Dating genomic variants and shared ancestry in population-scale sequencing data

Patrick K. Albers[1*] and Gil McVean[1]

[1]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Old Road Campus, Oxford OX3 7LF, United Kingdom

*To whom correspondence should be addressed;
E-mail: patrick.albers@bdi.ox.ac.uk

Abstract. The origin and fate of new mutations within species is the fundamental process underlying evolution. However, while much attention has been focused on characterizing the presence, frequency, and phenotypic impact of genetic variation, the evolutionary histories of most variants are largely unexplored. We have developed a non-parametric approach for estimating the date of origin of genetic variants in large-scale sequencing data sets. The accuracy and robustness of the approach is demonstrated through simulation. Using data from two publicly available human genomic diversity resources, we estimated the age of more than 45 million single nucleotide polymorphisms (SNPs) in the human genome and release the Atlas of Variant Age as a public online database. We characterize the relationship between variant age and frequency in different geographical regions, and demonstrate the value of age information in interpreting variants of functional and selective importance. Finally, we use allele age estimates to power a rapid approach for inferring the ancestry shared between individual genomes, to quantify genealogical relationships at different points in the past, as well as describe and explore the evolutionary history of modern human populations.

Introduction

Each generation, a human genome acquires an average of about 70 single nucleotide changes through mutation in the germ-line of its parents [1]. Yet while, at a global scale, many millions of new variants are generated each year, the vast majority are lost rapidly through genetic drift and purifying selection. Consequently, even though the majority of variants themselves are extremely rare, the majority of genetic differences between genomes result from variants found at global frequencies of 1% or more [2], which may have appeared thousands of generations ago. Genome sequencing studies [3] have catalogued the vast majority of common variation (estimated to be about 10 million variants [4]) and, at least within coding regions and particular ancestries, to date, more than 660 million variants genome-wide have been reported [5], many of them at extremely low frequency [2].

Despite the importance of genetic variation in influencing quantitative traits and risk for disease, as well as providing the raw material on which natural selection can act, relatively little attention has been paid to inferring the evolutionary history of the variants themselves; with notable exceptions of evolutionary importance, particularly those affecting geographically varying traits such as skin pigmentation, diet, and immunity [6–8]. Rather, attention has focused on the indirect use of genetic variation to detect population structure [9, 10], identify related samples [11, 12], infer parameters of models of human demographic history [13, 14], and estimate the time of ancestral population divergence [15]. The evolutionary history of classes of variants contributing to polygenic adaptation (for example those affecting height [6, 16, 17]) or causing potential loss of gene function [18] has received attention, though rarely at the level of specific variants. Previous work on rare variants has identified ancestral connections between individuals and populations [19–21] and demonstrated evidence for explosive population growth [22]. Nevertheless, to date, no comprehensive effort has been made to infer the age, place of origin, or pattern of spread for the vast majority of variants.

We have developed a method for estimating the age of genetic variants; that is, the time of origin of an allele through mutation at a single locus. Our approach, which we refer to as the Genealogical Estimation of Variant Age (GEVA), is similar to existing methods that involve coalescent modeling to infer the time to the most recent common ancestor (TMRCA) between individual genomes [13, 23, 24]. However, these methods typically operate on a discretized timescale [13], utilize only a fraction of the information available in larger sample data, or employ approximations to overcome computational complexity [14, 15, 25]. Alternate approaches, in particular those that have been used to indicate allele age, are often strongly parameterized and make assumptions about demographic history, selection, or the shape of genealogical trees [26–29]. Our method is non-parametric and makes no assumptions about the demographic or selective processes that shaped the underlying genealogy. Instead, we learn about the age of a mutation by combining probabilistic distributions of TMRCA between hundreds or thousands of genomes, on a continuous timescale, irrespective of contemporary allelic distributions or historic population boundaries.

A copy of the piece of the ancestral chromosome on which the mutation occurred is still present today in the individuals carrying the mutant allele (Figure 1A). Over time, additional mutations have accumulated along the inherited sequence (haplotype), and its length has been broken down by recombination during meiosis in each generation. We locate this ancestral segment (often referred to as identity by descent, IBD, region) relative to the position of a given target variant, using a hidden Markov model (HMM) constructed empirically from sequencing data to provide robustness to realistic rates of data error (Figure 1B). By measuring the impact of mutation and recombination on the segments shared between pairs of haplotypes, we infer the TMRCA using probabilistic models to accommodate the stochastic nature of the mutation and recombination processes (Figure 1C). Moreover, we make full use of the information available in whole genome sequencing data, to perform comparisons between pairs of chromosomes that both carry the mutation (concordant pairs) and pairs where one carries the mutation and the other carries the ancestral allele (discordant pairs); thereby considering genealogical relationships that are both younger and older, respectively, than the time of mutation. Information from hundreds or thousands of haplotype pairs is then combined within a composite likelihood framework to obtain an approximate posterior distribution on the age of the derived allele (Figure 1D). One benefit of our method is that we can increase the number

of pairwise TMRCA inferences incrementally to update age estimates, or to combine information across many data sources to improve the genealogical resolution from a wider distribution of independently sampled genomes. We additionally use a heuristic method for rejecting outlier pairs to improve robustness to low rates of data error and recurrent mutation. Full details are given in the Supplementary Text.

Simulation study

To validate our approach, we performed coalescent simulations under different demographic models. Using a standard coalescent model with constant mutation and recombination rates, we found low bias (relative error, ϵ=0.268; see Supplementary Text) for allele age estimates and high correlation between true and inferred age (Spearman's ρ=0.953; Figure 2A). We also compared our approach for estimating TMRCA to that obtained from the computationally more demanding pairwise sequentially Markovian coalescent (PSMC) methodology [13]. This estimates a model of the demographic history between pairs of chromosomes (over a discretized grid of time intervals) and can, for every position in the genome, return the inferred posterior distribution on the TMRCA. We implemented this methodology and found that PSMC-based age estimations performed similarly well (ρ=0.952), though the time discretization increased bias (ϵ=0.530) and, in particular, led to overestimation of the age for the youngest variants. We note that PSMC was not designed strictly for this purpose, hence is not optimized for estimating allele age. Conversely, pairwise estimates of TMRCA between concordant haplotypes were highly correlated with true TMRCA in both GEVA (ρ=0.922) and PSMC (ρ=0.919), but the correlation for discordant pairs was lower in GEVA (ρ=0.586) compared to PSMC (ρ=0.766); see Supplementary Figure S1. Such differences, in relation to estimating allele age with high accuracy, are tolerated, because the time of mutation is estimated from the composite distribution of TMRCA posteriors from many pairwise comparisons performed at a single locus, where events that occurred distant in time (relative to the time of mutation) may only have little influence on the estimation process.

Under a complex demographic model that recapitulates the human expansion out of Africa and with empirical and variable recombination rates, age estimation in GEVA maintained a similarly high level of accuracy (ϵ=0.198, ρ=0.937; Figure 2B). In this situation, although PSMC modeled the more dynamic demographic histories between haplotypes with higher accuracy (correlation of true and inferred TMRCA for discordant pairs: ρ=0.915) compared to GEVA (ρ=0.775), the time discretization resulted in artifacts at more recent times (TMRCA for concordant pairs: ρ=0.892) that were not present in GEVA (ρ=0.932), leading to worse performance when estimating the age (ϵ=0.409, ρ=0.900; Supplementary Figure S2), with the addition of substantial computational cost. We next introduced realistic data complications by reproducing empirically estimated genotype errors in simulated data (see Supplementary Text), as well as errors arising through *in silico* haplotype phasing (Figure 2B). We found age estimation in GEVA to remain largely unbiased and strongly correlated with true age after the inclusion of data error (ϵ=0.346, ρ=0.925; Supplementary Figure S3) and after phasing (ϵ=0.430, ρ=0.921; Supplementary Figure S4). The PSMC-based approach continued to show higher bias and reduced correlation at the same set of variants, both after error (ϵ=1.042, ρ=0.882) and after phasing (ϵ=1.009, ρ=0.880). Reduced data quality resulting

from sequencing errors may introduce false signals of pairwise differences seen between haplotypes, and phasing errors may lead to an underestimation of haplotype lengths at variants that are relatively young, for which we overestimate TMRCA and, hence, allele age (particularly for alleles younger than approximately 100 generations).

