Analysis of Population Structure

3

Per Sjödin, Lucie Gattepaille, Pontus Skoglund, Carina Schlebusch, and Mattias Jakobsson

Abstract

For humans, like any sexually reproducing diploid organism, mating may be random in the sense that individuals are equally likely to mate and produce offspring. Such a view of a population has been important in population genetics as a basis for modeling and analysis. Population structure denotes deviation from this panmixia, regardless of the cause. In this chapter, we will briefly discuss random mating, populations, population structure, and various methods and practices to infer population structure among individuals from empirical genome-wide data.

3.1 What Is Population Structure?

A loose definition of a "panmictic population" is any collection of randomly mating individuals living at a specific point in time. Strictly speaking, a population is "randomly mating" if the probability to produce offspring is larger than zero

P. Sjödin · L. Gattepaille ·
Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden

P. Skoglund
Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden

The Francis Crick Institute, London, UK

C. Schlebusch · M. Jakobsson (✉)
Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden

Science for Life Laboratory, Uppsala University, Uppsala, Sweden

Department of Anthropology and Development Studies, Centre for Anthropological Research, University of Johannesburg, Auckland Park, South Africa
e-mail: mattias.jakobsson@ebc.uu.se

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

K. E. Lohmueller, R. Nielsen (eds.), *Human Population Genomics*,
https://doi.org/10.1007/978-3-030-61646-5_3

and equals for all possible pairs of individuals drawn from the population (often, depending on the species, given that a pair consists of a female and a male).

Random mating in this sense is, however, rarely an accurate representation of reality since most populations have a spatial distribution that affects the mating probability of a random pair of individuals. Hence, if by "structured population," we refer to a population that is not randomly mating, then essentially all populations are structured, and a dual categorization of "structured populations" vs "unstructured populations" is not very useful. Instead, it may be better to think of any population as structured and attempt to quantify the degree of structure. The level of population structure can also (often) be accounted for in downstream analyses. In practice, there are often groups of individuals for which no structure is detectable (with the data and methods at hand), and these groups can be regarded as unstructured for most intents and purposes.

It should be noted that this way of thinking about populations does not reflect common practice in which the definition of populations is typically subjective, based on, for example, linguistic, cultural, ethnic, and/or the geographic location of sampled individuals. Moreover, almost all population structure analyses are necessarily based on samples, and detection of stratification within the sample does not necessarily reflect biological populations. For instance (following Pritchard et al. (2000)), imagine a species that lives on a continuous plane but has a low dispersal rate, so allele frequencies vary continuously across the plane. A few clustered sampling points will result in a signal of (a few) clustered genetic groups, which does not give an accurate description of the biological reality. This example illustrates that studies of populations are always indirect and limited to information contained in samples and sensitive to sampling biases.

Accounting for population structure is often crucial in order to reduce both type I and type II errors in statistical analyses of genetic data. For instance, it has been shown that not accounting for population structure can result in spurious signals in association mapping studies and will thus invalidate standard tests (Ewens and Spielman 1995; Pritchard et al. 2000). It is also important to account for population structure in forensic applications like DNA fingerprinting to estimate the probability of random individuals matching a particular profile (Balding and Nichols 1994, 1995; Foreman et al. 1997; Roeder et al. 1998).

Since the first historical opportunity of quantifying molecular genetic variation until today's (almost) complete genome sequencing, numerous molecular techniques have been used to genotype individuals, which in turn can be used to investigate population structure. Early strategies involved typing the human blood groups (Landsteiner and Weiner 1940), followed by the development of allozymes (Lewontin and Hubby 1966), and various forms of DNA fragment length assays, including microsatellites that are abundant in eukaryote genomes (Katti et al. 2001). The sequencing of the human genome (Venter et al. 2001) led to access to a large number of human single-nucleotide polymorphisms (SNPs) as well as the human genome sequence. Various studies have employed novel molecular techniques to investigate human population structure, including classical markers (Cavalli-Sforza et al. 1994), mitochondrial genome (Cann et al. 1987), microsatellites

(Rosenberg et al. 2002), SNPs (Jakobsson et al. 2008; Li et al. 2008), and complete genomes (1000 Genomes, Mallick et al. 2016). Although these different types of data have different properties, the general study of population structure follows straightforward principles of genetic variation, and the difference in datatype can be accounted for by using different assumptions on the mutation model (see e.g., Veeramah and Hammer 2014).

We will review some methods to quantify structure within a population. We will (primarily) assume biallelic markers, where only two states exist that give rise to three possible genotypes. These methods are often part of an initial exploratory step in order to get a better picture of how and to what extent the sample and the population have been influenced by population structure. First, we will present methods where all individuals are treated equally and no a priori information is used for clustering individuals or parts of individual genomes. We will then discuss some methods to contrast clusters of individuals and how this can be used in more explicit demographic modeling. To illustrate concepts and methods, we will analyze a subset of the HGDP data, which has been thoroughly investigated in previous studies (Cann et al. 2002; Li et al. 2008; Jakobsson et al. 2008). This example dataset consists of individuals from Africa—the West African Yoruba, the Southern African San, and the North African Mozabite—and from Europe (France).

3.2 Individual-Based and Unsupervised Methods for Inferring Population Structure

3.2.1 Tree Construction Methods at the Individual Level

Genetic distance is a traditional measure of differentiation in population genetics. Distances are calculated between each pair of individuals (can also be calculated between groups of individuals; see below) and are represented by a pairwise distance matrix. Distance matrices can be visualized by various approaches such as the multidimensional statistics discussed in the next section or in the form of graphs/trees.

A common distance measure between a pair of individuals is the identity by state (IBS) measure. IBS examines biallelic SNPs between two individuals and puts them into one of three categories: identical = 1 (e.g., for the genotypes AA and AA, BB and BB, and AB and AB, where A and B denote the two alleles), one-allele-shared = 0.5 (i.e., AA and AB; AB and BB), and no-allele-shared = 0 (i.e., AA and BB). The state-values are then averaged over all loci to provide genome-wide pairwise IBS similarity values between 1 and 0 for all individual pairwise comparisons. This is summarized in an individual similarity matrix of which 1-IBS will give the distance matrix.

Distance measures based on substitution models for DNA and protein sequence evolution have also been developed. The evolutionary distance between a pair of sequences is measured by the number of nucleotide (or amino acid) substitutions occurring between them. The p-distance is the simplest model and is based on the proportion (p) of nucleotide sites at which two sequences being compared

are different. The proportion is obtained by dividing the number of nucleotide differences by the total number of nucleotides compared. It does not make any correction for multiple substitutions at the same site, substitution rate biases (for example, differences in the transition and transversion rates), or differences in evolutionary rates among sites. More specialized measures (e.g., Jukes-Cantor, Kimura 2-parameter, Tamura 3-parameter, Tamura-Nei) take some/more of these complexities into account. The pairwise distances between diploid individuals are then obtained by averaging over the distances obtained for all pairwise comparisons of sequences between the two individuals.

