Title

Methods for Assessing Population Relationships and History Using Genomic Data.

Authors

Hellenthal, Garrett

Moorjani, Priya

Publication Date

2023-08-25

DOI

10.1146/annurev-genom-111422-025117

HHS Public Access

Author manuscript

Annu Rev Genomics Hum Genet. Author manuscript; available in PMC 2024 April 24.

Published in final edited form as:

Annu Rev Genomics Hum Genet. 2023 August 25; 24: 305–332. doi:10.1146/annurev-genom-111422-025117.

Methods for Assessing Population Relationships and History Using Genomic Data

Priya Moorjani[1], **Garrett Hellenthal**[2]

[1]Department of Molecular and Cell Biology and Center for Computational Biology, University of California, Berkeley, California, USA

[2]UCL Genetics Institute and Research Department of Genetics, Evolution, and Environment, University College London, London, United Kingdom

Abstract

Genetic data contain a record of our evolutionary history. The availability of large-scale datasets of human populations from various geographic areas and timescales, coupled with advances in the computational methods to analyze these data, has transformed our ability to use genetic data to learn about our evolutionary past. Here, we review some of the widely used statistical methods to explore and characterize population relationships and history using genomic data. We describe the intuition behind commonly used approaches, their interpretation, and important limitations. For illustration, we apply some of these techniques to genome-wide autosomal data from 929 individuals representing 53 worldwide populations that are part of the Human Genome Diversity Project. Finally, we discuss the new frontiers in genomic methods to learn about population history. In sum, this review highlights the power (and limitations) of DNA to infer features of human evolutionary history, complementing the knowledge gleaned from other disciplines, such as archaeology, anthropology, and linguistics.

Keywords

demographic inference; admixture; ancestry; effective population size; molecular clocks

1. INTRODUCTION

The emergence of genetic variation data from thousands of present-day and ancient DNA (aDNA) samples has made genomics a powerful tool for learning about human population history, complementing evidence from other sources, such as archaeology, anthropology, and linguistics (74, 98). The latest genomic analyses integrate data from millions of markers across the genome, either focusing on variants on a genotyping chip or capture array or using whole-genome sequencing (WGS). The latter provides a comprehensive catalog of variants within an individual's genome, including rare variants that are especially useful

moorjani@berkeley.edu .

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

for learning about recent population history (6). By contrast, genotyping and capture arrays contain mainly biallelic single-nucleotide polymorphisms (SNPs) that are selected based on genetic variation patterns found in previously sampled individuals, which can lead to ascertainment bias as they often contain common variants and miss most rare, population-specific variants in diverse groups (59). However, most current genomic datasets still contain approximately 500,000 to 2 million SNPs, which can potentially provide a detailed snapshot of genetic diversity over time.

Over the past few decades, many computational methods have been introduced to analyze large-scale genomic datasets. These methods use the insight that, since DNA is passed down from generation to generation, it contains information about our ancestors. In each generation, genetic variation can arise due to two main processes: mutations that lead to changes in the DNA sequence, and recombinations that shuffle DNA across homologous chromosomes (51). Due to recombination, the genome of an individual is a mosaic of chromosomal segments, with each segment inherited from one of their ancestors. For this reason, the genome contains several regions, or haplotypes, of independent information reflecting distinct evolutionary histories. Moreover, as alleles at nearby markers are often inherited together from the same ancestor (108), they are often correlated, exhibiting nonrandom association referred to as linkage disequilibrium (LD). The historical signatures of our past—including population bottlenecks and expansions, intermixing among populations, and natural selection driven by, e.g., disease or environmental adaptation—often leave traces in our genomes. Thus, by analyzing the genomic data of diverse individuals, one can make inferences about population history. As mutations and recombination events accrue steadily over time, they provide a record of the time elapsed and hence serve as molecular clocks for learning about the timing of past events using genomic data (51).

In this review, we describe some of the latest and most widely used statistical methodologies to study genetic relatedness and demographic history of populations. We illustrate the use of DNA to infer features of history, including past population size changes and periods of intermixing or admixture between distinct groups. We briefly describe the intuition behind commonly used approaches, their interpretation, and important limitations. For illustration, we apply many of these techniques to genome-wide data from 929 individuals representing 53 worldwide populations that are part of the Human Genome Diversity Project (HGDP) (6, 67). This widely used dataset is freely available (29), includes both WGS and SNP array genotypes, and allows for a detailed characterization of human population variation. Here, we consider both WGS data from Bergström et al. (6) and 474,491 SNPs on a genotyping array from Li et al. (67).

We note that this review is not meant to serve as a comprehensive and exhaustive catalog of all of the many interesting methods in the field. We focus our discussion on methods for characterizing population histories and relationships using human data, though the majority of these methods can be readily applied to other species (26, 94, 122). Furthermore, we discuss only SNPs from the 22 autosomes, ignoring the sex-based X and Y chromosomes and non-nuclear mitochondrial DNA. Analyses of non-autosomal regions provide deep insights about evolutionary history, particularly sex-biased demographic events and sex

chromosome evolution. These analyses are beyond the scope of this review but have been covered elsewhere (e.g., 45, 51, 87). Beyond demographic inference, many related methods also have widespread applications in disease mapping, scans of selection, and inference of other evolutionary parameters. However, we do not discuss these applications in detail; instead, we point interested readers to several reviews on these topics (36, 93, 119).

We also note that the application of genomic methods to aDNA samples offers unique opportunities and challenges compared with the use of data from present-day individuals. aDNA samples have transformed research into human evolutionary history. These time-stamped samples provide a direct snapshot of genetic variation in the past instead of requiring researchers to infer it retrospectively (74). However, aDNA samples often have high rates of DNA degradation, characteristic patterns of DNA damage (e.g., deamination of cytosine nucleotides), and contamination from living individuals and other sources (88). Such conditions lead to high rates of missing data, low and uneven sequence coverage across the genome, and difficulties in reliably inferring diploid genotype calls. To address this, sometimes pseudohaploid genotypes are generated, where the diploid genotype is determined by selecting a single random allele observed in the reads mapped at a particular site (88). Furthermore, some aDNA samples can be subject to a nonrandom ascertainment process [i.e., restricting analysis to sites that are heterozygous in some human population(s)] that can bias statistical analysis (92). Many methods described below do not specifically account for these features of ancient genomes, and thus results from such applications should be interpreted carefully. Where applicable, we have indicated methods that are robust to or specifically designed for analysis of aDNA samples. However, we encourage readers to investigate the reliability of different methods for application to specific samples, sequence coverage, and timescales that may be relevant to their data and application.

2. QUANTIFYING AND VISUALIZING PATTERNS OF POPULATION RELATIONSHIPS

Many statistical methods have been developed to describe and visualize genetic diversity among a set of sampled individuals. These methods are designed to capture significant components of variation in the data and highlight important trends. Such patterns can arise from an array of demographic processes, such as isolation and divergence, population bottlenecks, or admixture, warranting caution in the interpretation of the patterns (76). Nonetheless, such model-free, descriptive approaches are useful and widely applied tools for understanding how different groups relate to one another genetically.

2.1. Principal Component Analysis

Along with multidimensional scaling, one widely used technique is principal component analysis (PCA), which was first proposed by Cavalli-Sforza and colleagues (77). PCA is a nonparameteric, algebraic approach that, when applied to genetic variation data, can be used to project each individual onto orthogonal axes, referred to as principal components or eigenvectors, each of which is a weighted linear combination of its SNP data (93). The SNP weights are chosen such that the first eigenvector captures the largest amount of variation in the data. Likewise, the second eigenvector is chosen to explain as much of the remaining

variation not explained by the first eigenvector as possible, and so on. In this manner, the first two eigenvectors reduce each person's data from many thousands of SNPs to only two data dimensions. Typically, studies use the first few eigenvectors to represent samples; however, determining how many eigenvectors to use to most comprehensively capture all or most of the variation is not straightforward (93).

Figure 1 shows the first four eigenvectors generated by applying PCA to the HGDP array data using the program SMARTPCA, which is part of the EIGENSOFT software (93). PCA assumes that all SNPs are independent, so the array data were first filtered to remove SNPs in LD with each other (for details, see Figure 1). We observe a strong correlation between genetics and geography (in this case, major geographic regions), attesting to the amount of information in genome-wide autosomal SNPs. Eigenvector 1, which explains 5.5% of the overall variation in the dataset, separates Africans on the one end from East Asians and Americans on the other end, with other continental groups in between. Eigenvector 2, which explains 4.2% of the overall variation in the dataset, further separates people from Central South Asia, Europe, the Middle East, and Oceania. Eigenvector 3 separates Native American groups, while eigenvector 4 most notably separates the two Oceanian groups in the dataset (Melanesians and Papuans). Interestingly, while the top four eigenvectors each explain more variation in the data than any other eigenvectors, in sum they explain only a relatively modest amount (<13%) of the total variation across individuals. This highlights how a relatively small amount of overall genetic variation is attributable to differences among individuals from different continental regions.

While PCA efficiently summarizes key patterns in the data in a few dimensions, interpreting the underlying factors leading to these patterns is challenging. For example, groups that have been isolated from others, even if only relatively recently in the history of anatomically modern humans, can have an outsize influence, which may explain the strong separation of the island-dwelling populations of Melanesians and Papuans from other groups along eigenvector 4. Technically, populations with larger sample sizes may distort eigenvectors by making up a larger proportion of the overall variation. Batch effects in genotyping (e.g., when combining data from multiple sources) can also affect PCA results, as can ascertainment bias (76). Furthermore, as noted above, PCA assumes that SNPs are independent, which in practice can necessitate removing some SNPs based on varying criteria and may lose power over other approaches that explicitly model association among neighboring SNPs (93). Nonetheless, PCA applied to human populations typically reveals a strong correlation between genetics and geography, both at macroscales (6) (Figure 1) and within continents (14, 60, 64, 86, 109). Extensions of PCA, such as Procrustes transformation (115), principal component factor analysis (34), and PCA projection (64), where the eigenvectors are generated on a subset of samples and the rest of the samples are projected on those axes, work more robustly for sparse datasets such as aDNA samples.

2.2. Measures of Genetic Distance Across Populations

While dimensionality reduction techniques like PCA provide effective visualizations of genetic distance, other approaches attempt to quantify the genetic distance between two groups. Probably the most widely used genetic distance measure is the definition of pairwise

F_{ST}, which calculates the proportion of total observed allele frequency variance that is attributable to allele frequency differences between two groups, averaged across genome-wide SNPs (though note that other definitions of F_{ST} exist; for a review, see 7). F_{ST} ranges from 0 to 0.2 for most pairwise comparisons of human populations, which is notably lower than the maximum theoretical F_{ST} value of 1 (7). F_{ST} between groups from Africa and outside Africa in the HGDP is typically high, and in general, groups from the same major geographical region have lower pairwise F_{ST} (upper left triangle of Figure 2).

Alternative genetic distance metrics can readily leverage haplotype information. One such haplotype-based approach uses the software CHROMOPAINTER (62), which is based on the copying-with-recombination model described by Li & Stephens (68), to model each sampled individual's two haploid genomes as a mosaic of chromosomal segments from the genomes of other sampled individuals. Each region of an individual's mosaic genome reflects a chromosomal segment that shares an ancestor with a particular sampled genome more recently than with any other sampled genome in the dataset. The inferred proportions of genome-wide DNA for which each individual (or population) shares a most recent ancestor with individuals from each sampled group can be compared with those of other individuals (or populations) using total variation distance (TVD) (65); these values are depicted for HGDP populations in the lower right triangle of Figure 2. While pairwise F_{ST} approaches typically compare pairs of populations with each other irrespective of other sampled groups, CHROMOPAINTER attempts to determine which individuals share ancestors more recently with each other than with any other individuals from all sampled groups. In particular, individuals from groups that are relatively isolated from other sampled populations are more likely to share a most recent ancestor with other members of the same group and hence share fewer most recent ancestors with people from other sampled groups. As a consequence, TVD between isolated groups and other sampled groups is often relatively high, indicating a high genetic distance (65). For example, some Native American (Karitiana, Pima, and Surui) and African hunter–gatherer Pygmy populations, along with the Kalash (discussed below), show particularly high genetic distance with other sampled populations as measured by TVD, though such differences are not as apparent when instead using F_{ST} (Figure 2). Most of these populations show high levels of identity-by-descent (IBD) sharing (described below) due to recent founder events and endogamous marriages (122).