Age of selected variants

To evaluate the performance of GEVA on empirical data, we first considered variants affecting the well-studied lactase persistence (LP) trait, where numerous approaches, including the use of archaeological data, genetic data, and a biological understanding of the functional and evolutionary impact of previously associated variants has resulted in consensus expectations for the age. The *LCT* gene encodes the lactase enzyme, but is regulated by variants in an intron of the neighboring *MCM6* gene. We estimated the age of the derived T allele of the rs182549 variant (G/A-22018), which is at a frequency of approximately 50% in European populations and which forms part of a haplotype associated with LP [30]. Under a model that jointly considers mutational and recombinational information, we estimated the allele to be 688 generations old (Figure 3A); approximately 14,000 to 21,000 years ago, depending on assumptions about generation time in humans [31, 32]. Our estimate is based on data from two different sources, the 1000 Genomes Project (TGP) [2] and the Simons Genetic Diversity Project (SGDP) [33], which, when estimated separately, give very similar ages (692 and 687 generations, respectively). The full result data set for this variant is available online; https://human.genome.dating/snp/rs182549. We obtained a similar age estimate of 693 generations for the derived A allele of the rs4988235 (C/T-13910) variant, see https://human.genome.dating/snp/rs4988235, which is also strongly associated with LP and in near perfect association with rs182549; though we note that there is evidence for multiple origins of the variant [34]. Previous estimates of the age of these variants range between 2,200 and 21,000 years [35], putting our estimates on the higher end of this range. Multiple sources of information suggest that these variants only achieved high frequency in European populations within the last 10,000 years (∼400 generations) [36], and that LP alleles were rare until the advent of dairy farming in Europe [37]. Our results therefore suggest that the mutation conferring the strongly selected phenotype (estimated to have a selection coefficient of up to 15% in European and up to 19% in Scandinavian populations [36]) was present for hundreds of generations before its rapid sweep through the population.

We next considered the protein-coding missense variant rs3827760 in the *EDAR* gene, where the derived G allele (Val370Ala substitution) is found at high frequency in East Asian populations (87% in TGP, 82% in SGDP) and American populations (39% and 80%, respectively), and which is associated with sweat, facial and body morphology, and hair phenotypes [38, 39]. We estimated the variant to be 1,456 generations old, approximately 29,000 to 44,000 years (Figure 3B; https://human.genome.dating/snp/rs3827760), again with strong concordance between TGP (1,513 generations) and SGDP (1,346 generations). Our estimate is consistent with previous estimates and limited evidence from ancient DNA studies [7, 40]. Our results further show that most individuals carrying the allele share a common ancestor close to the time the allele arose through mutation. This may suggest that the variant rapidly rose in frequency following its origin,

which is consistent with previous findings of strong positive selection of this variant in East Asia [38]. We compared this result to variants of similar age (1,450±100 generations) in the Atlas of Variant Age (see further below); we found 1,043,376 variants dated in TGP, of which 2,073 have reached a frequency higher than 30% globally, and only 130 above 80% in East Asian populations, demonstrating how unusual such a rapid rise in frequency is.

Finally, we considered the variant rs80194531, where the derived allele causes an Asn78Thr substitution in the *ZEB1* gene. The variant is reported as pathogenic for corneal dystrophy [41], but is present at 6% in African ancestry samples within TGP or SGDP. We estimated the age of the variant to be 5,866 generations old (120,000 to 180,000 years), again with consistency between TGP and SGDP (5,879 and 5,854 generations, respectively; Figure 3C; https://human.genome.dating/snp/rs80194531). Such an ancient age seems inconsistent with the reported dominant pathogenic effect [41]. Moreover, of the 1.2 million variants found at comparable frequencies (5—7%) in African ancestry individuals within TGP, we found that 46% were estimated to be younger than the rs80194531 allele, suggesting that this variant is in no way unusual.

Distribution of allele age in the human genome

We next sought to characterize the age distribution of genetic variation across the human genome, for which we applied GEVA to more than 45 million variants identified in TGP or SGDP. More than 32 billion haplotype pairs were analyzed to estimate shared haplotype segments and TMRCA. For variants present in both data sources (13.7 million), we additionally estimated the age by combining pairwise TMRCA distributions that were inferred independently in each sample; after confirming that separately obtained age estimates agreed (Spearman's ρ=0.862; Supplementary Figure S5). We make this information, referred to as the Atlas of Variant Age for the human genome, publicly available as an online database (https://human.genome.dating). A breakdown by chromosome of the number of variants dated and haplotype pairs analyzed is given in Supplementary Table S1.

We find substantial variation in the relationship between estimated age and allele frequency, depending on the population in which frequency is measured and the geographical distribution of the variant (Figure 4A). Variants in African ancestry groups are typically older than in other groups and also have the greatest variance in age for a given frequency. For example, variants below 0.5% (within a given population) have a median age of 670 generations in African ancestry groups, 377 generations in East Asian ancestry groups, and 488 generations in Europeans (see Supplementary Table S2). The age distribution of variants restricted to a particular ancestry group (or shared between them) indicates the degree of connection between populations. For example, there are many variants up to 5,000 generations old (100,000 to 150,000 years) that are restricted to African ancestry groups, yet are observed at frequencies up to 10%, but variants in this frequency range that are restricted to East Asian or South Asian ancestry groups are typically under 1,000 generations (20,000 to 30,000 years) or 1,300 generations old (26,000 to 39,000 years), respectively. Conversely, cosmopolitan variants that are shared among every ancestry group are typically older than 2,000 generations (40,000 to 60,000 years) despite being observed at global frequencies below 0.5% (Supplementary Figure S6). Variants restricted to American ancestry groups are typically younger than 750 generations (15,000 to 22,500 years), consistent with existing knowledge about

the settlement of the Americas via the Bering land bridge that connected Asia and North America during the last glacial maximum around 15,000 to 23,000 years ago [42, 43]. We note, however, that recent admixture and the sampling strategies of the different data sets [44, 45] can have a strong impact on age distributions. For example, variants at high frequency within American populations, but which are nevertheless restricted to just American and African populations, are on average younger than lower frequency variants (within American populations) with the same geographical restriction (Supplementary Figure S6). These variants likely arose recently within Africa and entered American populations through admixture, rising to high frequency through population bottlenecks [46].

Such heterogeneity in the relationship between allele age and frequency, coupled with heterogeneous and unknown sampling strategies, complicates the use of frequency as a means of assessing variants for potential pathogenicity during the interpretation of individual genomes. The Atlas of Variant Age potentially offers a more direct approach for screening variants, given the high probability of elimination of non-recessive deleterious variants within a few generations [47]. To assess the value of allele age in the interpretation of potentially pathogenic variants, we estimated the ages of variants that had effects predicted as *damaging* by PolyPhen-2 [48] or *deleterious* by SIFT [49] in TGP (Figure 4B). Of the ∼70,000 variants analyzed, 50% of damaging and 49% of deleterious variants were estimated to have arisen within the last 500 generations (10,000 to 15,000 years), compared to 41% of *benign* (PolyPhen-2) and 42% of *tolerated* (SIFT) variants (Supplementary Figure S7). Compared with control sets of variants (those annotated as benign or tolerated and matched for allele frequency within the focal ancestry group), variants annotated as damaging or deleterious had a notable dearth of older variants (>1,000 generations) for a given frequency, consistent with theoretical expectations and previous findings [19, 50, 51]. Our results suggest that old alleles can largely be excluded from consideration of pathology (though recent origin is not evidence in favor of pathogenicity).

Shared ancestry

Finally, we investigated the extent to which patterns of sharing of variants of different ages could power approaches for learning about genealogical history. Previous work has highlighted the descriptive value of genetic variants in identifying individuals who share recent common ancestry and patterns of demographic isolation and migration [10, 52, 53], though has also highlighted the challenges of interpreting the output of approaches such as PCA [54, 55]. Conversely, numerous model-based approaches have been developed that use patterns of variant and haplotype sharing to infer underlying demographic parameters [14, 56–58], though these typically make strong simplifying assumptions about the space of possible histories. Patterns of sharing of variants of different ages provide a non-parametric approach for combining descriptive and inferential approaches by learning about connections between individual genomes and groups of individuals over time. Specifically, for any two haploid genomes, we can estimate the fraction that has reached a common ancestor at a given point in time (the cumulative coalescent function, CCF) through the fraction of variants of that age or less that are shared. We use dynamic programming to estimate a maximum likelihood CCF between any pair or group of individuals (Figure 5A; see Supplementary Text). Using simulated

data, we found that inference of the ancestral relationship between individuals (CCFs inferred from estimated allele ages) correctly reflected differences of relatedness among individuals within and between different ancestry groups, and revealed patterns qualitatively consistent with past demographic events (see Supplementary Figure S8). Though we note that uncertainty in variant age estimates may cause over-smoothing of coalescent profiles, in particular in the distant past (>20,000 generations ago).