There are several tree construction methods for distance data, two of the most common methods are unweighted pair group method with arithmetic mean (UPGMA) and neighbor joining (NJ) (Saitou and Nei 1987). UPGMA is a simple hierarchical clustering method that combines the nearest two units (individuals or grouped individuals) in a distance matrix into a higher-level cluster. The distance between any two units is the average of all distances between each element of each unit. UPGMA assumes a constant molecular clock model and produces an ultrametric tree (a tree where all the path lengths from the root to the tips are of equal length). Similar to UPGMA, neighbor joining is also a bottom-up clustering method; however, compared with UPGMA, neighbor joining has the advantage that it does not assume that all lineages evolve at the same rate. Both NJ and UPGMA are fast-clustering (tree-building) algorithms, but since only two elements of the distance matrix are considered at a time, they have no optimization criterion to fit the best solution (or tree) over all the data. An optimal criterion method that is commonly applied to distance data is minimum evolution (ME), which accepts the tree with the shortest sum of branch lengths, and thus minimizes the total amount of evolution assumed. Tree-building methods applicable to discrete characters such as nucleotides are also available (e.g., maximum parsimony and maximum likelihood), but these are more applicable to phylogenetic purposes and fall outside the scope of this chapter. Note that inferred trees based on non-recombining chromosomes, such as the mitochondrial genome or the Y chromosome, represent estimates of the genealogy of a specific chromosome (see Chap. 1 for a review of gene genealogies) that may poorly capture an individual's or a population's evolutionary history or structure. Inferred trees based on genome-wide data represent averages over the genealogical process across the genome, and such summary trees capture population structure in a more accurate way.

After an initial tree is produced from the distance matrix, a confidence measure can be calculated making use of procedures such as jackknifing or bootstrapping. The most widely used tool for confidence inference is a version of bootstrapping introduced by Felsenstein (1983). Each bootstrap sample consists of the same number of markers resampled (with replacement) from the original data set and is then subjected to the same distance calculation and tree reconstruction. From these trees produced by bootstrapping, a consensus tree can be constructed in which the confidence of the tree is noted on the nodes as a bootstrap value (the percentage of times the bootstrap procedure supported the specific node). Jackknife is a similar resampling procedure, but in this case, the estimate is systematically recomputed by leaving out one or more observations at a time from the sample set. The bootstrap

and jackknife assume near-independence of the markers used, and so the linkage between nearby markers can be accounted for by performing the bootstrap or jackknife procedure over larger blocks of the genome or whole chromosomes.

Various software packages that can handle different types of genetic data are able to calculate many of the distance measures discussed above and give a pairwise distance matrix as output. The distance matrix can then be used for clustering using, e.g., a tree construction algorithm. In certain software, both matrix calculation and tree construction algorithms are available. Some examples of commonly used software are Mega (Kumar et al. 2016), Arlequin (Excoffier and Lischer 2010), and Plink (Purcell et al. 2007).

An example tree of West African Yoruba, European French, North African Mozabites, and Southern African San is given in Fig. 3.1. To construct the tree,

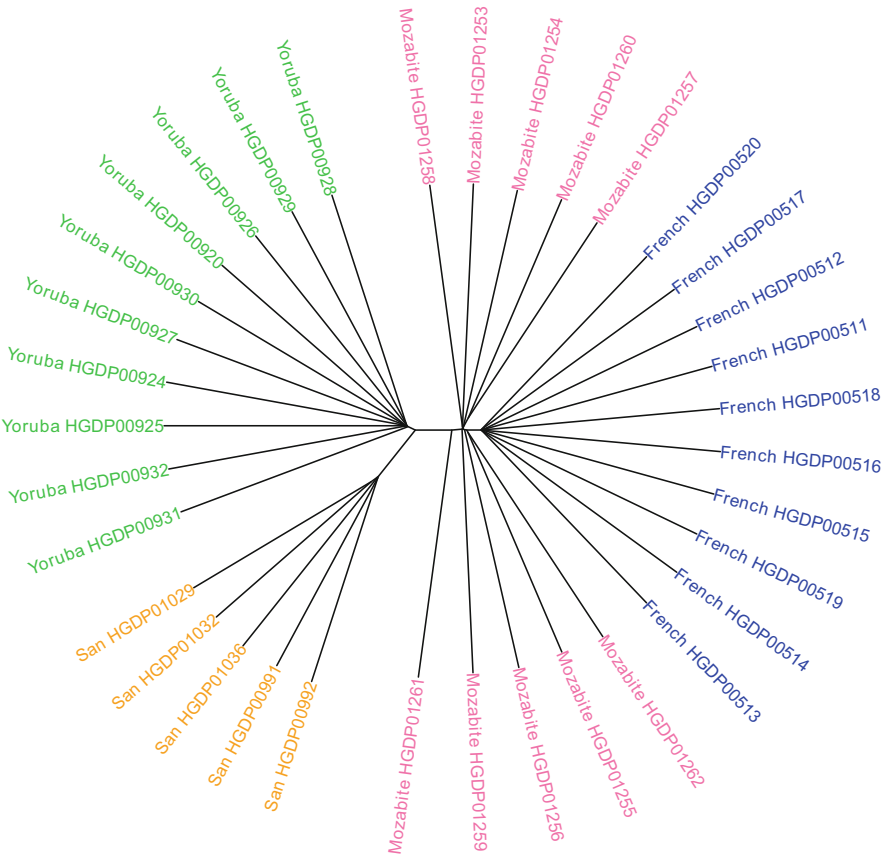

Fig. 3.1 An example neighbor-joining tree of individuals from West Africa (Yoruba), Europe (French), North Africa (Mozabite), and Southern African (San). The NJ tree was built from an IBS distance matrix (computed in Plink) and the R-package *ape*. The individuals cluster according to their sample locations, except for the Mozabites that contain the French sample as a subset

we created an IBS distance matrix in Plink and used the R-package *ape* to construct an NJ tree. We see that the individuals cluster quite well according to their sample locations, except for the Mozabites that contain the French sample as a subset.

3.2.2 Principal Component Analysis and Related Approaches

A single individual can be represented by a position in a multidimensional space where each locus characterizes one dimension. The number of dimensions is therefore very large if we have access to information from many sites; oftentimes, however, many of these dimensions are correlated. A number of methods based on linear algebra aim at finding the best way to summarize and visualize the data in a reduced number of dimensions that capture the greatest axes of variation. These methods, which are typically agnostic to the underlying model of genetic variation, can potentially reveal inherent population structure in a set of sampled individuals. We will describe one of the most widely used method for initial data exploration—principal component analysis (PCA)—and then give a brief overview of a few other related approaches.