Like PCA, both F_{ST} and TVD are sensitive to sample size and number of SNPs, with F_{ST} also potentially affected by SNP array ascertainment (7). Moreover, different demographic features can lead to similar patterns under either measure.

3. CLUSTERING ALGORITHMS

Another widely used approach applied in many genetic studies is to cluster individuals into some number of groups that are relatively genetically homogeneous. This can demonstrate latent substructure in the dataset and highlight which groups are relatively genetically differentiated from each other. Here, we discuss two classes of clustering approaches that differ based on whether they directly model LD information among SNPs.

3.1. Using Allele Sharing

One of the most popular algorithms for clustering individuals based on genetic variation patterns is the Bayesian method STRUCTURE (31, 101). While computationally expensive, its central methodology has been implemented in several other approaches that scale to current levels of genome-wide SNP data in large cohorts, including ADMIXTURE (2), FRAPPE (67), and fastSTRUCTURE (104). In each of these approaches, individuals are clustered together based on their relative amounts of allele sharing across SNPs, which is assumed to be independent, with the user typically prespecifying the number of clusters K to use. Importantly, for many approaches an individual can have partial membership in multiple clusters, which in practice can enable identifying admixed individuals (101).

Figure 3 shows results when applying ADMIXTURE with $K = 2$–8 to the HGDP array data. With $K = 2$, Africans are clustered separately from East Asians, Americans, and Oceanians, with people from other geographic regions assigned to both clusters. With $K = 3$, people from Central South Asia, Europe, and the Middle East more clearly cluster separately from East Asians, Americans, and Oceanians. Subsequently, individuals from populations in the Americas largely cluster separately ($K = 4$), individuals from Oceanic populations largely cluster separately ($K = 5$), and individuals from Central South Asian ($K = 6$) and Middle Eastern ($K = 8$) populations partially cluster separately from other groups.

These results illustrate how ADMIXTURE is capturing genetic patterns strongly related to geography, similar to PCA (Figure 1). Furthermore, because each individual can be assigned to multiple clusters, ADMIXTURE can more clearly highlight intermixing among groups. For example, some individuals from the Maya population from Mexico have a high proportion of SNPs assigned to the light blue cluster predominantly seen in Native American groups with $K = 8$, but also a smaller proportion consistently assigned to the dark blue cluster that is predominantly seen in Europeans (Figure 3). This pattern likely reflects admixture in the Maya from Native American and European sources that began in the colonial era (47, 107).

A drawback of STRUCTURE and related methods is that choosing an appropriate K is a notoriously challenging problem. While efforts have been made to choose K automatically based on the data [e.g., using cross-validation (1) or a Bayesian approach (50)], it is not likely that any given K has a biological interpretation (e.g., as ancestral populations that intermixed at some time in the past). Moreover, depending on the set of SNPs, composition of the individuals in the dataset, and K value, the clusters can differ; thus, caution is warranted in interpreting the results (63). For example, relying only on STRUCTURE and ADMIXTURE for inferring admixture proportions in admixed individuals is fraught with challenges and currently not best practice (63, 123). In our HGDP application, specific populations tend to dominate some clusters—e.g., with nearly all Kalash assigned entirely to the yellow cluster (from $K = 6$ onward) and most Yakut assigned entirely to the purple cluster (from $K = 7$ onward) (Figure 3). While other individuals outside these groups are also assigned partially to the yellow and purple clusters, this likely does not reflect admixture between Kalash-related and Yakut-related sources with other groups. Instead, sampled Kalash and Yakut are likely assigned primarily to one cluster each because they

are more isolated (e.g., due to endogamy) relative to other sampled populations, making interpretation of the purple and yellow patterns in other groups less straightforward.

3.2. Using Haplotype Sharing

Building upon STRUCTURE and ADMIXTURE, fineSTRUCTURE (62) leverages LD among neighboring SNPs to increase power to identify latent substructure. To do so, it first uses CHROMOPAINTER (62) to generate a coancestry matrix that contains the inferred number of segments genome-wide for which each person shares an ancestor most recently with each other individual in the dataset. As an illustration, the inferred coancestry matrix for the HGDP array data is provided in Figure 4a. Next, fineSTRUCTURE uses a Bayesian approach to cluster individuals who have similar patterns of summarized recent genetic sharing. In essence, fineSTRUCTURE clusters the columns of Figure 4a, choosing K automatically based on a procedure first described by Huelsenbeck et al. (50). It then merges these clusters one at a time, using a greedy approach that at each step minimizes the decrease in posterior probability according to its Bayesian model. This process creates a bifurcating tree, or dendrogram, relating the clusters, as illustrated at the top of Figure 4a, with the final inferred clusters given at the bottom of the tree (62).

fineSTRUCTURE can also be applied to a coancestry matrix that has been inferred while ignoring LD among neighboring SNPs—i.e., instead treating each SNP independently, which has been shown to capture the same information used in PCA (62). Figure 4b shows, for all HGDP populations, the proportion of individuals that are assigned by fineSTRUCTURE to a cluster that contains only other individuals from the same population. For most populations, a higher proportion of individuals from the same population are clustered together when using the so-called linked CHROMOPAINTER model, which leverages LD information, relative to the so-called unlinked model, which analyzes each SNP independently. This illustrates how using LD information can capture more subtle structure that is missed by ignoring LD information, which has been shown in practice most strikingly when clustering individuals from the same country (Figure 4; see also 65).

As with STRUCTURE and related methods described above, determining the demographic processes behind a fineSTRUCTURE clustering or dendrogram is not straightforward. For example, the dendrogram should not be interpreted as an average genealogy relating groups. An additional limitation of the fineSTRUCTURE model is that each individual must be assigned to a single cluster, so that highlighting whether an individual may be admixed is not as straightforward as it is in clustering models like ADMIXTURE. In theory, the programs GLOBETROTTER (47) and SOURCEFIND (20) can infer admixture by forming individuals' CHROMOPAINTER coancestry information as mixtures of those from multiple populations, though this reflects relative proportions of recent ancestor sharing rather than admixture explicitly. Nonetheless, signatures of admixture are suggested by the coancestry matrix in instances where an individual shares a relatively high inferred number of recent ancestors with people far away in the dendrogram, e.g., with some Native Americans showing relatively high recent relatedness to Africans and Europeans (107) (Figure 4a).

4. POPULATION PHYLOGENETIC TREES AND GRAPHS

While clustering methods are powerful for detecting population substructure, they do not provide any formal tests of population history or demography. To explicitly characterize population relationships, several methods have been recently introduced that fit a demographic model or phylogenetic tree or graph to population-level data. Below, we describe these methods and apply them to HGDP data.

4.1. Allele Frequency Correlation Statistics (*f* Statistics)

One class of formal methods, inspired by Cavalli-Sforza & Edwards (18), uses allele frequency differences (e.g., F_{ST}) across populations to build phylogenetic trees to model population relationships. These methods do not explicitly model each demographic parameter (details such as population size changes are captured by the branch lengths of the phylogenetic trees), making it feasible to examine many populations simultaneously. A limitation, however, is that gene flow between populations violates the tree assumption. Building on this idea, several studies have proposed the use of phylogenetic graphs (models that allow for migration edges leading to closed loops in the tree) to model allele frequencies and to provide formal tests for gene flow (5, 61, 71, 92, 97, 121, 124).

The most widely used methods currently are ADMIXTOOLS (92) and TreeMix (97) as they are computationally feasible for large population datasets. ADMIXTOOLS contains a suite of methods based on *f* statistics (e.g., f_3, f_4, D, and f_4 ratio) for characterizing population relationships and investigating signals of admixture across groups. *f* statistics measure variance in allele frequencies or genetic drift that has occurred on a lineage in a phylogenetic tree. By comparing the magnitude and the sign of the shared genetic drift across populations, *f* statistics provide quantitative expectations under different models of demographic histories (92, 109). For instance, the three-population test or f_3 test (92, 109) compares allele frequencies across three populations or groups. Briefly, it measures the difference in allele frequencies across the three groups (say, A, B, and C) as $(p_C - p_A)(p_C - p_B)$ averaged across multiple genome-wide SNPs, where p_A, p_B, and p_C are the allele frequencies in populations A, B, and C respectively. If C is set to be an outgroup population that is highly diverged from A and B, then the f_3 test infers the shared drift between A and B since their split from C. In this setup, referred to as the outgroup f_3 test, the populations B_i, among a panel of $i = 1, 2, \ldots, k$ populations, with higher standardized f_3 values can be inferred as being more closely related to A. The outgroup f_3 test was used by Raghavan et al. (103) to infer that the 24,000-year-old ancient Siberian Mal'ta sample was genetically closer to groups from the Americas that lived thousands of miles away than it was to the West Eurasian individuals that were sampled geographically closer to the specimen. Note that another standard approach, instead of outgroup f_3 tests, for finding the most closely related group to A is to use pairwise distance measures between A and B, such as F_{ST} or f_2 [equal to $(p_A - p_B)^2$ averaged across SNPs (92)]. These statistics are, however, much more sensitive to drift in A and B and tend to downweight B groups that have small effective population sizes.

The f_3 statistic (92, 109) can also provide a formal test for admixture. In a population phylogeny, the genome-wide value of $f_3(C; A, B)$ is expected to be ≥0. However, when there is admixture in C from populations related to A and B, the allele frequencies in C

will be intermediate between A and B, and thus the expected value of the f_3 statistic would be significantly negative (92). In many cases, positive values may reflect a lack of power, especially if there is a high degree of drift postadmixture in the target population C due to a recent population bottleneck or endogamy, and hence a positive f_3 value should not be interpreted as lack of evidence for admixture (92, 109). Surveying the f_3 values for all sets of three populations in the HGDP dataset, we find that there is significant evidence for admixture in 29 (out of 53) populations ($Z < -3$). This highlights how admixture is a recurrent theme in human history and prehistory (98). A strength of f statistics is that they can model the patterns of genetic sharing using surrogate or proxy populations that are related to, though fairly diverged from, the ancestral groups (92, 109). This feature, however, also makes it difficult to interpret the results historically, as the proxy populations may not obviously be informative of the ancestral admixing source populations.

Another widely used test is Patterson's D statistic or f_4 statistic (92, 109), which uses sets of four populations to study genetic sharing across groups. For any four populations (A, B, C, and D), it measures $(p_A - p_B)(p_C - p_D)$ averaged across genome-wide SNPs. $f(A, B; C, D)$ is expected to be 0 if (A, B) and (C, D) are clades in the population tree. However, if pairs of groups (A, C) or (B, D) and (A, D) or (B, C) are closer to each other due to gene flow, a significant deviation from 0 is observed (92). This approach is more powerful than f_3 statistics, since it is less sensitive to drift postadmixture. For example, f_3 statistics often fail to identify admixture in endogamous groups (e.g., many groups from India), though f_4 statistics find significant evidence for admixture in the same populations with the addition of one more outgroup (109). A modification of the f_4 statistic, Patterson's D statistic, which uses alleles that are polarized as ancestral or derived [by comparing with chimpanzee or the inferred human ancestral allele (41)], is widely used as a rooted, asymmetric, four-population topology test. It was first applied to quantify genomic sharing between Neanderthals and modern humans, uncovering that non-Africans share more alleles with Neanderthals than Africans do (41). Extensions of D statistics that compare estimates across sets of reference populations can be used to infer the minimum number of gene flow events (qpWave) (107), estimate admixture proportions (qpAdm) (15, 92), and infer the direction of gene flow with additional outgroups [D_{FOIL} (94) and partitioned D statistics (28)].

4.2. Modeling Population Relationships as Trees and Graphs

f statistics are also widely applied to build and test topologies detailing population relationships and gene flow events among set of n populations. For example, qpGraph in ADMIXTOOLS (92) compares how well the estimated f statistics fit the predicted model based on a user-input graph topology relating n populations. The graph can contain both present-day and aDNA samples, though in current practice aDNA samples are treated the same as present-day samples (i.e., there is no accounting for missing evolution in ancient genomes). Using estimated values of f_2, f_3, and f_4 statistics for all n populations and the graph topology, the method infers expected values of f statistics and admixture proportions and then compares how closely the estimated and predicted f statistics align with each other, providing a formal test for the hypothesized phylogeny relating the n populations (92).