To illustrate the value of this non-parametric approach in describing the history of individuals and groups, we first considered the coalescent history between a single individual of American (Puerto Rican) ancestry from TGP (ID: HG00733) and all others in the TGP sample, using GEVA age estimates for variants on Chromosome 5 (Figure 5B). As a positive control, we included the parents of HG00733 (HG00732 and HG00731), who reach a CCF of near one in the most recent epoch (though note that the parents were used for haplotype phasing, which estimates transmitted haplotypes, hence the CCF reaching one, rather than the expected one-half). We show the ancestry of HG00733 shared with every individual in the TGP sample in Movie 1; also see https://human.genome.dating/ancestry/HG00733. Within the first 100 generations, we see additional coalescence with the untransmitted parental chromosomes and other individuals from the Puerto Rican sample. The earliest common ancestry outside Puerto Rico is seen with a European individual in Spain (paternal side; ∼60 generations ago) and a Mexican ancestry individual (maternal side; ∼80 generations ago). Coalescence with individuals sampled from outside the Americas occurs further back in time (>100 generations ago); initially with European individuals (dominantly in a period around 300 to 600 generations ago), then uniformly with non-African individuals around 1,500 to 4,000 generations ago, more strongly with African individuals around 5,000 to 15,000 generations ago, and uniformly with everyone around 20,000 generations ago. Because of the impact of data errors on age estimation of recent variants (highlighted above), the absolute timings of the early events are likely substantially overestimated. We expect the relative ordering of events to be robust and consistent among sample comparisons (see Supplementary Text).

The CCFs to all other members of a reference panel (averaged across all chromosomes in both haploid genomes) provide an overview of the genealogical relationships for a target individual. As an example, we inferred the CCF profiles of a Siberian Eskimo to all other individuals in SGDP (Figure 5C, top), showing common ancestry to other Central Asian and Siberian individuals within a few hundred generations, substantial common ancestry with American individuals before 1,000 generations, and typically more recent common ancestry with East Asians than West Eurasians or Africans; see https://human.genome.dating/ancestry/S_Eskimo_Sireniki-1. Notably, relatively little additional coalescence is seen during a period around 2,000 to 10,000 generations ago, which is a pattern shared among non-African individuals (also see Movie 1) and agrees with previous findings of a period of reduced coalescence, peaking 100,000 to 200,000 years ago [14]. The CCF can also be represented as a coalescent intensity function (CIF; see Supplementary Text), which measures the rate of change of common ancestry over time (Figure 5C, middle). The CIF reveals additional structure; for example, around 3,000 to 20,000 generations ago, those parts of the Siberian Eskimo's genome that have not yet coalesced with other genomes sampled from the same ancestry group have a very low CIF, while the CIF to the African-ancestry samples (which have had very little coalescence until this point) is relatively high (though note the absolute rate remains very

low over this period). We can further summarize the coalescent profiles for the target individual by computing the maximum CIF over the sample. This statistic captures properties analogous to the effective population size parameter, N_e, in population genetics modeling, where the expected coalescent intensity is inversely related to population size. We refer to the maximum CIF in the following as the N_e equivalent (Figure 5C, bottom), though we note that (in the case where there are genuine populations) it is likely to be downward biased compared to existing methods [13, 14]. We therefore use the N_e equivalent to provide a relative (rather than absolute) summary of shared and divergent genealogical histories across the entire cohort.

We estimated the pairwise CIFs for every genome in TGP (Movie 2), which we aggregated across autosomes to generate a coalescent profile between each pair of individuals (see Supplementary Text). Likewise, we estimated the ancestry shared between each individual pair in SGDP (Movie 3); CIFs were further aggregated among the 130 population groups (Figure 6). These reveal how the rates and structure of coalescence have changed over time, with the most recent epoch (up to ∼200 generations ago) dominated by coalescence within populations, but also identifying recent connections between groups, such as between southern Siberian and northern East Asian populations (Figure 6A). Several populations, such as the Kusunda (Nepal), Saami (Finland), and Negev Bedouins (Israel), show strong within-group coalescence up to this point, and by 500 generations ago are coalescing primarily with other populations. The epoch around 800 generations ago is dominated by structure broadly corresponding to the continental level (Figure 6B); though some African populations, for example the Mbuti (Congo), but notably the Khomani San (South Africa) and Ju'hoansi (Namibia), remain isolated up to ∼1,500 generations ago (30,000 to 45,000 years), which overlaps with previous findings [59], and we see these two populations to be strongly connected for an extended period further back in time (Movie 3). Around 800 generations ago, there is very little remaining structure among West Eurasian populations, but many additional inter-continental connections are now identified. For example, we see a north-to-south gradient of decreasing coalescence between American populations and Siberian or East Asian populations. In particular, we identify strong coalescence of American ancestry individuals with Siberian Eskimos, Aleutian Islanders, and Tlingit people in a period between 500 and 1,000 generations ago, and very little structure among American, Siberian, and East Asian populations as a whole further back than 1,000 generations ago (Movie 3), which agrees with previous results regarding the human migration into the Americas, extended isolation, and subsequent dispersal across the continent [43]. By 4,000 generations ago, we see high levels of coalescence between non-African and African populations (Figure 6C), and essentially no structure in the epoch around 20,000 generations ago (Figure 6D). We note that the exact timings of demographic events (signified by periods of intense or reduced coalescence), while showing consistent patterns within and among population groups, may carry additional noise due to uncertainty in allele age estimates.

The maximum CIF profiles (N_e equivalents; Figure 6E) highlight several features, including differences among modern ancestry groups in the intensity of coalescence within the last 1,000 generations (particularly intense for American and Oceanic populations); a major period of intense coalescence among all non-African ancestry individuals around 1,000 to 2,000 generations ago, following the migration of modern humans out of Africa [60]; a weaker, but still marked increase in coalescent intensity for African ancestry samples around 2,000 generations ago; and an older

reduction in coalescent intensity, peaking around 5,000 to 8,000 generations ago, potentially driven by ancient population structure within Africa and (for non-African populations) possible admixture with archaic lineages [61–63]. We find minor quantitative, but not qualitative differences among chromosomes (Supplementary Figure S9). We note that the bottleneck and steep change in population size around 60,000 to 80,000 generations ago, although consistent across ancestry profiles inferred for all individuals in the sample, is likely to be an artifact; possibly caused by increased variability of age estimates in the distant past, along with a breakdown of assumptions regarding the statistical independence of very old mutations.

Discussion

We have demonstrated how allele age estimates can provide insight to a range of problems in statistical and population genetics. However, there are several important assumptions and limitations of the approach. First, a key assumption is that of a single origin for each allele. Given the size of the human population and the mutation rate, it is likely that every allele has arisen multiple times over evolutionary history. Yet, unless the mutation rate is extremely high, it is still probable that most individuals carrying the allele do so through common ancestry. Moreover, multiple origins can potentially be identified through the presence of the allele on multiple haplotype backgrounds, as has, for example, been seen for the rs4988235 variant at *LCT* [34] (though we note that [64] conclude that the allele of variant rs4988235 was brought into African populations through historic gene flow, possibly through the Roman Empire), the O blood group [65], or alleles in the Human Leukocyte Antigene (*HLA*) region [66]. A variant lying in a region with high rates of non-crossover (gene conversion) may similarly be found on multiple haplotype backgrounds [67]. However, for genomes with very high mutation rates, such as HIV-1 [68], recurrence is sufficiently high to make estimates of allele age meaningless. In addition, while we have shown GEVA to be robust to realistic levels of sequencing and haplotype phasing error, the actual structures of error found in reference data sources, such as TGP [69], have additional complexity whose effect is unknown.

Our approach also assumes a known and time-invariant rate of recombination. For most species, only indirect estimates of the per-generation recombination rate are available and, in humans [70] and mice [71], there is evidence for evolution in the fine-scale location of recombination hotspots through changes in the binding preferences of *PRDM9*. However, because broad-scale recombination rates evolve at a much lower rate than hotspot location [72], and because our approach for detecting recombination events is driven largely by the presence of recombinant haplotypes, we expect GEVA to be relatively robust for recent variants. Older variants may be more affected, but for such variants most information comes from the mutation clock, which is likely to have been more stable over time.

The Atlas of Variant Age has multiple applications beyond statistical and population genetics. For example, recent variants provide a natural index when searching for related samples in population-scale data sets. The CCF (and sample-wide CIF) profiles leverage allele age information to bestow a time dimension upon individual genomes (and whole-genome sequencing data sets) from which we can distinguish recent from past effects, thus demonstrating the value of our method for exploratory data analysis. Moreover, as demonstrated here, it is possible to combine information from multiple,

potentially even distributed data sets, by estimating coalescent time distributions for pairs of concordant and discordant haplotypes in each data resource separately, or to update age estimates by the inclusion of additional samples. Future extensions to infer location of origin or the ancestral haplotype, integrating the growing wealth of genome data from ancient samples, will be an important step towards reconstructing the ancestral history of the entire human species.