The principle of PCA is straightforward: finding and ordering orthogonal axes (or principal components, PCs) that capture the variation of the sample so that the first PC represents the axis of greatest variation in the data, the second PC represents the axis of greatest remaining variation when the data is projected orthogonally to the first PC, and so on, down to the last PC where all variation has been taken into account. Consider, for example, n SNP-loci. Each individual is represented by a vector of n values, with 0 if homozygous for the reference allele, 1 if homozygous for the alternative allele, and 0.5 if heterozygous. PCA performs a rotation of the original n-dimensional orthogonal base, where each of the n loci represents one dimension, into a new orthogonal base, formed by linear combinations of the loci, and defining directions called principal components. The first PC is the direction that maximally explains the variance among individuals when projected into a 1-dimensional space. Together with the first PC, the second PC defines the plane that maximizes the variance of the individuals when projected into a 2-dimensional orthogonal space, and more generally, together with the first k-1 PCs, PC k defines an orthogonal space that maximizes the variance of the individuals when projected into a k-dimensional space. Each PC explains a proportion of the total variance, with the first PC explaining the most variance and the last PC explaining the least.

Applied to our example data of four populations, PCA reveals differences between the groups (Fig. 3.2). We see that the first and second PCs explain 10.8% and 5.5% of the total sample variance, respectively. The first PC separates the sub-Saharan African populations from the European and the North African population; the second PC separates the Southern African San and the West African Yoruba, with the French and the Mozabite in-between. The first two PCs together show the Mozabite individuals in-between the European population and the West African population, suggesting historical models for the ancestry of the Mozabite such as (1) a history of admixture between European and West African groups, (2) shared

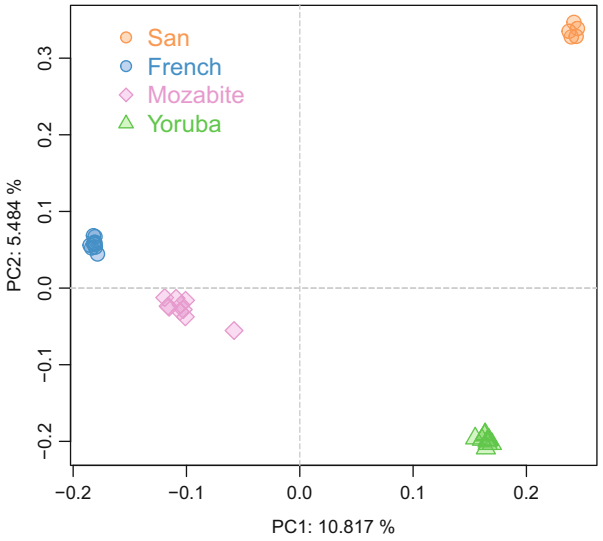

Fig. 3.2 Principal component analysis of our example HGDP data of four populations, with the first two PCs displayed, computed using EIGENSOFT. The first PC explains 10.8% of the total sample genetic variance, and the second PC explains 5.5%. The first PC separates the sub-Saharan African populations from the European and the North African population, and the second PC separates the Southern African San and the West African Yoruba, with the French and the Mozabite in-between. The first two PCs together show the Mozabite individuals distributed between the European population and the West African population, consistent with the Mozabite being an admixed group with a European and a West African source population

ancestry in a treelike population model (without direct admixture) with both these groups, or (3) a combination of the aforementioned (1) and (2).

Investigating outliers, which are easily identifiable using, e.g., PCA, is an important step in many applications. In this particular example, there are no obvious outlier individuals. Outliers are easily identified as individuals "far away" from any other cluster of individuals in PC space. Outliers can be due to low genotype quality (for particular individuals), recent migrants from unsampled populations, or displaying some, potentially unknown, level of population structure in the sample.

The number of PCs to visualize and investigate is arbitrary, but some rule of thumb has been utilized in past studies. Sometimes, this number can be decided by a threshold of variance to be explained (for example, investigate the K PCs that explain at least a fixed percentage, X, of the variance). However, for genome-wide data, the variance explained by each PC, including the first few ones, is typically small due to the high dimensionality (see e.g., Fig. 3.2). Another way to choose the number of PCs to investigate would be to use a "scree plot." A property of PCA is that each PC represents an eigenvector of the variance-covariance matrix of the data (or the correlation matrix if the data is scaled) and is associated with an eigenvalue. The first PC is associated with the highest eigenvalue λ_1, the second PC with the

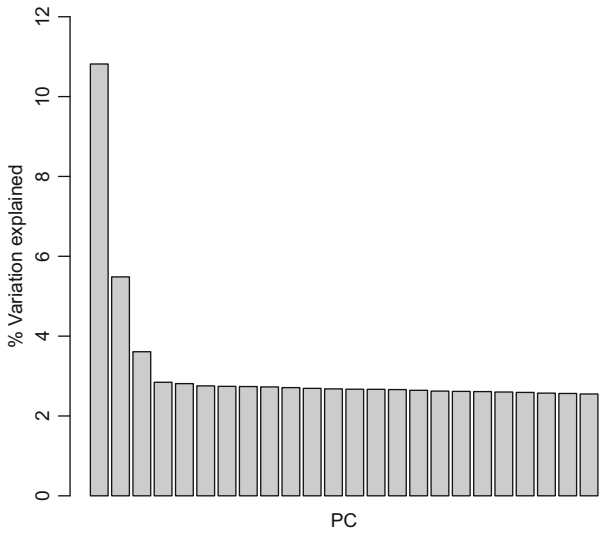

Fig. 3.3 The eigenvalues associated with each PC from our example PCA on HGDP data, sorted decreasingly. In a PCA, the eigenvalues of the variance-covariance matrix (or correlation matrix if the data is scaled) are directly linked to the variance explained by the PCs: the first PC is associated with the highest eigenvalue, the second PC to the second-highest eigenvalue, and so on

second-highest eigenvalue λ_2, and so on. The scree plot is the graph that displays the eigenvalues, sorted decreasingly. In a PCA, the eigenvalues of the variance-covariance matrix (or correlation matrix if the data is scaled) are directly linked to the variance explained by the PCs: λ_i divided by the sum of all eigenvalues gives the proportion of variance explained by the *i*th PC. In our HGDP example, there is a clear flattening of the difference between consecutive PCs from PC4 and onward suggesting that the most interesting patterns can be seen using the first three PCs (see Fig. 3.3). An alternative approach is to test each PC if there is significant evidence for structure (Patterson et al. 2006).

