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 previous efforts have 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 that can be applied to large-scale genomic variation data sets. We demonstrate the accuracy and robustness of the approach through simulation and apply it to over 16 million single nucleotide polymorphisms (SNPs) from two publicly available human genomic diversity resources. We characterize the differential relationship between variant frequency and age in different geographical regions and demonstrate the value of allele age in interpreting variants of known functional and selective importance. Finally, we use allele age estimates to power a rapid approach for inferring the genealogical history of a single genome or a group of individuals.

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, 2]. 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 [3], which may have appeared thousands of generations ago. Genome sequencing studies [3–7] have catalogued the vast majority of common variation (estimated to be about 10 million variants [8]) and, at least within coding regions and particular ancestries, to date, more than 660 million variants genome-wide have been reported [9], many of them at extremely low frequency [5, 10–13].

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 [14–16]. Rather, attention has focused on the indirect use of genetic variation to detect population structure [17–23], identify related samples [24–26], and estimate parameters of models of human demographic history [27–30]. The evolutionary history of classes of variants contributing to polygenic adaptation (for example those affecting height [14, 31–33], though see [34]) or causing potential loss of gene function [35] has received attention, though rarely at the level of specific variants. Previous work on rare variants has identified ancestral connections between individuals and populations [3, 36–38] and demonstrated evidence for explosive population growth [10–13]. 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.

Method

We have developed an integrated framework for estimating the age of genetic variants; the point in time when the mutation arose that is ancestral to all chromosomes that carry the allele observed at a single locus in sample data (Figure 1A). Our approach, which we refer to as the genealogical estimation of variant age (GEVA), borrows from coalescent modeling [27, 30, 39, 40], but makes no assumptions about the demographic or selective processes that influence genealogical history, or about relatedness among sampled individuals. Instead, we learn about the age of a mutation from the distribution of the time to the most recent common ancestor (TMRCA) between pairs of chromosomes. 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. 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 estimate this ancestral segment in a targeted manner, using a simple 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 distribution 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 between pairs where one carries the mutation and the other carries the ancestral allele (discordant pairs). Information from hundreds or thousands of haplotype pairs is then combined within a composite likelihood framework to obtain an approximate posterior distribution on the time of the ancestral mutation (Figure 1D). One benefit of our method is that we can increase the number of pairwise TMRCA inferences incrementally to update allele age estimates, or to combine information across many data sets to improve the genealogical resolution from a wider distribution of independently sampled chromosomes. 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 [41] under different demographic models. Using a standard coalescent model with constant mutation and recombination rates, we found low bias (relative error, $\epsilon = 0.268$; see Supplementary Text) for allele age estimates and high correlation between true and inferred age (Spearman's $\rho = 0.953$; Figure 2A, Supplementary Figure 1A). We also compared our approach for estimating pairwise TMRCA posteriors to that obtained from the computationally more demanding pairwise sequentially Markovian coalescent (PSMC) methodology [27] (Supplementary Figure 1B). This estimates a demographic model (over a grid of time intervals) for each pair separately and can, for every position in the genome, return the inferred posterior distribution on the TMRCA. We found that PSMC-based estimations perform similarly well ($\rho = 0.952$), though the time discretization of PSMC increases bias ($\epsilon = 0.530$) and, in particular, leads 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. Under a complex demographic model that recapitulates the human expansion out of Africa [42] and with empirical and variable recombination rates, GEVA maintained a similarly high level of accuracy ($\epsilon = 0.198$, $\rho = 0.937$; Figure 2B; Supplementary Figure 2). In this situation, although the PSMC methodology is expected to model the demographic history for pairs of individuals better, the time discretization still leads to worse performance overall, with the addition of a substantial computational cost. We next introduced realistic data complications (see Supplementary Text), including genotype error calibrated using data from the 1000 Genomes Project [3] compared to data from the Illumina Platinum Genomes Project [43], as well as errors arising through *in silico* haplotype phasing (Figure 2B). We found that GEVA estimates remain largely unbiased and strongly correlated with true age after the inclusion of data error ($\epsilon = 0.346$, $\rho = 0.925$; Supplementary Figure 3) and after phasing ($\epsilon = 0.430$, $\rho = 0.921$; Supplementary Figure 4). We found that the PSMC-based approach continued to show higher bias and reduced correlation at the same set of variants, both after error ($\epsilon = 1.042$, $\rho = 0.882$) and after phasing ($\epsilon = 1.009$, $\rho = 0.880$). Reduced data quality resulting from sequencing or phasing errors leads to an underestimation of haplotype lengths for 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 considered three loci. First, we 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 [44]. We estimate the variant to be 696 generations old (Figure 3A, Supplementary Figure 5); approximately 14,000 to 21,000 years ago, depending on assumptions about generation time in humans [45, 46]. Our estimate is based on data from two different sources, the 1000 Genomes Project (TGP) [3] and the Simons Genetic Diversity Project (SGDP) [4], which, when estimated separately, give very similar ages (696 and 699 generations respectively). We obtained a similar age estimate of 693 generations for the derived A allele of rs4988235 (C/T-13910; Supplementary Figure 6), 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 [47]. Previous estimates of the age of these variants range between 2,200 and 21,000 years [48], putting our estimate 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) [49]. 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 [49]) 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 (>80%) in East Asian and American populations, and which is associated with sweat, facial and body morphology, and hair phenotypes [50–52]. We estimated the variant to be 1,462 generations old (approximately 30,000 to 45,000 years; Figure 3B, Supplementary Figure 7), again with strong concordance between TGP (1,513 generations) and SGDP (1,350 generations). Our estimate is consistent with previous estimates and limited evidence from ancient DNA studies [15, 53]. Our results further suggest that the variant rapidly rose in frequency following its origin through mutation, which is consistent with previous findings of strong positive selection of this variant in East Asia [54]. Of the 430,568 variants estimated to have arisen between 1,300 and 1,500 generations ago (within SGDP, see below), only 3,052 variants have reached a frequency higher than 30% in frequency globally, and only 423 variants higher than 80% in frequency within 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 [55], but is present at 6% in African ancestry samples within TGP. We estimated the age of the variant to be 5,892 generations old (110,000 to 180,000 years), again with consistency between TGP and SGDP (5,879 and 5,905 generations respectively; Figure 3C, Supplementary Figure 8). Such an ancient age seems inconsistent with the reported dominant pathogenic effect [55]. Moreover, of the 1,142,335 variants found at comparable frequencies (5–7%) in African ancestry individuals within SGDP, 54% were estimated to be younger, 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, by applying GEVA to more than 16 million variants identified in TGP or SGDP, referred to as the atlas of variant age, after confirming that estimates of allele age obtained from the two sources agreed (Spearman's $\rho = 0.871$; Figure 4, Supplementary Figure 9). We find substantial variation in the relationship between variant frequency and age depending on the population on which frequency is measured and the geographical distribution of the variant (Figure 5A). 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 ancestry group) have a median age of around 600 generations in African ancestry groups, 350 generations in East Asian ancestry groups, and 400 generations in European ancestry groups. The age distribution of variants restricted to a particular ancestry group (or, conversely, shared between them) indicates the degree of connection between populations (Supplementary Figure 10). For example, there are many variants up to 10,000 generations old (0.2–0.3 million years) that are restricted to African ancestry groups, yet are observed at frequencies up to 10%, but the oldest variants in this frequency range that are restricted to East Asian ancestry groups are typically under 1,000 generations (20,000–30,000 years) old. Variants restricted to American ancestry groups are typically under 750 generations old (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 [56–58]. We note, however, that recent admixture and the sampling strategies of the different data sets [59, 60] 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 considerably younger (on average), than lower frequency variants (within American populations) with the same geographical restriction (Supplementary Figure 10). These variants likely arose recently within Africa and entered American populations through admixture, rising to high frequency through population bottlenecks [61].

Such heterogeneity in relationship between frequency and age, 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 [62]. To assess the value of allele age in the interpretation of potentially pathogenic variants we estimated the ages of >70,000 variants in TGP annotated by the Ensembl Variant Effect Predictor [63], by Polyphen-2 [64] and SIFT [65], as *damaging* or *deleterious* (Figure 5B). Of the variants analyzed, 50% of damaging (PolyPhen-2) and 49% of deleterious variants (SIFT) are estimated to have arisen within the last 500 generations (10,000 to 15,000 years), compared to 41% of benign and 42% of tolerated variants (Supplementary Figure 11). Compared with control sets of variants (those annotated as *benign* by PolyPhen-2 and *tolerated* by SIFT and matched for allele frequency within the focal ancestry group), variants annotated as damaging or deleterious have a notable dearth of older variants (>1,000 generations) for a given frequency, consistent with theoretical expectations and previous results [36, 66, 67]. These results suggest that old alleles can largely be excluded from consideration of pathology (though recent origin is not evidence in favour of pathogenicity).

