed number of descendant samples is approximately normal. We can thus use the method of moments to fit a lognormal distribution to the mean and variance of the age of each variant under the coalescent:

$$\sigma = \sqrt{\log \left(\frac{\text{Var}(\tau)}{E(\tau)^2} + 1 \right)};$$

$$\mu = \log(E(\tau)) - \frac{1}{2}\sigma^2.$$

Although the prior is valid at any given marginal tree topology in the tree sequence, recombination introduces two complications: nodes may not span the full tree sequence and a given node may have descendant subsamples of differing sizes in the various marginal trees in which it appears. To address these points, we first iterate through each marginal tree in the tree sequence to find the total length of sequence *spanned* by node u , which we will refer to as s_u . We then determine the “sample weights” for each node by finding the span of sequence where u has a particular number of descendent samples as a proportion of s_u . Since neighbouring marginal trees in a tree sequence are correlated, we use the difference in edges between each adjacent tree as a highly efficient method for calculating the sample weights.

Once we have evaluated sample weights for each node in the tree sequence, we construct a mixture distribution for the prior on node age. This is accomplished by finding the weighted average of expectation and variance of the prior distribution using the sample weights. With the weighted means and variances, we again use the method of moments to determine the parameters of the matched lognormal distribution.

We show that the lognormal approximation is well calibrated to data simulated under the coalescent both with and without recombination in fig. S13.

Time discretisation

Our inference approach requires a time grid for efficient computation. This is constructed by taking the union of the quantiles of the prior distribution of each ancestral node. The advantage of this approach is that inference is focused on times with greater probability under the prior, outperforming a naive, uniform grid. The density of the time grid is determined by the user-specified number of quantiles to draw from each ancestral node as well as a value, ϵ , which establishes the minimum time distance between points in the grid. The conditional coalescent

prior π_u for a node u allows us to find a probability $\pi_u(t)$ for each time-slice t in the grid.

Inside-Outside algorithm

With a prior in place for ancestral nodes in the tree sequence and a time grid, we infer the age of nodes using a belief propagation approach we call the inside-outside algorithm, based on an HMM where the age of nodes are hidden states. In the case of a single tree, this equates to the standard forward-backward algorithm. In the case of a tree sequence, we must also consider the relative genomic spans associated with edges and deal with cycles in the undirected graph underlying the tree sequence. Cycles occur whenever a node has multiple parents and present a general problem in belief propagation (65).

The algorithm is efficient because it uses dynamic programming and the tree sequence traversal methods implemented in `tskit`, the tree sequence toolkit (66). Scaling is linear with the number of edges in the tree sequence (fig. S5) and quadratic with the number of time slices used.

Inside pass

We seek to compute all values in the inside matrix I for all nodes and times in the discretised time grid. $I_u(t)$ is the probability of node u at time t , which encompasses the probability of all nodes and edges in the subgraph beneath u .

We initialise the prior probability of a sample node to be 1 at its sampled time and 0 elsewhere. We then proceed backwards in time, using the relationships between nodes encoded in the `tskit` edge table until we reach the most recent common ancestor (MRCA) nodes of the tree sequence. For each node, we visit every child as well as every time t in the time grid using

$$I_u(t) = \pi_u(t) \prod_{d \in C(u)} \sum_{t' \leq t} L_{du}(t - t' + \epsilon; D_{du}, \theta) I_d(t')^{w_{du}},$$

where $C(u)$ is the set of all child nodes of u , and as previously defined, $\pi_u(t)$ is the prior probability of node u at time t . $L_{du}(t - t' + \epsilon; D_{du}, \theta)$ is the mutation-based likelihood function of the edge from focal node u at time t to child node d at time t' . ϵ is an arbitrarily small value that is used to prevent parent and child nodes from existing at the same time slice. D_{du} is the data associated with the edge including the span of the edge and the number of mutations on it. θ is the population-scaled mutation rate. w_{du} is the span of the edge leading from u to d divided by s_d , the total span of node d in the tree sequence. Note that the inside probability of node d is geometrically scaled by w_{du} to address overcounting if d has multiple parents.

The likelihood function gives the probability of observing k mutations on an edge of length $\delta t = t - t' + \epsilon$ with span l_{du} . It is Poisson distributed with parameter $(\theta l_{du} \delta t)/2$

$$L_{du}(\delta t; D_{du}, \theta) = \frac{\left(\frac{\theta l_{du} \delta t}{2}\right)^k}{k!} e^{-\frac{\theta l_{du} \delta t}{2}}.$$

Note that this likelihood carries the assumption that all differences on a branch are the result of a single mutation (and conversely that identical bases have not arisen through back mutation). We can factorise the inside probability as

$$I_u(t) = \pi_u(t) G_u(t),$$

where $G_u(t)$ is

$$\prod_{d \in C(u)} g_d(t)$$

and $g_d(t)$ is

$$\sum_{t' \leq t} L_{du}(t - t' + \epsilon; D_{du}, \theta) I_d(t')^{w_{du}}.$$

This factorisation will be useful in describing the outside pass in the next section.

The equation terminates at the MRCA(s) of the tree sequence. The total likelihood of the tree sequence is obtained by taking the product of the inside matrix of each MRCA.

Outside pass

Once we have iterated up the tree sequence to find the inside matrix at every node, the inside probability of the MRCA(s) contain all of the information encoded in the tree sequence. To find the full posterior on node age, we now take account of the information in the tree sequence “outside” of the subgraph of each ancestral node. While this algorithm empirically performs well with a single inside and outside pass, any cycles in the underlying undirected graph (which occur when recombination causes a node to have more than one parent) will result in overcounting. The alternative “outside-maximisation” pass introduced below provides another approximate solution in these cases, though we find that the outside pass performs better empirically (fig. S1).

Beginning with the MRCA nodes in each marginal tree (the roots), we initialise the outside value of these nodes, O_{MRCA_s} , to be one at all non-zero time points. There is no information “outside” the MRCA(s) because all information in the tree has already been propagated to the node and is encoded in the MRCA(s)’ inside matrices. In a tree sequence, it is possible for a node u to be the MRCA in some of the marginal trees in which it appears but not in other trees. In these cases we find O_u by dividing the span of trees where the node is the MRCA by s_u , the total span of u in the tree sequence.

We then proceed down the tree sequence (forwards in time), again using the edge table sorted in descending order by the children’s ages. At every node we calculate:

$$O_u(t) = \prod_{p \in P(u)} \sum_{t' \geq t} O_p(t')^{w_{up}} L_{pu}(t' - t + \epsilon; D_{pu}, \theta) \left(\frac{I_p(t')}{g_{up}(t')} \right)^{w_{up}},$$

where $P(u)$ is the set of parents of node u and other terms are defined in the previous section

on the inside algorithm.

Once the inside and outside passes are complete, the approximate posterior can be calculated as

$$\phi_u(t) \propto I_u(t)O_u(t).$$

Importantly, the mean value of the posterior distribution may not be consistent with the tree sequence topology. We provide the option to “constrain” node age estimates by forcing each node to be older than the estimated age of its children. The unconstrained mean and variance of each node are retained as metadata in the tree sequence. The full posterior can also be retained separately if desired.

We observe that mutations mapping to edges descending from the single oldest root in tree sequences inferred by `tsinfer` are generally of lower quality, so in our implementation of the outside pass we include an option to avoid traversing such edges. We use this setting in all analyses using `tsdate` in this work. Additionally, results from `tsdate` do not include estimates for mutations appearing on these edges.

Outside-Maximisation Pass

Cycles in the undirected graph underlying a tree sequence occur when a node has more than one parent as a consequence of recombination (an example is shown in Fig. 1A). These cycles result in the over-counting of information in the outside pass of the inside-outside algorithm. To account for this, we introduce the outside-maximisation pass as an alternative. The rationale for this approach is that the inside value of the MRCA(s) of the sample has accumulated all of the data encoded in the tree sequence. In our model, the age of a node and the age of all non-descendant nodes are conditionally independent, given the age of a node’s parents. Thus, if we fix the age of the MRCA(s), we can then walk down the tree sequence fixing the age of

parent nodes before considering the age of their children.

First, we set the age of each MRCA in the tree sequence to the time slice with the maximum probability in the node’s inside matrix (equivalent to assigning the age with the greatest maximum posterior probability). Next, we establish a traversal path using the topology such that each node is only visited once all of its direct and indirect parents have been considered. For each node we visit on this traversal path, we know the age of all of a node’s ancestors as well as the relative likelihood of the subtree below each node for a given age. For any possible node age in the discretised time grid we can also compute the likelihood of the events on the branches to its direct and already-dated ancestors, as described in the inside pass. With this information we can compute the conditional posterior density for the age of the node. We set its age to the time slice with the maximum value.

Formally, for each node u we seek to calculate the time of the node, T_u as

$$T_u = \arg \max_{t \leq U_P} \left(I_u(t) \prod_{p \in P(u)} L_{pu}(T_p - t + \epsilon; D_{pu}, \theta) \right),$$

where U_P is the age of the youngest parent of node u and other variables are defined in the Methods section on the inside and outside passes.

5 Iterative approach for inferring tree sequences with ancient and modern samples

We combined `tsinfer` and `tsdate` in an iterative approach that allows for the incorporation of ancient samples and improves inference accuracy in many settings.

The first step of the iterative approach is to order derived alleles appearing in the sample by their frequency. `tsinfer` requires a relative ordering of derived alleles to both build ancestral haplotypes and infer copying paths. Frequency is a largely accurate and highly efficient means

of providing an ordering for these ancestors (13). Once alleles are ordered, it is possible to infer a tree sequence topology with `tsinfer` (Fig. 1B Step 1).

With an inferred tree sequence topology, we next estimate the age of inferred ancestral haplotypes with `tsdate` (Fig. 1B Step 2). If using `tsdate`'s outside pass, we do not constrain the resulting date estimates by the topology.

If ancient samples are present, we can use them to constrain the estimated age of derived alleles. The previous step (Fig. 1B Step 2) provides date estimates for the inferred ancestors as well as for mutations. Since we estimate the age of the ancestral nodes above and below a mutation, the child node of an edge hosting a mutation is constrained by the ancient sample-informed lower bound on derived allele age. This bound is determined by gathering the haplotypes of ancient samples (either sequenced or genotyped) and examining derived alleles that can be called in these ancient samples with high confidence. If multiple ancient samples carry the same derived allele, we use the oldest sample as the lower bound on its age. Once lower bounds have been collected for all derived alleles observed in ancient samples, we compare these with our statistically inferred lower bounds on allele age, adjusting our age estimates where necessary to ensure consistency with ancient samples. Any radiocarbon-dated ancient samples with high-confidence variant calls may provide constraints in this step, including unphased and/or low-coverage samples. Although a substantial fraction of radiocarbon dates are likely inaccurate, we note that there is a low probability that errors on the order of a few thousand years will meaningfully affect tree sequences inferred using this approach. Only a subset of erroneously dated alleles will be older than the true age of the mutation, which would affect allele age estimation accuracy, and still fewer will be older than the ancestral haplotype from which they descend, which would affect topological estimation accuracy. This is confirmed using simulations in Section S1.

With allele age estimates from step 2, possibly constrained by ancient samples in step 3,

we are now able to re-infer the tree sequence topology. The revised age estimates are used to order the age of derived alleles when re-estimating ancestral haplotypes with `tsinfer`; if they are more accurate than frequency in determining a relative ordering of mutations, topological inference accuracy should be improved. Indeed, we find that the iterative approach improves accuracy when re-inferring tree sequences from variation data simulated with a uniform recombination map and without error (Fig. 1D). When reinferring tree sequences from data simulated with error or with a variable recombination map, less improvement is observed (fig. S3). After reinferring the tree sequence topology, we may re-estimate the age of inferred ancestral haplotypes with `tsdate`. In the unified genealogy presented in this work, we do not run `tsdate` a second time, since the prior of `tsdate` does not currently incorporate ancient samples.

Ancient samples can be included in tree sequences that are (re)-inferred with estimated allele ages. This is accomplished by inserting ancient samples at their correct relative ordering among ancestors generated by `tsinfer`. Only phased ancient samples with an age estimate may be included, although we note that extending `tsinfer`'s HMM to handle diploid individuals may allow for phasing of ancient samples in this step.

In `tsinfer`, samples cannot directly descend from other samples (which is highly unlikely in reality). Instead we allow descent from ancient sequences, which represents shared ancestry with modern samples (29). This is technically accomplished by producing “proxy ancestors” associated with ancient samples at a slightly older time than the ancient sample. These are composed of all non-singleton sites carried by the ancient sample, and may serve as ancestors to younger ancestors and samples. Thus, when inferring a tree sequence of modern and ancient samples, the final step of `tsinfer` is to infer copying paths between ancestors (including proxy ancestors) and samples.

6 Inferring the Location of Ancestors in a Tree Sequence

We use a naive, non-parametric approach to gain insight into the geographic location of ancestral haplotypes based on the known locations of sampled genomes. We note that while the relationships between genealogies and spatial structure has been an active area of research in both phylogenetics and population genetics for many years (191–196), and a recent, more sophisticated method uses marginal trees in genome-wide genealogies (197), such approaches typically ignore recombination.

Geographic coordinates are available for samples from the SGDP, HGDP, Afanasievo, and Archaic datasets. The latitude and longitude coordinates of individual samples are provided for SGDP individuals, while the location of sampling centers were used for the HGDP individuals. No geographic information was provided for TGP individuals, so these were not used in location inference. We also used the coordinates of the archaeological sites associated with the Afanasievo and Archaic individuals.

The weighted center of gravity is determined for each ancestral node by iterating up the tree sequence, visiting child nodes before their parents using the same traversal pattern as for the previously described inside algorithm. At each focal ancestral node, we find the geographic midpoint between each of the children of that node. The following simple approach was used to find the geographic midpoints. For a node u , the latitude and longitude coordinates of the child node of each edge descending from u were converted to Cartesian coordinates. We find the average of the children’s coordinates and convert this back to latitude and longitude. We then continue up the tree sequence using this location to calculate the coordinates of u ’s parents.

This method is highly efficient, requiring less than one minute to compute on the combined tree sequence of chromosome 20.

Supplementary Text

S1 Simulation-based evaluation of methods

Code implementing all evaluations can be found at https://github.com/awohns/unified_genealogy_paper (67).

Evaluating the accuracy of topological inference

There are two phases of matching in `tsinfer`: firstly matching inferred ancestors against older ancestors, secondly matching the known sample haplotypes against all ancestors (13). Two mismatch ratios can therefore be specified, one in the ancestor-matching phase (ϕ_a) and the other in the sample-matching phase (ϕ_s). We expect optimal ratios to depend on the degree to which assumptions in the inference algorithm are met (for example, the assumption that frequency of the derived allele can be used as a proxy for ancestral age), and the amount of error in the analysed dataset (for example, a suitable value of ϕ_s should make inference reasonably robust to sequencing error). To find appropriate mismatch ratios, we performed inference for a range of 15 mismatch ratios from 10^{-5} to 10^4 in both ancestor- and sample-matching phases, using a variety of datasets. Results are shown in Figs. S11 (simulated datasets) and S12 (real datasets).

Simulated sequences were obtained from a standard three population “Out of Africa” model (198), as implemented in `stdpopsim` (199) version 0.1.2, with uniform mutation and recombination rates of 1.29×10^{-8} bp/gen and 1.72×10^{-8} bp/gen respectively. 500 chromosomes of 10 megabases (Mb) in size were sampled from each of the populations in the model, for a total sample size of $n = 1,500$. Accuracy of inference can be directly assessed by comparing the topology of the inferred trees along the genome with the ground-truth topologies. However, inferred trees contain polytomies (nodes whose number of direct children, i.e. arity, is greater than

two), which have different effects on different tree distance metrics. For this reason, we compare accuracy using three separate metrics: the Kendall-Colijn (KC) metric (200) with $\lambda = 0$, the same metric but with each inferred tree having polytomies resolved equiprobably into a randomly chosen bifurcating topology, and finally the Robinson-Foulds (RF) metric (201). The two KC metrics are normalised by dividing by the value obtained when comparing the ground truth tree with a “star” topology in the first case, and a random binary topology in the second. The RF metric, which is notoriously sensitive to minor tree changes but which has been shown to perform reasonably in practice (202), is normalized against the maximum possible number of disagreeing splits for two bifurcating trees ($2n - 4$): for this reason the RF with randomly split polytomies was used. We can also assess inference performance indirectly by looking at file size or numbers of edges and mutations, under the assumption that lower values provide a more parsimonious representation of evolutionary history. This is particularly important when assessing performance on real data, for which ground-truth topologies are not available.