PCA is an important tool for data exploration. For a richer mathematical description and illustrations in different contexts, see Jolliffe (2005). For population genetics, McVean (2009) showed that expected pairwise coalescent times is what determines the primary PCs in a PCA implying that it is impossible to distinguish models with the same expected coalescent times using a PCA approach. He also demonstrated how PCA can, under some models, be used to estimate divergence time between populations, as well as admixture proportions within individuals (McVean 2009). There are several software packages that compute PCA including the R prcomp package and Eigensoft (Patterson et al. 2006).

Multidimensional scaling (MDS) is a group of methods that use a matrix of dissimilarities between individuals and represent the individuals in a smaller number of dimensions, so that the pairwise distances between individuals in the plotting space are good approximations of the original dissimilarities (see Quinn and Keough

2002, for a review on some of these methods). The dissimilarity measure of the individuals in the original data is the scientist's choice, and some examples of useful dissimilarity measures were presented in the previous section. The number of dimensions in the plotting space is chosen in advance and is usually low, to facilitate observation and interpretation of the data.

3.2.3 Ancestry Component Estimation with Few Model Assumptions

The program STRUCTURE (Pritchard et al. 2000) implements a method that makes very few assumptions about the data, and it was one of the first tools that utilized the power of multiple markers for inference. This approach and the many ensuing approaches have become standard in population structure investigations and population-genetic studies in general. STRUCTURE-like methods infer a predefined number of ancestry components (K) among individuals, based on genotype frequencies. Each individual's genotype is assigned to one of K number of clusters with a certain probability. In the first implementation, STRUCTURE searched for the assignment of individuals that minimizes deviation from Hardy-Weinberg equilibrium (see Box 3.1) and linkage equilibrium in each of the K clusters and allowed individuals to be admixed and to have membership proportions to more than one of the K clusters (Pritchard et al. 2000). Population structure is then visible in the dataset as individuals that are closely related having a greater proportion of their genome assigned to the same cluster/s than individuals that are not. This approach analyzed single markers separately and then added up the information to produce a global estimate for each individual. Information about the relative positions of markers to each other was not used and was considered to be segregating independently. This approach works well for low-density marker sets but is less suitable for the high density and full genome datasets that are available today. In the 2003 update of the STRUCTURE algorithm (Falush et al. 2003), sites/markers are not required to be independent, and correlations between subsequent markers due to admixture events are explicitly modeled. This allowed for individual ancestry estimates, known as local ancestry estimates, where the ancestry of chromosomal chunks can be traced along the chromosomes. It also introduced a simplistic model (the F-model, originally described in Nicholson et al. (2002)) to account for correlations of allele frequencies between populations. Although a clearly unrealistic model, it improved the performance of the algorithm considerably (Falush et al. 2003).

Box 3.1 Hardy–Weinberg Equilibrium (HWE)
An assumption of random mating is that the probability to produce viable offspring is equal for all possible pairs of individuals drawn from the

(continued)

population (given that a pair consists of a female and a male). A consequence of this is that the probability of an allele contributed by the mother being of type A is equal to the population frequency of allele type A. The same is true for alleles contributed by the father. Given a population frequency p of allele A and $1-p$ of alleles of type a, the probability of an offspring being of type AA, Aa, and aa are p^2, $2p(1-p)$, and $(1-p)^2$, respectively. When this situation is true, the population is said to be in Hardy-Weinberg equilibrium (HWE).

STRUCTURE has been very popular for population structure inference. However, with the ever-increasing density of genome-wide markers, meeting the computational demands of the algorithm has become a challenge. The Markov chain Monte Carlo method that STRUCTURE employs places a high burden on computer resources for large datasets. This has led to the recent development of alternative approaches, using fast maximum-likelihood-based estimations (FRAPPE (Tang et al. 2005) and ADMIXTURE (Alexander et al. 2009)).

HAPMIX (Price et al. 2009) extends the local ancestry method implemented in the second version of STRUCTURE. It is based on the Li and Stephens (2003) model for patterns of linkage disequilibrium (Li and Stephens 2003) between markers and infers local ancestry estimates of unphased admixed individuals based on the phased haplotype data of exactly two populations. Modeling the full demographic process with recombination and mutation is a notoriously difficult and computationally intractable problem. The Li and Stephens (2003) approach models the $k + 1$ haplotype by (imperfectly) copying from the first k "parental" haplotypes where recombination events correspond to changing parental haplotype from which to copy from. Correlations of genealogies (due to linkage) across the sequence is surprisingly well captured by this approach, and importantly, it is sufficiently simple to permit even full genome analyses.

Recent developments along this line have led to a method (CHROMOPAINTER, Lawson et al. (2011)) that does not need discretely defined admixed and parental populations. Instead, each individual in a sample is considered, in turn, both as a recipient and a donor, and chromosomes are reconstructed using blocks of DNA donated by the individuals to each other. Each individual's chromosome is thus "painted" by markers donated by donor individuals in any number of other populations or within the same population. These "painted" chromosomes can be summarized as a co-ancestry matrix, which is proposed to fully capture the information provided by PCA and STRUCTURE-like methods (also for nonindependent sites). In addition, consecutive markers that are in linkage disequilibrium are combined into haplotypes, which increases the ability of the method to observe subtle population structure. A downstream model-based extension (fineSTRUCTURE, Lawson et al. (2011)) is then used to identify discrete populations using the inferred co-ancestry matrix.

A subclass of population admixture models can utilize the spatial coordinates of sampled individuals (BAPS (Corander et al. 2003), TESS (Chen et al. 2007), GENELAND (Guillot et al. 2005)). While in STRUCTURE-like methods, the assignment to a population is independent and identical among all individuals in the dataset, this class of methods takes into account (a priori) the spatial distribution of individuals and aims to detect genetic discontinuities in space. Since geographic spatial correlation is often present among individuals and populations, it can be useful to incorporate spatial coordinates into the population structure analysis. Recent attempts to incorporate geographic information into population structure estimations involve approaches that use Wishart distributions (a generalization of multidimensional gamma distributions) to model genetic similarity as a function of spatial distance. In uniform isolation-by-distance scenarios, genetic distances visualized in two dimensions should mirror the samples/individuals in geographic space. Migration and admixture and hinders to gene flow would disturb this correlation. In one approach, implemented in the software SpaceMix, a covariance model of genetic data is used to build maps of the geographic positions of the populations, but distances are distorted according to inferred rates of gene flow (Bradburd et al. 2016). Barriers to gene flow result in larger distances between groups, while migration and admixture can be identified as abnormal strong covariances over long distances. The inferred admixture is then estimated and represented as "arrows," on a generated map, from the source population to the recipient population. Another approach based on the Wishart distribution, EEMS (Petkova et al. 2016), uses pairwise genetic similarities among populations and estimates a surface map of effective migration rates. The effective migration rates are scaled by effective population sizes under an equilibrium model. These methods that incorporate spatial information may highlight important features of population structure that might have remained undetected using other, spatially "blind," methods for inferring population structure. Two other methods that use F_{ST} measures between populations to identify violations of isolation-by-distance patterns have also been developed (Duforet-Frebourg and Blum 2014; Jay et al. 2013).