A limitation of qpGraph, however, is that the model of population relationships must be specified a priori.

MixMapper (71) extends qpGraph by automating the process for generating the topology by first building a scaffold tree of all unadmixed populations, inferred using f_2 statistics, and then adding admixed populations onto this tree. TreeMix (97) is a related method that first builds a maximum likelihood tree of all or most of the n populations and then identifies populations that are poor fits to the tree model by comparing the covariance matrix implied by the tree with observed estimates in real data. It then adds migration edges between branches to account for admixture. While these methods are useful for characterizing population relationships, one limitation is that inferring the topology involves a combinatorial search of a vast search space of numerous possible tree topologies and admixture events. Therefore, while these approaches can provide solutions that show close concordance between predicted and inferred allele frequency correlation patterns among populations, there are many equally likely solutions that remain unexplored.

5. INFERRING POPULATION DEMOGRAPHY USING GENOME-WIDE SUMMARY INFORMATION

5.1. Patterns of Identity-by-Descent Sharing

IBD refers to the inheritance of two alleles or haplotypes that are identical and are inherited from a shared ancestor (12). Studying regions of IBD among individuals can be useful to infer close relationships such as siblings or cousins in the dataset, or shared population relationships from a distant common ancestor among unrelated individuals. Closely related individuals share many long regions of IBD. For example, half siblings share long regions of more than 50 cM (or approximately 50 Mb) with each other and cumulatively may share almost half the genome with each other on one chromosome (and not the other, which is inherited from the other parent). As the relationship becomes more distant in time, the average proportion of shared IBD in the genome decreases exponentially (12).

Within small isolated populations (referred to as founder populations), IBD among individuals may persist over long distances for many generations, as the majority of the individuals share relatively few recent ancestors. This key insight can be leveraged to learn about population size changes and bottleneck events over time (12). By measuring genome-wide total IBD shared across pairs of individuals within the same population, we find that many groups in the HGDP dataset—such as Kalash, African hunter–gatherers, and several groups from the Americas—share a large amount of total IBD with each other, as is characteristic of small isolated populations (Figure 5a). Recent methods, such as DoRIS (90) and IBDNe (13), leverage the distribution of IBD sharing among individuals within a group to infer effective population sizes over time. To reliably measure IBD segments, these methods focus on long IBD segments that are >2–4 cM long and hence are more informative of recent founder events. Application of IBDNe to the Kalash population shows a recent bottleneck in the past 20 generations (Figure 5b), consistent with historical records (4).

Most commonly used IBD-based methods use phased data that can be obtained from computational phasing of population data; however, this typically requires large numbers of high-quality samples. Errors in computational phasing (switch errors) or sparsity of data due to limited numbers of samples or missingness, as is characteristic of aDNA samples, can result in biased estimates of IBD segment lengths and, in turn, population size inference (90). Recent methods such as popshare (109) and ASCEND (122) instead propose to use allele-sharing correlation across individuals that can be readily measured in genotype data without phasing. By measuring allele sharing as a function of genomic distance, these methods infer the time (popshare and ASCEND) and strength (ASCEND) of founder events. Application of ASCEND to HGDP datasets reveals that more than 65% of the populations (35 out of 53) have evidence for a significant founder event in the past 200 generations, or 6,000 years [assuming 1 generation is 28 years (33, 81)] (Figure 5). This includes many populations from the Americas, African hunter–gatherers, and Northeast Asian indigenous groups that show, e.g., high TVD with other populations (Figure 2) as well as groups highlighted by previous surveys of other datasets (122).

A special case of IBD within an individual—i.e., the two chromosomes of an individual share IBD due to inheritance of identical haplotypes from a recent common ancestor—leads to runs of homozygosity (ROHs) (19). ROHs are ubiquitous among human populations and correlate with pedigree inbreeding and consanguinity (8). Outbred populations have fewer and shorter ROHs, though isolated or founder populations may have large proportions of their genomes in ROHs (23). Thus, characterizing ROHs can be useful for learning about population size changes in human history and prehistory (19, 110). Moreover, ROHs are often associated with recessive Mendelian diseases and a high burden of deleterious variants. Hence, mapping ROHs also offers useful insights about disease variants (19, 38, 114).

5.2. Site Frequency Spectrum Techniques

The site frequency spectrum (SFS) is a summary of the distribution of allele frequencies in a sample of individuals from a population (51). Different population events leave distinct signatures on the SFS. For example, a population that has undergone a recent founder event or bottleneck will have a reduced number of rare variants as compared with a constant-sized population. On the other hand, population expansion models will lead to an excess of rare variants compared with a constant-sized population. Similarly, deviation from neutral evolution and population structure will also leave characteristic imprints on the SFS of a population (51).

SFS-based inference methods compare the observed value of the SFS (and other summary statistics) with the expected SFS under a given demographic model. By doing so, they can infer a range of demographic parameters such as population sizes, times of population splits and expansions, and gene flow events across populations. Because the genealogy relating the samples is not directly observed, SFS-based methods such as BEAST (27) and LAMARC (57) use approximate or marginal genealogies by sampling over the range of possible genealogies, using sampling methods such as Markov chain Monte Carlo or simulating genealogies and using approximate Bayesian computation (9, 30). These approaches can be computationally demanding, as the parameter space is very large. The

expected SFS, in principle, can be efficiently computed and scale to hundreds of samples when the demographic history is a bifurcating tree (54). Admixture or gene flow, which is ubiquitous in human history (98), violates the assumption of a treelike population history. Joint composite likelihood methods such as *δaδi* (43) and momi2 (53) that use the allele frequency distributions across multiple populations can jointly infer demographic histories in a computationally tractable manner for a handful of populations. However, SFS approaches ignore LD across nearby sites, which can impact power and potentially bias inference. Another limitation of SFS-based methods is that the population history over time is inherently not identifiable from the SFS, as multiple histories can lead to similar genetic patterns (83).

6. INFERRING AND DATING ADMIXTURE EVENTS

While the approaches mentioned above can identify whether a population is admixed, most (excluding some SFS methods) do not infer the timing of admixture. In theory, the timing of admixture events can be inferred by modeling spatial patterns across the genome in an admixed population. In particular, the genome of an admixed individual is a mosaic of chromosomal segments inherited from distinct ancestral populations. Due to recombination, these ancestral segments get shuffled in each generation and become smaller and smaller over time (21). Therefore, recombination can serve as a clock to measure when admixture occurred. In particular, assuming a pulse model of admixture events whereby two or more populations intermix instantaneously (i.e., over a generation or short period of time), followed by random mating of people from the admixed population over time, the sizes (in morgans) of segments inherited from different admixing sources follow an exponential distribution, with a rate equal to the time since admixture in generations (21). This pulse model is assumed by the techniques described in this section.

Methods that characterize the spatial patterns along the genome in admixed individuals to infer and date admixture can be classified broadly into two categories [though with some overlap (111)]: (*a*) local ancestry–based methods that attempt to recreate the blocklike mosaic of each admixed genome [TRACTS (40), RFMIX (73), etc.] and (*b*) admixture LD–based methods that model patterns of LD among SNPs [ROLLOFF (79), ALDER (72), MALDER (96), and DATES (24)] or haplotypes [GLOBETROTTER (47) and fastGLOBETROTTER (126)] as a proxy to the ancestry tracts. Here, we refer to ancestry as sharing a most recent ancestor with an individual from a particular sampled population, noting that this depends strongly on the reference sample considered.

6.1. Local Ancestry–Based Methods

Local ancestry–based methods such as HAPMIX (100), RFMIX (73), ELAI (42), and MOSAIC (111) deconvolve each haplotype (ancestry tract) of an admixed genome to the ancestral source population it was inherited from. To do so, most methods compare the admixed genome with genomes from reference populations that are meant to reflect the true admixing sources. Recent methods (40, 99, 102) assume the ancestry tract lengths follow an exponential distribution to estimate the time of mixture, in addition to estimating the proportion of admixture inherited from each source.

A major drawback of this approach is that it requires accurate assignment of ancestry at each position in the genome. This limits its utility to cases where admixture is relatively recent, which results in long segments inherited from each source and/or involves intermixing among genetically different groups that are relatively easy to distinguish. In humans, applications typically involve dating admixture among different continental groups, with examples including admixture signals in Latin Americans (15), African Americans (15, 40), and other peoples with mixed ancestry related to, e.g., Africans and Europeans (44). In addition to their use in studies of admixture, such ancestry assignment techniques have been used to investigate associations between genetic variants and traits, potentially with population-specific effect sizes, when analyzing large-scale cohorts of admixed individuals (3, 91).

6.2. Admixture Linkage Disequilibrium–Based Methods

Admixture LD–based methods measure the extent of the allelic correlation or nonrandom association across loci inherited from ancestral sources to infer the time of admixture. Here, loci can be defined as SNPs or as haplotypes.

6.2.1. Using single-nucleotide polymorphisms.—

Chakraborty & Weiss (21) introduced the idea to characterize the extent of LD in an admixed population to infer the proportion and time of admixture. Moorjani and colleagues (79, 80, 92) implemented this in ROLLOFF, which measures the decay in weighted SNP correlations (or covariance) with genetic distance. The weight—typically the difference in allele frequencies between reference populations that serve as surrogates of the ancestral admixing populations—is chosen to enhance the signal of admixture LD over background LD, with SNPs assumed to be independent within each admixing population. ROLLOFF infers the time of admixture (in generations) by fitting an exponential distribution to the decay of weighted LD with genetic distance (79, 92). ALDER (72) extends this idea by describing precise mathematical properties of admixture LD statistics. These expectations can in turn be used to infer admixture proportions (using the amplitude of the exponential decay) and provide formal tests for admixture by comparing dates of admixture generated using one or two reference populations. Moreover, ALDER provides a substantial speedup for computing pairwise covariances across markers by implementing the fast Fourier transform that makes the approach computationally tractable for large datasets (72).

Both ALDER and ROLLOFF are applicable mainly to populations with two-way admixture. However, many worldwide populations show evidence for multiple pulses of gene flow (98). MALDER (96) generalizes the ALDER model to allow for multiple gene flow events by using allele frequencies in sets of more than two reference populations to model admixture LD in an admixed population. By fitting sums of exponential distributions to admixture LD with distance, it then infers the ancestry proportion and the timing of the various pulses of admixture (96). Simulations show that admixture LD–based methods that use SNPs provide robust inference for the timing of admixture, up to hundreds of generations in the past (79, 92). Furthermore, for highly divergent ancestral groups that have fixed differences, such as Neanderthals and modern humans, LD statistics are also applicable and are most robust when applied to ascertained SNPs that are informative for the gene flow (instead of using

allele frequencies in the reference populations) (112). Using this insight, Sankararaman et al. (112) inferred the timing of Neanderthal gene flow into the ancestors of modern humans as ~1,500–2,000 generations ago or ~50,000 years ago.

Methods introduced by Moorjani and colleagues (24, 81) measure the allelic covariance across neighboring SNPs within a single genome, rather than measuring LD across SNPs using multiple genomes as in ROLLOFF and ALDER. Application of this idea to Upper Paleolithic Eurasian aDNA samples using an ascertainment informative for Neanderthal ancestry (as in 112) recovers the timing of Neanderthal gene flow in individual aDNA specimens (35, 81). Building upon this idea, a new approach, DATES (24, 84), measures the weighted allelic covariance (or ancestry covariance) across the genome in a single admixed individual, leveraging the allele frequency difference between two reference populations (representing the ancestral source populations). When multiple individuals from an admixed population are available, DATES simply computes the ancestry covariance separately for each individual and then combines the results, which is in principle similar to running ROLLOFF or ALDER (72, 79). For sparse datasets that include aDNA, DATES outperforms other LD-based methods, as it works reliably with limited samples, large proportions of missing variants, and pseudohaploid genotypes (24).