As expected, for simulations with no injected error, the first KC plot and the RF plot indicate that the lower the mismatch ratio, the greater the accuracy (fig. S11A); the KC metric suggests this effect is stronger for ϕ_a , whereas the RF metric suggests it is (slightly) stronger for ϕ_s . In contrast, where inferred trees have had their polytomies split at random, the KC metric suggests that very low mismatch ratios, particularly in the ancestor matching phase, are suboptimal: this is also suggested from the size of the resulting tree sequence. This difference may be due to the effect of polytomy size (arity of nodes), as the KC metric returns lower values when comparing a resolved tree with a polytomy than the average value comparing the same tree with the polytomy split equiprobably into binary resolutions. Also as expected, as mismatch ratios tend to zero, the number of mutations tends to the minimum possible (i.e. the number of sites): however there is a notable decrease in the number of mutations (and concomitant increase in the number of edges) when mismatch ratios drop from 10^{-3} to 10^{-4} ; this is also associated with

an increase in file size.

In seeking optimal mismatch ratios for general use, it is more relevant to consider inference in which simulated sequences have had error added. fig. S11B gives results in which the sequences produced by simulation have had error added before inference. Genotyping error was added on the basis of estimates from the platinum genomes project (33) and ancestral states were flipped at a randomly chosen 1% of variant sites. As expected, with these injected errors, better results were obtained with higher mismatch ratios. The RF metric suggests an optimal ϕ_a between 10^{-4} and 10 but a much higher ϕ_s . The KC metric with polytomies retained suggests lower ϕ_a values and ϕ_s somewhere between 10^{-3} and 10^2 . However, both the file size and the count of edges plus mutations clearly indicate an optimal range between 10^{-3} and 10 for both ϕ_a and ϕ_s , a range that is mirrored in the case of ϕ_a by the KC metric with randomly resolved polytomies. From these results we deduce that none of the tree distance measures that we use are an ideal measure of inference accuracy, but that all agree that a mismatch ratio in the ancestor matching phase between 10^{-3} and 10 provides good inference. Moreover, within this range, the exact value is of minor importance. In the sample matching phase, the RF and plain KC metrics conflict over whether mismatch ratios higher than 10 are of benefit. However, such high values are not only contraindicated by measures of file size, but also because we would not expect error rates to be orders of magnitude higher than recombination rates. The simulated data therefore leads us to conclude that the value 1 is a reasonable mismatch ratio in both the ancestor and sample matching phases, that both file size and number of mutations plus edges are sensible proxies for inference accuracy, and that similar results would be obtained by using mismatch ratios anywhere between 0.1 and 10.

The proposed mismatch ratios of $\phi_a = \phi_s = 1$ are also supported by inference of real sequence data from two public datasets: the Thousand Genomes Project data (using build 37 of the human genome reference sequence), and the Human Genome Diversity Project (using build

38). The appropriate genetic map for chromosome 20 was used to specify transition probabilities via Equation 1, and for reasons of computational efficiency, inference was performed after subsetting down to sites 1,000,000 to 1,100,000 . fig. S12 shows that mismatch ratios between 0.1 and 10 provide the smallest file sizes and most parsimonious number of mutations and edges, with the newer and less error-prone HGDP dataset able to retain small sizes for slightly smaller values of ϕ_i . Reassuringly, the pattern in node arity reflects that in fig. S11B, with larger polytomies for low values of ϕ_a and ϕ_s . This indicates that, for the `tsinfer` algorithm, our method of injecting error into simulated datasets produces a similar effect to error in real data.

Optimal mismatch ratios around 1 might initially appear to be unreasonably high: recombination is expected to be common, recurrent mutations rare, and in modern datasets error should affect only a small fraction of the genotype matrix. However, simulations show that reasonable results are obtained even for mismatch ratios well above 1, where mismatches are substantially more probable than recombination events. This indicates that a correctly inferred recombination event can remove the need for several mismatches at nearby sites.

Multiple mutations in inferred tree sequence

Figure S14 shows the distribution of numbers of mutations at each site. Even using an infinite sites model, where each site has only a single true mutation, the addition of realistic sequencing error leads to greater than 10% of sites requiring multiple mutations to explain the observed pattern of variation. A promising area of future research is to use genealogical inference techniques to identify and remove errors in sequenced datasets. We do not attempt such error correction here, but we do note that when the true genealogy is known, the number of multiple mutation sites can be substantially diminished by removing mutations above sample nodes (green dotted line). As might be expected, this approach is less effective in the equivalent inferred tree

sequence (blue dotted line), since convergent sequencing errors above genealogically distant samples will tend to lead to these samples erroneously grouping together. Nevertheless we still find that in this inferred tree sequence, about 70% of the mutations above sample nodes are attributable to sequencing error rather than mutations in the underlying trees.

This distribution of mutations in tree sequences inferred from real data (red & orange lines) is noticeably less biased towards sites with 5 to 40 mutations per site than the tree sequence inferred from simulation with error, implying that our procedure for adding sequencing and ancestral state polarization error may be adding more error than expected in today's genomic datasets. Nevertheless, the real data has a long tail of sites with very large numbers of mutations (>100 per site). If inference is accurate, these must indicate either highly mutable sites, or (more likely) additional, complex sources of sequencing error such as alignment issues, paralogous sequences, or phasing issues. We investigate these sites in Figure S6.

Evaluating the accuracy of the `tsdate` prior

We evaluate the accuracy of the `tsdate` coalescent prior in fig. S13. As described in Section 4 of the Materials and Methods, `tsdate` uses results from the coalescent conditioned on the frequency of an ancestral haplotype (*190*) to determine the mean and variance of the age of each node in the tree sequence. Moment matching is then used to fit a lognormal distribution on the age of each node. We evaluate the accuracy of this prior distribution, and also an alternative using a gamma distribution approximation, in simulations with and without recombination. The results of `msprime` simulations under the neutral coalescent show that the 95% credible interval of the gamma prior includes the true node ages more frequently than the equivalent lognormal distribution. However, the credible interval of the gamma is much wider than the lognormal and skews towards younger ages. For this reason, the lognormal is the default setting in `tsdate`. Note also that the expectation of the prior exactly matches the moving average of

simulated node times in the case of no recombination; in simulations with recombination the prior expectation underestimates the true node times.

Accuracy of Age Estimation

Simulation-based evaluation of the accuracy and scaling properties of `tsinfer` and `tsdate` were performed with `msprime` version 1.0, (21) and `stdpopsim` (199). We evaluate neutral coalescent simulations with recombination as well as a more complex model based on the Out of Africa event (198) (also used to evaluate `tsinfer` in the previous section). The effects of genotype and ancestral state errors are modelled using an error model based on an empirical comparison of the Illumina Platinum Genomes Project to matched data from the TGP (33). Our methods are compared to `Relate` (34), a leading method for inferring dated genealogies and `GEVA` (33), a method for estimating allele age based on pairwise haplotype comparisons.

`Tsdate` was evaluated under a variety of inference settings. First, the “true” topologies produced by `msprime` were passed to `tsdate`. This is indicated in relevant figures as “`tsdate` (using simulated topology)”. The second method of evaluation was to run `tsdate` on tree sequence topologies inferred from the simulated variation data with `tsinfer`. This is referred to as “`tsdate` using `tsinfer` topologies”. Both the outside and outside-maximisation passes are assessed in fig. S1, as are various ratios of mutation to recombination rates. Since the outside pass empirically outperforms the outside-maximisation pass, only results using the former are used in subsequent simulations as well as in all results using real data. Variation data was converted to the required input formats for `GEVA` and `Relate` before running these programs using default settings. Mutation rates of 10^{-8} per base pair per generation are passed to all inference methods and recombination maps are passed to `tsinfer` and `Relate`. Ground truth allele ages were provided by `msprime`. For `tsdate` and `Relate`, allele age point estimates were determined using the arithmetic mean of the ages of the upper and lower bounding an-

cestors. The `tsdate`-estimated age of these ancestors was given by the mean of each node’s posterior distribution. The mean age of the joint clock with quality-filtered pairs was used for GEVA. Only sites that could be dated by all methods were evaluated. This means that singletons, $n - 1$ tons, and other sites not dated by one or more methods were excluded from the evaluation.

Root mean squared log error (RMSLE), Pearson’s r , Spearman’s ρ , and estimator bias were used to assess the accuracy of allele age estimates. Tree inference accuracy was measured by comparing simulated and inferred tree sequences using the KC tree distance metric (200), which includes a parameter, λ , determining the relative weight given to tree topology vs branch lengths when performing comparisons. We test both $\lambda = 0$, where all weight is given to the topology, and $\lambda = 1$, where all weight is given to branch lengths. See Section S1 for details on the advantages and disadvantages of this metric.

Figure S2 shows that our methods infer genealogies and allele ages with accuracy comparable or superior to the state-of-the-art in neutral coalescent simulations. When using the simulated topologies, `tsdate` has high accuracy in recovering allele age. When using inferred topologies from `tsinfer`, the accuracy of allele age estimation is comparable to GEVA and Relate.

Figure S3 shows the performance of `tsdate` on simulations of chromosome 20 using the Out of Africa model (198, 199) and the previously described error models. Demography impacts allele age estimates since `tsdate` uses a constant value of N_e and does not currently model population structure. Using mismatch terms and the recombination map in `tsinfer` results in improved topological inference in all simulations and improvements in allele age accuracy by most metrics. The panel in the third row and third column approximates the empirical tree sequences of modern samples produced in this work, since the simulated data was injected with genotype and ancestral state error and age estimates are the result of inferring the tree sequence topology with mismatch terms and dating it with `tsdate`. In this panel, we observe an inflation

in variant age of 15.7% (a bias of 783 generations). Since the results in Section S1 indicate that the real sequencing data we use is consistent with lower levels of error than that induced by the empirical error model used in these simulations, this is an upper-estimate of any bias in age estimation due to errors in the dataset. The panel in the fourth row and third column reflects the strategy we used to produce the unified tree sequences of modern and ancient samples. These tree sequences are not re-dated as the prior of `tsdate` does not currently accommodate ancient samples.

The iterative approach shown in the fifth row of fig. S3 shows variable effects depending on the simulation model used and evaluation metric being considered. Inference using the iterative approach provides the lowest values of KC distance with $\lambda = 1$ (corresponding to the highest accuracy in branch length estimation); however, KC distance with $\lambda = 0$ and the various allele age accuracy measures show that the iterative approach generally only offers comparable or slightly improved accuracy compared to a single pass of `tsinfer` with mismatch terms and dating with `tsdate`. However, Fig. 1D shows that when inferring genealogies using `tsinfer` without mismatch from data simulated with a uniform recombination map and without error injected, iteration improves accuracy. Inference using error-prone variation data leads to lower accuracy by most measures. These same patterns are observed in fig. S4, which compares the accuracy of our methods with GEVA and a version of `Relate` that re-estimates effective population size and branch lengths.

Scaling Simulations

We evaluated the scaling properties of `tsinfer` (without mismatch) and `tsdate` compared to `Relate` and GEVA using `msprime` neutral coalescent simulations (fig. S5). The CPU runtime and memory requirements are recorded using two evaluation schemes. First, the length of the simulated chromosomes is held constant while increasing sample sizes are used. Sec-

ond, sample size is fixed while sequence length increases. Default parameters are used for all methods. The maximum memory allowance for `Relate` was set to 32 Gb to allow the scaling analysis to proceed without error.

Ancient Samples and Allele Age Estimation Accuracy

Figure 1D shows a simulation-based evaluation of allele age inference accuracy when incorporating ancient samples into the previously described Out of Africa model (198, 199). We simulated 20 replicates of 5 Mb and sampled 1008 modern samples split evenly between the three populations of the Out of Africa model: Yorubans (YRI), Han Chinese (CHB), and Utah Residents with Northern and Western European Ancestry (CEU). 40 ancient samples were also sampled. Three ancient sampling schemes were evaluated: (1) pick samples at times randomly selected from the empirical distribution of the ages of published ancient samples, (2) pick samples from the human population immediately prior to the time of the Out of Africa event in this model (5,650 generations ago), and (3) pick samples from 10,000 generations ago. 20 simulation replicates were performed with each sampling scheme.

The first sampling scheme used the estimated ages of ancient samples from the The Reich laboratory's Allen Ancient DNA resource, version 42.4, available at https://reichdata.hms.harvard.edu/pub/datasets/amh_repo/curated_releases/index_v42.4.html. The ancient sample times ranged from 90 years to 90,000, with a mean of 4,553 years (standard deviation: 4,554 years). If the randomly sampled ages were younger than 21,200 years (the split time of CEU and CHB in the Out of Africa model), they were assigned to CEU or CHB with 50% probability. If older than the split time, they were assigned to the N_B population as described in Gutenkunst *et al.* (2009) (198). All mutations only carried by ancient samples were removed from the simulation.

We then inferred dated tree sequences using three approaches. First, we inferred and dated

a tree sequence using only the modern samples and evaluated the accuracy of the resulting allele age estimates. Second, we used these estimated allele ages to reinfer and redate tree sequences (still only using modern samples) and evaluated the result. Third, we again used the allele age estimates to reinfer tree sequences, but constrained the estimates with progressively larger numbers of ancient samples. We then redate the resulting tree sequence with only modern samples, since the prior of `tsdate` does not currently support ancient samples. Note that in the unified tree sequence presented in this work, we also do not run `tsdate` a second time using only modern samples as we wish to retain ancient samples in the final genealogy.

Allele age estimation accuracy was evaluated by comparing the true allele ages from the simulated tree sequence to the estimated allele ages using root mean squared log error (RMSLE) and Spearman’s ρ . Allele ages were determined as described in the previous section. While the initial iteration step without ancient samples increases allele age estimation accuracy, adding samples drawn from empirical distribution of ancient sample ages does not have a noticeable effect on overall accuracy. However, increasing numbers of samples from sampling schemes (2) and (3) consistently improve both MSLE and Spearman’s rank correlation. This is likely because older samples will “correct” larger numbers of derived alleles that are initially estimated to be too young.

More sophisticated simulations and sampling schemes are possible, but we expect modifying demographic parameters in these simulations will generally not bias allele age estimates or hamper the inclusion of high coverage ancient genomes in the tree sequence, though they may affect downstream inference of genetic relationships.

Misspecifying ancient sample ages

One assumption of the use of ancient samples in this algorithm is that radiocarbon dates associated with samples are accurate. To evaluate the effect of misspecified sample ages, we per-

formed simulations using a demographic model reflecting human history described above (198), with 1008 samples split evenly between the three modern populations. We then sampled ancient samples at five points uniformly in log-space, at 100, 299, 894, 2675, and 8000 generations. The first two samples were taken from the CEU population, the third and fourth from the Out of Africa population, and the last from the ancestral African population. The effects of age resampling are shown in Figure S15, where results using the true age of ancient samples are compared to resampled age.

This figure shows that that incorrect ages associated with ancient samples are unlikely to have a large effect on estimated age of variants carried by that sample. For ancient samples which are $\leq 2,675$ generations old, if the age of the ancient sample is 25% inflated, fewer than 0.66% of the sites carried by the ancient sample are *younger* than the inflated ancient sample age. Even an upward bias of 25% in the radiocarbon estimate of a 8,000 generation old ancient human sample, which would be more than four times older than the oldest sequenced human genomes, would only affect $< 2.5\%$ of the mutations carried by that sample. We thus conclude that errors in the estimated age of ancient samples are highly unlikely to negatively effect our age estimates.

We note that the rather limited impact of sample dating does not mean that ancient samples are uninformative. Indeed, our simulations demonstrate their value in dating variants. For example, an ancient sample might show that a variant that is only carried by a few related contemporary samples is, in fact, old. Given that branch lengths within genealogies increase rapidly back in time, a small amount of error in sample age is likely to have limited impact, though the added constraint can be highly informative.

We also note that, in future, it is relatively easy to introduce probabilistic dating as a way of accommodating uncertainty in sample age.