Ancestry sharing

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 with recent common ancestry and patterns of demographic isolation and migration [23, 68–72], though has also highlighted the challenges of interpreting the output of approaches such as PCA [73, 74]. Conversely, numerous model-based approaches have been developed that use patterns of variant and haplotype sharing to infer underlying demographic parameters [27–30, 75–79], 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 individuals and groups of people 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. More generally, we use dynamic programming to estimate a maximum likelihood CCF between any pair or group of individuals (Figure 6A; see Supplementary Text), though noting that uncertainty in variant age estimates and haplotyping error will tend to cause over-smoothing of coalescent profiles.

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 (Individual ID: HG00733) and all others in the TGP sample, using GEVA age estimates for variants on Chromosome 20 (Figure 6B, Supplementary Video 1). 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). 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 Colombian individual at around 90 generations ago (maternal side) and with a Peruvian individual at around 100 generations ago (paternal side). Coalescence with individuals sampled from outside the Americas occurs further back in time (>100 generations ago), initially with European individuals in a period around 300–600 generations ago, then uniformly with non-African individuals around 1,000–4,000 generations ago, and more strongly with African individuals around 6,000–10,000 generations ago. Because of the impact of data errors on rare variants discussed above, the absolute timings of the early events are likely substantially overestimated, though we expect the relative ordering of events to be robust.

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 6C), 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 than Africans. Notably, relatively little additional coalescence is seen during the period from c. 2,000 to c. 10,000 generations ago, which is a pattern shared among non-African individuals and agrees with previous findings of a period of reduced coalescence, peaking 100,000–200,000 years ago [30].

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 6C, middle panel), analogous (for a pair of individuals) to the effective population size, N_e, in population genetics modeling. The CIF reveals additional structure, for example around a 3,000 to 20,000 generation period; 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). Over time, the maximum CIF for the target individual across all others in the sample fluctuates between an N_e equivalent of one to two thousand until approximately 1,000 generations ago, before climbing to an N_e equivalent of 10^5 and then decreasing. Note that the N_e equivalent from the maximum CIF will tend to be lower than parametric estimates that assume exchangeability among individuals sampled from the same location.

More generally, patterns of allele sharing over time can be used across the entire cohort to summarize genealogical history. We estimated the pairwise CIFs for the 130 population groups defined in SGDP, after aggregating the CCFs across chromosomes and samples (see Supplementary Text), and show their ancestral relationships at different time periods (Figure 7; Supplementary Video 2). These reveal how the rates and structure of coalescence have changed over time, with the most recent epoch (around 200 generations) dominated by coalescence within each sampling group, but also identifying recent connections between groups, such as between southern Siberian and north-east Asians (Figure 7A; note that some populations, such as the Chaplin Eskimo, Balochi and Samaritans, show strong within-group coalescence prior to this point and by 800 generations are coalescing primarily with related populations). The epoch around 800 generations ago (Figure 7B) is dominated by structure broadly corresponding to the continental level, though some southern African populations (notably the Khomani San, Ju'hoan North, and the Mbuti) remain isolated up to around 1,500 generations ago (30,000–45,000 years), which overlaps with previous findings [80]. By this date, there is very little remaining structure among European 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 all American ancestry individuals with Siberian Eskimos, Aleutian Islanders, and Tlingit people in a period between 500 and 1,000 generations ago (Supplementary Video 2), and very little structure among American, Siberian, and East Asian populations as a whole prior to around 1,000 generations ago, which agrees with previous results regarding the human migration into the Americas, extended isolation, and subsequent dispersal across the continent [58]. By 4,000 generations ago, we see high levels of coalescence between non-African and African populations (Figure 7C), and essentially no structure in the epoch around 20,000 generations ago (Figure 7D). The Khomani San and Ju'hoan North remain largely isolated from other populations (apart from each other) until c. 5,000 generations ago.

The maximum CIF profiles, which provide a non-parametric equivalent to the effective population size (N_e) in population genetics modelling, (Figure 7E) 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 1,000-2,000 generations ago, following the migration of modern humans out of Africa [81, 82]; 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 [83–88]. We find minor quantitative, but not qualitative, differences among chromosomes (Supplementary Figure 12).

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. Nevertheless, unless the mutation rate is extremely high, it is still probable that most individuals with 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 allele at LCT [47, 89, 90] (though we note that [90] 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 [91], or alleles in the Human Leukocyte Antigene (*HLA*) region [92, 93]. A variant lying in a region with high rates of non-crossover (gene conversion) may similarly be found on multiple haplotype backgrounds [94]. However, for genomes with very high mutation rates, such as HIV-1 [95], 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 [96], 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 [97] and mice [98], 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 [99], 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.

An atlas of allele ages 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. 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] A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, W. S. W. Wong, G. Sigurdsson, G. B. Walters, S. Steinberg, H. Helgason, G. Thorleifsson, D. F. Gudbjartsson, A. Helgason, O. T. Magnusson, U. Thorsteinsdottir, and K. Stefansson, "Rate of de novo mutations and the importance of father's age to disease risk," *Nature*, vol. 488, pp. 471–475, Apr. 2013.

[2] R. Acuna-Hidalgo, J. A. Veltman, and A. Hoischen, "New insights into the generation and role of de novo mutations in health and disease," *Genome biology*, vol. 17, pp. 1–19, Nov. 2016.

[3] A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. G. Reid, Y. Zhu, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, N. Gupta, N. Gharani, L. H. Toji, N. P. Gerry, A. M. Resch, J. Barker, L. Clarke, L. Gil, S. E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, A. Thormann, I. Toneva, and B. Vaughan, "A global reference for human genetic variation," *Nature*, vol. 526, pp. 68–74, Sept. 2015.

[4] 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. Villems, 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, and D. Reich, "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations," *Nature*, vol. 538, pp. 201–206, Oct. 2016.

[5] A. Telenti, L. C. T. Pierce, W. H. Biggs, J. di Iulio, E. H. M. Wong, M. M. Fabani, E. F. Kirkness, A. Moustafa, N. Shah, C. Xie, S. C. Brewerton, N. Bulsara, C. Garner, G. Metzker, E. Sandoval, B. A. Perkins, F. J. Och, Y. Turpaz, and J. C. Venter, "Deep sequencing of 10,000 human genomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, pp. 11901–11906, Oct. 2016.

[6] D. F. Gudbjartsson, H. Helgason, S. A. Gudjonsson, F. Zink, A. Oddson, A. Gylfason, S. Besenbacher, G. Magnusson, B. V. Halldorsson, E. Hjartarson, G. T. Sigurdsson, S. N. Stacey, M. L. Frigge, H. Holm, J. Saemundsdottir, H. T. Helgadottir, H. Johannsdottir, G. Sigfusson, G. Thorgeirsson, J. T. Sverrisson, S. Gretarsdottir, G. B. Walters, T. Rafnar, B. Thjodleifsson, E. S. Bjornsson, S. Olafsson, H. Thorarinsdottir, T. Steingrimsdottir, T. S. Gudmundsdottir, A. Theodors, J. G. Jonasson, A. Sigurdsson, G. Bjornsdottir, J. J. Jonsson, O. Thorarensen, P. Ludvigsson, H. Gudbjartsson, G. I. Eyjolfsson, O. Sigurdardottir, I. Olafsson, D. O. Arnar, O. T. Magnusson, A. Kong, G. Masson, U. Thorsteinsdottir, A. Helgason, P. Sulem, and K. Stefansson, "Large-scale whole-genome sequencing of the Icelandic population," *Nature Publishing Group*, vol. 47, pp. 435–444, Mar. 2015.

[7] Genome of the Netherlands Consortium, "Whole-genome sequence variation, population structure and demographic history of the Dutch population.," *Nature Genetics*, vol. 46, pp. 818–825, Aug. 2014.

[8] 1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean, "A map of human genome variation from population-scale sequencing.," *Nature*, vol. 467, pp. 1061–1073, Oct. 2010.

[9] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation.," *Nucleic acids research*, vol. 29, pp. 308–311, Jan. 2001.

[10] A. Coventry, L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell, J. Crosby, J. E. Hixson, T. J. Rea, D. M. Muzny, L. R. Lewis, D. A. Wheeler, A. Sabo, C. Lusk, K. G. Weiss, H. Akbar, A. Cree, A. C. Hawes, I. Newsham, R. T. Varghese, D. Villasana, S. Gross, V. Joshi, J. Santibanez, M. Morgan, K. Chang, W. Hale IV, A. R. Templeton, E. Boerwinkle, R. Gibbs, and C. F. Sing, "Deep resequencing reveals excess rare recent variants consistent with explosive population growth," *Nature communications*, vol. 1, pp. 131–6, Nov. 2010.