For our example dataset, we ran 10 iterations in the program ADMIXTURE at $K = 2$ to $K = 5$. The iterations at each value of K were then compared to detect different clustering solutions using the program CLUMPP (Jakobsson and Rosenberg 2007). For $K = 2$ to $K = 4$, all 10 iterations arrived at very similar solutions, and the combined output is shown in Fig. 3.4 (visualized with the program DISTRUCT (Rosenberg 2004)). The analysis shows clustering of sub-Saharan Africans (orange component) and clustering of Europeans and North Africans (blue component) at $K = 2$, although Yoruba individuals show a small fraction of ancestry from the blue component, and the Mozabites show a small ancestry fraction from the orange component. At $K = 3$, the Yoruba obtains its own cluster (green), while the Mozabite still clusters with the French but showing some ancestry fraction from the green component. The Mozabite forms its own group at $K = 4$, but some individuals showed shared ancestry with Yoruba and French. For $K = 5$, there was no common solution, and each of the 10 iterations had a different solution (this pattern can be seen clearly from the similarity matrix output of CLUMPP). The lack of a "common

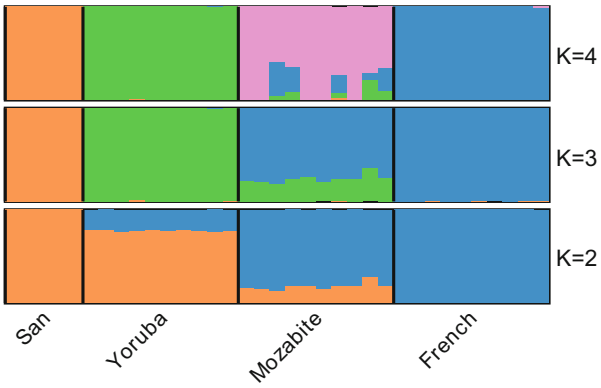

Fig. 3.4 Combined output from 10 iterations of running ADMIXTURE with our example HGDP data analysis for $K = 2$ to $K = 4$ visualized by DISTRUCT. For $K = 5$, there was no common solution, and each of the 10 iterations had a different solution. The analysis show clustering of sub-Saharan Africans (orange component) and clustering of Europeans and North Africans (blue component) at $K = 2$, although Yoruba individuals show a small fraction of ancestry from the blue component and the Mozabite show a small ancestry fraction from the orange component. At $K = 3$, the Yoruba obtains its own cluster (green), while the Mozabite still clusters with the French but showing some ancestry fraction from the green component. The Mozabite forms its own group at $K = 4$, but some individuals showed shared ancestry with Yoruba and French. Note that these algorithms, like the ADMIXTURE algorithm, is typically set up to only utilize the genetic information and to be agnostic to all other information (and other information like self-identified ancestry/ethnicity, geographic sample location, and/or language can be added onto the results for visualization purposes)

mode" in the iterations at $K = 5$ is an indication that there is no additional level of structure to reveal by dividing the individuals' genomes into additional ancestry components. In summary, we note that these three choices of assumed number of clusters ($K = 2$, 3, and 4) all reveal interesting patterns of population structure that are related to a hierarchical ancestry relationship among the four populations (e.g., Jakobsson et al. 2008; Schlebusch et al. 2012). We further note the interesting pattern for the Mozabite that display ancestry components related to Europeans and West Africans at $K = 3$ but that form their own cluster (to a large extent) at $K = 4$—a pattern consistent with a population with mainly Eurasian ancestry, followed by some level of admixture with West Africans. This admixture likely happened some time ago since the Mozabites make up their own cluster at $K = 4$, which is consistent with subsequent genetic drift in the Mozabites since the admixture.

3.3 Population-Based and Supervised Methods

Once an overview of the signals in the data has been obtained using PCA, STRUCTURE-like analyses, and tree construction at the individual level, a natural next step is to assign individuals to predefined populations. We may then want to

quantify the amount of structure among groups in order to learn something about the demographic past of the full collection of individuals. Although these populations are ideally identified with the aid of the previously presented methods, it is not uncommon to have assessments based on geography-only defined populations. It may still be of value to contrast these predefined populations in order to affirm that they do correspond to separate biological populations.

3.3.1 Genetic Differentiation at the Population Level

3.3.1.1 F_{ST}

Introduced by Sewall Wright (Wright 1949), F_{ST} is one of the first measures of genetic differentiation (sometimes referred to as "genetic distance") among or between populations. There are many variations on the original definition, and the usefulness of F_{ST} and relatives is still a debated topic (e.g., Holsinger and Weir 2009; Rousset 2013; Jost 2008; Ryman and Leimar 2009). F_{ST} was originally defined as the correlation between gametes chosen randomly from within the same subpopulation relative to the entire population or, equivalently, as the departure of genotype frequencies from Hardy–Weinberg expectations relative to the entire population (see Holsinger and Weir 2009 for a thorough review of F_{ST}). A common definition for more practical purposes is

$$F_{\mathrm{ST}} = \frac{\mathrm{Var}\,(p_i)}{E\,[p_i]\,(1 - E\,[p_i])},$$

(e.g., Holsinger and Weir 2009) where $\mathrm{Var}(p_i)$ is the variance in allele frequencies across subpopulations and $E[p_i]$ is the expected allele frequency. There are other formulations, including (Nei 1973) in terms of heterozygosity,

$$G_{\mathrm{ST}} = \frac{H_{\mathrm{T}} - H_{\mathrm{S}}}{H_{\mathrm{T}}},$$

where H_{T} is the total (pooled) heterozygosity and H_{S} is the mean heterozygosity across subpopulations. Note that these definitions are all coined in terms of genetic variation. If demography is the primary interest, these definitions may not be ideal since the distribution of genetic variation depends on the mutational process as well as demography—via the genealogical process. Slatkin (1995) isolated the purely genealogical aspect of F_{ST} by studying the limit as the mutation rate approached zero and showed that F_{ST} can be expressed in terms of expected coalescent times

$$F_{\mathrm{ST}} = \frac{t_s - t_w}{t_s},$$

where t_s is the average time for two randomly picked genes—from the whole population—to find a common ancestor and t_w is the average time for two genes

from the same subpopulation to find a common ancestor. The intuition for this definition is that it measures the relatively longer time it takes for two genes situated in individuals from different subpopulations to find a common ancestor compared to when they are situated in individuals from the same subpopulation. From this definition, it is clear that if the expected coalescent time for two genes does not depend on which population the genes are drawn from ($t_s = t_w$), then $F_{ST} = 0$. In contrast, if genes from different populations take much longer to find a common ancestor than genes from the same population, F_{ST} tends toward 1.