Archaic Descent Simulation

Tree sequences encode the genetic relationships between modern and ancient samples at every point in the genome. We used a published simulation model approximating the Out of Africa Event with archaic introgression into the ancestors of Eurasians and Papuans (37) to evaluate how well our inferred tree sequences reflect these relationships. Using an implementation of the simulation model from `stdpopsim` (199), we simulated 10 replicates of 15 Mb from chromosome 20, sampling 200 chromosomes each from the African, West Eurasian, East Asian, and Papuan populations. We slightly modified the ancient sampling scheme from the model by sampling three archaic individuals: the Denisovan (two chromosomes from the sampled Denisovan lineage, 2,203 generations ago), the Vindija (two chromosomes from the introgressing Neanderthal lineage, 1,725 generations ago), and the Altai Neanderthal (two chromosomes from the ancestral Neanderthal lineage, 3,793 generations ago). See Jacobs *et al.* (2019) (37) fig. S5 for a schematic summarising the model. This model defines a generation to be 29 years (compared to 25 years in our other analyses).

We used migration records from the simulated tree sequences to determine patterns of archaic local ancestry. This provided a “ground truth” set of introgressed spans of sequence in modern samples. Migrations from Denisovan and Neanderthal lineages to the ancestors of modern, non-African samples in the simulated trees provide the locations of these introgressed spans of sequence in each modern sample. Importantly, in this simulation setting introgressed tracts exist regardless of whether or not they carry derived alleles and thus may be undetectable.

We examined how well patterns of common ancestry between *sampled* archaic and modern individuals reflect archaic introgression. The simulation from Jacobs *et al.* (2019) (37) modelled introgression as “pulses” of admixture from archaics to moderns at discrete points in time. Consequently, the time to most recent common ancestor (TMRCA) between a modern and archaic sample is only observed to be more recent than $T_{ArchaicModern}$, the split time of modern

and archaic lineages occurring 20,225 generations ago, at marginal trees where introgression occurred in the ancestors of that modern sample. However, the TMRCA between sampled archaics and moderns may not be more recent than $T_{ArchaicModern}$ at all introgressed tracts: the TMRCA is older than the split time at marginal trees where the sampled and introgressing archaics are distantly related. TMRCAs that are more recent than $T_{ArchaicModern}$, but older than T_{DenNea} , the split time of Neanderthals and Denisovans occurring 15,090 generations ago, can be attributed to introgression from *either* Neanderthals or Denisovans. Signals of introgression from these two populations may be disentangled at marginal trees where the TMRCA of sampled archaics and moderns is more recent than T_{DenNea} .

We record tracts of common ancestry in this final category as well as the full set of introgressed tracts from the migration records using $n \times c$ matrices, where n is the sample size of modern individuals and c is the number of 1 kilobase (kb) chunks in the simulated chromosome. Two introgression matrices were constructed, one each for the introgressing Denisovan and Neanderthal populations. At each cell in an introgression matrix, a “1” indicates introgression from the relevant archaic population in a modern individual in any marginal tree overlapping the 1 kb chunk. Three common ancestry matrices were constructed, one each for the Vindija, Altai, and Denisovan samples. In each common ancestry matrix, a “1” indicates that the TMRCA of the modern individual and *either* of the archaic individual’s sampled chromosomes at any tree overlapping the chunk is less than T_{DenNea} .

Since TMRCAs of moderns and archaics more recent than T_{DenNea} reflect introgression from Denisovan or Neanderthal lineages, values of “1” in the common ancestry matrices are guaranteed to only exist where “1”s are observed in the corresponding introgression matrix. Common ancestry between moderns and the Denisovan more recent than T_{DenNea} encompasses $\geq 99.9\%$ of introgressed Denisovan tracts (Std Dev 0.04%). This is likely attributable to the simulated population size of only $N_e = 100$ in the introgressing Denisovan D1 and D2 lineages

from Jacobs *et al.* (2019) (37). The Vindija Neanderthal shares common ancestry with humans more recently than T_{DenNea} for 61.3% (Std Dev 4.07%) of introgressed Neanderthal tracts. For the Altai, the value is 55.0% (Std Dev 1.99%).

Next, to investigate how well our inferred tree sequences reflect these observed patterns of introgression and common ancestry, we ran our iterative approach to infer a tree sequence of modern and archaic samples from the simulated data. We used an N_e value of 10,000 and a mutation rate of 10^{-8} per base pair per generation when running `tsdate`. We then examined the inferred tree sequences, noting the spans of genome where the two “proxy ancestors” associated with each archaic individual have direct descendants among modern samples.

Observed descent from proxy archaic ancestors should overlap with signals of introgression where sampled archaic haplotypes are a better approximation of ancestral material than `tsinfer`’s inferred ancestors at a given epoch. This approach enables an understanding of the relationships between sampled archaic and modern individuals without needing to identify a split time between moderns and archaics. Results are expected to be heavily influenced by such factors as the sampling times of ancient individuals, relationships between modern and ancient populations, error rates, and the quality of ancestors estimated by `tsinfer`. Relationships between sampled and introgressing ancient lineages may particularly affect the amount of introgressed material that can be recovered from the inferred tree sequences since, as noted previously, in this simulation model more recent common ancestry is observed between the sampled and introgressing Denisovans than between the sampled and introgressing Neanderthals.

We used the inferred tree sequences to construct three additional $n \times c$ matrices that record direct descent from proxy ancestors associated with the three sampled archaic individuals. In each cell in a matrix, a “1” records where a modern individual descends from the relevant proxy archaic sample in a marginal tree overlapping the 1 kb chunk. Comparing the binary matrices recording ground truth tracts of introgression or common ancestry from the simulations with the

matrices recording inferred descent from archaic proxy ancestors yielded true positives (TP), corresponding to introgression or common ancestry in the simulation and descent in inferred tree sequence, false positives (FP), where no introgression or common ancestry is noted in the simulation but descent is inferred, and false negatives (FN), where introgression or common ancestry is noted in the simulation but no descent is inferred. Table S2 shows the resulting rates of precision and recall, which are defined as follows:

$$Precision = \frac{TP}{TP + FP}, \quad (4)$$

$$Recall = \frac{TP}{TP + FN}. \quad (5)$$

The proportion of all modern genetic material that is introgressed from each archaic population, the proportion of modern genetic material sharing common ancestry with archaic samples more recently than T_{DenNea} , and the proportion of modern genetic material that is inferred to descend from archaic samples are also given in Table S2.

The simulation results indicate that descent from the Vindija and Denisovan proxy ancestors, and to a lesser extent the Altai, recovers ground truth introgression and shared ancestry with high precision. In the case of the Neanderthals, where shared ancestry tracts incompletely represent introgressed tracts, inferred descent recovers more shared ancestry tracts than introgressed tracts.

It is possible to increase recall by examining patterns of common ancestry between sampled archaics and moderns occurring more recently than a given split time in the inferred tree sequence, as was performed in the simulated tree sequence. The precision and recall for recovering ground truth introgressed and shared ancestry tracts at different split times is analysed in detail in fig. S16, where observed precision and recall values are obtained by testing split times in intervals between the time of each archaic sample and T_{DenNea} . These results show that progressively older split times recover nearly all shared ancestry tracts, though at lower levels of

precision. While it is possible to use these results to choose time cutoffs for each sample which maximise recall at a given level of precision, the optimal values derived from simulations will doubtlessly differ from the real data due to the imperfect modelling of demographic history as well as varying levels of error. In this study we chose to report tracts that directly descend from these samples in order to avoid choosing a cutoff time while likely maximising precision.

Geographic Inference Evaluation

We evaluated the accuracy of our method for inferring the location of ancestral haplotypes using a coalescent simulation framework with `msprime` version 1.0 (21) and `stdpopsim` (199). Using the previously described “Out of Africa” demographic model (198), we simulated 10 replicates of 5 Mb of chromosome 20 using the GRCh37 recombination map, drawing 300 modern samples split evenly between the YRI, CHB, and CEU populations. As described in Section S1, `msprime` records the population associated with both leaf (sample haplotypes) and internal nodes (ancestral haplotypes) in simulated tree sequences. We assigned latitude and longitude values of 1° N, 32° E to YRI, 37° N, 84° E to CHB, and 52° N, 0° E to CEU. These correspond to points in the continents of Africa, Asia, and Europe which are roughly equidistant from one another. The ground truth location of a simulated YRI ancestor, for instance, is thus assigned to 1° N, 32° E. Ancestral haplotypes may also be assigned to population “B”, the Out of Africa population in the model. Note these locations are not meant to represent the actual sampling locations of the corresponding HapMap populations, rather are chosen to reflect continental-level origin. The results of these simulations are plotted in fig. S9.

We estimated the coordinates of each ancestral haplotype in the tree sequence using the method described above and plotted the results for each population in fig. S9A. Taking the estimated latitude and longitude coordinates of each ancestor, we assess whether it is closer to the YRI, CEU, or CHB coordinates. For instance, if a YRI ancestral haplotype is estimated to

exist in Egypt at 24° N, 26° E, this is counted as a successful comparison as the coordinates are closer to the YRI coordinates than to CEU or CHB. We find that our geographic estimator places 98.0% of YRI ancestors closest to the African location. The figures for CEU and CHB are 98.8% and 96.7%, respectively. 41.9% of ancestral haplotypes from population “B” are located closer to CHB or CEU than YRI; we note that population “B” has no defined geographic location and many ancestors from this population have descendants restricted to a single population.

Next, we inferred and dated tree sequences describing the variation data produced by the above described simulations. This data was injected with error using the empirical error model (33) and 1% ancestral state error which are described in detail in Section S1. We passed the GRCh37 recombination map to `tsinfer` and used a precision value of 15 and mismatch ratios of 1. Assessing the accuracy of geographic inference in inferred tree sequences is more challenging as inferred ancestral haplotypes do not necessarily have a one-to-one correspondence to simulated ancestors. Instead, we assess how often ancestors which we estimate to be older than the Out of Africa event, which occurs at 5,600 generations in the Gutenkunst *et al.* (2009) model (198), are closer to the CEU or CHB population coordinates than to YRI and plot the results in fig. S9B. We find that the inferred locations of ancestors older than 5,600 generations are closer to the European or Asian population than the African population 16% of the time. This value decreases when older ancestors are considered (fig. S17).

This suggests that the oldest ancestors shown in Figure 4C. may be more widely dispersed than would be expected. However, a fully spatial, and highly parametric simulation framework is needed for a more complete evaluation of the results produced by this estimator.

S2 Empirical data-based analyses

Duplicated samples in modern datasets

154 modern individuals appear in more than one dataset. This consists of 130 individuals which appear in both SGDP and HGDP, as well as 24 individuals which appear in both SGDP and the TGP GRCh38 release. Duplicated individuals are retained in all analyses for the following reasons: the scalability of our methods mean that removal of these samples would result in negligible savings in computational resources, none of the analyses we conducted would be adversely affected by duplicated samples, and inspecting the duplicated samples revealed disagreements in genotyping and phasing. Across all 154 duplicated samples on chromosome 20, the mean proportion of genotype calls that are incompatible is 0.04% (standard deviation 0.26%). This figure was found by examining the variant sites shared between each pair of duplicated individuals and computing what proportion of these variants had differing numbers of derived alleles. This discrepancy can be explained by the fact that some of the duplicated samples were sequenced using different libraries, that different variant calling pipelines were used, and also potentially by the effects of lifting over SGDP individuals from GRCh37 to GRCh38.

We determined the number of switches between phasing configurations for each pair of duplicated samples. The mean across all pairs is 1420.38 switches (standard deviation 1028.65). Since individuals have different numbers of heterozygous sites, and thus more or fewer opportunities for switches between phasing configurations, we divided the number of switch errors by the number of heterozygous sites in each individual. This provides the proportion of heterozygous sites where the phasing configuration switches. The average value across pairs is 3.73% (standard deviation 1.90%), with a maximum of 10.7% in individual HGDP01032 from the San population in Africa and a minimum of 1.48% in individual HG00190 from the Finnish population in Europe. The greatest level of phasing inconsistency was observed in the 30 du-

plicated African samples, where the phasing configuration switched at an average of 6.93% heterozygous sites, while the 11 samples from the Americas showed the greatest consistency with a switch of phasing configuration at 2.05% of heterozygous sites.

Multiple mutations in the unified genealogy

To investigate whether sites with higher mutation rates are enriched for multiple mutations inferred by `tsinfer`, we examined CpG dinucleotides and transitions vs. transversions. Deamination at CpGs is known to have an approximately order of magnitude higher mutation rate than other mutation types. We defined CpG dinucleotides using alleles from the Ensembl ancestral human genome (release 100) (183), identifying when a cytosine nucleotide was followed by a guanine nucleotide. C to T mutations at the first nucleotide or G to A mutations at the second nucleotide were counted as “CpG” mutations. We observed that 12.7% of the 90,776,900 biallelic SNPs in our unified genealogy of HGDP, TGP, SGDP, and eight high coverage ancient samples were the result of such mutations at CpG sites. 50.0% of these sites contained greater than one mutation, compared to 36.7% for non-CpG sites, a 1.72 fold enrichment. When only considering non-singleton sites, the values for CpGs and non-CpGs are, respectively, 72.6% and 60.2%: a 1.75 fold enrichment. While this is significantly less than the expected enrichment of 10x, it shows that `tsinfer` is capturing meaningful biological signal without knowledge of a differential mutation rate at CpG sites.

We also evaluated whether transitions were enriched for multiple mutations compared to transversions. We find that the biallelic SNPs in the unified genealogy have an overall transition/transversion ratio of 2.02. Furthermore, 39.7% of transitions and 35.7% of transversions are inferred to have more than one mutation: a 1.19 fold enrichment. When only considering non-singleton sites, the values are 63.3% and 59.3%, constituting a 1.18 fold enrichment. Along with the evidence from CpG sites, these values suggest that c. 20% of sites inferred to

have multiple mutations may require a biological (as opposed to technical) explanation.

Multiple mutations and ancient samples

We observe elevated numbers of transitions, and specifically variants in CpG dinucleotides in the variants which are “corrected” by ancient samples (see 5). We find that 7.03% of the biallelic variants are observed in ancient samples. The ti/tv at these sites is 2.44 and 14.4% are in CpG dinucleotides. Because these figures are comparable to those seen in modern samples, we believe that they most likely reflect genuine segregating variants rather than errors (note that we do not analyze sites exclusive to ancient samples).

Among the 559,431 variants whose age is affected by inclusion of the ancient sample (because the lower age bound provided by the estimated archaeological date of the oldest ancient sample in which the derived allele is found are *older* than the `tsdate` estimated age of the haplotype), the transition to transversion ratio is 2.88, and 27.2% are in CpG dinucleotides. That these figures are substantially higher does suggest that transition errors in ancient samples may affect variant dating (because a truly variant appears “ancient”). This enrichment of transitions and CpGs in the sites where an ancient sample is older than the `tsdate` estimated age of a site suggests the presence of errors and/or recurrent mutation in the ancient samples used in this analysis, though we cannot easily distinguish between these explanations.

To quantify the effect of this observed excess ti/tv ratio at variants whose age is modified when comparing ancient samples and `tsdate` estimates, we evaluated the effect of only adjusting transversions. Of the 2,090,402 variant sites on chromosome 20, 654,880 (31.3%) are transversions. We find 158,401 variant sites where a derived allele is observed in both modern and ancient samples. 43,155 (27.2%) of these sites are transversions. Of all the sites appearing in modern and ancient samples, 16,908 have an ancient lower bound that is older than the estimated age from `tsdate`. 4,020 (23.8%) of these sites are transversions. We adjusted these

ages to be consistent with the radiocarbon dates of the ancient samples. The average age estimate of the haplotypes on which mutations appear is only 0.45% younger when considering transversions than when using all sites. The tree sequences re-inferred using haplotype ages corrected by all ancient sites and after only correcting transversions are highly similar to one another. When all sites are used, the unified genealogy of chromosome 20 includes 665,680 nodes, 6,131,667 edges, 5,765,475 mutations, and 133,191 trees. When only using transversions, the inferred tree sequence has 665,382 nodes, 6,129,716 edges, 5,769,403 mutations, and 133,269 trees. All inferences presented in this work which are derived from the unified tree sequence of chromosome 20 are unchanged.

Multiple mutations and dating accuracy

The analysis in Section S1 indicates that sequencing errors are responsible for many of the additional mutations inferred by `tsinfer`. One possible consequence of dating tree sequences with large numbers of additional mutations is an upward bias in age estimates. We sought to evaluate whether including 2,044 sites carrying over 100 mutations on the long arm of chromosome 20 (0.17% of all sites on the chromosome arm) creates bias in our estimates. [fig. S6](#) strongly suggests that the majority of these sites are not genuine variants. Collectively, these sites contain 422,196 mutations, 13.34% of the total number of inferred mutations on the long arm of chromosome 20.