[11] A. Keinan and A. G. Clark, "Recent explosive human population growth has resulted in an excess of rare genetic variants.," *Science*, vol. 336, pp. 740–743, May 2012.

[12] M. R. Nelson, D. Wegmann, M. G. Ehm, D. Kessner, P. S. Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zoellner, J. C. Whittaker, S. L. Chissoe, J. Novembre, and V. Mooser, "An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People," *Science*, vol. 337, no. 6090, pp. 100–104, 2012.

[13] J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, and J. M. Akey, "Evolution and functional impact of rare coding variation from deep sequencing of human exomes," *Science*, vol. 336, pp. 64–69, July 2012.

[14] J. J. Berg and G. Coop, "A Population Genetic Signal of Polygenic Adaptation," *PLoS Genetics*, vol. 10, pp. e1004412–25, Aug. 2014.

[15] I. 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. M. 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, and D. Reich, "Genome-wide patterns of selection in 230 ancient Eurasians," *Nature*, vol. 528, pp. 499–503, Dec. 2015.

[16] N. G. Crawford, D. E. Kelly, M. E. B. Hansen, M. H. Beltrame, S. Fan, S. L. Bowman, E. Jewett, A. Ranciaro, S. Thompson, Y. Lo, S. P. Pfeifer, J. D. Jensen, M. C. Campbell, W. Beggs, F. Hormozdiari, S. W. Mpoloka, G. G. Mokone, T. Nyambo, D. W. Meskel, G. Belay, J. Haut, NISC Comparative Sequencing Program, H. Rothschild, L. Zon, Y. Zhou, M. A. Kovacs, M. Xu, T. Zhang, K. Bishop, J. Sinclair, C. Rivas, E. Elliot, J. Choi, S. A. Li, B. Hicks, S. Burgess, C. Abnet, D. E. Watkins-Chow, E. Oceana, Y. S. Song, E. Eskin, K. M. Brown, M. S. Marks, S. K. Loftus, W. J. Pavan, M. Yeager, S. Chanock, and S. A. Tishkoff, "Loci associated with skin pigmentation identified in African populations," *Science*, vol. 358, pp. eaan8433–16, Nov. 2017.

[17] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, pp. 945–959, June 2000.

[18] D. Falush, M. Stephens, and J. K. Pritchard, "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.," *Genetics*, vol. 164, pp. 1567–1587, Aug. 2003.

[19] H. Gao, S. Williamson, and C. D. Bustamante, "A Markov Chain Monte Carlo Approach for Joint Inference of Population Structure and Inbreeding Rates From Multilocus Genotype Data," *Genetics*, vol. 176, pp. 1635–1651, May 2007.

[20] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante, "Genes mirror geography within Europe," *Nature*, vol. 456, pp. 98–101, Aug. 2008.

[21] K. Bryc, C. Velez, T. Karafet, A. Moreno-Estrada, A. Reynolds, A. Auton, M. Hammer, C. D. Bustamante, and H. Ostrer, "Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations.," *Proceedings of the National Academy of Sciences*, vol. 107 Suppl 2, pp. 8954–8961, May 2010.

[22] D. J. Lawson, G. Hellenthal, S. Myers, and D. Falush, "Inference of population structure using dense haplotype data.," *PLoS Genetics*, vol. 8, p. e1002453, Jan. 2012.

[23] P. Ralph and G. Coop, “The Geography of Recent Genetic Ancestry across Europe,” *PLoS biology*, vol. 11, pp. e1001555–20, May 2013.

[24] B. S. Weir, A. D. Anderson, and A. B. Hepler, “Genetic relatedness analysis: modern data and new challenges,” *Nature Reviews Genetics*, vol. 7, pp. 771–780, Oct. 2006.

[25] E. A. Thompson, “Identity by descent: variation in meiosis, across genomes, and in populations.,” *Genetics*, vol. 194, pp. 301–326, June 2013.

[26] D. Speed and D. J. Balding, “Relatedness in the post-genomic era: is it still useful?,” *Nature Publishing Group*, vol. 16, pp. 1–12, Nov. 2014.

[27] H. Li and R. Durbin, “Inference of human population history from individual whole-genome sequences,” *Nature*, vol. 475, no. 7357, pp. 493–U84, 2011.

[28] S. Sheehan, K. Harris, and Y. S. Song, “Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach,” *Genetics*, vol. 194, pp. 647–661, July 2013.

[29] M. Steinrücken, J. S. Paul, and Y. S. Song, “A sequentially Markov conditional sampling distribution for structured populations with migration and recombination,” *Theoretical population biology*, vol. 87, pp. 51–61, Aug. 2013.

[30] S. Schiffels and R. Durbin, “Inferring human population size and separation history from multiple genome sequences.,” *Nature Publishing Group*, vol. 46, pp. 919–925, Aug. 2014.

[31] J. K. Pritchard and G. Coop, “The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation,” *Current Biology*, vol. 20, pp. R208–R215, Feb. 2010.

[32] M. C. Turchin, C. W. Chiang, C. D. Palmer, S. Sankararaman, D. Reich, and J. N. Hirschhorn, “Evidence of widespread selection on standing variation in Europe at height-associated SNPs,” *Nature Genetics*, vol. 44, pp. 1015–1019, Aug. 2012.

[33] M. R. Robinson, G. Hemani, C. Medina-Gomez, M. Mezzavilla, T. Esko, K. Shakhbazov, J. E. Powell, A. Vinkhuyzen, S. I. Berndt, S. Gustafsson, A. E. Justice, B. Kahali, A. E. Locke, T. H. Pers, S. Vedantam, A. R. Wood, W. Van Rheenen, O. A. Andreassen, P. Gasparini, A. Metspalu, L. H. Van den Berg, J. H. Veldink, F. Rivadeneira, T. M. Werge, G. R. Abecasis, D. I. Boomsma, D. I. Chasman, E. J. C. de Geus, T. M. Frayling, J. N. Hirschhorn, J. J. Hottenga, E. Ingelsson, R. J. F. Loos, P. K. E. Magnusson, N. G. Martin, G. W. Montgomery, K. E. North, N. L. Pedersen, T. D. Spector, E. K. Speliotes, M. E. Goddard, J. Yang, and P. M. Visscher, “Population genetic differentiation of height and body mass index across Europe,” *Nature Publishing Group*, vol. 47, pp. 1357–1362, Sept. 2015.

[34] Y. Field, E. A. Boyle, N. Telis, Z. Gao, K. J. Gaulton, D. Golan, L. Yengo, G. Rocheleau, P. Froguel, M. I. McCarthy, and J. K. Pritchard, “Detection of human adaptation during the past 2000 years,” *Science*, vol. 354, no. 6313, pp. 760–764, 2016.

[35] H. J. Oh, D. Choi, C. J. Goh, and Y. Hahn, “Loss of gene function and evolution of human phenotypes,” *BMB Reports*, vol. 48, pp. 373–379, July 2015.

[36] I. Mathieson and G. McVean, “Demography and the Age of Rare Variants,” *PLoS Genetics*, vol. 10, p. e1004528, Aug. 2014.

[37] T. D. O’Connor, W. Fu, J. C. Mychaleckyj, B. Logsdon, P. Auer, C. S. Carlson, S. M. Leal, J. D. Smith, M. J. Rieder, M. J. Bamshad, D. A. Nickerson, and J. M. Akey, “Rare Variation Facilitates Inferences of Fine-Scale Population Structure in Humans,” *Molecular Biology and Evolution*, vol. 32, pp. 653–660, Nov. 2014.

[38] S. Schiffels, W. Haak, P. Paajanen, B. Llamas, E. Popescu, L. Loe, R. Clarke, A. Lyons, R. Mortimer, D. Sayer, C. Tyler-Smith, A. Cooper, and R. Durbin, “Iron Age and Anglo-Saxon genomes from East England reveal British migration history,” *Nature communications*, vol. 7, p. 10408, Jan. 2016.

[39] G. A. T. McVean and N. J. Cardin, “Approximating the coalescent with recombination,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, pp. 1387–1393, July 2005.

[40] P. Marjoram and J. D. Wall, “Fast "coalescent" simulation,” *BMC Genetics*, vol. 7, p. 16, Mar. 2006.

[41] J. Kelleher, A. M. Etheridge, and G. McVean, "Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes," *PLoS computational biology*, vol. 12, pp. e1004842–22, May 2016.

[42] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, "Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data," *PLoS Genetics*, vol. 5, pp. e1000695–11, Oct. 2009.