References

1. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome biology* **17,** 1–19 (2016).
2. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
3. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Publishing Group* **17,** 333–351 (2016).
4. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
5. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29,** 308–311 (2001).
6. Berg, J. J. & Coop, G. A Population Genetic Signal of Polygenic Adaptation. *PLoS Genetics* **10,** e1004412–25 (2014).
7. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528,** 499–503 (2015).
8. Crawford, N. G. *et al.* Loci associated with skin pigmentation identified in African populations. *Science* **358,** eaan8433–16 (2017).
9. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genetics* **8,** e1002453 (2012).
10. Ralph, P. & Coop, G. The Geography of Recent Genetic Ancestry across Europe. *PLoS biology* **11,** e1001555–20 (2013).
11. Thompson, E. A. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* **194,** 301–326 (2013).
12. Speed, D. & Balding, D. J. Relatedness in the post-genomic era: is it still useful? *Nature Publishing Group* **16,** 1–12 (2014).
13. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–U84 (2011).
14. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature Publishing Group* **46,** 919–925 (2014).
15. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Publishing Group* **49,** 303–309 (2016).
16. Turchin, M. C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics* **44,** 1015–1019 (2012).
17. Robinson, M. R. *et al.* Population genetic differentiation of height and body mass index across Europe. *Nature Publishing Group* **47,** 1357–1362 (2015).
18. Oh, H. J., Choi, D., Goh, C. J. & Hahn, Y. Loss of gene function and evolution of human phenotypes. *BMB Reports* **48,** 373–379 (2015).
19. Mathieson, I. & McVean, G. Demography and the Age of Rare Variants. *PLoS Genetics* **10,** e1004528 (2014).
20. O'Connor, T. D. *et al.* Rare Variation Facilitates Inferences of Fine-Scale Population Structure in Humans. *Molecular Biology and Evolution* **32,** 653–660 (2014).
21. Schiffels, S. *et al.* Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nature communications* **7,** 10408 (2016).
22. Gao, F. & Keinan, A. Explosive genetic evidence for explosive human population growth. *Current opinion in genetics & development* **41,** 130–139 (2016).
23. McVean, G. A. T. & Cardin, N. J. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360,** 1387–1393 (2005).
24. Marjoram, P. & Wall, J. D. Fast "coalescent" simulation. *BMC Genetics* **7,** 16 (2006).

25. Palamara, P. F., Terhorst, J., Song, Y. S. & Price, A. L. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics* **50,** 1311–1317 (2018).

26. Ormond, L., Foll, M., Ewing, G. B., Pfeifer, S. P. & Jensen, J. D. Inferring the age of a fixed beneficial allele. *Molecular Ecology* **25,** 157–169 (2015).

27. Chen, H., Hey, J. & Slatkin, M. A hidden Markov model for investigating recent positive selection through haplotype structure. *Theoretical population biology* **99,** 18–30 (2015).

28. Nakagome, S. *et al.* Estimating the Ages of Selection Signals from Different Epochs in Human History. *Molecular Biology and Evolution* **33,** 657–669 (2016).

29. Smith, J., Coop, G., Stephens, M. & Novembre, J. Estimating Time to the Common Ancestor for a Beneficial Allele. *Molecular Biology and Evolution* **35,** 1003–1017 (2018).

30. Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nature Genetics* **30,** 233–237 (2002).

31. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology* **128,** 415–423 (2005).

32. Matsumura, S. & Forster, P. Generation time and effective population size in Polar Eskimos. *Proceedings of the Royal Society B: Biological Sciences* **275,** 1501–1508 (2008).

33. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538,** 201–206 (2016).

34. Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* **39,** 31–40 (2006).

35. Gerbault, P. *et al.* Evolution of lactase persistence: an example of human niche construction. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **366,** 863–877 (2011).

36. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics* **74,** 1111–1120 (2004).

37. Burger, J., Kirchner, M., Bramanti, B., Haak, W. & Thomas, M. G. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 3736–3741 (2007).

38. Kamberov, Y. G. *et al.* Modeling Recent Human Evolution in Mice by Expression of a Selected EDAR Variant. *Cell* **152,** 691–702 (2013).

39. Tan, J. *et al.* The adaptive variant EDARV370A is associated with straight hair in East Asians. *Human genetics* **132,** 1187–1191 (2013).

40. Unterländer, M. *et al.* Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. *Nature communications* **8,** 1–10 (0001).

41. Riazuddin, S. A. *et al.* Missense mutations in TCF8 cause late-onset Fuchs corneal dystrophy and interact with FCD4 on chromosome 9p. *American journal of human genetics* **86,** 45–53 (2010).

42. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488,** 370–374 (2012).

43. Raghavan, M. *et al.* Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349,** aab3884–aab3884 (2015).

44. Shringarpure, S. & Xing, E. P. Effects of sample selection bias on the accuracy of population structure and ancestry inference. *G3 (Bethesda, Md.)* **4,** 901–911 (2014).

45. Risso, D. *et al.* Estimating Sampling Selection Bias in Human Genetics: A Phenomenological Approach. *PloS one* **10,** e0140146–13 (2015).

46. Moreno-Estrada, A. *et al.* The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* **344,** 1280–1285 (2014).

47. Glémin, S. How are deleterious mutations purged? Drift versus nonrandom mating. *Evolution; international journal of organic evolution* **57,** 2678–2687 (2003).

48. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Publishing Group* **7,** 248–249 (2010).

49. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* **4,** 1073–1081 (2009).

50. Maruyama, T. The age of a rare mutant gene in a large population. *The American Journal of Human Genetics* **26,** 669–673 (1974).

51. Kiezun, A. *et al.* Deleterious Alleles in the Human Genome Are on Average Younger Than Neutral Alleles of the Same Frequency. *PLoS Genetics* **9,** e1003301–12 (2013).

52. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519,** 309–314 (2015).

53. Sun, M. *et al.* On the use of dense SNP marker data for the identification of distant relative pairs. *Theoretical population biology* **107,** 14–25 (2016).

54. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics* **40,** 646–649 (2008).

55. François, O. *et al.* Principal Component Analysis under Population Genetic Models of Range Expansion and Admixture. *Molecular Biology and Evolution* **27,** 1257–1268 (2010).

56. Harris, K. & Nielsen, R. Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genetics* **9** (2013).

57. Sheehan, S., Harris, K. & Song, Y. S. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. *Genetics* **194,** 647–662 (2013).

58. Steinrücken, M., Paul, J. S. & Song, Y. S. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theoretical population biology* **87,** 51–61 (2013).

59. Tishkoff, S. A. *et al.* History of Click-Speaking Populations of Africa Inferred from mtDNA and Y Chromosome Genetic Variation. *Molecular Biology and Evolution* **24,** 2180–2195 (2007).

60. Campbell, M. C. & Tishkoff, S. A. African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annual review of genomics and human genetics* **9,** 403–433 (2008).

61. Yang, M. A., Malaspinas, A. S., Durand, E. Y. & Slatkin, M. Ancient Structure in Africa Unlikely to Explain Neanderthal and Non-African Genetic Similarity. *Molecular Biology and Evolution* **29,** 2987–2995 (2012).

62. Sankararaman, S., Patterson, N., Li, H., Pääbo, S. & Reich, D. The Date of Interbreeding between Neandertals and Modern Humans. *PLoS Genetics* **8,** e1002947–9 (2012).

63. Vattathil, S. & Akey, J. M. Small Amounts of Archaic Admixture Provide Big Insights into Human History. *Cell* **163,** 281–284 (2015).

64. Ranciaro, A. *et al.* Genetic origins of lactase persistence and the spread of pastoralism in Africa. *American journal of human genetics* **94,** 496–510 (2014).

65. Blumenfeld, O. O. & Patnaik, S. K. Allelic genes of blood group antigens: A source of human mutations and cSNPs documented in the Blood Group Antigen Gene Mutation Database. *Human mutation* **23,** 8–16 (2003).

66. De Bakker, P. I. W. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genetics* **38,** 1166–1172 (2006).

67. McVean, G., Awadalla, P. & Fearnhead, P. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* (2002).

68. Cuevas, J. M., Geller, R., Garijo, R., López-Aldeguer, J. & Sanjuán, R. Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS biology* **13,** e1002251–19 (2015).

69. Belsare, S. *et al.* Evaluating the quality of the 1000 Genomes Project data. *bioRxiv,* 1–20 (2018).

70. Myers, S. *et al.* Drive against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. *Science* **327,** 876–879 (2010).

71. Smagulova, F. *et al.* Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* **472,** 375–378 (2011).

72. Auton, A. *et al.* A fine-scale chimpanzee genetic map from population sequencing. *Science* **336,** 193–198 (2012).

73. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).