In order to estimate F_{ST}, we need genetic variation from individuals drawn from predefined populations. Depending on the type of genetic data, different assumptions of the mutational process can be used. For sequence data, the mutational process is well approximated by the infinite sites model, for which at most one mutation is allowed for each site. The mutation rate per site is typically very low and the influence of branch-specific, novel mutations when estimating F_{ST} will be assumed to be negligible compared to demographic factors that affect sites polymorphic in the ancestral population to the predefined populations. Alternatively, an outgroup can be utilized to delimit the data to SNPs that were polymorphic prior to the time period of interest. In order to account for the sample variance (due to limited sample sizes), Weir and Cockerham (1984) developed a robust (and commonly used) estimator for F_{ST} (see also Weir 1996; Bhatia et al. 2013).

Model-specific demographic parameters such as migration rate and/or divergence time can often be directly related to F_{ST}, although caution is warranted for directly equating an F_{ST} estimate with a specific demographic parameter as there are many different factors that influence estimates of F_{ST}. For instance, in a two-population divergence model, the relationship between F_{ST} and the divergence time t is (Slatkin 1995):

$$F_{ST} = \frac{t}{t + 8N_e},$$

while in an infinite island model, the migration rate m is related to F_{ST} as (see Fig. 3.5 and Box3.2)

$$F_{ST} = \frac{1}{1 + 4N_e m}.$$

Note that estimates of F_{ST}, like many other population genetic parameters, depend on genetic drift and include the term of effective population size. Hence, estimates of F_{ST} transformed into estimates of other population genetic parameters, such as divergence time or migration rate, are typically estimates of the scaled (in terms of N_e) parameter.

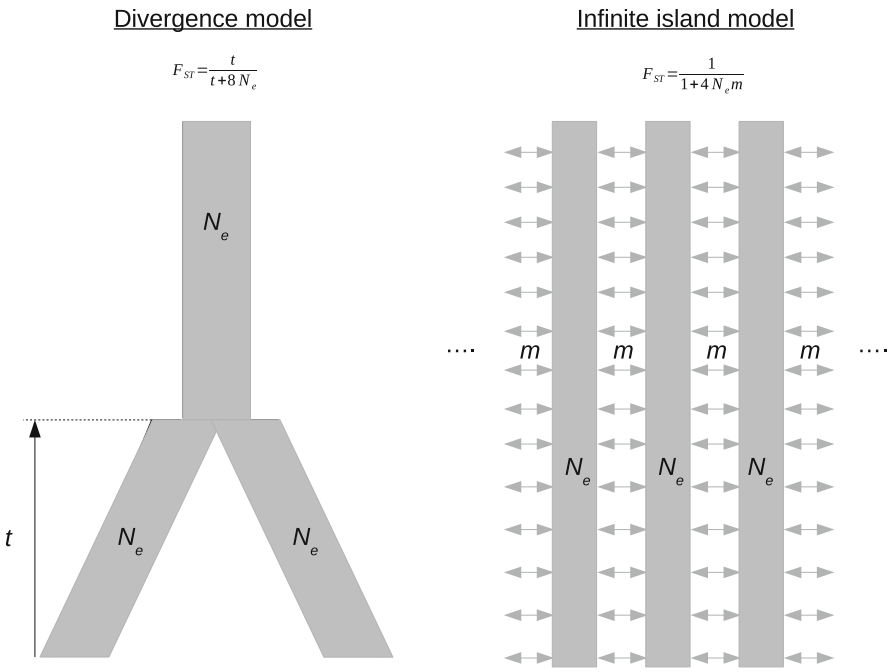

Fig. 3.5 Two simple demographic models and their relation to F_{ST} (in the limit as the mutation rate approaches 0 (Slatkin 1995)). The left model is a two-population split model with no migration, and the scaled divergence time $t/2N_e$ under this model can be estimated by $4F_{ST}/(1-F_{ST})$. The right model illustrates the infinite island model with an infinite number of equally sized subpopulations with a constant migration rate between neighboring subpopulations. Under this model, the scaled migration rate $4N_em$ is estimated by $(1-F_{ST})/F_{ST}$

Box 3.2 F_{ST}

Applied to the example data, the pairwise F_{ST} values, estimated using Weir and Cockerham's (1984) estimator, are displayed in the table below. The largest estimate of F_{ST} is found between the San and the French leading to the largest estimated divergence time (scaled by effective population sizes) or, alternatively, the smallest estimated migration rate, between these two populations. Note that because San ancestors may have diverged earlier than any of the other populations (Schlebusch et al. 2012), we may expect that the divergence time between San and any of the other population would be equal, but because time is scaled in N_e, this is not necessarily the case. The large divergence between San and French is, for instance, likely to be a consequence of a relatively small N_e for the French. Interestingly, F_{ST} is markedly smaller between the French and the Mozabite than between the Mozabite and the

(continued)

Yoruba. Under a pure migration model, this would suggest more gene flow between the Mozabite and the French compared to between the Mozabite and the Yoruba. An alternative interpretation (under a simple divergence model) would state that the first population split was between Yoruba and the ancestor population to the French and the Mozabite. As both these models are highly unlikely to be a good approximation to human demographic history in this particular case, caution is warranted about such direct interpretations when the underlying model is largely unknown. From the PCA and STRUCTURE analyses, we see indications of more complicated demographic scenarios that incorporate both gene flow and population divergence, which can be reconciled with these pairwise F_{ST} values.

	F_{st}
San vs. French	0.105
San vs. Mozabite	0.091
Yoruba vs. French	0.0905
Yoruba vs. Mozabite	0.073
San vs. Yoruba	0.0511
French vs. Mozabite	0.0185

3.3.1.2 Other Measures of Genetic Distance

There are many alternative measures of genetic distance between populations. The simplest distance measure is the Euclidian distance between two points in a multidimensional space. Many variations of this distance are available such as Rogers's (1972) scaled Euclidian distance, Prevosti et al.'s (1975) distance, Cavalli-Sforza and Edwards' (1967) chord distance, and Nei et al.'s (1983) D_A distance. These are all geometric distances and do not involve any evolutionary models.

Other, more model-based, distance measures are Cavalli-Sforza's chord distance (1969); Reynolds, Weir, and Cockerham's θ_W (1983), and Nei's (1972) D_s. The first two measures utilize existing variation (without modeling the possibility of new mutations), while Nei's D_s includes the possibility of new mutations occurring in an infinite allele mutation model.