6.2.2. Using haplotypes.—Methods such as GLOBETROTTER (47), fastGLOBETROTTER (126), and MOSAIC (111) model LD among pairs of haplotype segments (rather than SNPs, which ALDER and DATES use) to infer the time of admixture. Both GLOBETROTTER and fastGLOBETROTTER first paint the phased genomes of individuals from a putative admixed population, as well as from surrogate populations that may represent admixing source populations, against haplotypes from a set of reference populations using CHROMOPAINTER (62). They then infer the timing of admixture by fitting an exponential model to the probability that two segments in an admixed genome are matched to particular surrogate populations relative to the genetic distance between the two segments. By contrast, MOSAIC merges concepts from Sections 6.1 and 6.2.1, automatically inferring which segments within each admixed genome were inherited from each admixing source and fitting an exponential model to segments from different sources. In contrast to DATES and similar approaches (6.2.1), GLOBETROTTER, fastGLOBETROTTER, and MOSAIC do not rely on predefining reference populations to act as surrogates to each admixing source. Instead, they automatically infer the genetic makeup of each admixing population, i.e., as a mixture of haplotypes found in the reference populations. They can also report the single surrogate population that best represents the inferred genetic makeup of each admixing source using a distance metric (e.g., F_{ST}; see Section 2.2), which we report here for applications to HGDP populations (Figure 6).

These haplotype-based methods are highly sensitive, often able to date admixture between closely related reference populations [e.g., between different European sources (65)]. A drawback, however, is that they are not as straightforward to use as SNP-based techniques such as ALDER and DATES. For example, they require prephasing of data, with an additional CHROMOPAINTER step for GLOBETROTTER and fastGLOBETROTTER, and potentially require large reference samples to reliably model admixed genomes.

6.3. Comparison of Methods for Admixture Inference

To illustrate the power and complexities of some current methods for admixture inference, we applied five widely used methods—the f_3 test, ROLLOFF, ALDER, GLOBETROTTER, and DATES—to the HGDP array data (Figure 6). Both the f_3 test and ALDER examine whether a target population can be modeled as a mixture of two sources related to prespecified reference (i.e., surrogate) populations. For both approaches, we considered every possible combination of three populations (i.e., one target and two reference populations) in HGDP to test for admixture in the target population. By contrast, for computational simplicity we ran ROLLOFF and DATES once for each target population, using the pair of reference populations that produced the most significant (i.e., most negative) f_3 score. Finally, we applied GLOBETROTTER (which relies on CHROMOPAINTER output) once for each target population, allowing all other populations as potential surrogates of the admixing sources. We applied each method's recommendation to identify significant evidence for admixture in the target population. We inferred admixture in 29 out of 53 populations using the f_3 test. By contrast, ALDER and GLOBETROTTER each found significant evidence for admixture in 34 populations, with 30 of these groups overlapping. The apparent increased power of ALDER and GLOBETROTTER is in part due to the fact that f_3 tests often fail to detect admixture in populations that are more drifted (109). For example, the f_3 test inferred admixture in only 1 of 14 HGDP populations with inferred median within-population genome-wide IBD sharing of >60 cM, indicating high relatedness, compared with GLOBETROTTER inferring admixture in 9 of these 14 populations (Figure 5). This may also reflect how methods that model spatial correlations attributable to admixture (i.e., local ancestry–based methods) are more sensitive for detecting recent admixture.

We next compared the surrogate populations chosen by the f_3 test, ALDER, and GLOBETROTTER to best represent the admixing sources. [We note that the authors of the f_3 test and ALDER do not have specific recommendations for inferring the best admixing sources, though in practice many studies use a similar approach to the one we apply here (96)]. We focused on the 15 groups for which GLOBETROTTER inferred only a single date of admixture and for which the other approaches showed significant evidence of admixture. We found unexpected inconsistencies in the inferred best surrogates. In all but two cases, one or more of the surrogates chosen across methods were not even from the same major geographic region. However, in most of these cases (8/13), the discrepancy may be partially explained by some methods picking a reference population that showed evidence of previous admixture involving sources related to references chosen by one or more other methods. For example, for Hazara, Makrani, and Uygurs, the f_3 test and ALDER chose a West Eurasian group as the best reference population, while GLOBETROTTER favored a South Asian population such as Pathan, who in turn have some inferred West Eurasian ancestry (109). Moreover, f_3 statistics have more power to capture distal admixture events (as the ancestral populations are likely genetically more diverged), while GLOBETROTTER uses haplotypes and is more sensitive to picking geographically proximal populations as the best admixing source population. Finally, the f_3 test and ALDER tended to favor choosing drifted populations as best representing the admixing sources, while GLOBETROTTER tended to disfavor drifted populations, even when choosing populations from the same major

geographic regions (Figure 6b). Illustrating this, among the 15 populations, the f_3 test chose surrogate populations whose individuals have, on average, an inferred median pairwise genome-wide IBD sharing of 174.5 cM, which is approximately 8 times higher than the average in populations chosen by GLOBETROTTER (22.3 cM).

The confidence intervals for inferred admixture times overlap in most cases across the four dating methods, though in some cases GLOBETROTTER inferred a more recent date—for example, in the case of the Burusho, Cambodian, and Mongola populations (Figure 6a). In each of these three populations, the surrogates chosen by GLOBETROTTER are more geographically proximal to the target population. Therefore, it is plausible that the admixture inferred by GLOBETROTTER here involves one source population that was previously admixed, with this previous admixture occurring around the dates inferred by the other approaches (Figure 6). Supporting this, previous work indicated that GLOBETROTTER inferred dates can get older if excluding geographically proximal populations as surrogates (123). For Brahui, inferred dates and sources of admixture differed more strongly between GLOBETROTTER and other approaches, suggesting that the methods were capturing different signals of admixture, as the population may have a history of multiple gene flow events (89). Overall, these analyses highlight the ubiquity of admixture in human history and the complexities of making admixture inference from genomic patterns in present-day samples. Future aDNA studies could be particularly revealing in this regard as they can help fill in the gaps in the sampling over time and provide reference data for more closely related admixing sources.

7. INFERENCE OF THE ANCESTRAL RECOMBINATION GRAPH

In theory, WGS data enable much more precise insights into demographic events and their timings than SNP array data. For example, capturing all SNPs allows one to pinpoint which individuals share rare alleles, which is indicative of their being very recently related. Furthermore, ascertainment of SNP data no longer needs to be considered. It is particularly useful that the number of mutations separating a pair of genomes in a genetic region can be used as a clock to infer the number of generations back when those genomes shared a common ancestor. This clock can be converted to calendar years using an estimate of the mutation rate.

Recent approaches leverage WGS data to apply coalescent theory (49, 56, 85, 125), which attempts to reconstruct the entire genealogy relating the genomes of all sampled individuals back to their shared ancestors (55, 105, 117). In essence, historical recombination events separate the genome into regions with varying genealogies, with the collection of all such genome-wide genealogies referred to as the ancestral recombination graph (ARG) of the sample. An accurate description of the ARG is a major goal in population genetics, since it represents the sum of information on the genetic history of a sample that is obtainable (105). In this section, we describe a subset of ARG-inference approaches that attempt to reconstruct time-stamped genealogies, illustrating some of their applications.

7.1. Sequential Markovian Coalescent Models

One of the first ARG-inference approaches applied to WGS data, the pairwise sequentially Markovian coalescent (PSMC) model, infers the time to most recent common ancestor (TMRCA) across the two haploid genomes of a single individual, which should each have their own unique history (66). As noted above, historical recombination events separate the genome into regions where the TMRCA between the two haploids differs, with each region's TMRCA inferred using the number of heterozygotes (i.e., mutations) in the region. Several regions having a similar inferred TMRCA is indicative of a population bottleneck around that TMRCA, as it suggests there were relatively few ancestors alive at that TMRCA. Using this insight, PSMC is able to infer changes in effective population size, a measure of genetic diversity, going back in time, by analyzing only a single, unphased diploid genome (66).

Extensions to the PSMC model to incorporate genomes from multiple individuals include the multiple sequentially Markovian coalescent (MSMC) model (113), which for simplicity considers only the first coalescent event among a few (typically <10) phased sampled genomes in each genetic region, and SMC++ (120), which incorporates additional information from the distribution of allele frequencies across multiple people in each genetic region. Jointly analyzing more genomes is particularly informative for recent coalescent (and hence demographic) events, as the expected time to the first coalescence decreases by a factor of $(n-2)/n$ when increasing the sample size from n to $n+1$ genomes. In addition, by jointly analyzing genomes of individuals from different populations, each of these methods can also infer the so-called split times when populations became isolated from one another. However, these methods can be computationally demanding and hence scale to only a handful of populations and/or individuals.

7.2. Inferring Genealogies for Thousands of Individuals

Recent techniques such as RELATE (117) and tsinfer (55) infer genealogies relating hundreds of genomes across sequence data in a computational tractable way. Once accurately reconstructed, information on populations' split times and size changes can be extracted from these genealogies. Figure 7a gives the inferred effective population sizes over time for five HGDP populations from different major geographic regions. While most populations exhibit a recent increase in inferred effective population size starting approximately 20 kya, matching rapid population growth worldwide, the effective population size of the Surui has remained relatively constant or decreased since that time, consistent with their relatively high inferred amount of IBD sharing (Figure 5). Cross-coalescent rates showing the inferred rates of coalescence between individuals from different populations, relative to the inferred rates of coalescence between individuals from the same population, are illustrated in Figure 7b. Periods of time where cross-coalescent rates decline may reflect when these populations became isolated from one another. For example, going from past to present, cross-coalescent rates between Bantu-speaking peoples from Kenya and either French or Japanese individuals start to decrease approximately 50–100 kya, likely corresponding to the major out-of-Africa event of modern humans (6). Analogously, cross-coalescent rates between French and Japanese begin to rapidly decrease approximately

20–50 kya, likely reflecting when the ancestors of these populations became isolated from one another.

While an accurately reconstructed ARG for large sample sizes contains a wealth of information on processes such as bottlenecks, expansions, gene flow, and selection that lead to observed genetic variation patterns, extracting this information from ARGs is not always straightforward. Nonetheless, several approaches have attempted to leverage genealogies inferred by these new methods to increase power to identify SNPs under selection (118), learn about the history of populations represented by low-coverage aDNA (116), infer changes in mutation patterns over time (37, 117), and identify regions of introgression from archaic human groups, such as Neanderthals and Denisovans, into anatomically modern humans (117). However, several limitations remain, with these approaches each affected to varying degrees by the accuracy of inferred genealogies. Current ARG-inference methods rely on simplifying assumptions to maintain the efficiency to analyze large sample sizes, such as assuming an absence of population structure. They may output inaccurate genealogies due to, e.g., lack of informative data, hidden or missed recombination events, and issues with the input sequence data, such as missing data and errors or uncertainties in base calling or ancestral state inference (10). In principle, genotype imputation (69) may help with erroneous or missing input data, though it is unclear to what extent this may introduce bias, e.g., toward populations overrepresented in the imputation reference panel. Future research in this rapidly advancing field surrounding ARG-inference methodology and applications likely will evaluate the impact of, and hopefully in many cases overcome, these concerns.

8. CONCLUSIONS AND PERSPECTIVES

In this review, we have surveyed some of the widely used statistical methods in population genetics, providing some applications to HGDP data to illustrate their utility (for a summary of the datasets and software used in the analysis, see Table 1). While we have highlighted the scope and some limitations throughout, there remain many more challenges to overcome. Some of these challenges are computational, with most current approaches often not yet scalable to the level of biobank data that contain hundreds of thousands of individuals sequenced or genotyped at millions of markers, e.g., the UK Biobank (16) or FinnGen (58). Other challenges are inherent in the complicated natures of both biological processes and human interactions. For example, most genomic methods rely on using molecular clocks of mutation and recombination that are assumed to be constant over time, but actually both show strong evidence of rapid evolution over time and among human populations (46, 48, 78). Given that our understanding of rates in both is based largely on studying extant populations, this is likely particularly problematic for understanding more ancient demographic events. Dating ancient admixture events is further complicated by the small size of ancestry tracts left behind in present-day individuals. One approach is to use aDNA samples closer in time to the admixture event, in which the ancestry tracts will be larger, to more reliably infer the time of admixture (81). Reliably dated aDNA specimens can further be used to infer evolutionary rates of mutation and recombination to reset and calibrate the molecular clock (35, 81). In addition, while results from spatial population models (86, 95) show evidence of isolation by distance, whereby neighboring populations steadily

or continuously intermix over time, many tree-based and admixture techniques mentioned here assume instantaneous population splits and/or that admixture occurs in discrete pulses, which is an oversimplification of the population history (64, 70, 113). Thus, more detailed models that capture the complexities in real data are needed.