74. Eberle, M. A. *et al.* A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research* **27,** 1–9 (2016).

Acknowledgements. We thank members of the McVean group for comments and discussion. We thank Robert Esnouf (Oxford Big Data Institute) for computational support. Funded by the Wellcome Trust (100956/Z/13/Z to GM, 099685/Z/12/Z to PKA) and the Li Ka Shing Foundation (to GM). Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Data sources. Estimation of allele age and shared ancestry was conducted on publicly available data sets; the 1000 Genomes Project (TGP) [2] and the Simons Genetic Diversity Project (SGDP) [33]. We used phased haplotype data of Chromosomes 1-22 from the final release TGP panel (Phase 3; GRCh37), available for 2,504 individuals from 26 populations worldwide (five continental population groups). Additional data was available from TGP for 31 related individuals which we included in our shared ancestry analysis. We used phased haplotype data of Chromosomes 1-22 from the publicly available SGDP panel (PS2; GRCh37), consisting of 278 individuals from 130 populations worldwide (seven continental population groups). Recombination rates were determined for each chromosome using the genetic maps available from the International HapMap Project (Phase 2; GRCh37) [73]. Genotype data from the Illumina Platinum Genomes Project [74] (GRCh37; Chromosomes 1-22) was used as a reference to measure genotype error in a matched subsample from TGP. We used information from the Ensembl data base (human assembly GRCh37; release 92 version 20180221) to determine the ancestral and derived allelic states for variants in both TGP and SGDP panels, as predicted through multi-species alignments in the Ensembl EPO pipeline.

Availability of results. The Atlas of Variant Age is publicly available as an online database: https://human.genome.dating. It contains the age estimation results for more than 45 million variants in the human genome, including the results of all pairwise analyses conducted at each focal site to estimate shared haplotype segments and infer TMRCA between haplotype pairs (more than 32 billion). Further, we performed shared ancestry analyses for every pair of individual chromosomes in TGP (including related individuals) and SGDP, which are also available at https://human.genome.dating. The interface of the web portal provides functionality to search, download, and visualize allele age estimation profiles and the inferred ancestry shared between individual genomes using conventional internet browsers.

Source code availability. The source code of the GEVA method is available online at https://github.com/pkalbers/geva. The source code of the CCF dynamic programming algorithm is available at https://github.com/pkalbers/ccf. We modified the original source code of MSMC2 to optimize the performance of the PSMC algorithm in our simulation analysis; available at https://github.com/pkalbers/msmc2.

Figures

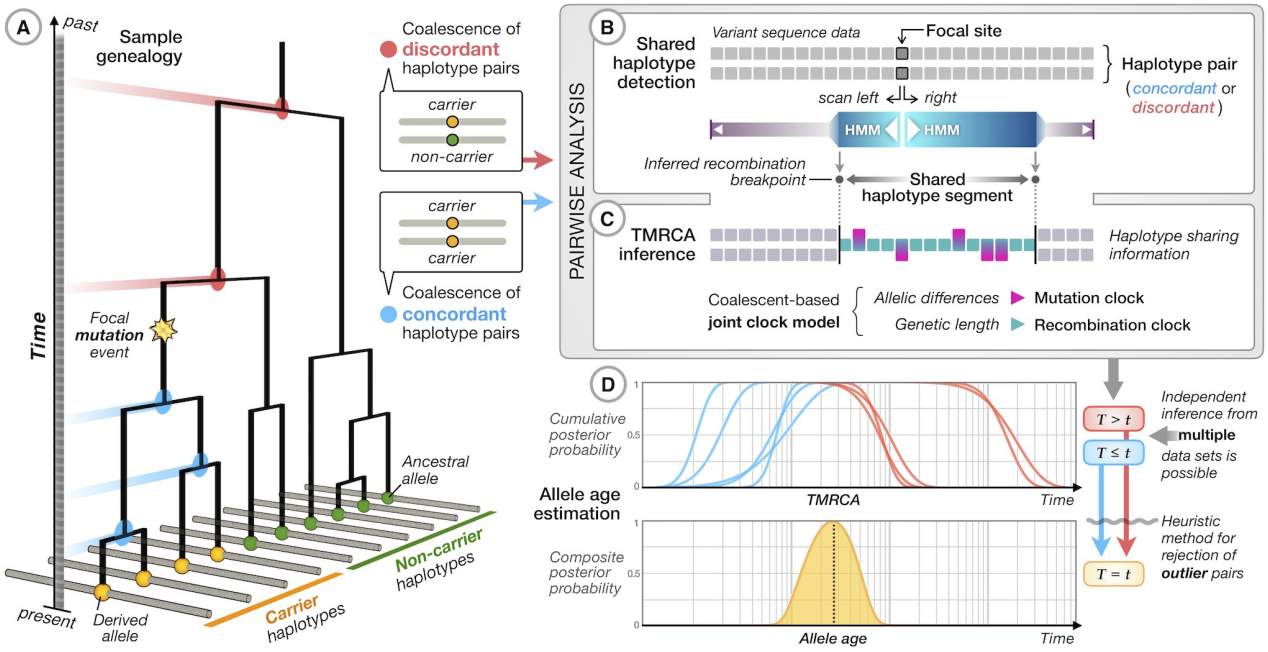

Figure 1. Overview of the genealogical estimation of variant age (GEVA) method. (A) At the chromosomal location of a variant, there exists an underlying (and unknown) genealogical tree describing the relationship between the samples. We assume that the derived allele (inferred by comparison to outgroup sequences) arose once in the tree. For concordant pairs of carrier chromosomes (*yellow terminal nodes*), their most recent common ancestors (MRCAs; *blue nodes*) occur more recently than the focal mutation event. For discordant pairs of chromosomes, between the ancestral allele (*green terminal nodes*) and the derived allele, the MRCAs (*red nodes*) are older than the focal mutation. (B) For each pair of chromosomes (concordant and discordant), we use a simple hidden Markov model (HMM) with an empirically calibrated error model to estimate the region over which the MRCA does not change; i.e. the distance to the first detectable recombination event either side of the focal position along the sequence. From the inferred ancestral segment, we obtain the genetic distance and the number of mutations that have occurred on the branches leading from the MRCA to the sample chromosomes. (C) For each pair of chromosomes, we use probabilistic models (see Supplementary Text) to estimate the posterior distribution of the time to the MRCA (TMRCA), represented as cumulative distributions of having coalesced for concordant pairs (*blue*) and of having not coalesced for discordant pairs (*red*). (D) An estimate of the composite posterior distribution for the time of origin of the mutation is obtained by combining the cumulative distributions for concordant and discordant pairs. Informally, the mutation is expected to be older than concordant and younger than discordant pairs. In practice, this composite likelihood-based approach results in approximate posteriors that are over-confident, hence they are summarized by the mode of the distribution. Additional filtering steps are carried out to remove inconsistent pairs of samples (see Supplementary Text).