We reran `tsdate` on the inferred tree sequence of the long arm of chromosome 20 after removing these sites and compared the estimated age of all sites with less than 100 mutations with estimates from the tree sequence with all mutations included. [fig. S10](#) shows that dating the tree sequence after removing sites with greater than 100 mutations produces similar results to dating the tree sequence with all mutations. We observe that the average age of a mutation decreases 4.65% from 3,833 to 3,655 generations.

In this work, we chose to retain all mutations in the inferred tree sequences so that our inferred tree sequences are lossless representations of the original data sources. Further work is necessary to confidently distinguish mutations reflecting error from genuine recurrent mutation. In addition to providing important information about the extent of genotype error, the quality of variants, and estimates of recurrence in population sequencing datasets, this will also likely improve allele age estimates.

Pairwise Time to Most Recent Common Ancestor

We use the unified and dated tree sequence to calculate the pairwise TMRCA for sampled haplotypes from all 215 populations in the combined tree sequence of HGDP, SGDP, TGP and ancient samples. This was accomplished by iterating over the trees in the tree sequence and at each tree calculating the TMRCA between pairs of chromosomes using the `mrca` function implemented in `tskit`. For efficiency, we down-sampled the chromosomes by randomly selecting up to ten samples from each population. In populations with fewer than 10 samples, all samples are used. We then find the weighted logarithmic average of all TMRCAs associated with each population combination, weighted by the span of the tree associated with each TMRCA. The logarithmic average is used as we expect the variance of TMRCAs to increase substantially with age: all date values are log transformed before analysis for this reason. The constrained age estimates produced by `tsdate` were used in this analysis (see Methods).

The histograms of TMRCAs for each population combination can be found in Interactive fig. S1. Three of these histograms are also shown in Fig. 2.

Empirical Estimates of Thousand Genome Project Variant Ages on Chromosome 20

We compared the estimated ages of GRCh37 chromosome 20 TGP Variants from `tsdate`, `Relate` and `GEVA`. We used `tsinfer` to infer a tree sequence of TGP individuals with the

GRCh37 recombination map, mismatch ratios of 1, and a precision setting of 15. The tree sequence was dated using `tsdate` with N_e set to 10,000 and a mutation rate of 10^{-8} mutations per base pair per generation. Default settings were used for all other parameters. Allele ages were estimated by using the arithmetic mean of the nodes above and below the oldest mutation. Allele age estimates from GEVA were gathered from <https://human.genome.dating/download/index>. We used the mean allele age estimate from the joint clock as a point estimate of allele age and the upper limit of the 95% confidence interval from the joint clock as an upper bound estimate (33). `Relate` TGP allele age estimates were downloaded from <https://zenodo.org/record/3234689>. Since the publicly available `Relate` age estimates are calculated separately in each TGP population, we averaged age estimates and upper bounds for alleles appearing in more than one population. The point estimate of allele age was obtained by taking the mean of the upper and lower age bounds.

Empirical lower bounds were provided by radiocarbon dates of ancient samples. These dates were derived from the Reich Laboratory dataset for all ancient individuals with the exception of the Afanasievo family (which we assigned to 4.6 kya), Chagyrskyaya Neanderthal (Chagyrskaya 8) (80 kya) (26), Vindija Neanderthal (Vindija 33.19) (50 kya) (25), Altai Neanderthal (Denisova 5) (110 kya), Denisovan (Denisova 3) (63.9 kya) (203), Ust'-Ishim (45 kya) (28), Loschbour (8 kya) (5) and LBK-Stuttgart (7 kya) (5). If multiple ancient samples carried a derived allele, the oldest sample carrying the allele was used as the lower bound on the age of that allele. This resulted in a combined dataset of 3,734 ancient individuals (including samples which appear in multiple publications and are thus duplicated in the Allen Ancient DNA resource).

A set of 659,804 sites for which age estimates can be found from all three methods was assembled. This excluded singletons and $n - 1$ tons, which GEVA did not estimate, sites where the derived and alternate alleles were inconsistent (as GEVA estimates the age of alternate alle-

les), indels, and sites with low quality ancestral states. The relationship of allele age estimates from each method to allele frequency is shown in fig. S18, the distribution of ages is shown in fig. S19, and the comparisons of allele ages from the differing methods to one another is shown in fig. S20. The combined set of ancient samples carried derived alleles at 76,889 of these sites. At each site, the oldest ancient sample carrying a derived allele was used as the empirical lower bound on the age of the site. A comparison of the estimated allele ages from the three methods with these bounds is shown in Fig. 3A.

It is important to note that the distribution of age estimates varies between the three methods (fig. S19), with `Relate` showing a higher mean estimated age compared to the other two methods. The `tsdate` mean estimated allele age at the 659,804 comparable sites is 5,919 generations and the mean upper bound on allele age is 9,012 generations. For `Relate`, the equivalent values are 6,816 and 11,732 generations. Since `Relate` provides an estimate for each population, these values were first found by averaging the available population-based estimates for each site, then averaging all the sites. For `GEVA` the equivalent values are 5,192 and 6,147 generations.

Descent from Ancient Haplotypes

We used the unified, inferred tree sequence to investigate the genealogical relationships of ancient and modern samples. A number of studies have sought to infer introgressed haplotypes from archaic individuals (204–207) and one has inferred an ancestral recombination graph from modern and archaic individuals (208). As shown in S1, our approach identifies tracts of introgressed sequence, particularly regions where sampled and introgressing archaic individuals share more recent common ancestry, and requires no assumptions about the nature of introgression.

We evaluate descent from the haplotypes of the Afanasievo family, Denisovan, and Vin-

dija Neanderthal among younger samples in Fig. 3 and Figs. S7, S8, S21. The simplified tree sequence of 3,754 individuals on chromosome 20 (with unary nodes retained) is used for this analysis. Regions of over 1 Mb without variant sites were trimmed from the inferred tree sequence, as well as regions before the first variant site and after the last variant site. We establish descent using the *proxy nodes* associated with each ancient sample in the inferred tree sequence, as detailed in the Materials and Methods.

Descent from ancient proxy nodes is described using two statistics. First, the genomic descent statistic defined in Scheib *et al.* 2019 (57) is used to evaluate the overall amount of genetic material in each population which descends from the proxy ancestors associated with ancient individuals. The results of this analysis are shown for the Denisovan (Fig. 3B and fig. S21), Afanasievo (fig. S7A), and Vindija (fig. S8). Second, we split chromosomes into 1 kb chunks and assess descent from the ancient proxy ancestors in each chunk, as described in section S1. Pearson product-moment correlation coefficients were calculated using `Numpy` (209) for the $n \times c$ matrices of descent, where n is the number of descendants and c is the number of 1 kb blocks. The correlation coefficients were then hierarchically clustered using the UPGMA algorithm implemented in `Scipy` (210). The results of this analysis are shown for the Denisovan (Fig. 3C) and Afanasievo (fig. S7B).

Guidelines for use of Ancient samples

Ancient samples may be added to tree sequences produced by `tsinfer` and `tsdate` to improve age estimates and to infer genetic relationships. The net gain of adding ancient samples depends on the sample's age, coverage, level of genotype error, and population of origin.

In the previous sections we found that older samples provided improved age estimates (Fig. 1D) and that misspecified ages are highly unlikely to bias age inference (fig. S15). We also found strong enrichment for CpG sites among variants corrected by ancient samples (see

section S1), suggesting that care should be taken in such correction, particularly for transitions at CpG sites. Table S3 summarizes these findings and provides guidelines for incorporating ancient samples into inferred tree sequences.

As noted in Materials and Methods Section 5, ancient samples without modern descendants are suitable to use with this method since fragments of the DNA sequences of ancient samples may be highly related to ancestral material of modern samples

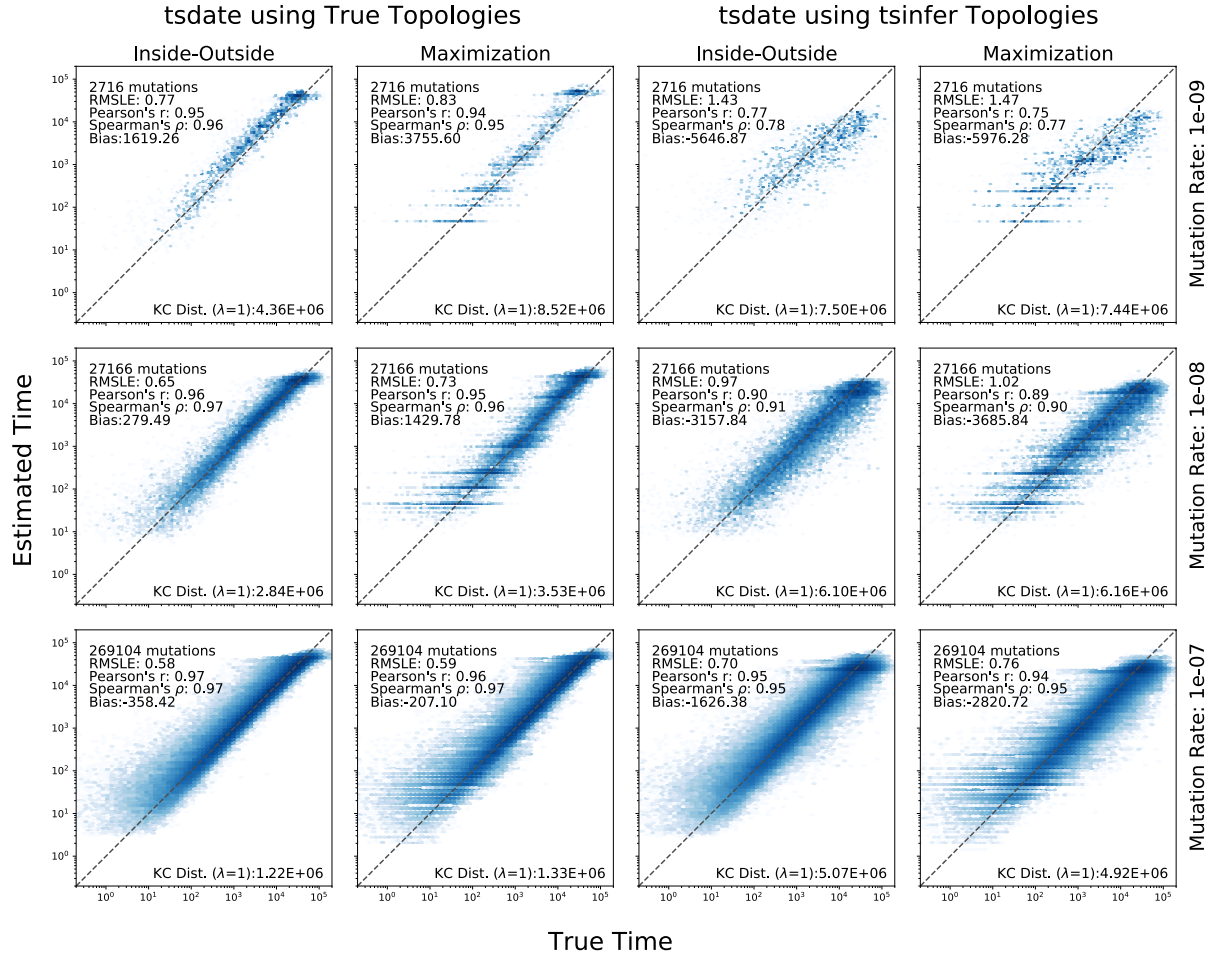


Fig. S1. Accuracy of `tsdate` at various mutation rate settings. Each row shows the results of ten `msprime` simulations of 1 Mb with 500 samples, $N_e = 10,000$, and $r = 10^{-8}$. Three different values of μ were used, 10^{-9} , 10^{-8} , and 10^{-7} . In each subplot simulated allele ages are on the x-axis and estimated allele ages are on the y-axis. The first two columns show the results of running `tsdate` on topologies simulated by `msprime`. The third and fourth columns show the results when running `tsdate` on topologies inferred by `tsinfer` from the simulated genotype data. The results of the inside pass followed by either an outside pass or outside-maximisation pass are shown.

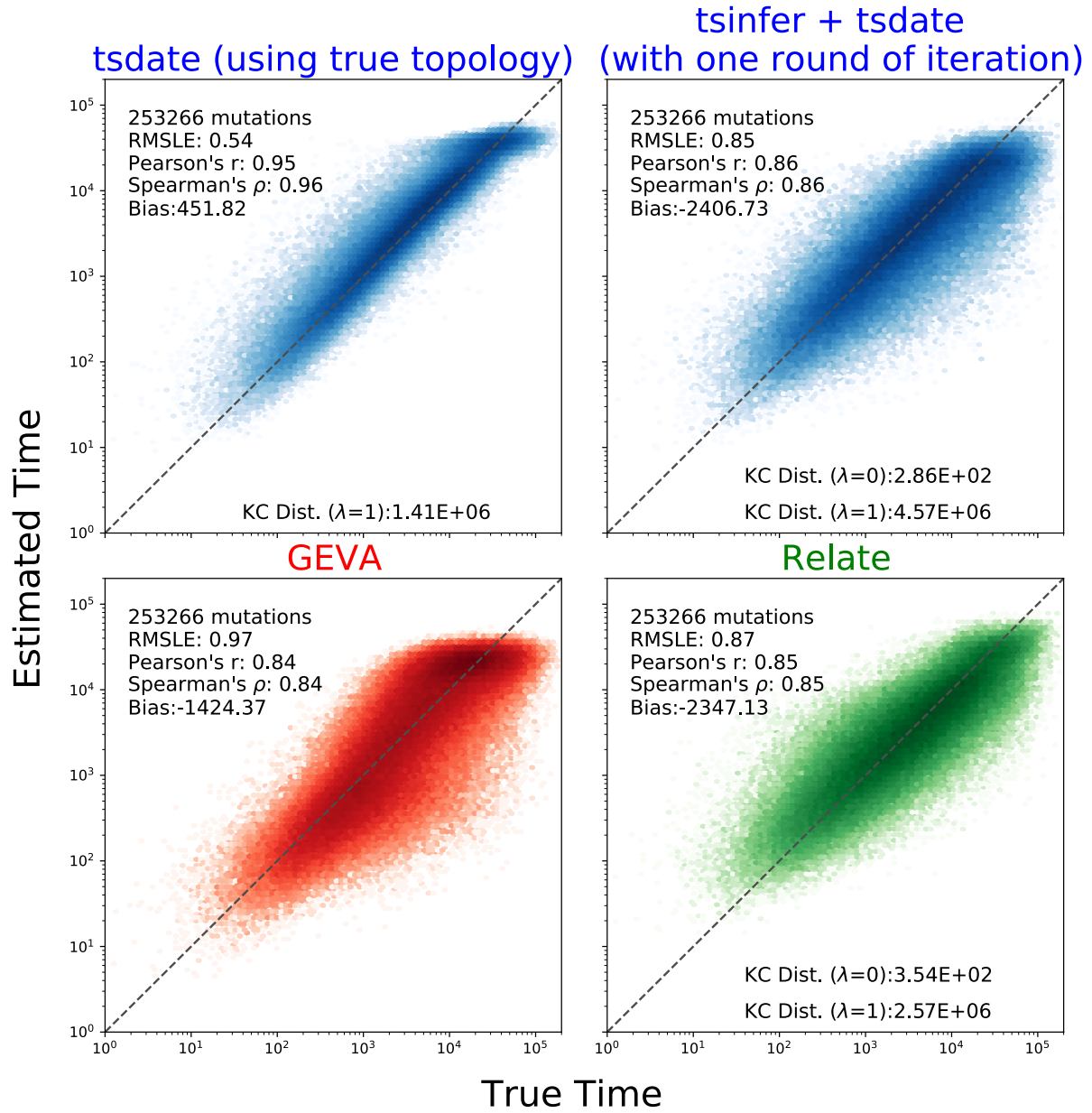


Fig. S2. Comparison of the accuracy of tdate, GEVA, and Relate on simulated tree sequence topologies. Thirty msprime simulations with 250 samples, 5 Mb of sequence, $N_e = 10,000$, and $\mu = r = 10^{-8}$ were used. The x-axis in all plots is the age of derived alleles from the simulation; estimated allele ages from the labeled method are on the y-axis. Top left subplot shows derived allele age estimates from running tdate on simulated topologies. Top right subplot shows results from inferring a tree sequence topology with tsinfer, dating the tree sequence with tdate, and using the resulting date estimates to reinfer and redate the tree sequence. Lower subplots show allele age estimates from GEVA and Relate. Only alleles dated by all methods are shown: this excludes singletons, $n - 1$ tons, and alleles that were deemed to map poorly by tdate or Relate.