A key limitation in population genetic studies is the bias in current data collections. For present-day populations, most of the world remains underrepresented in genetic studies relative to European populations (32). A notable example is the relative lack of available genetic variation data resources representing African populations, even though Africa harbors the largest amount of genetic diversity across all continents (6). For example, while several studies have focused on the fine-scale genetic relatedness and history of people sampled within a particular European country, relatively few analogous studies have been performed in African countries (17). While these deficiencies are being addressed through large-scale collaborative efforts for present-day samples (22, 39, 82), the aDNA collections reflect a similar lack in diversity of non-Eurasian samples (74). While the challenges to diversify genomics datasets may be real [e.g., preservation and environmental conditions make it difficult to obtain DNA from temperate regions and older time periods (88)], the disparity across human populations requires systematic investment of resources and involvement of local communities to fill this critical gap (32).

In addition to collecting data, another important consideration is how to communicate and share results from genomic surveys carefully and thoughtfully with the scientific community and the general public (17). Given the rise in direct-to-consumer testing companies and use of genomic data in clinics (106), it is important to ensure that individuals understand how their genome data can be used and the limitations of what knowledge can be gained from it. In this regard, classification of data into, e.g., population labels or assignment of geographic and genetic ancestry affiliation can have important implications for how results are interpreted and relayed (25). For example, the concept of genetic ancestry is debated (25, 75) and is taken here to refer to information about individuals or groups that a particular individual is biologically related to, including close and distant relationships. Genetic ancestry thus differs from genealogical ancestry because it encompasses relationships beyond the identifiable ancestors in a family tree or pedigree. In particular, it includes a set of paths through the human family tree, and it is variable across timescales as DNA is inherited from specific ancestors (55, 117).

Genomic methods measure genetic relationships by comparing the genetic data for a particular individual with those for other individuals or populations. In effect, they measure genetic similarity between the alleles and haplotypes of an individual and those of others in the sampled dataset. If some ancestral group is missing or unsampled, some population relationships will be missed or misinterpreted. Conversely, if individuals or populations are mislabeled or present-day samples do not reflect historical population structure, as we are learning from many aDNA surveys (64, 98), inferences and conclusions could be impacted. Thus, it is important to note that population relationships measured using genomic methods are dependent on the structure of the sampled populations in the dataset (e.g., here the HGDP) and timescale under consideration and should be interpreted carefully. As more comprehensive samples and methods become available, these data, with other sources of

evidence, will help to build a more complete picture of human evolutionary history over time.

ACKNOWLEDGMENTS

We thank Nick Patterson for helpful discussions. We thank Yulin Zhang and Leo Speidel for sharing the output of RELATE for the HGDP dataset. P.M. was supported by a Burroughs Wellcome Fund Career Award at the Scientific Interface and the National Institutes of Health (grant R35GM142978). G.H. was supported by the Wellcome Trust and the Royal Society (grants 098386/Z/12/Z and 224575/Z/21/Z) and by the National Institute for Health and Care Research/University College London Hospitals Biomedical Research Centre.

LITERATURE CITED

1. Alexander DH, Lange K. 2011. Enhancements to the admixture algorithm for individual ancestry estimation. BMC Bioinform 12:246
2. Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–64 [PubMed: 19648217]
3. Atkinson EG, Maihofer AX, Kanai M, Martin AR, Karczewski KJ, et al. 2021. Tractor uses local ancestry to enable the inclusion of admixed individuals in gwas and to boost power. Nat. Genet 53:195–204 [PubMed: 33462486]
4. Ayub Q, Mezzavilla M, Pagani L, Haber M, Mohyuddin A, et al. 2015. The Kalash genetic isolate: ancient divergence, drift, and selection. Am. J. Hum. Genet 96:775–83 [PubMed: 25937445]
5. Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. PNAS 98:4563–68 [PubMed: 11287657]
6. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, et al. 2020. Insights into human genetic variation and population history from 929 diverse genomes. Science 367:eaay5012 [PubMed: 32193295]
7. Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting F_{ST}: the impact of rare variants. Genome Res 23:1514–21 [PubMed: 23861382]
8. Bittles AH, Black ML. 2010. Consanguinity, human evolution, and complex diseases. PNAS 107(Suppl. 1):1779–86 [PubMed: 19805052]
9. Boitard S, Rodríguez W, Jay F, Mona S, Austerlitz F. 2016. Inferring population size history from large samples of genome-wide molecular data-an approximate Bayesian computation approach. PLOS Genet 12:e1005877 [PubMed: 26943927]
10. Brandt DYC, Wei X, Deng Y, Vaughn AH, Nielsen R. 2022. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. Genetics 221:iyac044 [PubMed: 35333304]
11. Browning BL, Zhou Y, Browning SR. 2018. A one-penny imputed genome from next generation reference panels. Am. J. Hum. Genet 103:338–48 [PubMed: 30100085]
12. Browning SR, Browning BL. 2012. Identity by descent between distant relatives: detection and applications. Annu. Rev. Genet 46:617–33 [PubMed: 22994355]
13. Browning SR, Browning BL. 2015. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. Am. J. Hum. Genet 97:404–18 [PubMed: 26299365]
14. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, et al. 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. PNAS 107:786–91 [PubMed: 20080753]
15. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. 2015. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am. J. Hum. Genet 96:37–53 [PubMed: 25529636]
16. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, et al. 2017. Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv 166298. 10.1101/166298

17. Carlson J, Henn BM, Al-Hindi DR, Ramachandran S. 2022. Counter the weaponization of genetics research by extremists. Nature 610:444–47 [PubMed: 36261568]
18. Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet 19:233–57 [PubMed: 6026583]
19. Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. 2018. Runs of homozygosity: windows into population history and trait architecture. Nat. Rev. Genet 19:220–34 [PubMed: 29335644]
20. Chacon-Duque J, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuna-Alonzo V, et al. 2018. Latin Americans show wide-spread Converso ancestry and the imprint of local Native ancestry on physical appearance. Nat. Commun 9:5388 [PubMed: 30568240]
21. Chakraborty R, Weiss KM. 1988. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. PNAS 85:9119–23 [PubMed: 3194414]
22. Chen Z, Chen J, Collins R, Guo Y, Peto R, et al. 2011. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int. J. Epidemiol 40:1652–66 [PubMed: 22158673]
23. Chiang CW, Ralph P, Novembre J. 2016. Conflation of short identity-by-descent segments bias their inferred length distribution. G3 6:1287–96 [PubMed: 26935417]
24. Chintalapati M, Patterson N, Moorjani P. 2022. The spatiotemporal patterns of major human admixture events during the European Holocene. eLife 11:e77625 [PubMed: 35635751]
25. Coop G 2022. Genetic similarity and genetic ancestry groups. arXiv:2207.11595 [q-bio.PE]
26. Dagilis AJ, Peede D, Coughlan JM, Jofre GI, D'Agostino ER, et al. 2021. 15 years of introgression studies: quantifying gene flow across eukaryotes. bioRxiv 2021.06.15.448399 10.1101/2021.06.15.448399
27. Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol 7:214 [PubMed: 17996036]
28. Eaton DAR, Ree RH. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). Syst. Biol 62:689–706 [PubMed: 23652346]
29. EMBL-EBI (Eur. Mol. Biol. Lab.–Eur. Bioinform. Inst.). 2020. Human Genome Diversity Project. IGSR: The International Genome Sampling Resource https://www.internationalgenome.org/data-portal/data-collection/hgdp
30. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. PLOS Genet 9:e1003905 [PubMed: 24204310]
31. Falush D, Stephens M, Pritchard J. 2003. Inference of population structure from multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–87 [PubMed: 12930761]
32. Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbaecker K. 2022. A roadmap to increase diversity in genomic studies. Nat. Med 28:243–50 [PubMed: 35145307]
33. Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol 128:415–23 [PubMed: 15795887]
34. François O, Jay F. 2020. Factor analysis of ancient population genomic samples. Nat. Commun 11:4661 [PubMed: 32938925]
35. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514:445–49 [PubMed: 25341783]
36. Fu W, Akey JM. 2013. Selection and adaptation in the human genome. Annu. Rev. Genom. Hum. Genet 14:467–89
37. Gao Z, Zhang Y, Cramer N, Przeworski M, Moorjani P. 2023. Limited role of generation time changes in driving the evolution of the mutation spectrum in humans. eLife 12:e81188 [PubMed: 36779395]
38. Garrod A 1902. The incidence of alkaptonuria: a study in chemical individuality. Lancet 160:1616–20
39. GenomeAsia100K Consort. 2019. The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature 576:106–11 [PubMed: 31802016]
40. Gravel G 2012. Population genetics models of local ancestry. Genetics 191:607–19 [PubMed: 22491189]

41. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. 2010. A draft sequence of the Neandertal genome. Science 328:710–22 [PubMed: 20448178]
42. Guan Y 2014. Detecting structure of haplotypes and local ancestry. Genetics 196:625–42 [PubMed: 24388880]
43. Gutenkunst R, Hernandez R, Williamson S, Bustamante C. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLOS Genet 5:e1000695 [PubMed: 19851460]
44. Hamid I, Korunes K, Beleza S, Goldberg A. 2021. Rapid adaptation to malaria facilitated by admixture in the human population of Cabo Verde. eLife 10:e63177 [PubMed: 33393457]
45. Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. PLOS Genet 4:e1000202 [PubMed: 18818765]
46. Harris K, Pritchard JK. 2017. Rapid evolution of the human mutation spectrum. eLife 6:e24284 [PubMed: 28440220]
47. Hellenthal G, Busby G, Band G, Wilson J, Capelli C, et al. 2014. A genetic atlas of human admixture history. Science 343:747–51 [PubMed: 24531965]
48. Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, et al. 2011. The landscape of recombination in African Americans. Nature 476:170–75 [PubMed: 21775986]
49. Hudson R 1990. Gene genealogies and the coalescent process. In Oxford Surveys on Evolutionary Biology, Vol. 7, ed. Futuyma D, Antonovics J, pp. 1–44. New York: Oxford Univ. Press
50. Huelsenbeck JP, Andolfatto P, Huelsenbeck ET. 2011. Structurama: Bayesian inference of population structure. Evol. Bioinform. Online 7:55–59 [PubMed: 21698091]
51. Jobling M, Hurles M, Tyler-Smith C. 2004. Human Evolutionary Genetics: Origins, Peoples and Disease New York: Garland Sci.
52. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, et al. 2017. Parental influence on human germline de novo mutations in 1,548 trios from iceland. Nature 549:519–22 [PubMed: 28959963]
53. Kamm JA, Terhorst J, Durbin R, Song YS. 2020. Efficiently inferring the demographic history of many populations with allele count data. J. Am. Stat. Assoc 115:1472–87 [PubMed: 33012903]
54. Kamm JA, Terhorst J, Song YS. 2017. Efficient computation of the joint sample frequency spectra for multiple populations. J. Comput. Graph. Stat 26:182–94 [PubMed: 28239248]
55. Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. 2019. Inferring whole-genome histories in large population datasets. Nat. Genet 51:1330–38 [PubMed: 31477934]
56. Kingman JF. 1982. On the genealogy of large populations. J. Appl. Probab 19:27–43
57. Kuhner MK. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22:768–70 [PubMed: 16410317]
58. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, et al. 2022. FinnGen: unique genetic insights from combining isolated population and national health register data. medRxiv 2022.03.03.22271360 10.1101/2022.03.03.22271360
59. Lachance J, Tishkoff SA. 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. BioEssays 35:780–86 [PubMed: 23836388]
60. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, et al. 2008. Correlation between genetic and geographic structure in Europe. Curr. Biol 18:1241–48 [PubMed: 18691889]
61. Lathrop G 1982. Evolutionary trees and admixture: phylogenetic inference when some populations are hybridized. Ann. Hum. Genet 46:245–55 [PubMed: 7125596]
62. Lawson D, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. PLOS Genet 8:e1002453 [PubMed: 22291602]
63. Lawson D, van Dorp L, Falush D. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nat. Commun 9:3258 [PubMed: 30108219]
64. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409–13 [PubMed: 25230663]
65. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, et al. 2015. The fine scale genetic structure of the British population. Nature 519:309–14 [PubMed: 25788095]

66. Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475:493–96 [PubMed: 21753753]
67. Li J, Absher D, Tang H, Southwick A, Casto A, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–4 [PubMed: 18292342]
68. Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–33 [PubMed: 14704198]
69. Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype imputation. Annu. Rev. Genom. Hum. Genet 10:387–406
70. Liang M, Nielsen R. 2014. The lengths of admixture tracts. Genetics 197:953–67 [PubMed: 24770332]
71. Lipson M, Loh PR, Levin A, Reich D, Patterson N, Berger B. 2013. Efficient moment-based inference of admixture parameters and sources of gene flow. Mol. Biol. Evol 30:1788–802 [PubMed: 23709261]
72. Loh P, Lipson M, Patterson N, Moorjani P, Pickrell J, et al. 2013. Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193:1233–54 [PubMed: 23410830]
73. Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet 93:278–88 [PubMed: 23910464]
74. Marciniak S, Perry GH. 2017. Harnessing ancient genomes to study the history of human adaptation. Nat. Rev. Genet 18:659–74 [PubMed: 28890534]
75. Mathieson I, Scally A. 2020. What is ancestry? PLOS Genet 16:e1008624 [PubMed: 32150538]
76. McVean G 2009. A genealogical interpretation of principal components. PLOS Genet 5:e1000686 [PubMed: 19834557]
77. Menozzi P, Piazza A, Cavalli-Sforza L. 1978. Synthetic maps of human gene frequencies in Europeans: These maps indicate that early farmers of the Near East spread to all of Europe in the Neolithic. Science 201:786–92 [PubMed: 356262]
78. Moorjani P, Gao Z, Przeworski M. 2016. Human germline mutation and the erratic evolutionary clock. PLOS Biol 14:e2000744 [PubMed: 27760127]
79. Moorjani P, Patterson N, Hirschhorn J, Keinan A, Hao L, et al. 2011. The history of African gene flow into southern Europeans, Levantines, and Jews. PLOS Genet 7:e1001373 [PubMed: 21533020]
80. Moorjani P, Patterson N, Loh PR, Lipson M, Kisfali P, et al. 2013. Reconstructing Roma history from genome-wide data. PLOS ONE 8:e58633 [PubMed: 23516520]
81. Moorjani P, Sankararaman S, Fu Q, Przeworski M, Patterson N, Reich D. 2016. A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. PNAS 113:5652–57 [PubMed: 27140627]
82. Mulder N, Abimiku A, Adebamowo S, de Vries J, Matimba A, et al. 2018. H3Africa: current perspectives. Pharmacogenom. Pers. Med 10:59–66
83. Myers S, Fefferman C, Patterson N. 2008. Can one learn history from the allelic spectrum? Theor. Popul. Biol 73:342–48 [PubMed: 18321552]
84. Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, et al. 2019. The formation of human populations in South and Central Asia. Science 365:eaat7487 [PubMed: 31488661]
85. Nordborg M 2019. Coalescent theory. In Handbook of Statistical Genomics, ed. Balding DJ, Moltke I, Marioni J, pp. 145–76. Hoboken, NJ: Wiley. 4th ed.
86. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko A, et al. 2008. Genes mirror geography within Europe. Nature 456:98–101 [PubMed: 18758442]
87. Novembre J, Ramachandran S. 2011. Perspectives on human population structure at the cusp of the sequencing era. Annu. Rev. Genom. Hum. Genet 12:245–74
88. Orlando L, Allaby R, Skoglund P, Der Sarkissian C, Stockhammer PW, et al. 2021. Ancient DNA analysis. Nat. Rev. Methods Primers 1:14
89. Pagani L, Colonna V, Tyler-Smith C, Ayub Q. 2017. An ethnolinguistic and genetic perspective on the origins of the Dravidian-speaking Brahui in Pakistan. Man India 97:267–78 [PubMed: 28381901]

90. Palamara PF, Lencz T, Darvasi A, Pe'er I. 2012. Length distributions of identity by descent reveal fine-scale demographic history. Am. J. Hum. Genet 91:809–22 [PubMed: 23103233]
91. Patterson N, Hattangadi N, Lane B, Lohmueller K, Hafler D, et al. 2004. Methods for high-density admixture mapping of disease genes. Am. J. Hum. Genet 74:979–1000 [PubMed: 15088269]
92. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, et al. 2012. Ancient admixture in human history. Genetics 192:1065–93 [PubMed: 22960212]
93. Patterson N, Price A, Reich D. 2006. Population structure and eigenanalysis. PLOS Genet 2:e190 [PubMed: 17194218]
94. Pease JB, Hahn MW. 2015. Detection and polarization of introgression in a five-taxon phylogeny. Syst. Biol 64:651–62 [PubMed: 25888025]
95. Petkova D, Novembre J, Stephens M. 2016. Visualizing spatial population structure with estimated effective migration surfaces. Nat. Genet 48:94–100 [PubMed: 26642242]
96. Pickrell JK, Patterson N, Loh P, Lipson M, Berger B, et al. 2014. Ancient west Eurasian ancestry in southern and eastern Africa. PNAS 111:2632–7 [PubMed: 24550290]
97. Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. Nat. Preced 10.1038/npre.2012.6956.1
98. Pickrell JK, Reich D. 2014. Toward a new history and geography of human genes informed by ancient DNA. Trends Genet 30:377–89 [PubMed: 25168683]
99. Pool JE, Nielsen R. 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181:711–19 [PubMed: 19087958]
100. Price A, Tandon A, Patterson N, Barnes K, Rafaels N, et al. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLOS Genet 5:e1000519 [PubMed: 19543370]
101. Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotypes data. Genetics 155:945–59 [PubMed: 10835412]
102. Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M. 2011. Dating the age of admixture via wavelet transform analysis of genome-wide data. Genome Biol 12:R19 [PubMed: 21352535]
103. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, et al. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505:87–91 [PubMed: 24256729]
104. Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197:573–89 [PubMed: 24700103]
105. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-wide inference of ancestral recombination graphs. PLOS Genet 10:e1004342 [PubMed: 24831947]
106. Regalado A 2019. More than 26 million people have taken an at-home ancestry test. MIT Technol. Rev, Feb. 11. https://www.technologyreview.com/2019/02/11/103446/more-than-26-million-people-have-taken-an-at-home-ancestry-test
107. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, et al. 2012. Reconstructing Native American population history. Nature 488:370–74 [PubMed: 22801491]
108. Reich D, Schaffner SF, Daly MJ, McVean G, Mullikin JC, et al. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. Nat. Genet 32:135–42 [PubMed: 12161752]
109. Reich D, Thangaraj K, Patterson N, Price A, Singh L. 2009. Reconstructing Indian population history. Nature 461:489–94 [PubMed: 19779445]
110. Ringbauer H, Novembre J, Steinrücken M. 2021. Parental relatedness through time revealed by runs of homozygosity in ancient DNA. Nat. Commun 12:5425 [PubMed: 34521843]
111. Salter-Townshend M, Myers S. 2019. Fine-scale inference of ancestry segments without prior knowledge of admixing groups. Genetics 212:869–89 [PubMed: 31123038]
112. Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. 2012. The date of interbreeding between Neandertals and modern humans. PLOS Genet 8:e1002947 [PubMed: 23055938]
113. Schiffels S, Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. Nat. Genet 46:919–25 [PubMed: 24952747]

114. Scott EM, Halees A, Itan Y, Spencer EG, He Y, et al. 2016. Characterization of greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat. Genet 48:1071–76 [PubMed: 27428751]
115. Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, et al. 2012. Origins and genetic legacy of neolithic farmers and hunter-gatherers in Europe. Science 336:466–69 [PubMed: 22539720]
116. Speidel L, Cassidy L, Davies RW, Hellenthal G, Skoglund P, Myers SR. 2021. Inferring population histories for ancient genomes using genome-wide genealogies. Mol. Biol. Evol 38:3497–511 [PubMed: 34129037]
117. Speidel L, Forest M, Shi S, Myers S. 2019. A method for genome-wide genealogy estimation for thousands of samples. Nat. Genet 51:1321–29 [PubMed: 31477933]
118. Stern AJ, Wilton PR, Nielsen R. 2019. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLOS Genet 15:e1008384 [PubMed: 31518343]
119. Sun N, Zhao H. 2020. Statistical methods in genome-wide association studies. Annu. Rev. Biomed. Data Sci 3:265–88
120. Terhorst J, Kamm JA, Song YS. 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat. Genet 49:303–9 [PubMed: 28024154]
121. Thompson EA. 1975. Human Evolutionary Trees Cambridge, UK: Cambridge Univ. Press
122. Tournebize R, Chu G, Moorjani P. 2022. Reconstructing the history of founder events using genome-wide patterns of allele sharing across individuals. PLOS Genet 18:e1010243 [PubMed: 35737729]
123. van Dorp L, Balding D, Myers S, Pagani L, Tyler-Smith C, et al. 2015. Evidence for a common origin of blacksmiths and cultivators in the Ethiopian Ari within the last 4500 years: lessons for clustering-based inference. PLOS Genet 11:e1005397 [PubMed: 26291793]
124. Waddell P 1996. Evolutionary trees of apes and humans from DNA sequences. In Handbook of Human Symbolic Evolution, ed. Lock A, Peters CR, pp. 53–73. Oxford, UK: Blackwell
125. Wakeley J 2009. Coalescent Theory: An Introduction New York: Macmillan Learn.
126. Wangkumhang P, Greenfield M, Hellenthal G. 2022. An efficient method to identify, date, and describe admixture events using haplotype information. Genome Res 32:1553–64 [PubMed: 35794007]
127. Zhou Y, Browning S, Browning B. 2020. A fast and simple method for detecting identity-by-descent segments in large-scale data. Am. J. Hum. Genet 106:426–37 [PubMed: 32169169]

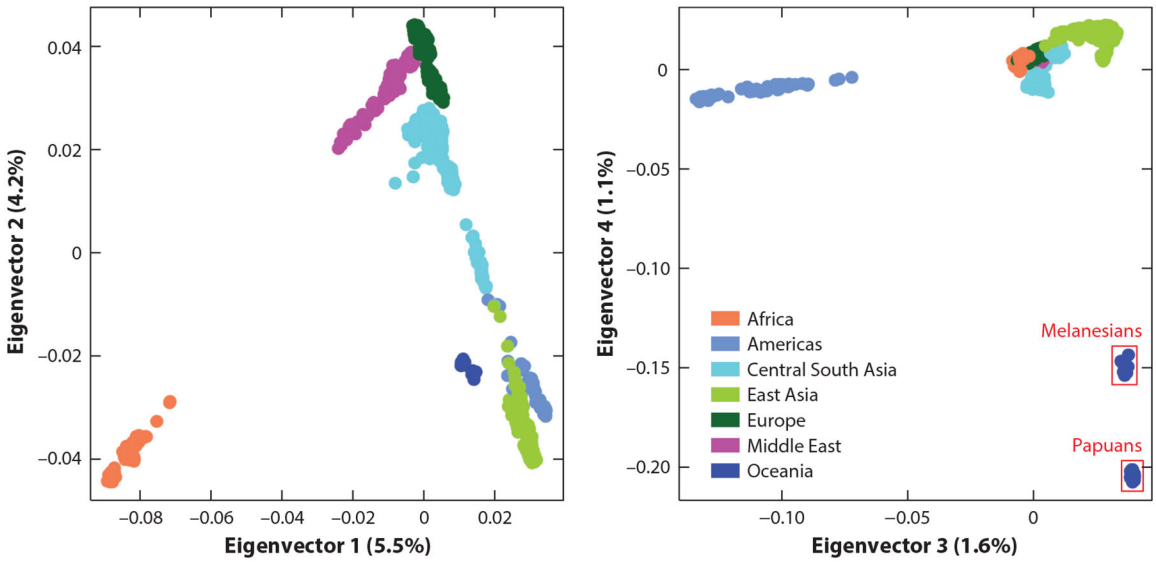

Figure 1.