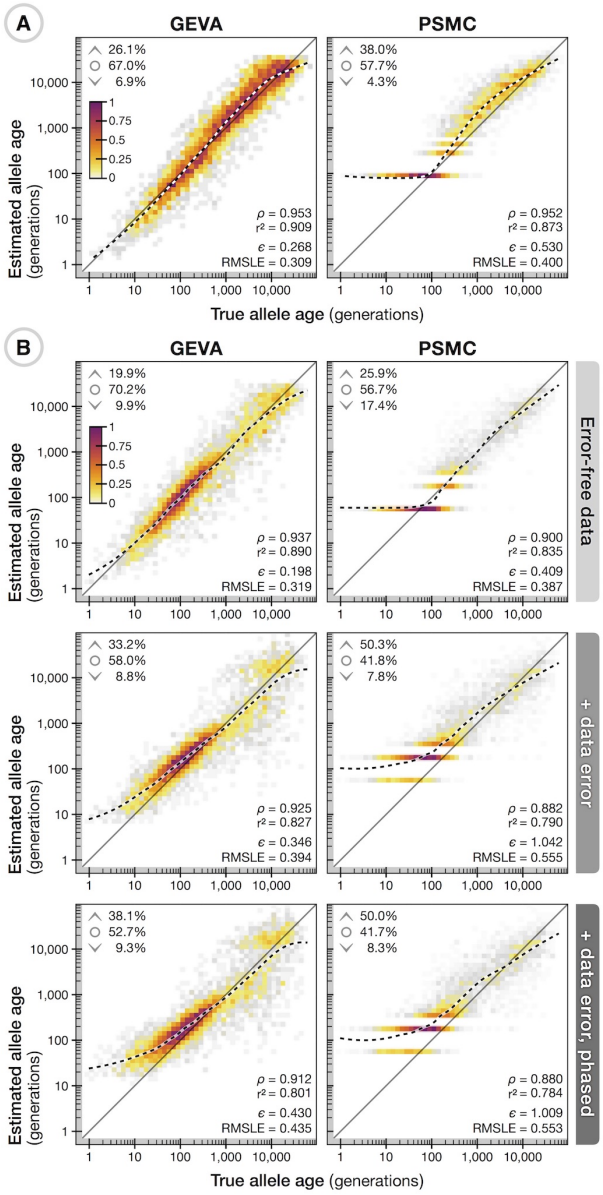

Figure 2. Validation of GEVA through coalescent simulations. (A) Density scatterplots showing the relationship between true allele age (geometric mean of lower and upper age of the branch on which a mutation occurred; *x-axis*) and estimated allele age (*y-axis*), using GEVA with the in-built HMM methodology (*left*) and PSMC (*right*) for the same set of 5,000 variants. Data were simulated under a neutral coalescent model with sample size N=1,000, effective population size N_e=10,000, and with constant and equal rates of mutation ($\mu = 1 \times 10^{-8}$) and recombination ($r = 1 \times 10^{-8}$) per site per generation. Variants were sampled uniformly from a 100 Mb chromosome, with allele count $1 < x < N$. Colors indicate relative density (scaled by the maximum per panel). Upper inserts indicate the fraction of sites where the point estimate (mode of the composite posterior distribution) of allele age lies above the upper age of the branch on which it occurred ($\wedge$), below the lower age ($\vee$), or within the age range of the branch ($\circ$). Lower inserts indicate the Spearman rank correlation statistic, ρ, squared Pearson correlation coefficient (on log-scale), r^2, interval-adjusted bias metric (see Supplementary Text), ϵ, and root mean squared $\log_{10}$ error, RMSLE. Also shown is a LOESS fit (2nd degree polynomials, neighborhood proportion $\alpha = 0.25$; *dashed line*). (B) The relationship between true and inferred ages for 5,000 variants sampled uniformly from a simulation of Chromosome 20 (63 Mb), simulated under a complex demographic model with N=1,000, N_e=7,300, $\mu = 2.35 \times 10^{-8}$, and variable recombination rates from human Chromosome 20. Allele age was estimated on haplotype data as simulated and without error (*top*), with error generated from empirical estimates of sequencing errors (*middle*), and with additional error arising from *in silico* haplotype phasing; see Supplementary Text. Allele age was estimated using scaling parameters as specified for each simulation. A further breakdown of results using mutation and recombination clocks alone, as well as the inferred pairwise TMRCAs, is shown in Supplementary Figures S1–S4.

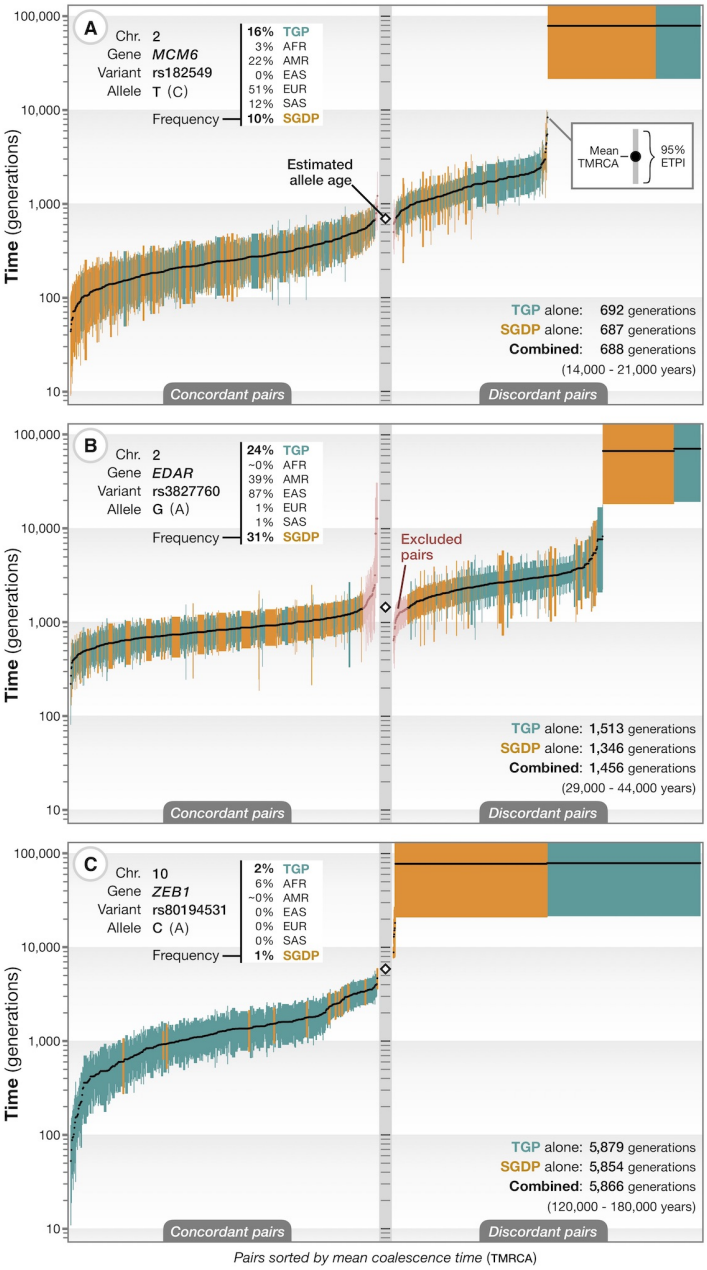

Figure 3. Application of GEVA to three variants of phenotypic and selective importance. (A) Estimated TMRCAs for concordant (*left*) and discordant (*right*) pairs of chromosomes for the derived T allele at rs182549, which lies within an intron of *MCM6* and affects regulation of *LCT* [30], which encodes lactase. Each bar reflects the approximate 95% credible interval (equal-tailed probability interval, ETPI) for a pair, ordered by posterior mean (*black dots*). Data from the 1000 Genomes Project (TGP; *green*) [2] and the Simons Genome Diversity Project (SGDP; *orange*) [33] were used. The frequency of the variant in SGDP, TGP, and the different population groups in TGP is shown (*top left*). The inferred allele age in generations from each data source and the combined estimate is shown (*bottom right*) and converted to an approximate age in years, assuming 20–30 years per generation. Age estimates are based on jointly modeling mutational and recombinational information to infer TMRCA from haplotype data. See https://human.genome.dating/snp/rs182549 for additional results. (B) As for part (A), for the derived G allele of rs3827760, which encodes the Val370Ala variant in *EDAR*, and which is associated with sweat, facial and body morphology [38, 39]; also see https://human.genome.dating/snp/rs3827760. Our filtering approach is to remove the smallest number of concordant and discordant pairs necessary (shown in *pink*) to obtain concordant and discordant sets with non-overlapping mean posterior TMRCAs. (C) As for part (A), for the derived C allele of rs80194531, which encodes the Asn78Thr substitution in *ZEB1*, reported as pathogenic for corneal dystrophy [41]; also see https://human.genome.dating/snp/rs80194531. Abbreviations refer to ancestry groups; AFR: African; AMR: American; EAS: East Asian; EUR: European; SAS: South Asian.