Distance measures have also been designed to handle the stepwise mutation model (SMM) that can be useful for microsatellite data. Goldstein et al.'s (1995) $(\delta\mu)^2$ distance is commonly used, as is the closely related average square distance (ASD) (Slatkin 1995). Two other commonly employed distances for microsatellites are Shriver et al.'s (1995) distance and the shared allele distance D_{SA} (Chakraborty and Jin 1993).

3.3.2 Formal Tests for Admixture Under a Population Tree-Model

Once we have some proposed demographic model/s, we can start estimating demographic parameters in these models. An alternative, but not mutually exclusive, way to proceed is to construct formal tests of these models. The proposed model is used as a null model to make predictions, and then these predictions are compared to the observed data in order to test if the data is consistent with the model. A recent suite of tests along these lines is the 3-population test, the 4-population test (Reich et al. 2009), and the *D*-test (Green et al. 2010). These have been successfully applied to test for admixture among human populations as well as for identifying a significant level of admixture from archaic humans (Neandertals and Denisovans) among human populations (Green et al. 2010; Reich et al. 2010). These methods, collectively referred to as *f*-statistics (in contrast to Wright's *F*-statistics), relate the expected covariances in allele frequencies between not only 2 but also 3 and 4 populations in a bifurcating population phylogeny with the possibility of punctual admixture events.

The f_3 statistic, or 3-population test, is computed as the product $(p_X - p_A)(p_X - p_B)$, where p_X, p_A, and p_B are the allele frequencies at each locus in population A, B, and X. The expected value of this product is positive under a tree model, but the estimate from data can be negative under certain admixture scenarios (which violate the tree model), and negative f_3-statistics can only occur due to admixture events.

The f_4 statistic, or 4-population test, is computed as the product $(p_A - p_B)(p_X - p_Y)$, where p_A, p_B, p_X, and p_Y are the allele frequencies at each locus in population *A*, *B*, *X*, and *Y*. This product is expected to be 0 if the 4 populations are related by an unrooted phylogeny of the form (*A*, *B*), (*X*, *Y*) without admixture. Violations of this assumption can create (significantly) positive or negative values where the sign of the statistic contains information on the direction of the admixture.

The *D*-test is a version of the f_4 statistic with a denominator that includes a term for heterozygosity. Jackknife or bootstrap permutation tests of chromosomes or blocks of the genome can be used to assess statistical uncertainty and perform hypothesis tests using these statistics (Reich et al. 2009).

We illustrate these methods by performing the *D*-test on our data (Table 3.1). Using San as the outgroup, we see that the single-tree hypothesis with the smallest deviation from $D = 0$ has the Mozabite and the French as the closest related populations. However, the negative *D*-value for this tree suggests gene flow from the Yoruba into the Mozabite. This result is consistent with the Mozabite having ancestry related to both Yoruba and the French with more gene flow from the French than from the Yoruba. This result closely mirrors the analysis based on F_{ST}, and it is (supposedly) robust to effects of genetic drift (e.g., from different effective population sizes in the different populations), which could impact F_{ST} results. However, we cannot rule out alternative models without a more detailed model of genetic drift in the population history model.

Table 3.1 *D*-test for different topologies (*W*, (*X*, (*Y*, *Z*)))

W	*X*	*Y*	*Z*	*D*-stat	*Z*-score
San	Yoruba	Mozabite	French	−0.0089	−8.018
San	Mozabite	Yoruba	French	0.1358	67.627
San	French	Yoruba	Mozabite	0.1445	76.227

A large absolute value of the *Z*-score indicates a poor fit of the observed data and the proposed topology. If gene flow involving the outgroup lineage can be ignored, a negative *D* value suggests gene flow between *X* and *Y*, while a positive *D* value indicates gene flow between *X* and *Z*

3.3.3 More Advanced Modeling

3.3.3.1 Population Graph Fitting

The *f*-statistic framework in Reich et al. (2009) where the 3- and 4-population tests were introduced also in a more complex model fitting framework where a multi-population model can be fitted so that population topology, admixture events, and genetic drift along lineages are fitted to the observed *f*-statistics. This approach has been implemented in the package qpgraph (Patterson et al. 2012) and in MixMapper (Lipson et al. 2013). A similar approach of fitting ancestry graphs to genetic data was attained by Pickrell and Pritchard (2012). Their method, implemented in the software TREEMIX, finds the tree structure with potential admixture events between populations that best explains the observed matrix of allele frequency covariances between populations.

3.3.3.2 Isolation-Migration Models

Several methods that attempt to co-estimate effective population sizes, divergence time, and migration rates in a 2-population "isolation-migration" (IM) model setting have been developed. In one line of approaches, the estimates are based on haplotype information using a Bayesian framework (Nielsen and Wakeley 2001; Hey and Nielsen 2004, 2007). Here the haplotypes are assumed to be known and that there is no intra-locus recombination. The latter assumption has been relaxed in the implementation of MIMAR (Becquet and Przeworski 2007). These methods are usually too computationally intensive to be applied to full genomic data. Alternatively, loci are assumed to be independent, and the full joint frequency spectrum is utilized in a composite likelihood approach to estimate the migration rates and divergence time (Gutenkunst et al. 2009). ABC methods (see below) have also been developed specifically for the IM model (Lopes et al. 2009; see also Tellier et al. (2011)).

3.3.3.3 Approximate Bayesian Computation

Approximate Bayesian computation (ABC) is a powerful and extremely flexible approach to fit and compare models to real data that does not rely on calculating the full likelihood of the data given a model (see, for instance, Beaumont et al. 2002; Csilléry et al. 2010). Instead, some (well-chosen) summary statistics calculated for

simulated data are compared to the values of these summary statistics observed from the real data. The simulations are performed by drawing model parameters from prior distributions and then choosing those simulations that best mimic the real data. The distribution of the model parameters in this chosen set of best fitting simulations can then be used to estimate the model parameters. Different models can then be contrasted using Bayes factors. The ABC approach has proven very flexible for inferring model parameters, and there is an active community developing novel and faster algorithms (e.g., Pudlo et al. 2016; Csilléry et al. 2012).