First four eigenvectors of a PCA on HGDP array data. Each point is an individual, represented by a linear combination of their genome-wide SNP data and colored by the major geographic region they were sampled from. In the right plot, Melanesians and Papuans are highlighted with red boxes. The parentheses on each axis give the proportion of overall data variation explained by each eigenvector. To limit the effects of LD, we ran PCA on 116,142 SNPs with a minor allele frequency of >5%, where all SNPs within 250 kb of each other had a Pearson correlation (r^2) of ≤0.2. However, we note that the results are very similar when analyzing all 474,491 SNPs. Abbreviations: HGDP, Human Genome Diversity Project; LD, linkage disequilibrium; PCA, principal component analysis; SNP, single-nucleotide polymorphism.

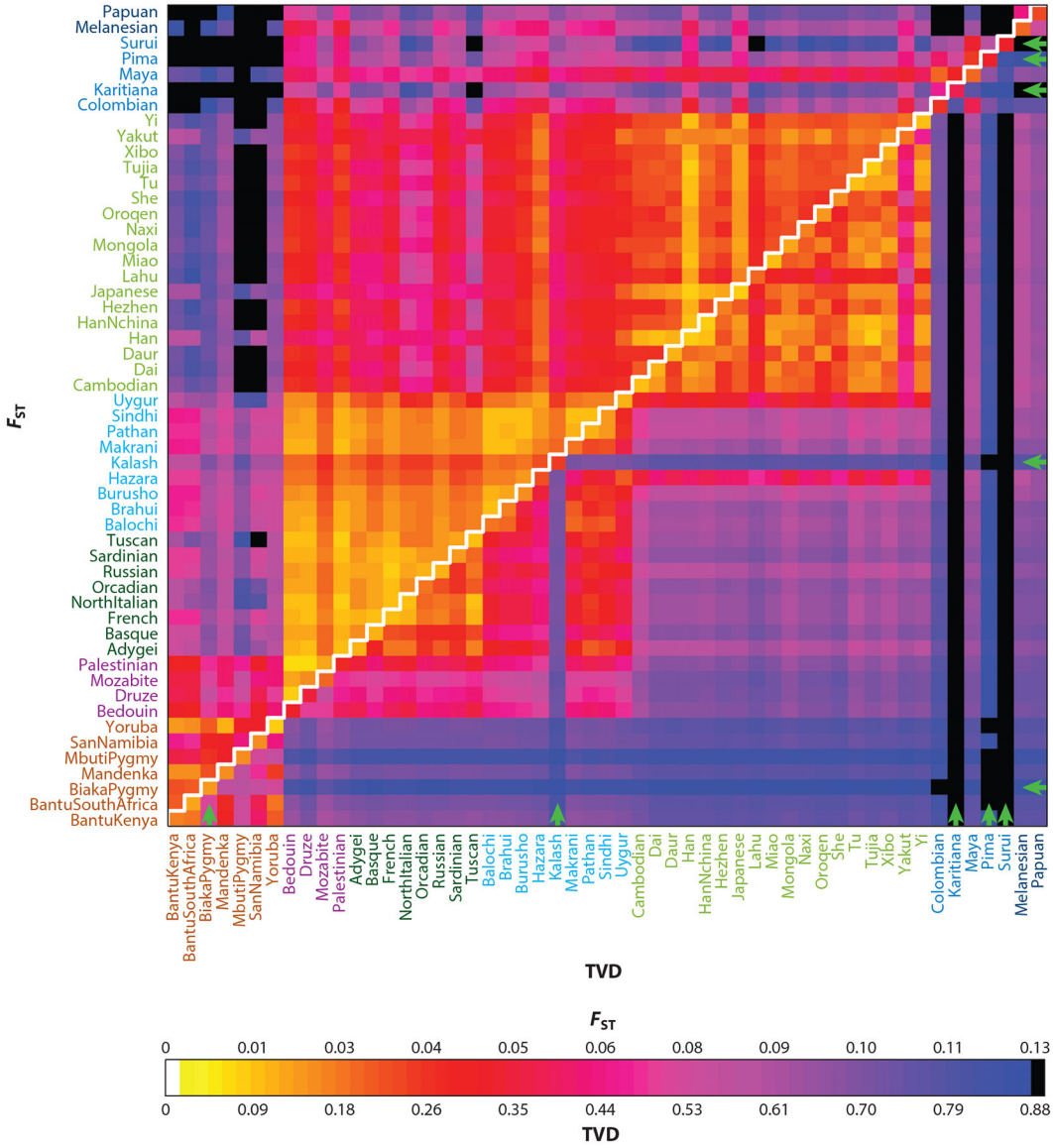

Figure 2.

F_{ST} (*upper left triangle*) and TVD (*lower right triangle*) between individuals from different HGDP groups, calculated using the array data. The diagonal gives the average TVD among individuals from the same group. The color bar shows the scales for both F_{ST} (*top*) and TVD (*bottom*). The population labels on the axes are colored according to major geographic region as in Figure 1. Populations highlighted with green arrows show evidence of endogamy under TVD (i.e., exhibiting relatively high TVD with other populations). Abbreviations: HGDP, Human Genome Diversity Project; TVD, total variation distance.

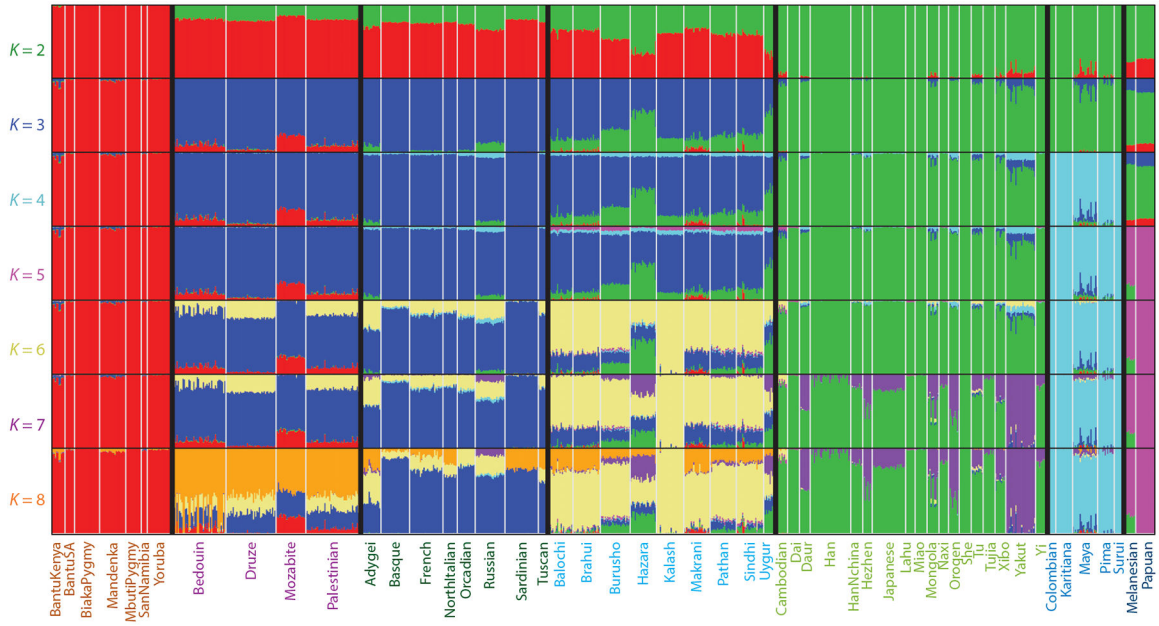

Figure 3.

ADMIXTURE results for HGDP array data, assuming $K = 2–8$ clusters (*rows*), when analyzing 116,142 filtered SNPs, as in the PCA analysis shown in Figure 1. Each vertical bar is a person, and the bars' colors represent the different clusters. Black vertical bars separate major geographic regions, and gray bars separate populations. The population labels at the bottom are colored by major geographic region as in Figure 1. The *y* axis (K) labels are colored according to the new cluster color that emerges for that K. Abbreviations: HGDP, Human Genome Diversity Project; PCA, principal component analysis; SNP, single-nucleotide polymorphism.

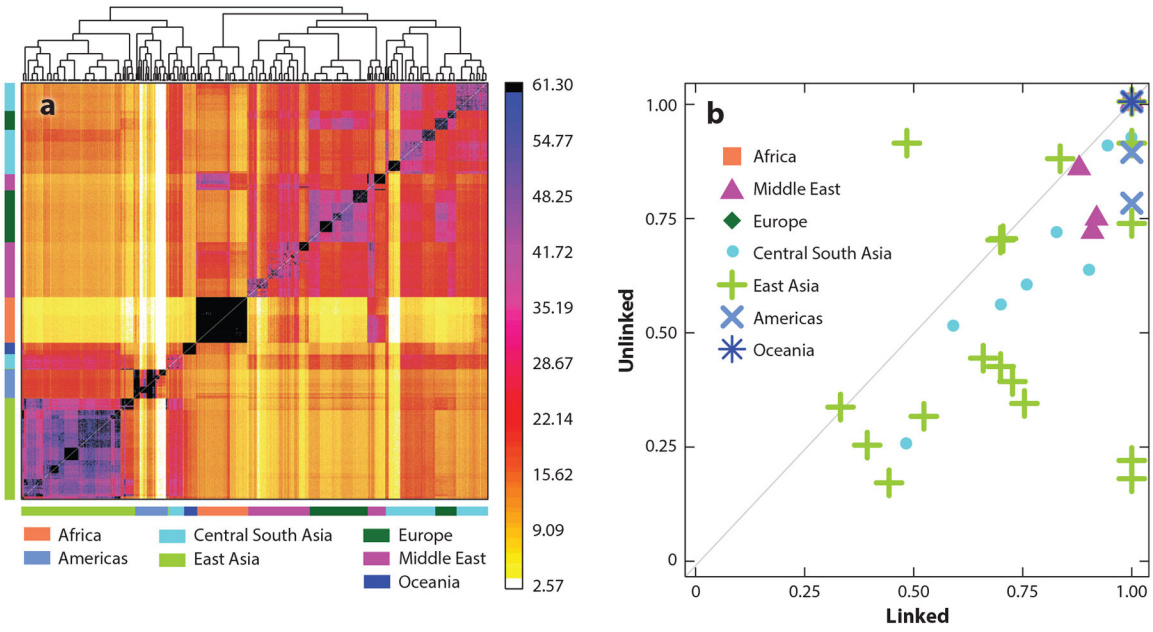

Figure 4.

(*a*) Number of genome-wide haplotype segments for which each individual (*column*) is inferred to share an ancestor most recently with each other person (*row*), when applying CHROMOPAINTER (62) to the HGDP array data. Here, we analyze all 474,491 SNPs, since the model does not assume that SNPs are independent and indeed leverages correlations among dense SNP data. Axes are colored by individuals' major geographic regions. The dendrogram at the top shows the merging of fineSTRUCTURE's (62) inferred clusters of individuals with the final clusters given at the bottom of the tree. (*b*) Proportion of individuals from each HGDP population (*symbols*) that are assigned to a cluster that contains only other individuals from the same population, as inferred by fineSTRUCTURE under the linked model that leverages LD information (which inferred $K = 129$ clusters) versus the unlinked model that ignores it (which inferred $K = 120$ clusters). Abbreviations: HGDP, Human Genome Diversity Project; LD, linkage disequilibrium; SNP, single-nucleotide polymorphism.