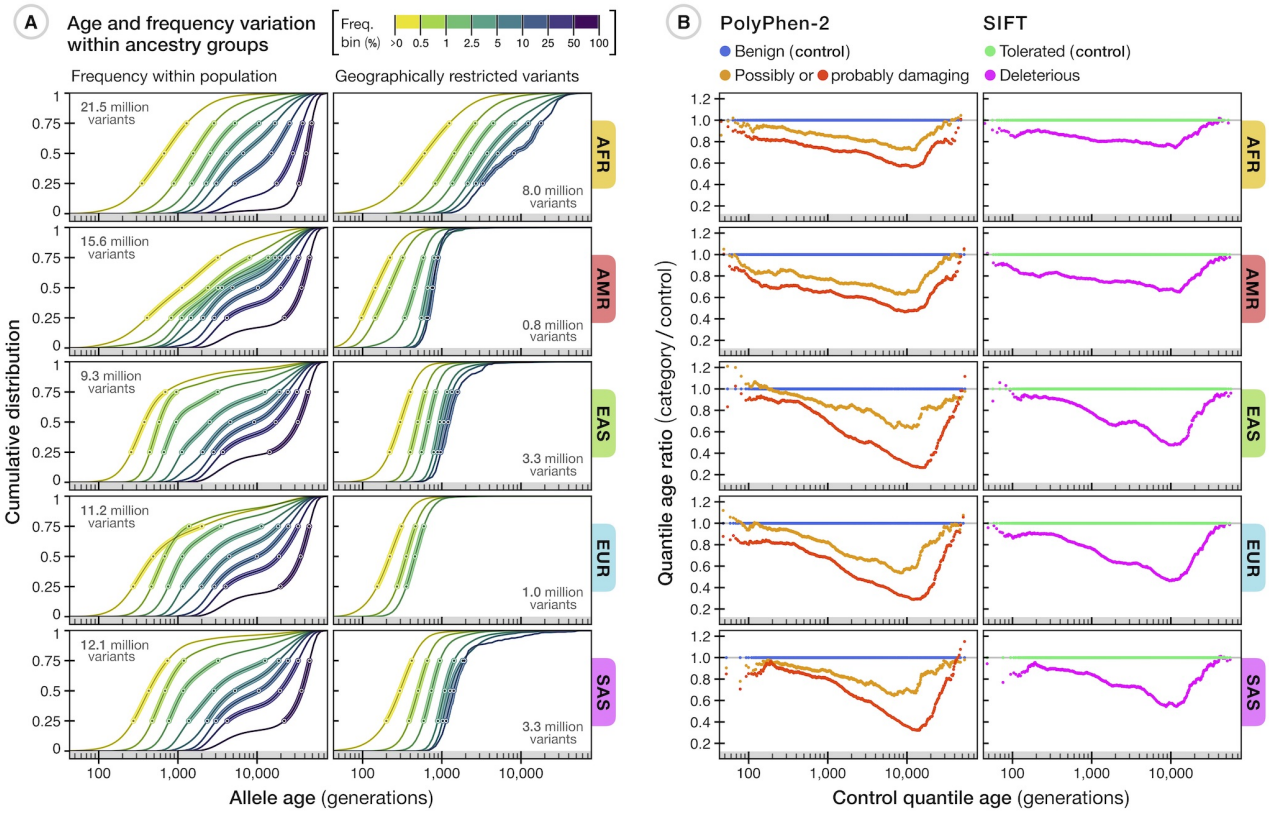

Figure 4. Age distribution of variants among different human populations. (A) The relationship between estimated allele age and frequency as observed within a given population group in the 1000 Genomes Project (TGP) sample. Of the 45.4 million variants available in the Atlas of Variant Age, 43.2 million were dated using TGP data alone; we excluded variants with low estimation quality and inconsistent ancestral allele information (see Supplementary Text), retaining 34.4 million variants. Each line shows the cumulative age distribution of variants within a given frequency bin (see legend) within a population group; circles indicate median and inter-quartile range. Panels on the left show the frequency-stratified cumulative distribution of estimated age for variants at non-zero frequencies as observed within a given ancestry group. The number of variants available per group is shown (*top left*). Panels on the right show the distributions of geographically restricted variants that only segregate within a group (number of available variants shown on *bottom right*). A summary of variants shared between different ancestry groups in TGP is provided in Supplementary Figure S6. (B) Differences in allele age distributions for ∼70,000 variants in TGP that are annotated as impacting protein function by PolyPhen-2 (*left*) and SIFT (*right*), compared to a reference set of variants (those annotated as *benign* by PolyPhen-2 or *tolerated* by SIFT), matched for allele frequency within a given ancestry group. These results are presented in more detail in Supplementary Figure S7. All age estimates are based on jointly modeling time from mutational and recombinational information.

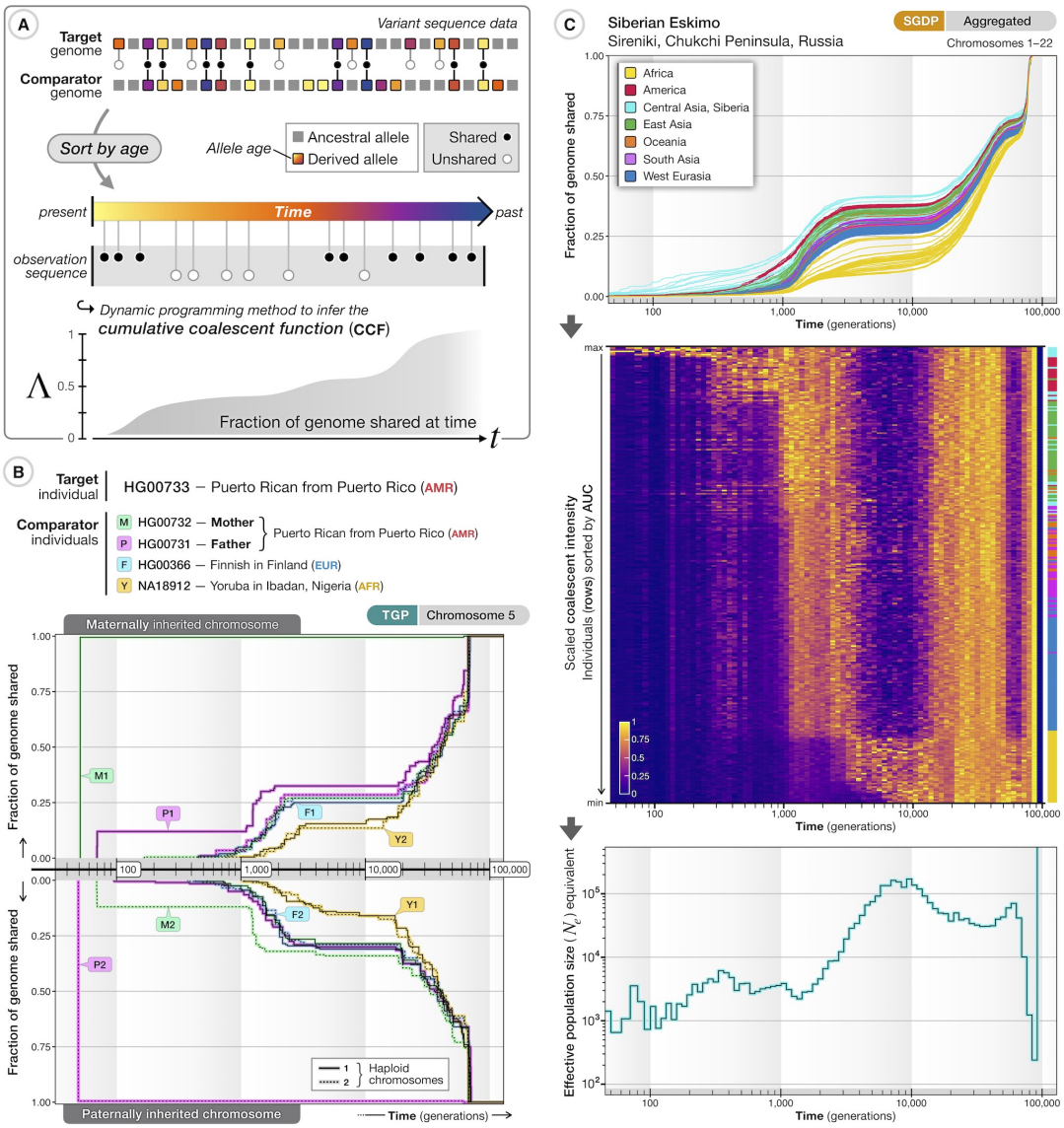

Figure 5. Age-stratified variant sharing to characterize ancestral relatedness. (A) Overview of approach for estimating the cumulative coalescent function (CCF) for a pair of haplotypes; the fraction of the genomes of the two samples that has coalesced by a given time. Derived variants within a target genome are identified, their estimated ages are obtained from the Atlas of Variant Age, and their presence (*black circles*) or absence (*white circles*) in another (comparator) genome of interest is recorded. Variants are sorted by allele age (indicated by color), t, to obtain a naive maximum likelihood estimate of the CCF, $\Lambda(t)$, using dynamic programming (assuming independence of variants and ignoring error in variant age estimates). (B) Selected pairwise CCFs for the two haploid genomes (Chromosome 5) of individual HG00733 (*top*: maternally derived; *bottom*: paternally derived) of a Puerto Rican individual from TGP, compared to the eight haplotypes from four individuals, including their mother and father. Maternal and paternal genomes were used for phasing, hence the inferred parental genomes are the transmitted (and untransmitted) genomes. The CCFs inferred with genomes from the entire TGP sample is shown in Movie 1. The full result data set for HG00733 is available at https://human.genome.dating/ancestry/HG00733. (C) Inferred genome-wide CCFs (averaged per diploid individual across autosomes) for a Siberian Eskimo from SGDP (ID: S_Eskimo_Sireniki-1) to all other sampled individuals (*top panel*). Colors indicate ancestry by geographic region (see legend). The CCF can also be expressed as a coalescent intensity function (CIF; *middle panel*) to reflect the increase in shared ancestry within a given time period. Each row represents an individual from SGDP, ordered by the area under the curve (AUC) of the CCF and scaled such that the maximum per column is equal to one. The color bar (*right*) indicates the ancestry group of sorted individuals. The CIF within any time epoch can be expressed as an effective population size (N_e equivalent) from the maximum over reference samples providing a summary of the rate at which common ancestor events occurred (*bottom panel*). The full result data set for this individual is available at https://human.genome.dating/ancestry/S_Eskimo_Sireniki-1.