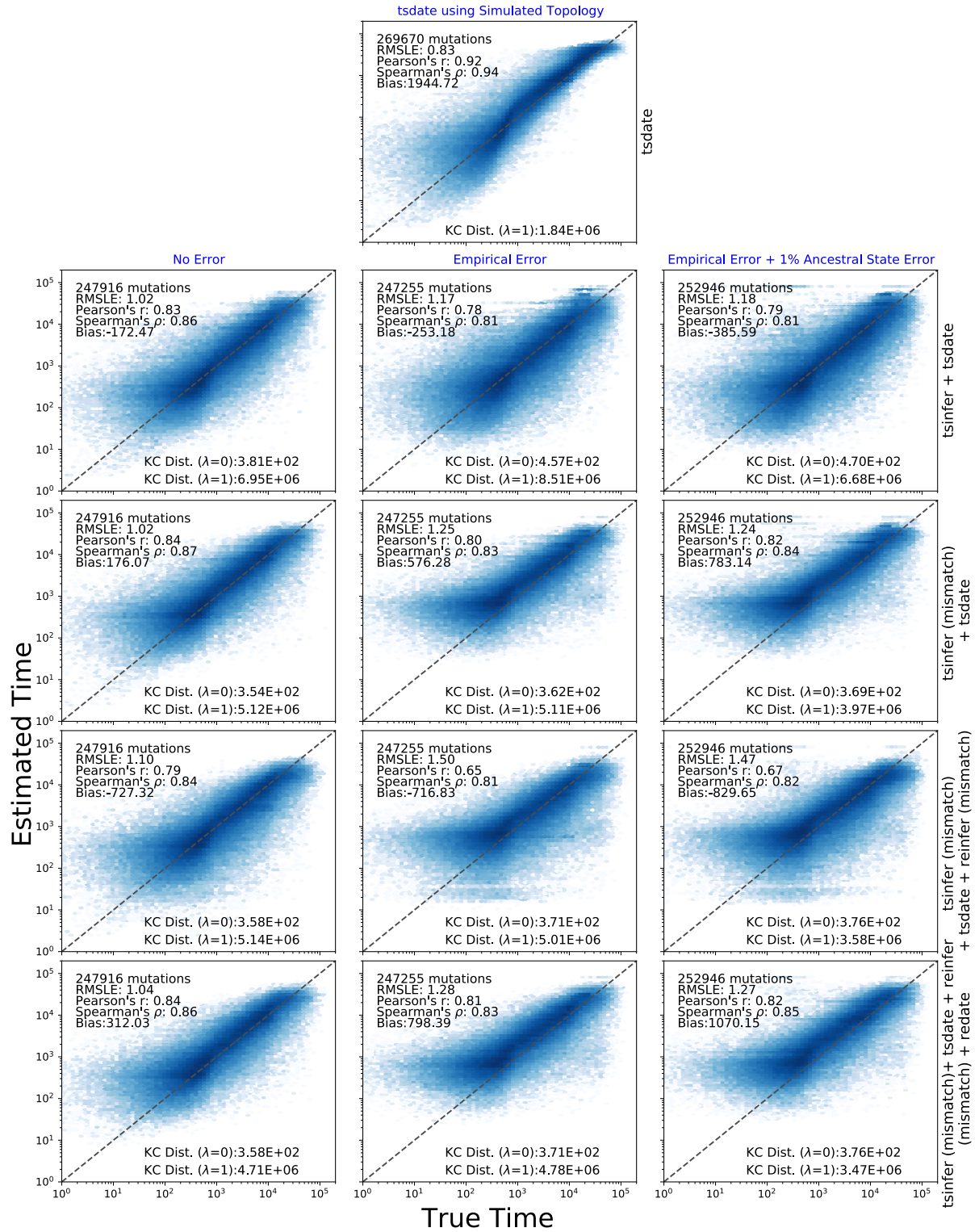


Fig. S3. Evaluation of `tdate` on a simulation of human chromosome 20 from `msprime` and `stdpopsim` using the GRCh37 recombination map, Out of Africa model (198) and 100 samples from each population. Row one shows the age of variants after running `tdate` on the simulated topology. Row two shows inferred topologies from `tsinfer` using the variation data from this simulation dated by `tdate`. Row three shows dated, inferred topologies from `tsinfer` with mismatch. Row four feeds the estimated dates from row three back into `tsinfer` without redating. Row five shows the results of redating tree sequences from row four. The first column in rows 2-4 shows results using variation data from the simulation without error, the second shows results after injecting error with an empirical genotype error model (33), and the third includes both the genotype error model and 1% ancestral state assignment error.

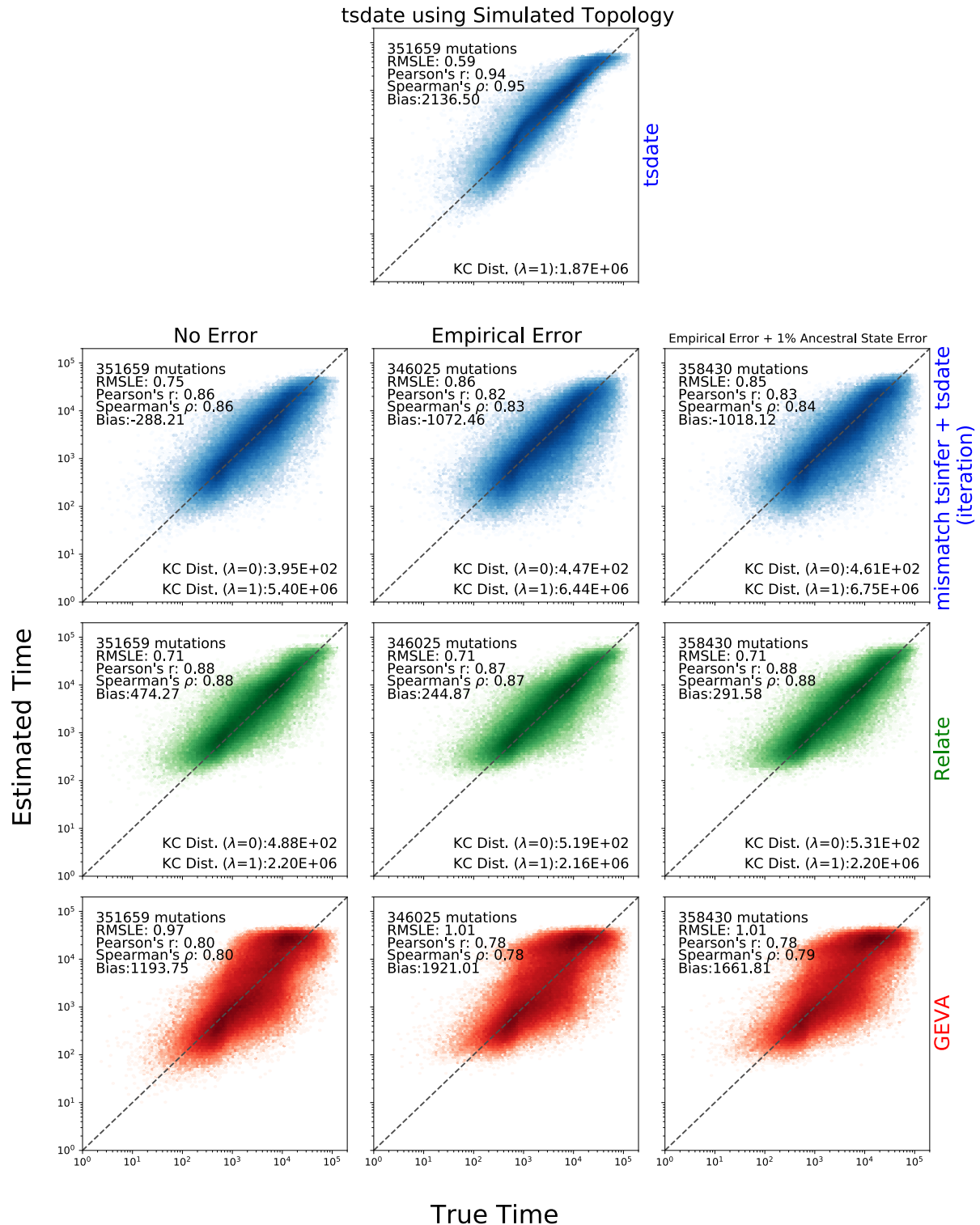


Fig. S4. Evaluation of the accuracy of *tsdate*, *tsinfer*, GEVA and Relate on 30 replicates of 5 Mb sections (positions 10-15 Mb) of chromosome 20 simulated as described in fig. S3. The top row shows the accuracy of *tsdate* on the simulated topology. The second row shows the results of inferring tree sequences with *tsinfer*, dating the tree sequence with *tsdate*, and then re-inferring and re-dating the tree sequence (using the chromosome 20 recombination map and a mismatch ratio of 1). The third row shows the results of Relate using a script provided by the authors to re-infer branch lengths and N_e . The fourth row shows the results of GEVA with default parameters. The columns use error models as described in fig. S3. In each column, only sites dated by all three methods are shown.

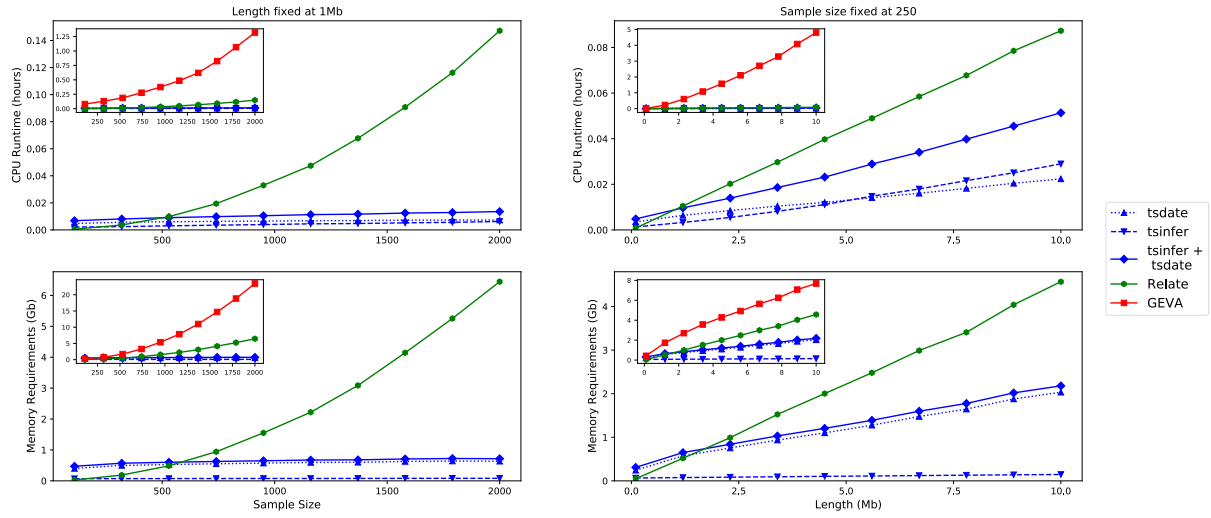


Fig. S5. Scaling properties of *tsdate* compared to *tsinfer* (without mismatch), *Relate*, and *GEVA*. The left column shows the CPU and memory requirements for inference using the three methods on *msprime* simulations of 1 Mb, with $N_e = 10^4$, $\mu = r = 10^{-8}$ and sample sizes from 10 to 2000 (*Relate* continues to scale quadratically with larger sample sizes). Five replicates were performed at each sample size. The column on the right shows results of ten *msprime* simulations with sample size fixed at 250 and simulated lengths of 100 kb to 10 Mb. The main axes compare results for *tsdate*, *tsinfer*, and *Relate*. The inset plots show the same data with the addition of *GEVA*, which otherwise obscures the differences between other methods.

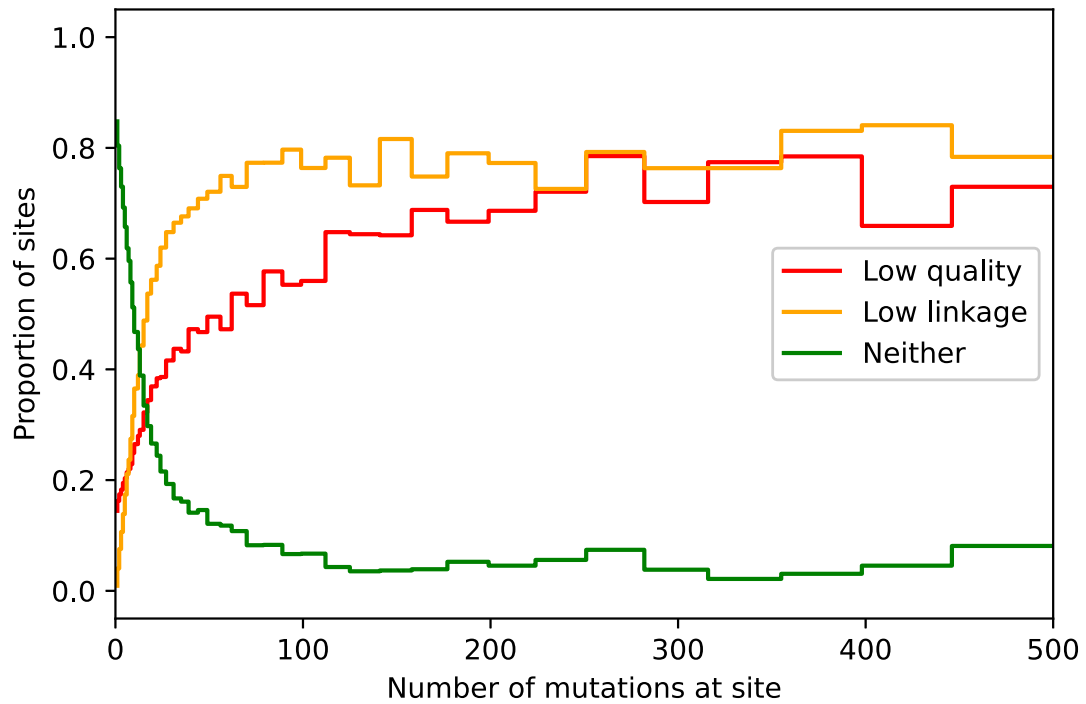
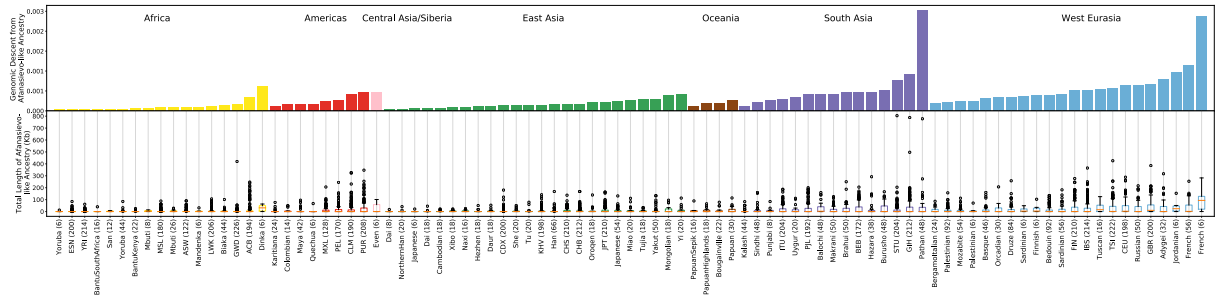
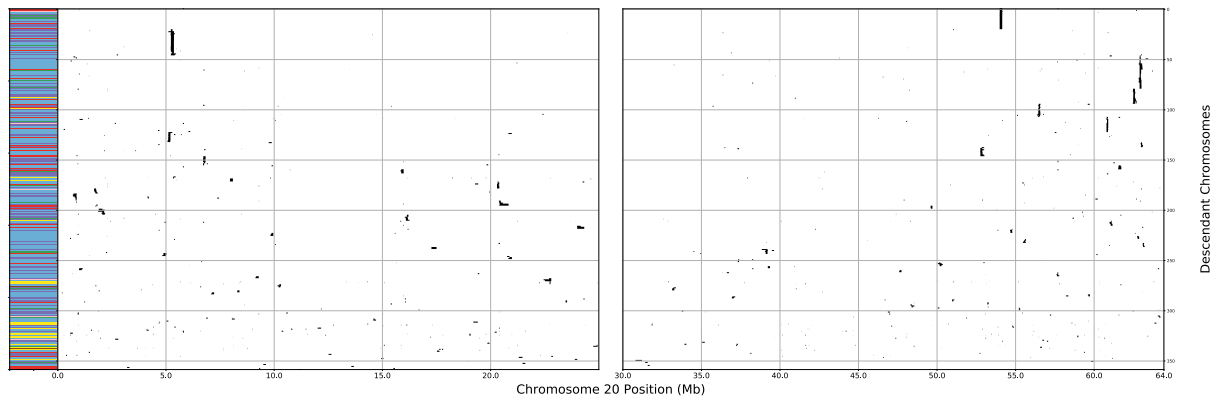


Fig. S6. The relationship between metrics of variant-calling error and the number of mutations on the inferred tree sequence. For the combined tree sequence of HGDP, TGP, SGDP, and ancient samples, the proportion of sites in two categories - low quality and low linkage, binned by the number of mutations at the site. Low quality is defined by the TGP strict accessibility mask. Sites that fail any of the accessibility mask filters such as low coverage or low mapping quality are marked as low quality. Low linkage is defined by summing the linkage disequilibrium (r^2) for the 50 sites either side of the focal site. If this quantity is less than 10 the site is marked as low linkage.