[43] M. A. Eberle, E. Fritzilas, P. Krusche, M. Kallberg, B. L. Moore, M. A. Bekritsky, Z. Iqbal, H.-Y. Chuang, S. J. Humphray, A. L. Halpern, S. Kruglyak, E. H. Margulies, G. McVean, and D. R. Bentley, "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*, vol. 27, pp. 1–9, Nov. 2016.

[44] N. S. Enattah, T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen, and I. Järvelä, "Identification of a variant associated with adult-type hypolactasia," *Nature Genetics*, vol. 30, pp. 233–237, Jan. 2002.

[45] J. N. Fenner, "Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies," *American Journal of Physical Anthropology*, vol. 128, pp. 415–423, Oct. 2005.

[46] S. Matsumura and P. Forster, "Generation time and effective population size in Polar Eskimos," *Proceedings of the Royal Society B: Biological Sciences*, vol. 275, pp. 1501–1508, July 2008.

[47] S. A. Tishkoff, F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, and P. Deloukas, "Convergent adaptation of human lactase persistence in Africa and Europe," *Nature Genetics*, vol. 39, pp. 31–40, Dec. 2006.

[48] P. Gerbault, A. Liebert, Y. Itan, A. Powell, M. Currat, J. Burger, D. M. Swallow, and M. G. Thomas, "Evolution of lactase persistence: an example of human niche construction.," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 366, pp. 863–877, Mar. 2011.

[49] T. Bersaglieri, P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn, "Genetic signatures of strong recent positive selection at the lactase gene.," *The American Journal of Human Genetics*, vol. 74, pp. 1111–1120, June 2004.

[50] A. Fujimoto, R. Kimura, J. Ohashi, K. Omi, R. Yuliwulandari, L. Batubara, M. S. Mustofa, U. Samakkarn, W. Settheetham-Ishida, T. Ishida, Y. Morishita, T. Furusawa, M. Nakazawa, R. Ohtsuka, and K. Tokunaga, "A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness," *Human molecular genetics*, vol. 17, pp. 835–843, Dec. 2007.

[51] M. L. Mikkola, "Molecular aspects of hypohidrotic ectodermal dysplasia," *American Journal of Medical Genetics Part A*, vol. 149A, pp. 2031–2036, Sept. 2009.

[52] J. Tan, Y. Yang, K. Tang, P. C. Sabeti, L. Jin, and S. Wang, "The adaptive variant EDARV370A is associated with straight hair in East Asians," *Human genetics*, vol. 132, pp. 1187–1191, June 2013.

[53] 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, and J. Burger, "Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe," *Nature communications*, vol. 8, pp. 1–10, 1.

[54] Y. G. Kamberov, S. Wang, J. Tan, P. Gerbault, A. Wark, L. Tan, Y. Yang, S. Li, K. Tang, H. Chen, A. Powell, Y. Itan, D. Fuller, J. Lohmueller, J. Mao, A. Schachar, M. Paymer, E. Hostetter, E. Byrne, M. Burnett, A. P. McMahon, M. G. Thomas, D. E. Lieberman, L. Jin, C. J. Tabin, B. A. Morgan, and P. C. Sabeti, "Modeling Recent Human Evolution in Mice by Expression of a Selected EDAR Variant," *Cell*, vol. 152, pp. 691–702, Feb. 2013.

[55] S. A. Riazuddin, N. A. Zaghloul, A. Al-Saif, L. Davey, B. H. Diplas, D. N. Meadows, A. O. Eghrari, M. A. Minear, Y.-J. Li, G. K. Klintworth, N. Afshari, S. G. Gregory, J. D. Gottsch, and N. Katsanis, "Missense mutations in TCF8 cause late-onset Fuchs corneal dystrophy and interact with FCD4 on chromosome 9p.," *American journal of human genetics*, vol. 86, pp. 45–53, Jan. 2010.

[56] D. H. O. Rourke and J. A. Raff, "The Human Genetic History of the Americas: Review The Final Frontier," *Current Biology*, vol. 20, pp. R202–R207, Feb. 2010.

[57] D. Reich, N. Patterson, D. Campbell, A. Tandon, S. Mazieres, N. Ray, M. V. Parra, W. Rojas, C. Duque, N. Mesa, L. F. García, O. Triana, S. Blair, A. Maestre, J. C. Dib, C. M. Bravi, G. Bailliet, D. Corach, T. Hünemeier, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, V. Acuña-Alonzo, C. Aguilar-Salinas, S. Canizales-Quinteros, T. Tusié-Luna, L. Riba, M. Rodríguez-Cruz, M. Lopez-Alarcón, R. Coral-Vazquez, T. Canto-Cetina, I. Silva-Zolezzi, J. C. Fernandez-Lopez, A. V. Contreras, G. Jimenez-Sanchez, M. J. Gómez-Vázquez, J. Molina, Á. Carracedo, A. Salas, C. Gallo, G. Poletti, D. B. Witonsky, G. Alkorta-Aranburu, R. I. Sukernik, L. Osipova, S. A. Fedorova, R. Vasquez, M. Villena, C. Moreau, R. Barrantes, D. Pauls, L. Excoffier, G. Bedoya, F. Rothhammer, J.-M. Dugoujon, G. Larrouy, W. Klitz, D. Labuda, J. Kidd, K. Kidd, A. Di Rienzo, N. B. Freimer, A. L. Price, and A. Ruiz-Linares, "Reconstructing Native American population history," *Nature*, vol. 488, pp. 370–374, Aug. 2012.

[58] M. Raghavan, M. Steinrucken, K. Harris, S. Schiffels, S. Rasmussen, M. DeGiorgio, A. Albrechtsen, C. Valdiosera, M. C. Avila-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. d. B. Damgaard, M. E. Allentoft, J. Lindo, E. Metspalu, R. Rodriguez-Varela, J. Mansilla, C. Henrickson, A. Seguin-Orlando, H. Malmstrom, T. Stafford, S. S. Shringarpure, A. Moreno-Estrada, M. Karmin, K. Tambets, A. Bergstrom, Y. Xue, V. Warmuth, A. D. Friend, J. Singarayer, P. Valdes, F. Balloux, I. Leboreiro, J. L. Vera, H. Rangel-Villalobos, D. Pettener, D. Luiselli, L. G. Davis, E. Heyer, C. P. E. Zollikofer, M. S. Ponce de Leon, 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, and E. Willerslev, "Genomic evidence for the Pleistocene and recent population history of Native Americans," *Science*, vol. 349, pp. aab3884–aab3884, Aug. 2015.

[59] S. Shringarpure and E. P. Xing, "Effects of sample selection bias on the accuracy of population structure and ancestry inference.," *G3 (Bethesda, Md.)*, vol. 4, pp. 901–911, Mar. 2014.

[60] D. Risso, L. Taglioli, S. De Iasio, P. Gueresi, G. Alfani, S. Nelli, P. Rossi, G. Paoli, and S. Tofanelli, "Estimating Sampling Selection Bias in Human Genetics: A Phenomenological Approach," *PloS one*, vol. 10, pp. e0140146–13, Oct. 2015.

[61] A. Moreno-Estrada, C. R. Gignoux, J. C. Fernandez-Lopez, F. Zakharia, M. Sikora, A. V. Contreras, V. Acuña-Alonzo, K. Sandoval, C. Eng, S. Romero-Hidalgo, P. Ortiz-Tello, V. Robles, E. E. Kenny, I. Nuño-Arana, R. Barquera-Lozano, G. Macín-Pérez, J. Granados-Arriola, S. Huntsman, J. M. Galanter, M. Via, J. G. Ford, R. Chapela, W. Rodriguez-Cintron, J. R. Rodríguez-Santana, I. Romieu, J. J. Sienra-Monge, B. del Rio Navarro, S. J. London, A. Ruiz-Linares, R. Garcia-Herrera, K. Estrada, A. Hidalgo-Miranda, G. Jimenez-Sanchez, A. Carnevale, X. Soberón, S. Canizales-Quinteros, H. Rangel-Villalobos, I. Silva-Zolezzi, E. G. Burchard, and C. D. Bustamante, "Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits.," *Science*, vol. 344, pp. 1280–1285, June 2014.

[62] S. Glémin, "How are deleterious mutations purged? Drift versus nonrandom mating.," *Evolution; international journal of organic evolution*, vol. 57, pp. 2678–2687, Dec. 2003.

[63] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham, "The Ensembl Variant Effect Predictor," *Genome biology*, vol. 17, pp. 1–14, June 2016.

[64] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature Publishing Group*, vol. 7, pp. 248–249, Apr. 2010.

[65] N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, "SIFT web server: predicting effects of amino acid substitutions on proteins," *Nucleic acids research*, vol. 40, pp. W452–W457, June 2012.

[66] T. Maruyama, "The age of a rare mutant gene in a large population.," *The American Journal of Human Genetics*, vol. 26, pp. 669–673, Nov. 1974.