3.4 Summary and Guidelines

Good practice when investigating a population-genetic dataset for population structure is to start by visualizing the data in a way that reveals the inherent characteristics of the data. To get an indication of whether or not the data contains different groups; PCA, simple tree-building methods, and "STRUCTURE-like" approaches are all good tools for initial data exploration. Once some overview of the data is obtained, we can start building simple models to investigate additional hierarchical patterns and characteristics of the underlying demographic history of the sample and population/s. More detailed hypotheses can subsequently be scrutinized using more explicit and advanced models. The more accurate models of demography we can infer, the better we understand the underlying processes shaping the population genetic patterns of variation. Ultimately, this understanding may allow in-depth analysis of the genetic architecture of traits and patterns of selection impacting the genome (Li et al. 2012), after controlling for patterns caused by demographic history manifesting as population structure.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664

Balding DJ, Nichols RA (1994) DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 64:125–140

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96:3–12

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162:2025–2035

Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. Genome Res 17:1505–1519

Bhatia G, Patterson N, Sankararaman S, Price AL (2013) Estimating and interpreting FST: the impact of rare variants. Genome Res 23:1514–1521

Bradburd GS, Ralph PL, Coop GM (2016) A spatial framework for understanding population structure and admixture. PLoS Genet 12:e1005703

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V et al (2002) A human genome diversity cell line panel. Science 296:261–262

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis -models and estimation procedures. Am J Hum Gen 19:233–257

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The History and Geography of Human Genes. Princeton University Press, Princeton, NJ

Chakraborty R, Jin L (1993) A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances. In: DNA fingerprinting: state of the science. Birkhäuser, Basel, pp 153–175

Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Mol Ecol Notes 7:747–756

Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. Genetics 163:367–374

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian computation in practice. Trends Ecol Evol 25:410–418

Csilléry K, François O, Blum MGB (2012) abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol 3:475–479

Duforet-Frebourg N, Blum MGB (2014) Nonstationary patterns of isolation-by-distance: inferring measure of local genetic differentiation with Bayesian kriging. Evolution 68:1110–1123

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455–464

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Res 10:564–567

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Felsenstein, J (1983) Parsimony in systematics: biological and statistical issues. Annu Rev Ecol Syst 14:313–333

Foreman L, Smith A, Evett I (1997) Bayesian analysis of DNA profiling data in forensic identification applications. J R Stat Soc A 160:429–469

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92:6723–6727

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U et al (2010) A draft sequence of the Neandertal genome. Science 328:710–722

Guillot G, Estoup A, Mortier F, Cosson JF (2005) A spatial statistical model for landscape genetics. Genetics 170:1261–1280

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167:747–760

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci USA 104:2785–2790

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet 10:639–650

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM et al (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998–1003

Jay F, Sjödin P, Jakobsson M, Blum MGM (2013) Anisotropic isolation by distance: the main orientations of human genetic differentiation. Mol Biol Evol 30:513–525

Jolliffe I (2005) Principal component analysis. Wiley, New York
Jost L (2008) G(ST) and its relatives do not measure differentiation. Mol Ecol 17:4015–4026
Katti MV, Rajekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–1167
Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874
Landsteiner K, Weiner AS (1940) An agglutinable factor in human blood recognized by immune sera for rhesus blood. Proc Soc Exp Biol NY 43:223
Lawson DJ, Hellenthal G, Myers S, Falush D (2011) Inference of population structure using dense haplotype data. PLoS Genet 8:e1002453
Lewontin RC, Hubby JL (1966) A molecular approach to the study of genetic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics 54:595–609
Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233
Li JZ, Absher DM, Tang H, Southwick AM, Casto AM et al (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104
Li J, Li H, Jakobsson M, Li S, Sjödin P, Lascoux M (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? Mol Ecol 21:28–44
Lipson M, Loh PR, Levin A, Reich D, Patterson N, Berger B (2013) Efficient moment-based inference of population admixture parameters and sources of gene flow. Mol Biol Evol 30:1788–1802
Lopes JS, Balding D, Beaumont MA (2009) PopABC: a program to infer historical demographic parameters. Bioinformatics 25:2747–2749
Mallick S, Li H, Lipson M, Mathieson I, Gymrek M et al (2016) The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature 538:201–206
McVean G (2009) A genealogical interpretation of principal components analysis M. PLoS Genetics 5:e1000686
Nei M (1972) Genetic distance between populations. Am Nat 106:283–292
Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323
Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. II Gene frequency data. J Mol Evol 19:153–170
Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K, Donnelly P (2002) Assessing population differentiation and isolation from single nucleotide polymorphism data. J R Stat Soc B 64:695–715
Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158:885–896
Patterson N, Price AL, Reich D (2006) Population structure and eigen analysis. PLoS Genetics 2:e190
Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N et al (2012) Ancient admixture in human history. Genetics 192:1065–1093
Petkova D, Novembre J, Stephens M (2016) Visualizing spatial population structure with estimated effective migration surfaces. Nat Genet 48:94–100
Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 8:e1002967
Prevosti A, Ocana J, Alonzo G (1975) Distances between populations for Drosophila subobscura based on chromosome arrangement frequencies. Theor Appl Genet 45:231–241
Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N et al (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet 5:e1000519
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP (2016) Reliable ABC model choice via random forests. Bioinformatics 32:859–866
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR et al (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am Hum Genet 81:559–575
Quinn GP, Keough MJ (2002) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge
Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. Nature 461:489–494
Reich D, Green RE, Kircher M, Krause J, Patterson N et al (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060
Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767–779
Roeder K, Escobar M, Kadane JB, Balazs I (1998) Measuring heterogeneity in forensic databases using hierarchical Bayes models. Biometrika 85:269–287
Rogers JS (1972) Measures of similarity and genetic distance. In: Studies in genetics VII. University of Texas Publication 7213. Austin, Texas, pp 145−153
Rosenberg NA (2004) Distruct: a program for the graphical display of population structure. Mol Ecol Notes 4:137–138
Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK et al (2002) Genetic structure of human populations. Science 298:2381–2385
Rousset F (2013) Exegeses on maximum genetic differentiation. Genetics 194:557–559
Ryman N, Leimar O (2009) G(ST) is still a useful measure of genetic differentiation – a comment on Jost's D. Mol Ecol 18:2084–2087
Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425
Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D et al (2012) Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science 338:374–379
Shriver M, Jin L, Boerwinkle E, Deka R, Ferrell RE et al (1995) A novel measure of genetic distance for highly polymorphic tandem repeat loci. Mol Biol Evol 12:914–920
Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462
Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol 28:289–301
Tellier A, Pfaffelhuber P, Haubold B, Naduvilezhath L, Rose LE et al (2011) Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum. PLoS One 6:e18155
Veeramah KR, Hammer MF (2014) The impact of whole-genome sequencing on the reconstruction of human population history. Nat Rev Genet 15:149–162
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al (2001) The sequence of the human genome. Science 291(5507):1304–51. https://doi.org/10.1126/science.1058040. Erratum in: Science 292(5523):1838 (2001). PMID: 11181995.
Weir BS (1996) Genetic data analysis II. Sinauer Associates, Sunderland
Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370
Wright S (1949) The genetical structure of populations. Ann Hum Gen 15:323–354