(A) Descent from proxy Afanasievo ancestors on chromosome 20 among HGDP, SGDP, and TGP populations. The upper panel shows the proportion of genetic material in each population that descends from the proxy Afanasievo ancestors (using the genomic descent statistic (57)). Only populations with at least two individuals and a mean genomic descent value of $\geq 0.01\%$ are shown. Box plots in the lower panel show the distribution of descendant material among samples from the corresponding population. The box indicates the inter-quartile range (IQR) of the total length of descendant material within each population. The center line indicates the median value. The whiskers extend to the minimum and maximum values within 1.5 IQR above the third quartile and below the first quartile. Diamonds indicate outlier data points greater than 1.5 IQR above the third quartile.



(B) Patterns of descent among samples with at least 100 kb of material descending from Afanasievo proxy ancestors. Each row is a sample and each column is a 1 kb section of chromosome 20. The region of origin of each sample is colored on the left side of the row.

Fig. S7. Inferred patterns of descent from Afanasievo proxy ancestors on chromosome 20. The color scheme for each region is the same as in Fig. 2.

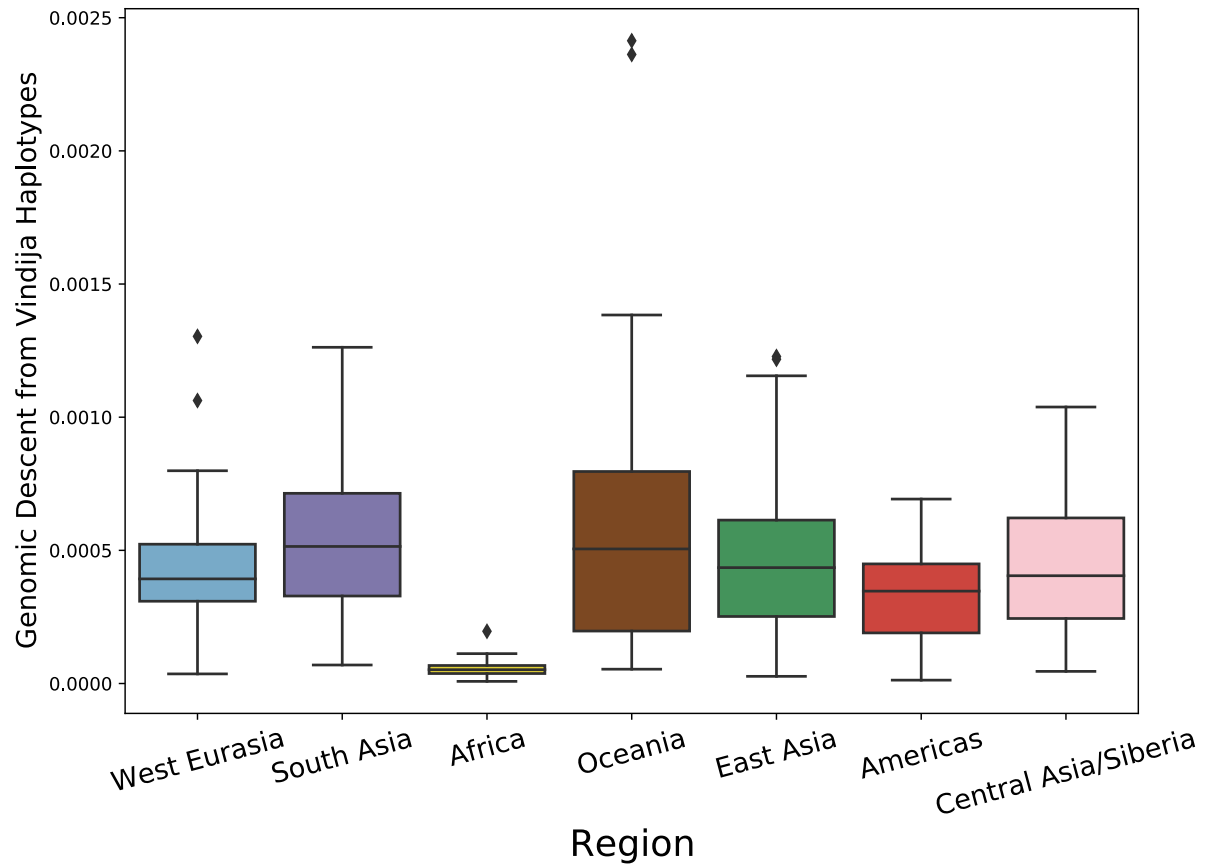
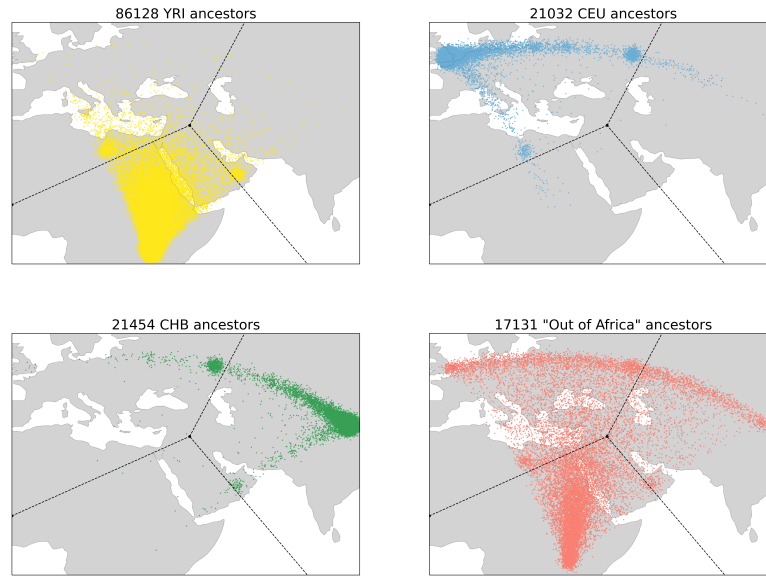
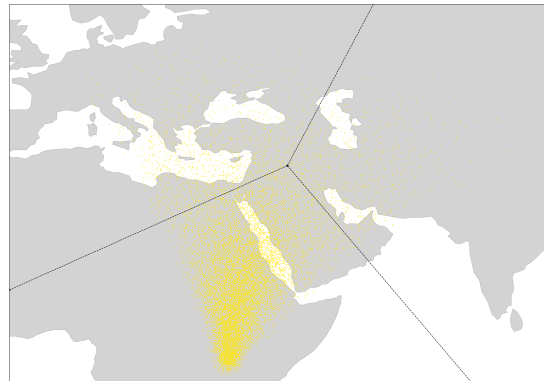


Fig. S8. Descent from Vindija Neanderthal proxy haplotypes. Each data point is a genomic descent statistic (57) calculated from the inferred tree sequence of the arms of each autosome. The statistic gives the proportion of genetic material in modern samples from each region that descends from Vindija proxy haplotypes. Box plot elements are defined in fig. S7A



(A) Inferred location of simulated ancestral haplotypes



(B) Inferred location of 17,246 inferred ancestral haplotypes predating the Out of Africa Event

Fig. S9. Evaluating the accuracy of our non-parametric estimator of ancestor geographic location. 10 replicates of 5 Mb of Chromosome 20 was simulated using a demographic model approximating human history (198) with 100 samples from YRI, CEU, and CHB. Samples are placed in Africa, Europe, and Asia at points roughly equidistant from one another (the locations are not meant to equate to the sampling locations for the three HapMap populations). Each subplot is subdivided into the regions closest to each point. (A) The inferred location of simulated ancestral haplotypes, colored by their population of origin. (B) The estimated locations of inferred ancestral haplotypes from a tree sequence produced from the simulated data by *tsinfer* and *tsdate* with empirical and ancestral state error injected. Only haplotypes older than 5,600 generations (the time of the Out of Africa event in the demographic model) are shown. Jitter is added to all coordinates plotted in this figure.

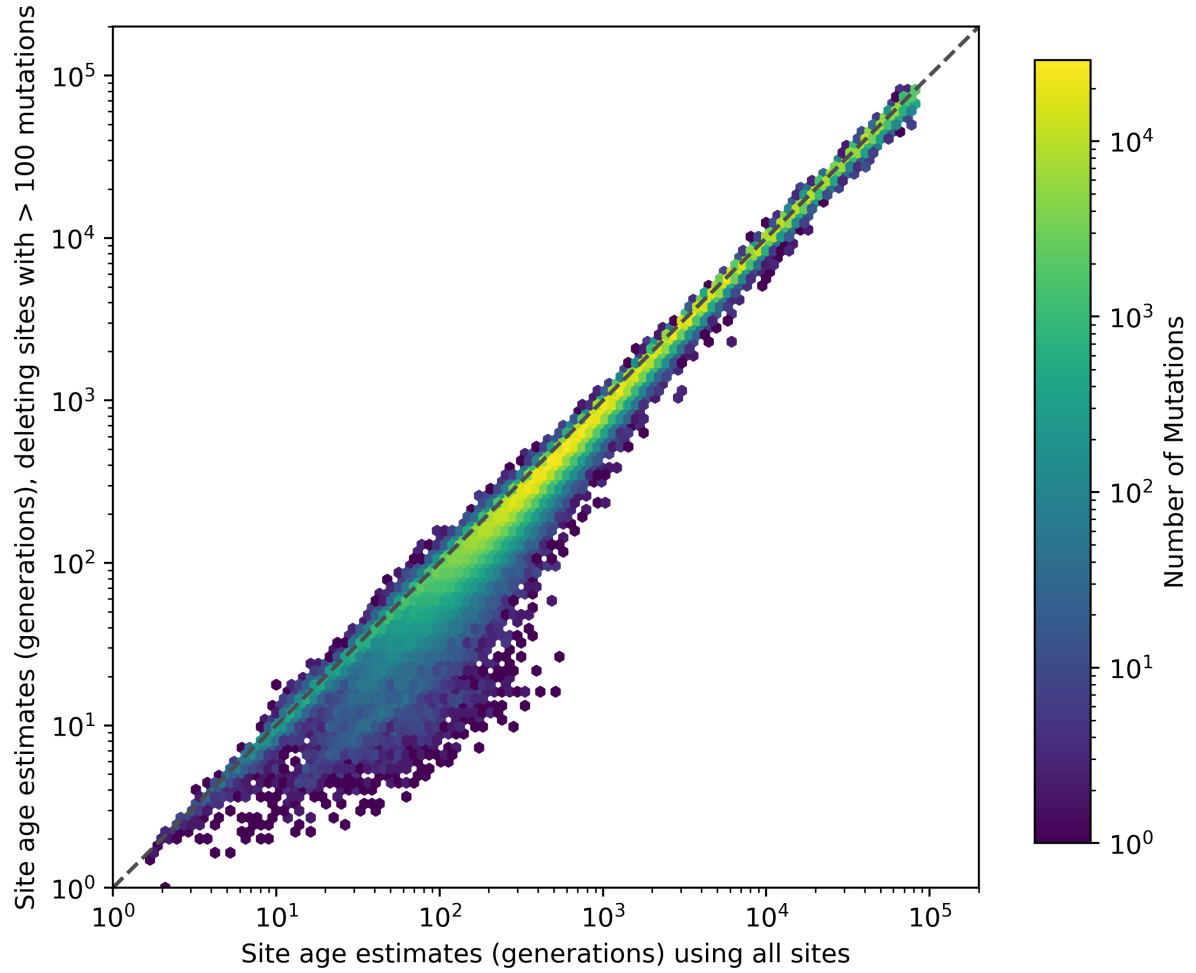


Fig. S10. Evaluating the effect of sites with greater than 100 mutations on the `tsdate` estimated age of variants on the long arm of Chromosome 20 in the unified genealogy of HGDP, TGP, and SGDP. Age estimates for 1,176,305 sites are shown, where age is assigned using the arithmetic mean of the upper and lower bounds of the constrained `tsdate` estimate of the oldest mutation at each site. Values on the x axis reflect age estimates from a tree sequence containing all 1,178,349 sites in the union of the HGDP, TGP, and SGDP on Chromosome 20. Values on the y axis show the result of dating the same tree sequence topology, but where 2,044 sites with greater than 100 mutations have been deleted.