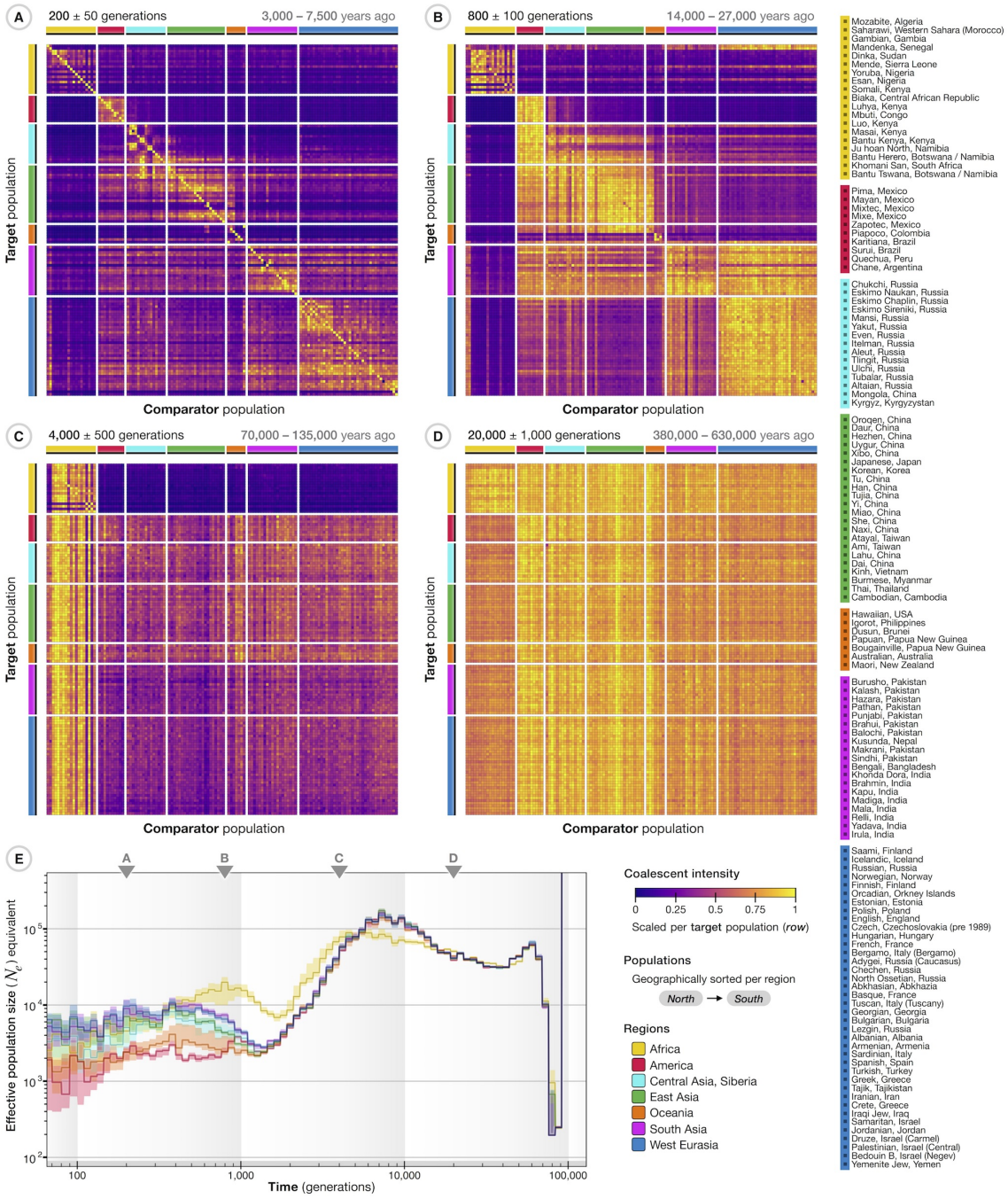

Figure 6. Age-stratified connections between ancestry groups in SGDP. The cumulative coalescent function (CCF) was inferred for all 556 haploid target genomes with all other comparator genomes in the SGDP sample and then aggregated by ancestry group (mean of CCFs from individuals within a population) and across chromosomes, with populations as defined in SGDP (see legend on the right). (A–D) The ancestry shared between populations is indicated by the coalescent intensity function (CIF) over a given time interval (epoch); shown as matrix with populations sorted from North to South within continental regions. Intensities were computed from aggregated CCFs to summarize relationships between populations; colors indicate intensity scaled per target population (rows) by the maximum over comparator populations. Ancestral connections are shown at different epochs back in time; around 200 generations ago (A), 800 generations (B), 4,000 generations (C), and 20,000 generations (D). The conversion (*top right*) assumes 20–30 years per generation. A more detailed summary, showing the ancestry shared between individuals, over a sliding time window (epoch) is shown in Movie 3. (E) The maximum CIF for individuals from different ancestry groups (continental regions) expressed as effective population size (N_e) equivalents over time; estimated from CCFs aggregated per diploid individual and summarized by the median and inter-quartile range per group. Triangles indicate the epochs shown in (A-D). A further breakdown of N_e equivalents estimated from non-aggregated CCFs per chromosome is shown in Supplementary Figure S9.

Movies

Movies are publicly available online at the links provided below. The runtime of each movie is 1 minute. For convenience, a web portal is provided in which all movies listed here can be watched: https://human.genome.dating/ancestry/movie

https://www.youtube.com/watch?v=ESMeiG7xpzU

Movie 1. Ancestry of individual HG00733. The change in coalescence rate (intensity) over time for the two haploid genomes (Chromosome 5) of target individual HG00733 (Puerto Rican in the American ancestry group) with all other genomes available through the 1000 Genomes Project (TGP) as comparator individuals; 5,070 haploid genomes in total, including related individuals that had been removed in the final release Phase 3 panel. Computations of the cumulative coalescent function (CCF) and coalescent intensities are based on 2.3 million variants dated (joint clock model) on Chromosome 5 from TGP, retained after excluding variants of low estimation quality and inconsistent ancestral allele information (see Supplementary Text). The movie shows the intensity scaled by the maximum over the TGP sample (*top two panels*), where chromosomes are distinguished by maternal and paternal modes of inheritance (determined from coalescent intensity inferred with both parental genomes). The scaling factor (maximum of coalescent intensity across the sample) is shown separately (*bottom left*). An estimate of the effective population size (N_e equivalent; *bottom right*) is obtained from the scaling factor and the number of generations within a given time interval (epoch). The two lines show the N_e equivalent inferred for the maternally and paternally inherited chromosomes of the target individual. Epochs are shown on a sliding window, moving back in time, with epoch size as indicated (*bottom*). The full result data set for HG00733 is available online at https://human.genome.dating/ancestry/HG00733.

https://www.youtube.com/watch?v=-uOqRBoE1CE

Movie 2. Ancestry shared back in time for all individuals from the 1000 Genomes Project (TGP). The sample consists of 2,535 individuals from 26 population groups, including related individuals that had been removed in the final release Phase 3 panel. We first computed the cumulative coalescent function (CCF) for each of the 5,070 haploid genomes in turn with every other genome, separately for Chromosomes 1-22. This was based on age estimates of 34.4 million variants dated using TGP data (joint clock model), retained after excluding variants of low estimation quality and inconsistent ancestral allele information; CCFs were then aggregated per diploid individual and across chromosomes (see Supplementary Text). We then calculated the coalescent intensity function (CIF) from the aggregated CCF of each pair of individuals. Ancestral relationships are shown in a matrix (*left*), where colors indicate the coalescent intensity between each pair of individuals; scaled per target individual (rows) by the maximum over the sample per time interval (epoch). Labels indicate the continental population as defined in TGP, which can be further subdivided into smaller population groups (see legend at end of movie). The movie shows ancestry backwards in time, where coalescent intensity is calculated in a sliding window with epoch size as indicated (*top right*). An estimate of the effective population size (N_e equivalent) is obtained per target individual from the maximum coalescent intensity across the sample and the time covered by a given epoch (*bottom right*); colors indicate an individual's ancestry group. Full result data sets for each individual are available online at https://human.genome.dating/ancestry/index#TGP.

https://www.youtube.com/watch?v=ffJRsbc9VLU

Movie 3. Ancestry shared back in time for all individuals from the Simons Genome Diversity Project (SGDP). The sample consists of 278 individuals from 130 populations in the publicly available SGDP sample. We estimated the coalescent intensity function (CIF) from the aggregated cumulative coalescent function (CCF) for each pair of individuals. CCFs were inferred separately per chromosome between each of the 556 haploid genomes as target in turn with all others as comparator genomes, based on 11.7 million variants dated under the joint clock model in SGDP (Chromosomes 1-22); retained after excluding variants of low estimation quality and inconsistent ancestral allele information. These were then aggregated per diploid individual across chromosomes (see Supplementary Text). As for Movie 2, ancestral relationships are shown in a matrix (*left*), moving back in time, where colors show the scaled coalescent intensity per epoch (as indicated on the time line; *top right*). Individuals are labelled according to their population group as defined in the SGDP sample (see legend at end of movie). Full result data sets for each individual are available online at https://human.genome.dating/ancestry/index#SGDP.