[67] A. Kiezun, S. L. Pulit, L. C. Francioli, F. van Dijk, M. Swertz, D. I. Boomsma, C. M. van Duijn, P. E. Slagboom, G. J. B. van Ommen, C. Wijmenga, Genome of the Netherlands Consortium, P. I. W. de Bakker, and S. R. Sunyaev, "Deleterious Alleles in the Human Genome Are on Average Younger Than Neutral Alleles of the Same Frequency," *PLoS Genetics*, vol. 9, pp. e1003301–12, Feb. 2013.

[68] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, pp. 904–909, July 2006.

[69] J. Cussens, M. Bartlett, E. M. Jones, and N. A. Sheehan, "Maximum Likelihood Pedigree Reconstruction Using Integer Linear Programming," *Genetic Epidemiology*, vol. 37, pp. 69–83, Oct. 2012.

[70] N. A. Sheehan, M. Bartlett, and J. Cussens, "Improved maximum likelihood reconstruction of complex multi-generational pedigrees," *Theoretical population biology*, vol. 97, pp. 11–19, Nov. 2014.

[71] S. Leslie, B. Winney, G. Hellenthal, D. Davison, A. Boumertit, T. Day, K. Hutnik, E. C. Royrvik, B. Cunliffe, Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, D. J. Lawson, D. Falush, C. Freeman, M. Pirinen, S. Myers, M. Robinson, P. Donnelly, and W. Bodmer, "The fine-scale genetic structure of the British population.," *Nature*, vol. 519, pp. 309–314, Mar. 2015.

[72] M. Sun, M. A. Jobling, D. Taliun, P. P. Pramstaller, T. Egeland, and N. A. Sheehan, "On the use of dense SNP marker data for the identification of distant relative pairs," *Theoretical population biology*, vol. 107, pp. 14–25, Feb. 2016.

[73] J. Novembre and M. Stephens, "Interpreting principal component analyses of spatial population genetic variation," *Nature Genetics*, vol. 40, pp. 646–649, Apr. 2008.

[74] O. François, M. Currat, N. Ray, E. Han, L. Excoffier, and J. Novembre, "Principal Component Analysis under Population Genetic Models of Range Expansion and Admixture," *Molecular Biology and Evolution*, vol. 27, pp. 1257–1268, May 2010.

[75] A. L. Price, A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers, "Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations," *PLoS Genetics*, vol. 5, pp. e1000519–18, June 2009.

[76] J. S. Paul, M. Steinrücken, and Y. S. Song, "An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination.," *Genetics*, vol. 187, pp. 1115–1128, Apr. 2011.

[77] P. F. Palamara, T. Lencz, A. Darvasi, and I. Pe'er, "Length distributions of identity by descent reveal fine-scale demographic history.," *American journal of human genetics*, vol. 91, pp. 809–822, Nov. 2012.

[78] P. F. Palamara and I. Pe'er, "Inference of historical migration rates via haplotype sharing," *Bioinformatics*, vol. 29, pp. i180–i188, June 2013.

[79] K. Harris and R. Nielsen, "Inferring Demographic History from a Spectrum of Shared Haplotype Lengths," *PLoS Genetics*, vol. 9, June 2013.

[80] S. A. Tishkoff, M. K. Gonder, B. M. Henn, H. Mortensen, A. Knight, C. Gignoux, N. Fernandopulle, G. Lema, T. B. Nyambo, U. Ramakrishnan, F. A. Reed, and J. L. Mountain, "History of Click-Speaking Populations of Africa Inferred from mtDNA and Y Chromosome Genetic Variation," *Molecular Biology and Evolution*, vol. 24, pp. 2180–2195, July 2007.

[81] S. A. Tishkoff and B. C. Verrelli, "P ATTERNS OFH UMANG ENETICD IVERSITY: Implications for Human Evolutionary History and Disease," *Annual review of genomics and human genetics*, vol. 4, pp. 293–340, Sept. 2003.

[82] M. C. Campbell and S. A. Tishkoff, "African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping," *Annual review of genomics and human genetics*, vol. 9, pp. 403–433, Sept. 2008.

[83] M. A. Yang, A. S. Malaspinas, E. Y. Durand, and M. Slatkin, "Ancient Structure in Africa Unlikely to Explain Neanderthal and Non-African Genetic Similarity," *Molecular Biology and Evolution*, vol. 29, pp. 2987–2995, Sept. 2012.

[84] S. Sankararaman, N. Patterson, H. Li, S. Pääbo, and D. Reich, “The Date of Interbreeding between Neandertals and Modern Humans,” *PLoS Genetics*, vol. 8, pp. e1002947–9, Oct. 2012.

[85] 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, and S. Pääbo, “The complete genome sequence of a Neanderthal from the Altai Mountains,” *Nature*, vol. 505, pp. 1–12, Dec. 2013.

[86] S. Vattathil and J. M. Akey, “Small Amounts of Archaic Admixture Provide Big Insights into Human History,” *Cell*, vol. 163, pp. 281–284, Oct. 2015.

[87] 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, and S. Pääbo, “Genetic history of an archaic hominin group from Denisova Cave in Siberia,” *Nature*, vol. 468, pp. 1053–1060, Dec. 2010.

[88] 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, and M. Stoneking, “Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania.,” *American journal of human genetics*, vol. 89, pp. 516–528, Oct. 2011.

[89] N. S. Enattah, A. Trudeau, V. Pimenoff, L. Maiuri, S. Auricchio, L. Greco, M. Rossi, M. Lentze, J. K. Seo, S. Rahgozar, I. Khalil, M. Alifrangis, S. Natah, L. Groop, N. Shaat, A. Kozlov, G. Verschubskaya, D. Comas, K. Bulayeva, S. Q. Mehdi, J. D. Terwilliger, T. Sahi, E. Savilahti, M. Perola, A. Sajantila, I. Järvelä, and L. Peltonen, “Evidence of Still-Ongoing Convergence Evolution of the Lactase Persistence T-13910 Alleles in Humans,” *The American Journal of Human Genetics*, vol. 81, pp. 615–625, Sept. 2007.

[90] A. Ranciaro, M. C. Campbell, J. B. Hirbo, W.-Y. Ko, A. Froment, P. Anagnostou, M. J. Kotze, M. Ibrahim, T. Nyambo, S. A. Omar, and S. A. Tishkoff, “Genetic origins of lactase persistence and the spread of pastoralism in Africa.,” *American journal of human genetics*, vol. 94, pp. 496–510, Apr. 2014.

[91] O. O. Blumenfeld and S. K. Patnaik, “Allelic genes of blood group antigens: A source of human mutations and cSNPs documented in the Blood Group Antigen Gene Mutation Database,” *Human mutation*, vol. 23, no. 1, pp. 8–16, 2003.

[92] P. I. W. de Bakker, G. McVean, P. C. Sabeti, M. M. Miretti, T. Green, J. Marchini, X. Ke, A. J. Monsuur, P. Whittaker, M. Delgado, J. Morrison, A. Richardson, E. C. Walsh, X. Gao, L. Galver, J. Hart, D. A. Hafler, M. Pericak-Vance, J. A. Todd, M. J. Daly, J. Trowsdale, C. Wijmenga, T. J. Vyse, S. Beck, S. S. Murray, M. Carrington, S. Gregory, P. Deloukas, and J. D. Rioux, “A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC,” *Nature Genetics*, vol. 38, pp. 1166–1172, Sept. 2006.

[93] S. Leslie, P. Donnelly, and G. McVean, “A Statistical Method for Predicting Classical HLA Alleles from SNP Data,” *The American Journal of Human Genetics*, vol. 82, pp. 48–56, Jan. 2008.

[94] G. McVean, P. Awadalla, and P. Fearnhead, “A coalescent-based method for detecting and estimating recombination from gene sequences,” *Genetics*, 2002.

[95] J. M. Cuevas, R. Geller, R. Garijo, J. López-Aldeguer, and R. Sanjuán, “Extremely High Mutation Rate of HIV-1 In Vivo,” *PLoS biology*, vol. 13, pp. e1002251–19, Sept. 2015.

[96] S. Belsare, M. Sakin-Levy, Y. Mostovoy, S. Durinck, S. Chaudhry, M. Xiao, A. S. Peterson, P.-Y. Kwok, S. Seshagiri, and J. D. Wall, “Evaluating the quality of the 1000 Genomes Project data,” *bioRxiv*, pp. 1–20, Aug. 2018.

[97] S. Myers, R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman, T. S. MacFie, G. McVean, and P. Donnelly, “Drive against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination,” *Science*, vol. 327, pp. 876–879, Feb. 2010.