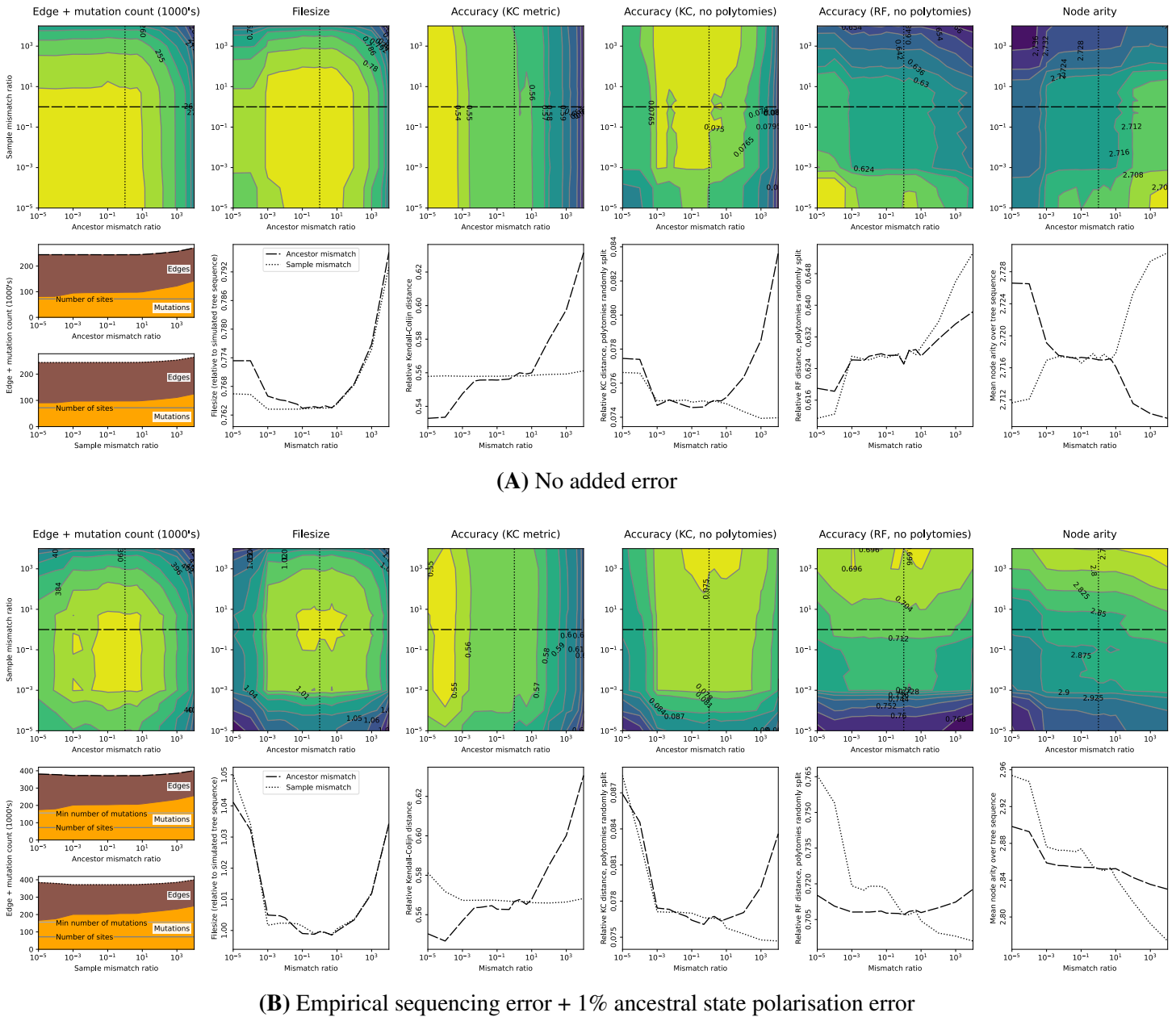
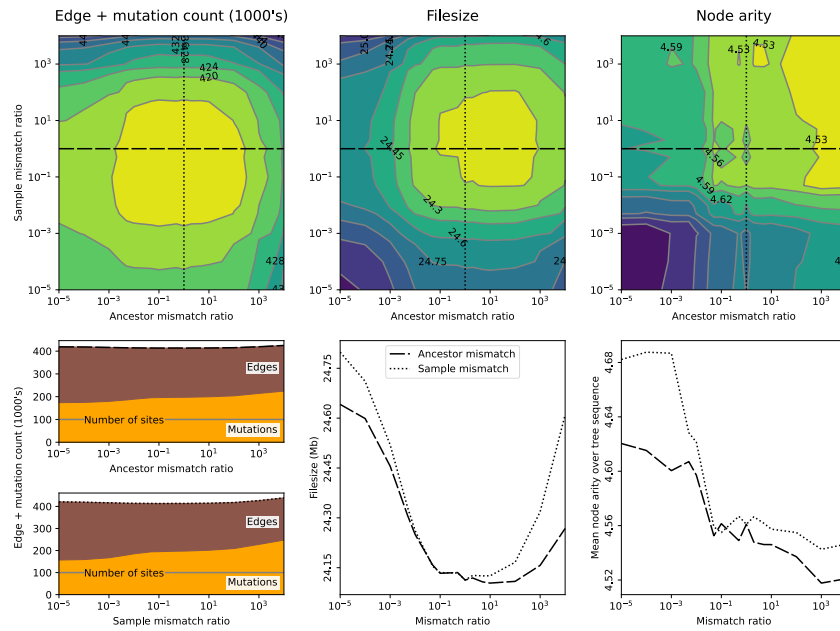
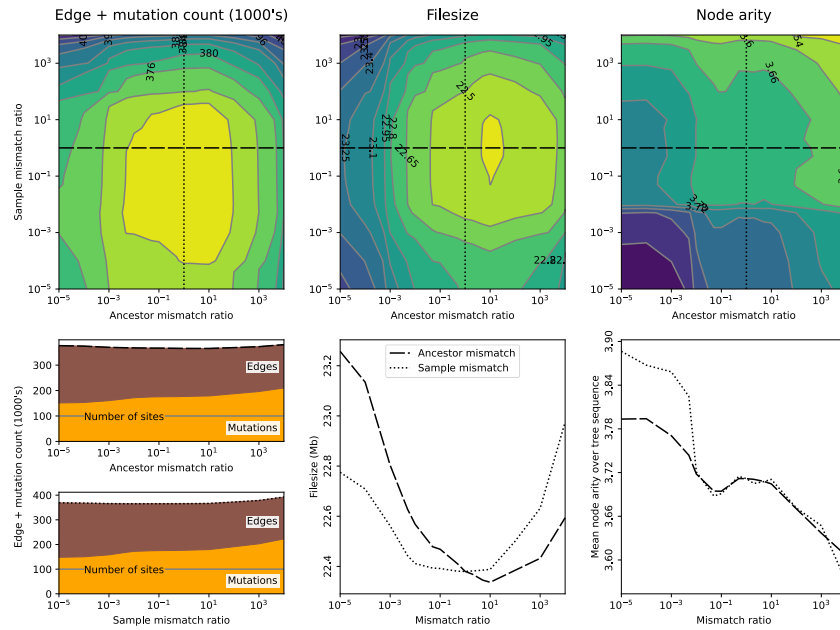


Fig. S11. The effect of varying the mismatch ratio on accuracy metrics for *tsinfer*. Results from 1,500 simulated human-like genome sequences of 10 Mb in length. (A) Simulations without error. (B) Simulations with an empirically calibrated genotyping error model and 1% error in ancestral state assignment. For each panel, the upper (colored contour) plots show accuracy metrics as a function of the mismatch ratio in ancestor matching (x-axis) and in sample matching (y-axis) algorithms. Slices through contour plots indicated by the dashed and dotted lines are plotted in the lower (line) plots. The total number of edges plus mutations, and file-size relative to the simulated tree sequence (first 2 columns) are indirect measures of accuracy (the minimum number of mutations required to explain error is calculated by overlaying mutations onto the known trees using parsimony). Direct measures of inference accuracy provided via the Kendall-Colijn (KC) or Robinson-Foulds (RF) tree-distance metrics (middle columns) which can, however, be influenced by polytomy size (i.e. node arity: last column); breaking polytomies at random may reduce this influence. Metrics are normalised against maximum expected distances. See Supplementary Text S1 for further details.



(A) 1000 Genomes Project (TGP), inference performed on a 4 Mb region from 5,008 genomes with no missing data



(B) Human Genome Diversity Project (HGDP), inference performed on 4.6 Mb region from 1,858 genomes with 1.2% missing data

Fig. S12. Effect of mismatch ratio parameter on `tsinfer` results from empirical sequence data (based on a subset of 100,000 sites on the short arm of chromosome 20). Dotted and dashed lines as for fig. S11. See Supplementary Text S1 for further details.

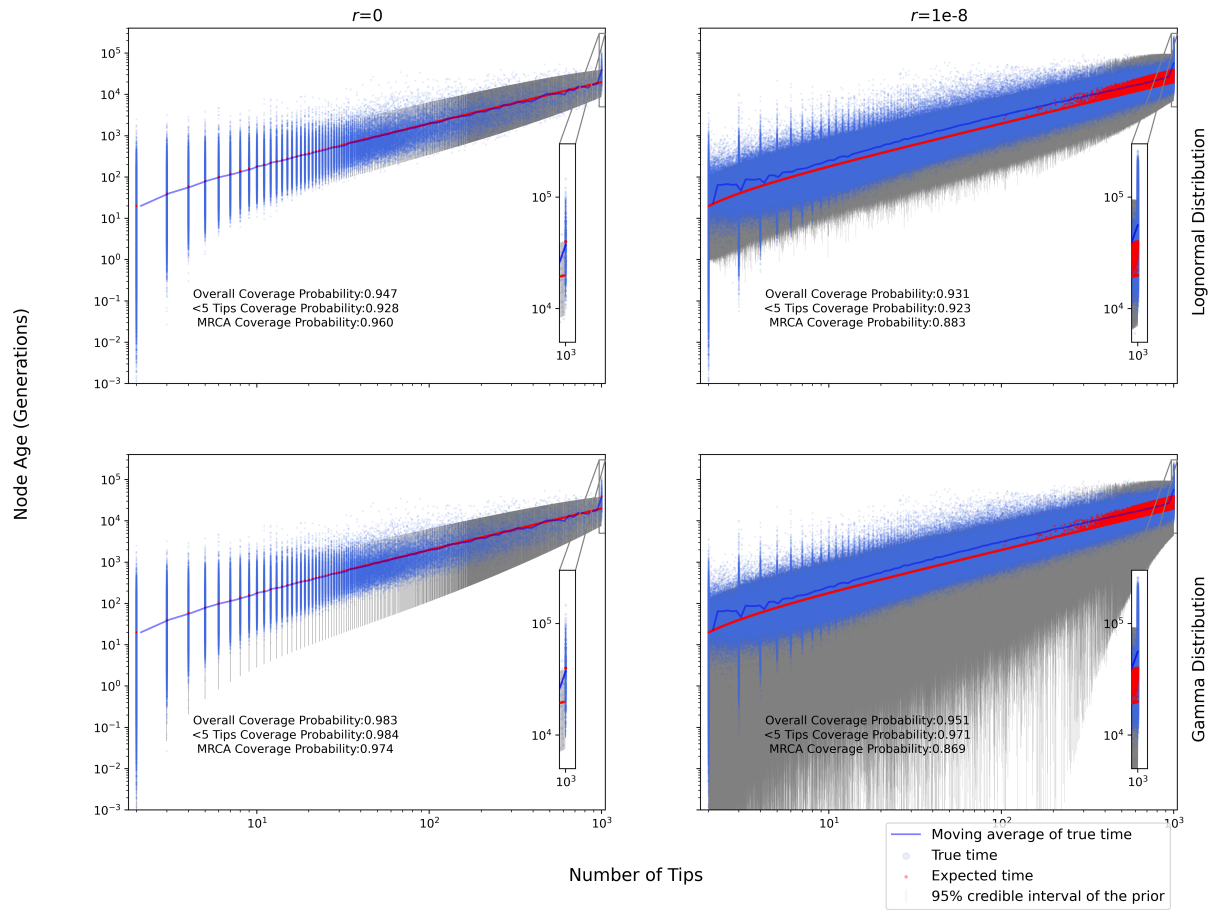


Fig. S13. Accuracy of the `tsvdate` node age prior distribution. The subplots in the left column show node ages from ten `msprime` simulations of length 500 kb with 1000 samples, $N_e = 10,000$, and $r = 0$. The right column shows results of ten simulations using $r = 10^{-8}$ and the same parameters otherwise. The top plots show the accuracy of the lognormal approximation to the conditional coalescent and the bottom show the gamma approximation.

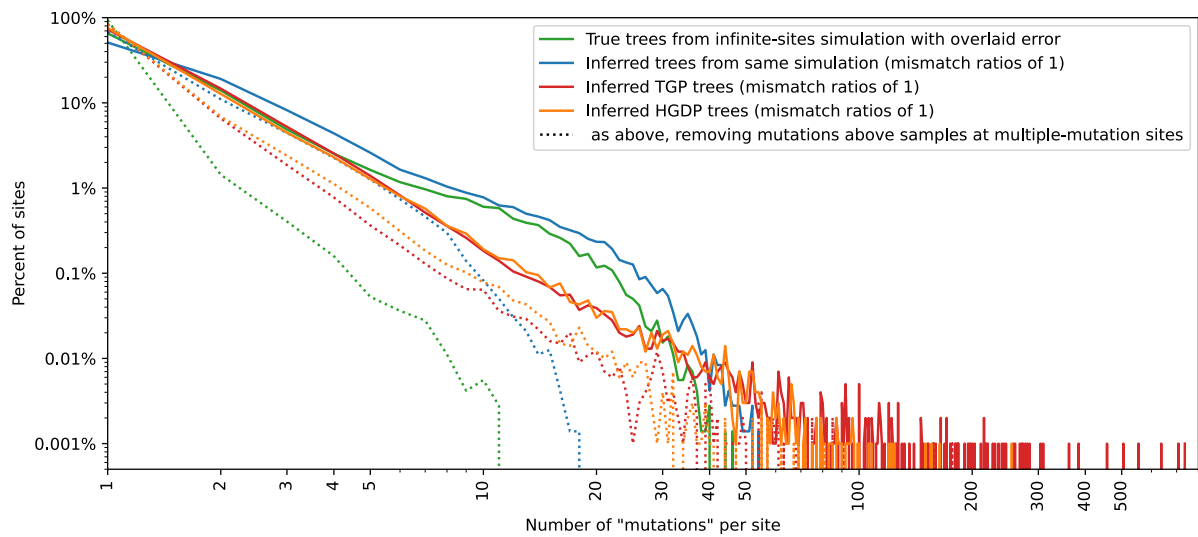


Fig. S14. The distribution of number of mutations at each site required to explain observed genetic variation, when including sequencing and ancestral state polarization error. Simulated (blue & green) data sources as in fig. S11, real (red & orange) data sources as in fig. S12. Minimum number of mutations required to explain error in the infinite-sites simulation (green) was calculated from the true genealogy by overlaying variant data on the known trees by parsimony. Other lines indicate tree sequences and associated mutations inferred using `tsinfer` with a mismatch ratio of 1. Dotted lines give the distribution of mutations per site when sites with greater than 1 mutation have mutations above sample nodes removed: these are predominantly indicative of mutations due to error.

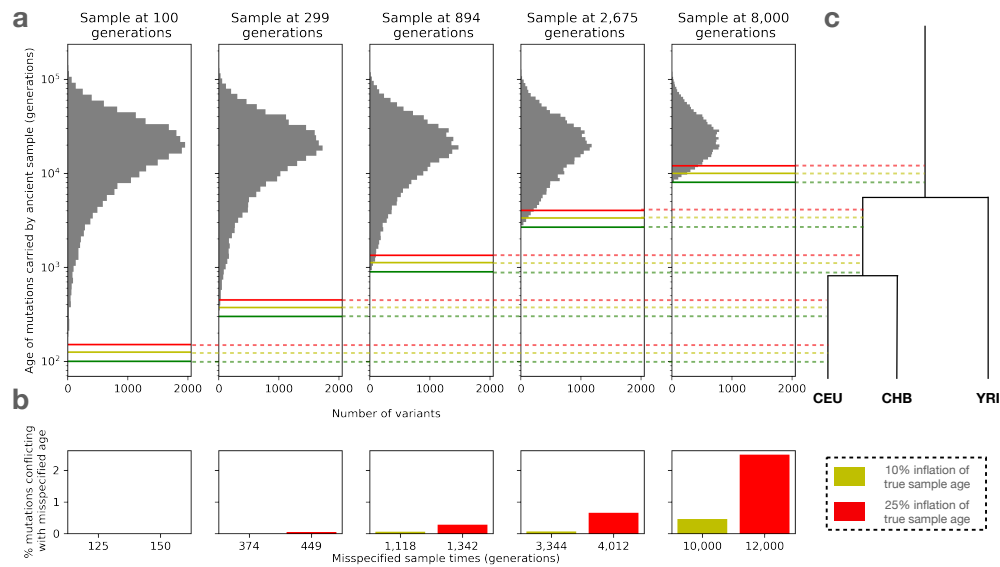


Fig. S15. The effect of misspecifying ancient sample age on mutation age. Ten replicates of the out of Africa coalescent simulation (21, 198, 199) were performed with ancient samples drawn at each of the indicated times (further details on the simulation can be found in the Supplementary Text S1). The top row of subplots show histograms of the age of mutations carried by ancient samples at the given times. Only mutations shared with modern samples are shown. Green horizontal lines indicate the sampling time of the ancient sample, yellow lines indicate a 10% inflation of the sampling time, and red lines indicate a 25% inflation. Dotted lines show the population in the demographic model from which each sample was drawn. The bottom row shows the percentage of sites carried by each ancient sample which are *younger* than the misspecified times, with colors corresponding to the horizontal lines above.