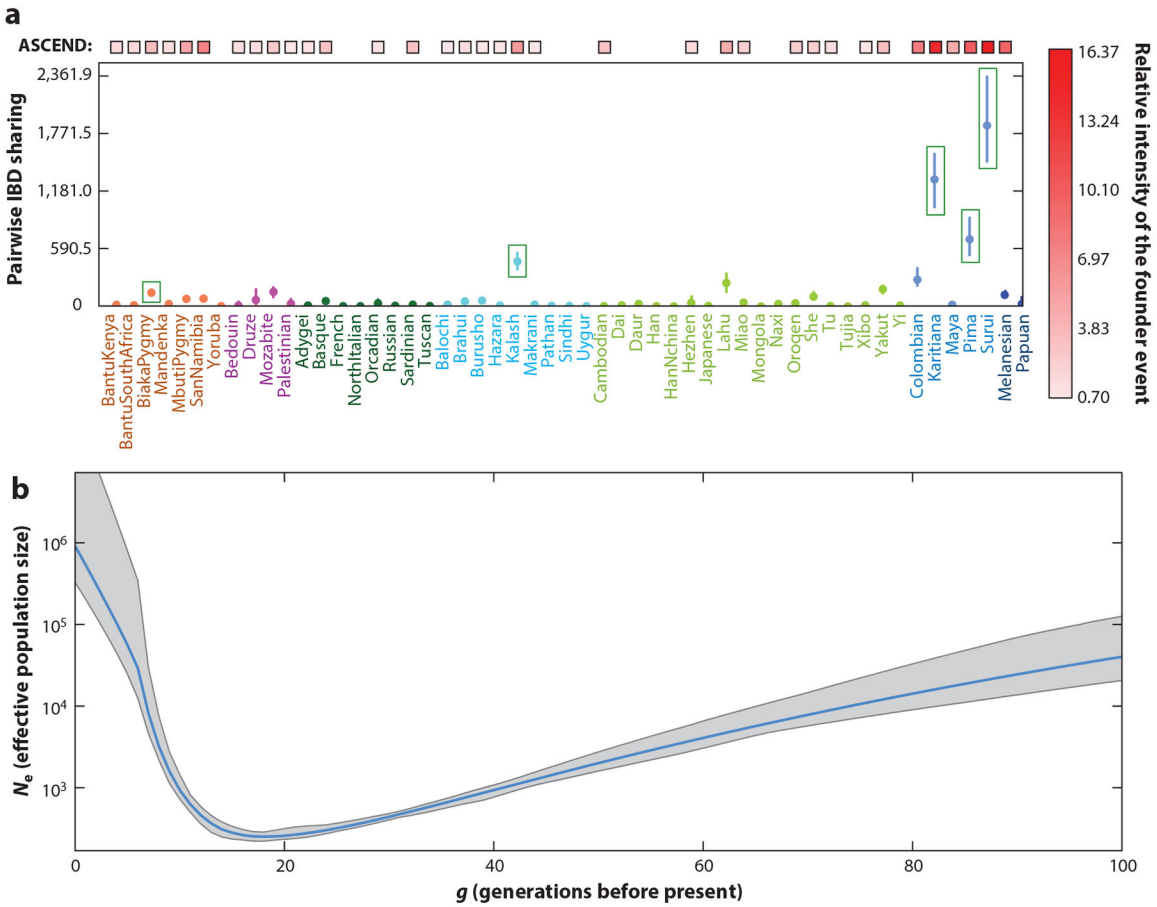

Figure 5.

Patterns of genome-wide IBD sharing in the HGDP dataset. (*a*) Median (*dots*) and interquartile range (*lines*) of IBD sharing across pairs of individuals in a population, measured as the total summed lengths of shared IBD segments that are ≥2 cM as inferred by applying hap-ibd (127). Green rectangles correspond to populations highlighted with arrows in Figure 2. For populations where ASCEND (122) infers evidence of a founder event, red boxes at the top are shaded according to relative intensity of the founder event (*color bar*), with strong events inferred in Native American cohorts. (*b*) Estimated effective population size over the past 100 generations in the Kalash population inferred using IBDNe (13). Abbreviations: HGDP, Humane Genome Diversity Project; IBD, identity by descent.

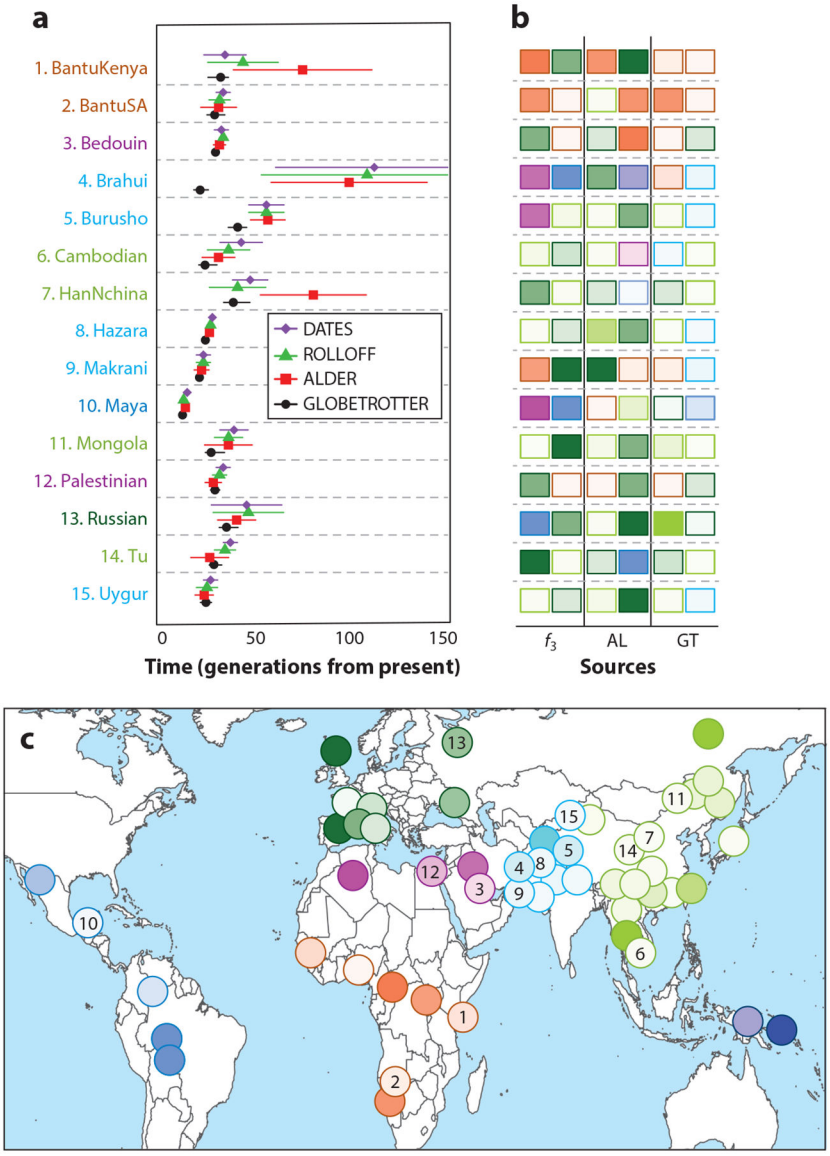

Figure 6.

Inferred admixture events in 15 HGDP populations. (*a*) Inferred dates (*symbols*) and 95% confidence intervals (*lines*) for four approaches. The population labels at left are colored by major geographic region as in Figure 1. GLOBETROTTER's inferred confidence intervals are based on bootstrap resampling, while all other approaches are based on chromosome-level jackknife resampling where one chromosome is removed in each run, which likely at least partially explains GLOBETROTTER's smaller intervals. (*b*) Surrogate populations chosen by f_3 test, ALDER (AL), and GLOBETROTTER (GT) to best represent the two admixing sources. We report the pair of surrogates with the most negative f_3 score, the highest-amplitude value in the decay curves when using both surrogates in ALDER, and the single surrogate population inferred to best represent each admixing source in GLOBETROTTER. The border color indicates the major geographic region of each chosen reference population; the internal color indicates the median amount of IBD sharing among

people in that chosen reference population, divided by the maximum such median IBD sharing among all populations from the same geographic region. (*c*) Indicative geographic locations of HGDP populations (*circles*), with circles' colors indicating the relative amount of IBD sharing (as in panel *b*) and their borders colored according to their major geographic region. The numbers correspond to target populations in panel *a*. Abbreviations: HGDP, Human Genome Diversity Project; IBD, identity by descent.

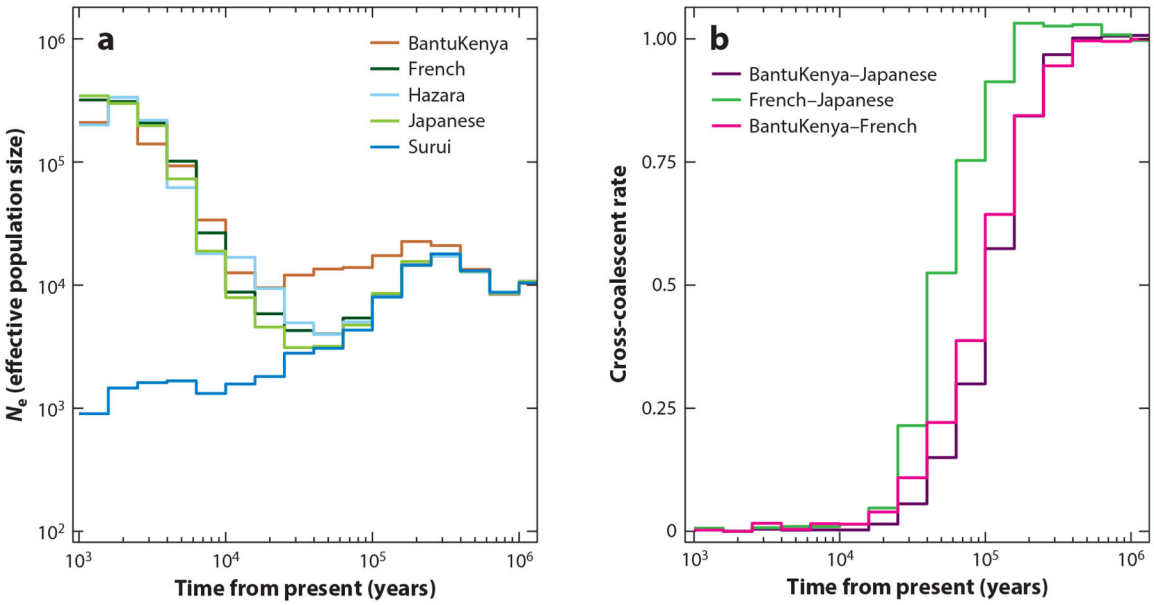

Figure 7.

Demographic inference in five HGDP populations, inferred by RELATE (117) using chromosome 1 and 4–5 people per HGDP population (262 people total) for computational simplicity. (*a*) Inferred effective population size (N_e) for each population over time [assuming 28 years per generation (81) and a mutation rate of 1.25×10^{-8} per base pair per generation (52)]. (*b*) Inferred cross-coalescent rate over time for individuals from different populations, which is the inferred rate of coalescence between individuals from different populations divided by the average of the two within-population coalescent rates. Before RELATE was run, BEAGLE 5.3 (11) was used to fill in missing genotypes, after which sites with >5% missingness were set as missing in the RELATE mask file. Abbreviation: HGDP, Human Genome Diversity Project.

Table 1

Datasets and software used for the analysis in this review

Data/software	Source	Reference(s)
Datasets		
HGDP SNP dataset	https://data.mendeley.com/datasets/ckz9mtgrjj73	
HGDP whole-sequence dataset	https://www.internationalgenome.org/data-portal/data-collection/hgdp	6
Quantifying and visualizing patterns of population relationships		
EIGENSOFT	https://www.hsph.harvard.edu/alkes-price/software	93
CHROMOPAINTER	https://people.maths.bris.ac.uk/~madjl/finestructure-old/chromopainter_info.html	62
Clustering algorithms		
ADMIXTURE	https://dalexander.github.io/admixture	2
fineSTRUCTURE	https://people.maths.bris.ac.uk/~madjl/finestructure/finestructure_info.html	62
Formal tests of admixture		
f statistics (including f_3, f_4, *D*, f_4 ratio test, qpGraph, qpWave, and qpAdm)	https://github.com/DReichLab/AdmixTools	92, 109
Inferring IBD sharing		
hap-ibd	https://github.com/browning-lab/hap-ibd	127
Dating admixture		
ROLLOFF		
Original statistic	https://github.com/DReichLab/AdmixTools	79, 92
Unbiased statistic	https://github.com/priyamoorjani/rolloff	80
ALDER	https://cb.csail.mit.edu/cb/alder	72
DATES	https://github.com/MoorjaniLab/DATES_v4010	24, 84
GLOBETROTTER	https://people.maths.bris.ac.uk/~madjl/finestructure/globetrotter.html	47
Characterizing population size changes		
ASCEND	https://github.com/sunyatin/ASCEND	122
IBDNe	https://faculty.washington.edu/browning/ibdne.html	13
Inferring genealogies		
RELATE	https://myersgroup.github.io/relate	117

Abbreviations: HGDP, Human Genome Diversity Project; IBD, identity by descent; SNP, single-nucleotide polymorphism.