[98] F. Smagulova, I. V. Gregoretti, K. Brick, P. Khil, R. D. Camerini-Otero, and G. V. Petukhova, "Genome-wide analysis reveals novel molecular features of mouse recombination hotspots.," *Nature*, vol. 472, pp. 375–378, Apr. 2011.

[99] A. Auton, A. Fledel-Alon, S. Pfeifer, O. Venn, L. Ségurel, T. Street, E. M. Leffler, R. Bowden, I. Aneas, J. Broxholme, P. Humburg, Z. Iqbal, G. Lunter, J. Maller, R. D. Hernandez, C. Melton, A. Venkat, M. A. Nobrega, R. Bontrop, S. Myers, P. Donnelly, M. Przeworski, and G. McVean, "A fine-scale chimpanzee genetic map from population sequencing.," *Science*, vol. 336, pp. 193–198, Apr. 2012.

[100] K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, 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, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, 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. Waye, 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, 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, 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, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, I. Yakub, R. C. Onofrio, B. W. Birren, 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, and J. Stewart, "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, pp. 851–861, Oct. 2007.

[101] O. Delaneau, J.-F. Zagury, and J. Marchini, "Improved whole-chromosome phasing for disease and population genetic studies.," *Nature methods*, vol. 10, pp. 5–6, Jan. 2013.

Acknowledgements

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). We thank members of the McVean group for comments and discussion.

Data sources and code availability

Estimation of allele age and shared ancestry was conducted on publicly available data sets; the 1000 Genomes Project (TGP) [3] and the Simons Genetic Diversity Project (SGDP) [4]. 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) [100]. Genotype data from the Illumina Platinum Genomes Project [43] (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.

Data availability

Atlas of variant age for the human genome: (temporary link)
`https://www.dropbox.com/sh/hkrrj7sopmvkjrx/AAAQFBwdhBUTR0xUvm8-72Lka?dl=0`
Shared ancestry in TGP and SGDP: (temporary link)
`https://www.dropbox.com/sh/h60yjoznqgvhre3/AAA56rAj0wZPGj9T0Ui-8l06a?dl=0`

Source code availability

GEVA: `https://github.com/pkalbers/geva`
CCF: `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: `https://github.com/pkalbers/msmc2`

Figures

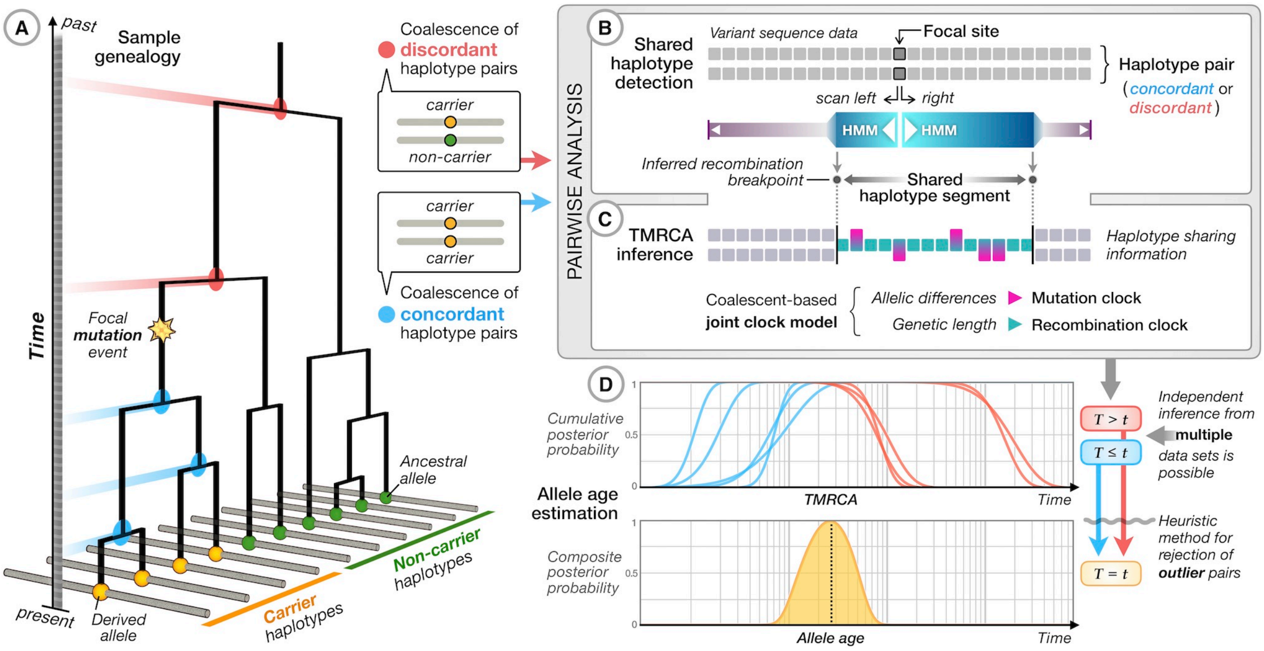

Figure 1. Overview of the genealogical estimation of variant age (GEVA) method. (A) At the chromosomal location of a variant there exists an underlying (and unknown) genealogical tree describing the relationship between the samples. We assume that the derived allele (inferred by comparison to outgroup 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 time to the MRCA does not change; i.e. the distance to the first detectable recombination event either side of the focal mutation. We obtain the genetic distance of the ancestral segment and the number of mutations that have occurred within this interval on the branches leading from the MRCA to the sample chromosomes. (C) For each pair of chromosomes, we use a probabilistic model (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 approximate 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, the composite likelihood approach results in approximate posteriors that are over-confident, hence they are summarized by the mode of the composite posterior distribution. Additional filtering steps are carried out to remove inconsistent pairs of samples (see Supplementary Text).

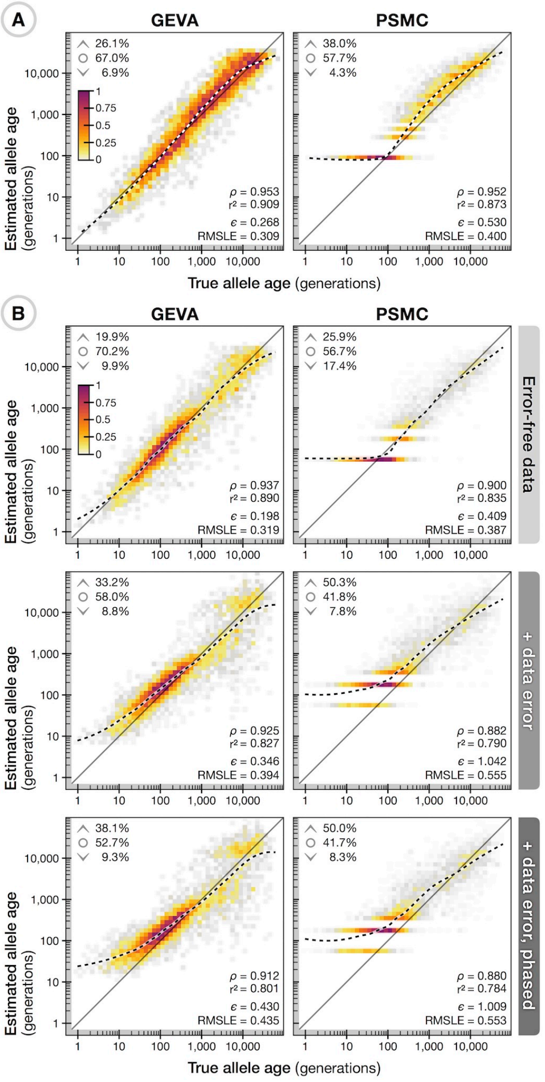

Figure 2. Validation of GEVA through coalescent simulations. (A) Density scatterplots showing the relationship between the true allele age (geometric mean of lower and upper age of the branch on which a mutation occurred; x-axis) and estimated allele age (y-axis), using GEVA with the in-built HMM methodology (left) and PSMC (right) for the same set of 5,000 variants. Data were simulated under a neutral coalescent model with a 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 $1 <$ allele count $< N$. 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$_1$ 0 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 [42], with N =1,000, N_e =7,300, $\mu = 2.35 \times 10^{-8}$, and variable recombination rates from HapMap (Phase 2; GRCh37; Chromosome 20) [100]. Allele age was estimated on haplotype data as simulated and without error (top), with error generated from empirical estimates of sequencing errors in 1000 Genomes Project data (see Supplementary Text; middle), and with additional error arising from *in silico* haplotype phasing using `SHAPEIT2` [101]. 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 1–4.