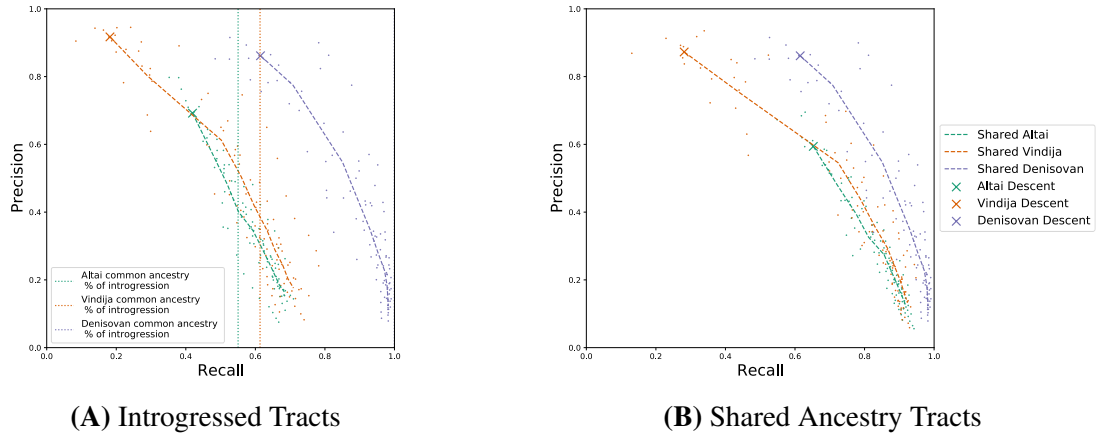


Fig. S16. Precision-recall curves for recovering relationships between archaic and modern samples using inferred tree sequences. (A) Accuracy of recovering introgressed tracts from archaic populations. Dotted vertical lines indicate total amount of modern genetic material where modern samples and each archaic individual share common ancestry in the *simulated* tree sequence more recently than T_{DenNea} as a proportion of total introgressed material. Common ancestry in the *inferred* tree sequences occurs when the TMRCA of modern samples and an archaic sample occurs more recently than a specified time cutoff. Cutoffs ranging from the time of each archaic sample to T_{DenNea} are tested, scatter plots show the results from each archaic sample at each cutoff and each simulation replicate. Dashed lines show the average precision and recall across replicates at each cutoff. Older time cutoffs result in progressively lower precision and higher recall. Precision is maximised at points marked with an “X”, where the cutoff is equal to the time of the sample, i.e. where only material directly descending from archaic proxy samples is used. (B) Precision-recall curves for recovering only the shared ancestry tracts for the three sampled archaics indicated by the dotted vertical lines in (A).

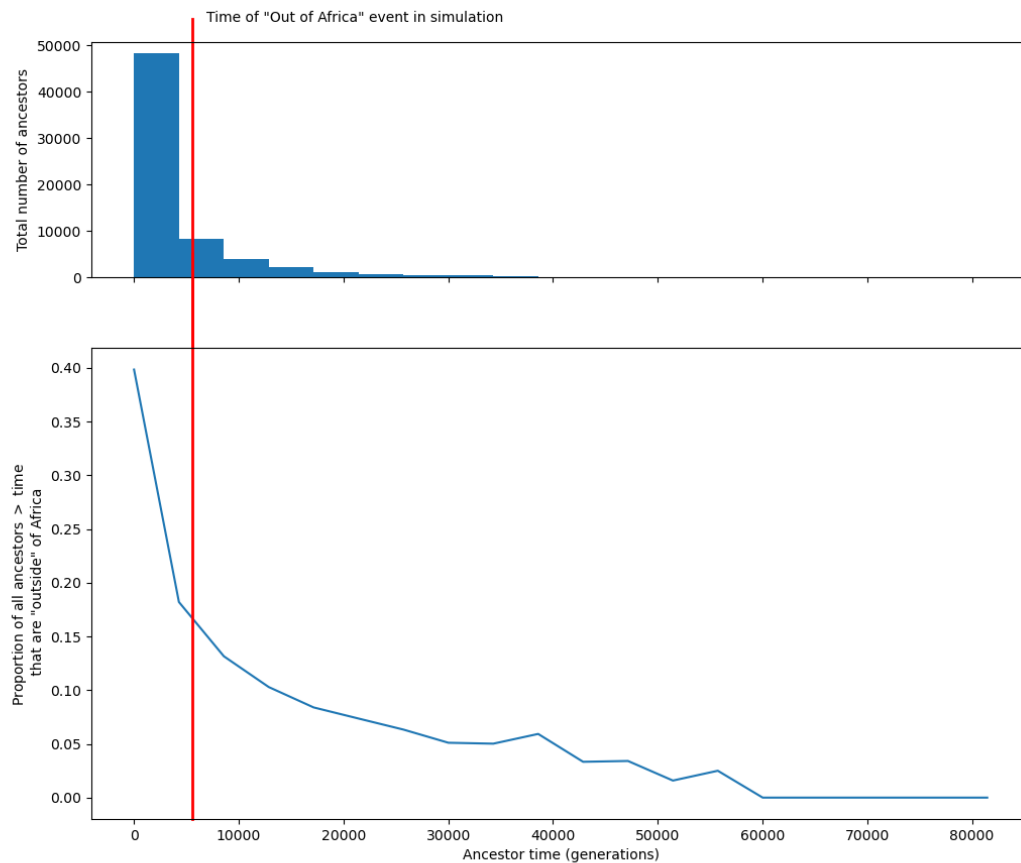


Fig. S17. Evaluating the proportion of ancestors located closer to Eurasia than Africa in geographic simulations. The same data from fig. S9 is shown in this figure. (A) Histogram showing the number of inferred ancestors at different times in the simulation. (B) The proportion of inferred ancestors older than the given time which are closer to the points in Europe or Asia than to the point in Africa.

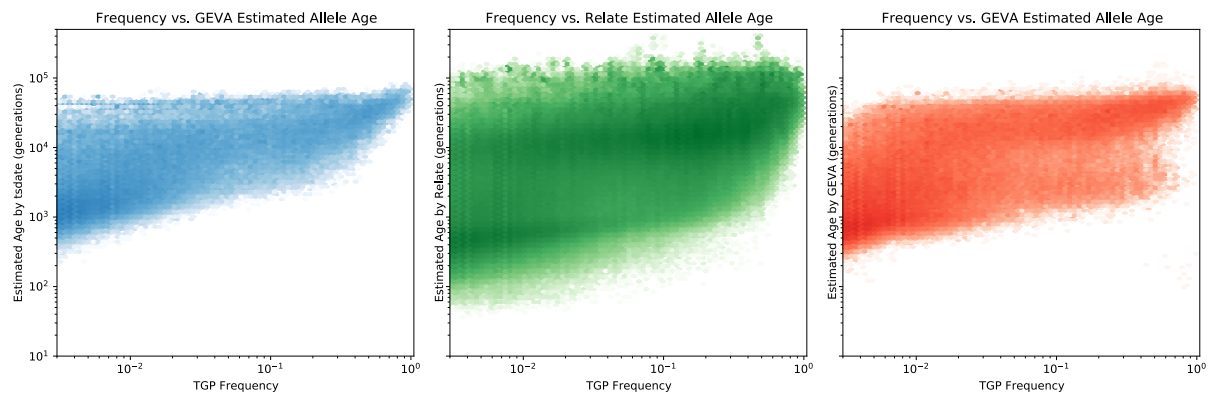


Fig. S18. Estimated age of TGP variants from `tsdate`, GEVA and `Relate` compared to allele frequency. Estimates of allele age are found using the arithmetic mean of the node ages above and below the mutation for `tsdate` and `Relate`. For GEVA, the mean age of the joint clock estimate is used. The same set of alleles is also used in Figs. 3A, S19, and S20

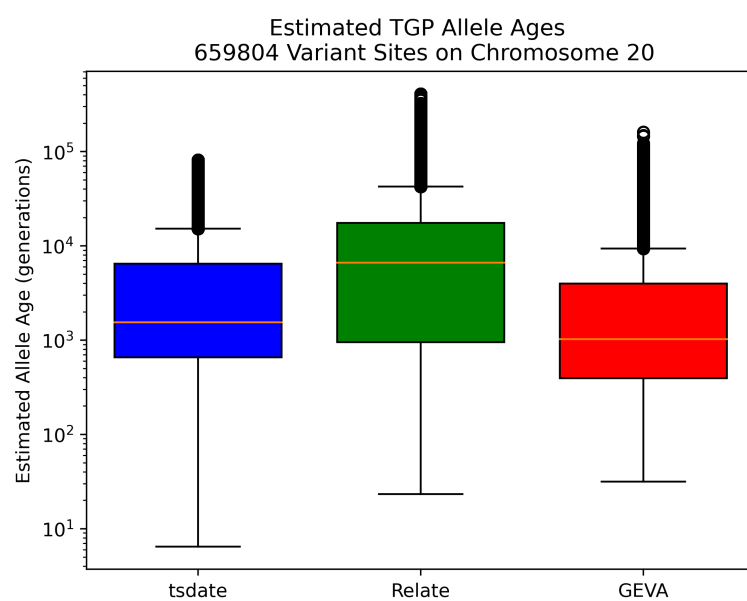


Fig. S19. Average age of TGP variants from *tsdate*, *GEVA* and *Relate*. Box plot elements are defined in fig. S7A. Note, the age range of *tsdate* depends on the number of time slices specified by the user (default settings were used for this paper).

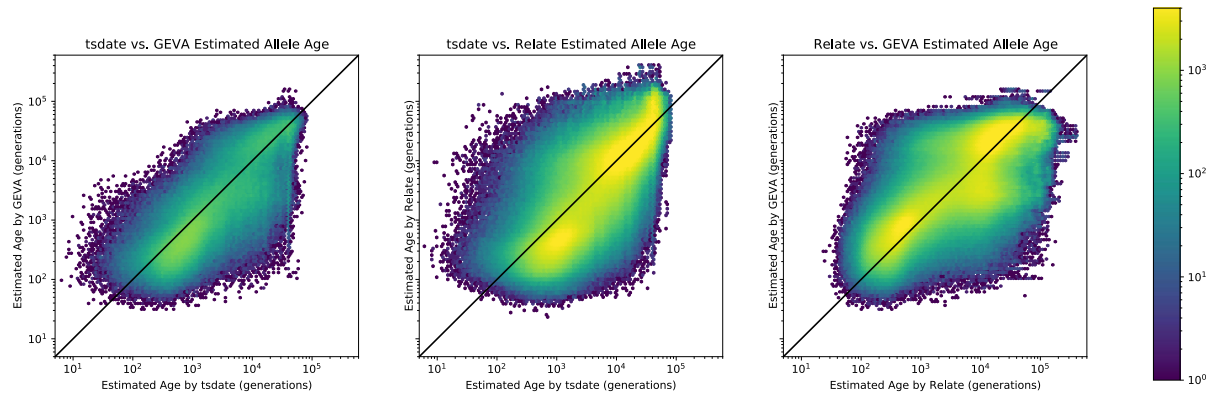
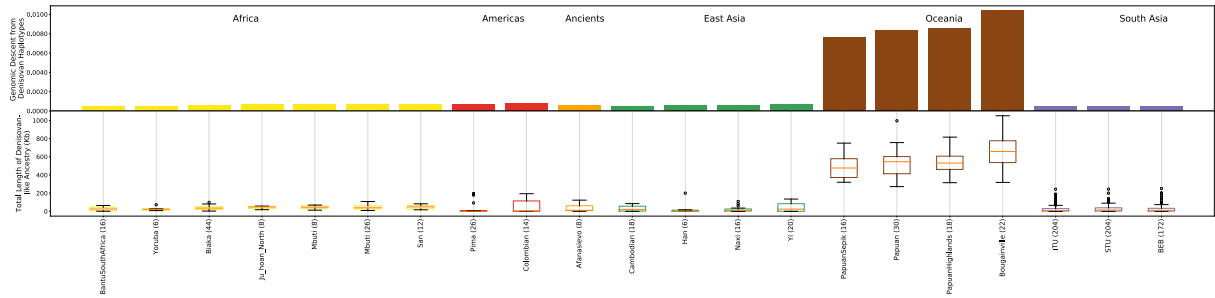


Fig. S20. Comparison of TGP derived allele age estimates from `tsdate`, `GEVA` and `Relate`. Each hexbin plot shows estimates of variant age by two different methods.



Master ID	Skeletal code	LibraryID(s)	Family position	Date (BCE)	Culture	Location	Country	Sex	mtDNA (restrict to >2x)	Y chrom. calls	HiSeq X10 lanes	% Endogenous	Mean length	mapped fragments	fragments after removing duplicates	Mean coverage on autosomal SNP targets
I3388	StPet53, collection 6612, individual 6	S3388.E1.L2	mother	2800-2600 BCE	Russia_Afanasievo	Altai Mountains, Yenisey River, left bank of Karasuk tributary, Karasuk III	Russia	F	U5a1d2b	..	8	59%	45	2,208,684,117	808,547,889	10.81666
I3950	StPet48, collection 6612, individual 2	S3950.E1.L1	father	2879-2632 calBCE (4160±25 BP, PSUAMS-1955)	Russia_Afanasievo	Altai Mountains, Yenisey River, left bank of Karasuk tributary, Karasuk III	Russia	M	U5b2a1a +16311	Q1b2a1a~	5	90%	45	2,106,410,130	1,657,567,059	25.79713
I6714	StPet49, collection 6612, individual 3	S6714.E1.L1	son2	2618-2468 calBCE (4020±25 BP, PSUAMS-3909)	Russia_Afanasievo	Altai Mountains, Yenisey River, left bank of Karasuk tributary, Karasuk III	Russia	M	U5a1d2b	Q1b	7	69%	41	2,066,920,322	1,494,595,877	21.22681
I3949	StPet47, collection 6612, individual 1	S3949.E1.L1	son1	2844-2496 calBCE (4075±20 BP, PSUAMS-2292)	Russia_Afanasievo	Altai Mountains, Yenisey River, left bank of Karasuk tributary, Karasuk III	Russia	M	U5a1d2b	Q1b	5	85%	44	1,898,514,533	1,582,864,152	25.31653

Table S1. Sequencing information for the Afanasievo family.

	Vindija		Altai		Denisovan	
	Introgressed	Common Ancestry	Introgressed	Common Ancestry	Introgressed	Common Ancestry
Precision	0.92 (0.03)	0.87 (0.05)	0.69 (0.08)	0.59 (0.08)	0.86 (0.04)	0.86 (0.04)
Recall	0.18 (0.05)	0.28 (0.07)	0.42 (0.04)	0.65 (0.06)	0.61 (0.11)	0.61 (0.11)
Simulated % of moderns	2.86% (1.12%)	1.77% (0.71%)	2.86% (1.12%)	1.57% (0.60%)	0.99% (0.24%)	0.98% (0.24%)
Inferred % of moderns	0.53% (0.15%)		1.65% (0.31%)		0.69% (0.12%)	

Table S2. Results of archaic descent simulations evaluating how well direct descent from proxy archaic haplotypes in inferred tree sequences recovers ground truth introgressed tracts of sequence as well as tracts where modern samples and sampled archaics share common ancestry more recently than T_{DenNea} . Definitions of precision and recall are given in Equations 4 and 5. “Simulated % of moderns” refers to the percentage of genetic material from all modern individuals in each simulation that is introgressed from archaics or shares recent common ancestry with archaics. “Descendant % in moderns” is the percentage of modern genetic material in each inferred tree sequence which descends from each archaic individual. In each cell, the mean values from 10 simulation replicates is given with the standard deviation in brackets.

	Improving allele age estimates	Inferring genetic relationships
Coverage	Any coverage acceptable (sequenced or genotyped)	High coverage ($\sim 15\times$) sequencing data required
Errors	Sites which are more highly mutable or prone to error (such as CpG sites) can upwardly bias results	Methods should be robust to occasional genotype errors, but high quality haplotypes over 10s to 100s of kb are required (comparable to modern data)
Sample age	Older samples provide more information	Value of older samples depends on proximity to age of ancestral populations of interest
Sample dating quality	Overestimating sample age by 25% will likely not bias results	Large errors could impact estimates of the timing of ancestral relationships

Table S3. Guidelines for use of ancient samples with `tsinfer` and `tsdate`. The impact of four features of ancient samples, rows, on two use cases of ancient samples, columns. The guidelines presented here are informed by simulations and empirical analyses presented in this work, specifically Figs. 1D, 3, S15, and S16, as well as Sections S1 and S2.

Movie S1.

Spatio-temporal dynamics in human history. This movie shows the estimated geographic locations of ancestors of Human Genome Diversity Project, Simons Genome Diversity Project, Neanderthal, Denisovan, and Afanasievo samples over time. Each dot represents an edge in the tree sequence of chromosome 20, where the time and geographic location of the parent and child nodes of the edge have been estimated. The locations of edges at each point in time are plotted along the great circle between the parent and child nodes. Edges are colored by the region of the descendants of the child node. If an ancestral lineage has ancestors in multiple regions, its color is the average of the respective colors of each region.

Interactive Figure S1.

Time to the most recent common ancestor on chromosome 20 between samples in the unified genealogy of modern and ancient genomes. Interactive Figure S1 is available at: https://awohns.github.io/unified_genealogy/interactive_figure.html. This dynamic version of Fig. 2 shows the span-weighted TMRCA histogram relating each pair of the 215 populations in the integrated tree sequence of chromosome 20. Hovering over each cell displays the relevant histogram, with a vertical bar indicating the logarithmic mean TMRCA.