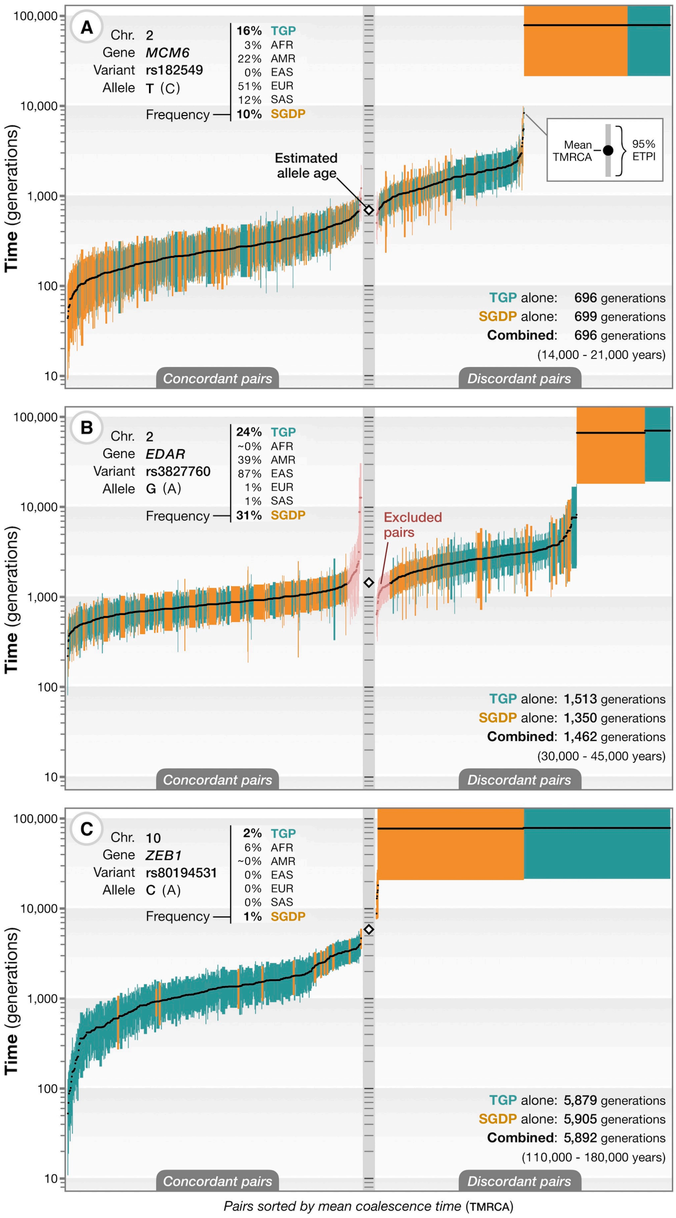

Figure 3. Application of GEVA to three variants of phenotypic and selective importance. (A) Estimated TMRCAs for concordant (left) and discordant (pairs) of chromosomes for the derived T allele at rs182549, which lies within an intron of *MCM6* and affects regulation of *LCT* [44], which encodes lactase. Each bar reflects the approximate 95% credible interval (equal-tailed probability interval, ETPI) for a pair, ordered by posterior mean (black dots). Data from two sources, the 1000 Genomes Project (TGP; blue) [3] and the Simons Genome Diversity Project (SGDP; orange) [4], were used. The top insert summarizes the frequency of the variant in SGDP, TGP, and the different population groups in TGP. The bottom insert summarizes the inferred allele age in generations from each data source and the combined estimate. The combined estimate is converted to age in years for a generation time range of 20–30 years. (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. 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 [55]. Abbreviations; AFR: African ancestry; AMR: American ancestry; EAS: East Asian ancestry; EUR: European ancestry; SAS: South Asian ancestry. A further breakdown of results using mutation and recombination clocks alone is shown in Supplementary Figures 5, 7, and 8.

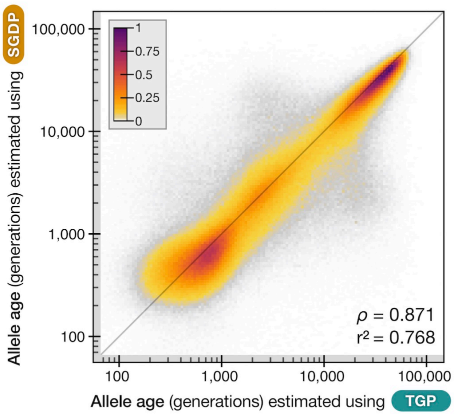

Figure 4. The relationship between estimates of allele age from TGP (x-axis) and SGDP (y-axis) for one million variants sampled from all autosomes and matched between TGP and SGDP. Colors indicate the relative density (see legend). Lower inserts indicate the Spearman rank correlation statistic, ρ, and the square of the Pearson correlation coefficient (calculated on log-scaled allele ages), r^2. A further breakdown of results using mutation and recombination clocks alone is shown in Supplementary Figure 9.

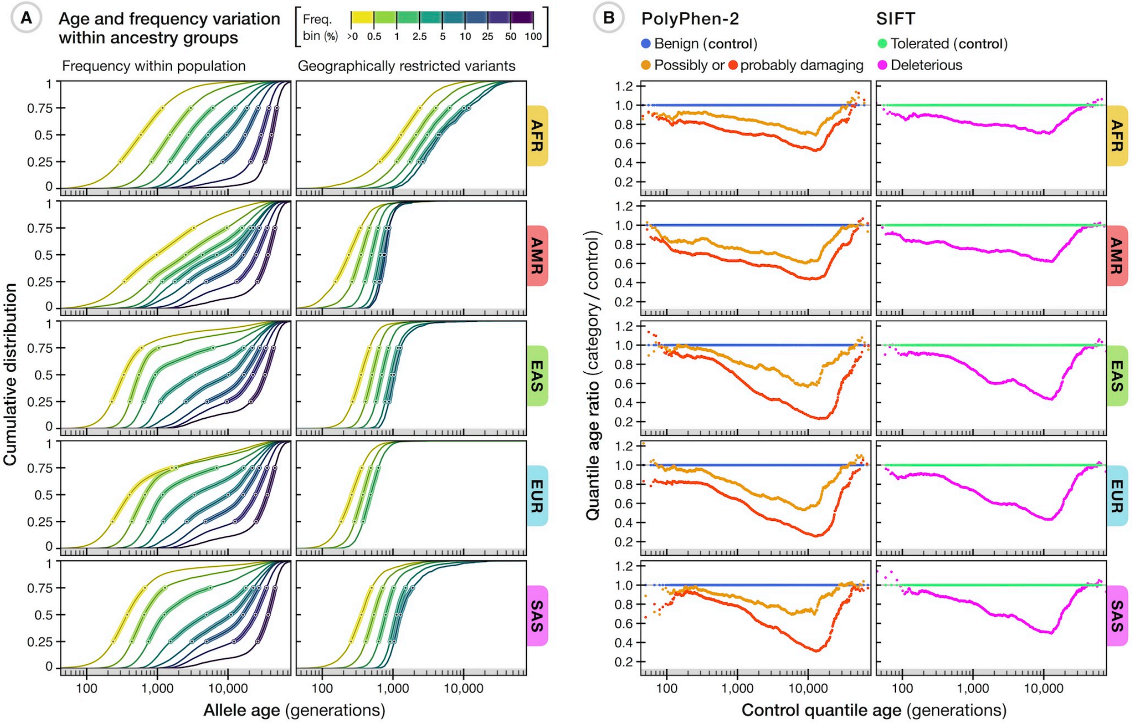

Figure 5. Age distribution of variants among different human populations. (A) The relationship between estimated allele age and allele frequency as observed within a given population group; for all dated variants on Chromosome 20 (>600,000 variants; left) and geographically restricted variants that only segregate within a population group (20,000 variants uniformly sampled from all autosomes per population group; right). Each line shows the cumulative age distribution of variants within a given frequency bin (see legend) within that population group; circles indicate median and inter-quartile range. A more detailed breakdown of variants restricted to multiple population groups is shown Supplementary Figure 10. (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 population group. These results are presented in more detail in Supplementary Figure 11.

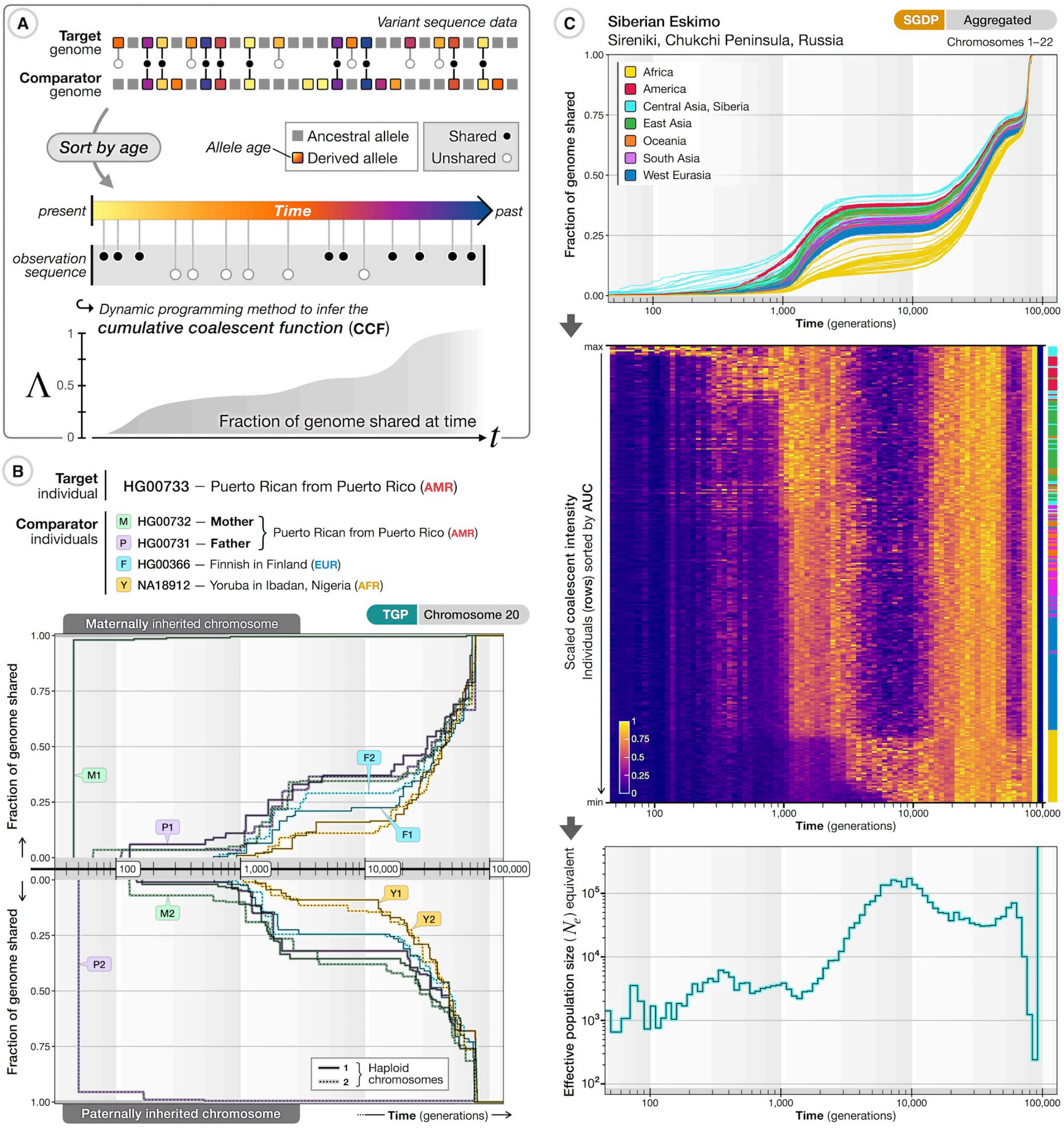

Figure 6. Age-stratified variant sharing to characterize ancestral relatedness. (A) Overview of approach for estimating the cumulative coalescent function (CCF) for a pair of haplotypes; the fraction of the genomes of the two samples that has coalesced by a given time. Derived variants within a target genome are identified, their estimated ages are obtained from the genome-wide atlas of allele 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_i, and a naive maximum likelihood estimate of the cumulative coalescent function, $\Lambda(t_i)$, is obtained using dynamic programming (assuming independence of variants and ignoring error in variant age estimates). (B) Selected pairwise CCFs for the two haploid Chromosome 20 genomes (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, a Finnish individual, and a Yoruba individual from Nigeria. Note that the maternal and paternal genomes were used for phasing, hence the inferred parental genomes are the transmitted (and untransmitted) genomes. (C) Inferred genome-wide CCFs (averaged over pairs of haplotypes and 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 represented as a coalescent intensity function (CIF; middle panel), which reflects the increase in shared ancestry that occurs within a given time period (see Supplementary Text). Each row represents an individual from SGDP, ordered by the area under the curve (AUC) of the CCF (ancestry of individuals indicated by color bar on the right) and scaled such that the maximum per column is one for visualization. The CIF within any time epoch can be expressed as an effective population size, N_e equivalent, with the maximum over reference samples providing a summary of the rate at which common ancestor events occurred (bottom panel).

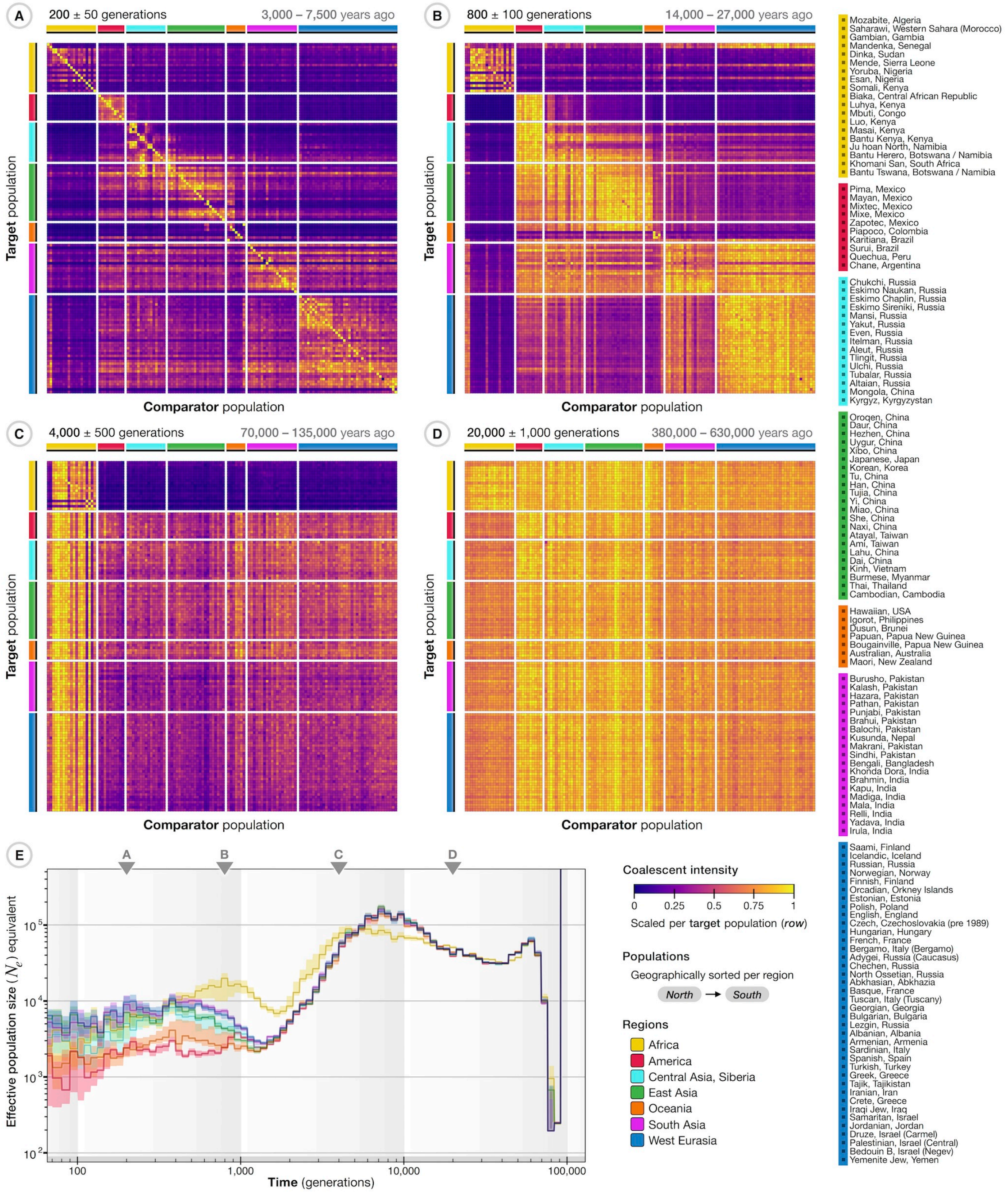

Figure 7. Age-stratified connections between ancestry groups in SGDP. The cumulative coalescent function (CCF) was inferred for all 556 haploid target genomes with all other comparator genomes in the SGDP sample and then aggregated by ancestry group (mean of CCFs from individuals within a population) 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 over a given time interval (epoch); shown as matrix with populations sorted from North to South within each continental region. Intensities were computed from the CCFs aggregated between a target and each comparator population; 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 detailed result of the ancestry shared between individuals over consecutive time intervals is shown in Supplementary Video 2. (E) The maximum coalescent intensity 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 12.