$F_{\mathbf{ST}}$ and kinship for arbitrary population structures I: Generalized definitions

Alejandro Ochoa[1,2] and John D. Storey[1,2,*]

[1]Lewis-Sigler Institute for Integrative Genomics, and [2]Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08544, USA

∗ Corresponding author: `jstorey@princeton.edu`

Abstract: F_{ST} is a fundamental measure of genetic differentiation and population structure currently defined for subdivided populations. F_{ST} in practice typically assumes the "island model", where subpopulations have evolved independently from their last common ancestral population. In this work, we generalize the F_{ST} definition to arbitrary population structures, where individuals may be related in arbitrary ways. Our definitions are built on identity-by-descent (IBD) probabilities that relate individuals through inbreeding and kinship coefficients. We generalize F_{ST} as the mean inbreeding coefficient of the individuals' local populations relative to their last common ancestral population. This F_{ST} naturally yields a useful pairwise F_{ST} between individuals. We show that our generalized definition agrees with Wright's original and the island model definitions as special cases. We define a novel coancestry model based on "individual-specific allele frequencies" and prove that its parameters correspond to probabilistic kinship coefficients. Lastly, we study and extend the Pritchard-Stephens-Donnelly admixture model in the context of our coancestry model and calculate its F_{ST}. Our probabilistic framework provides a theoretical foundation that extends F_{ST} in terms of inbreeding and kinship coefficients to arbitrary population structures.

1 Introduction

A population is structured if its individuals do not mate randomly, in particular, if homozygozity differs from what is expected when individuals mate randomly [1]. F_{ST} is a parameter that measures population structure [2, 3], which is best understood through homozygosity. $F_{\mathrm{ST}} = 0$ for an unstructured population, in which genotypes have Hardy-Weinberg proportions. At the other extreme, $F_{\mathrm{ST}} = 1$ for a fully differentiated population, in which every subpopulation is homozygous for some allele. Current F_{ST} definitions assume a partitioned or subdivided population into discrete, non-overlapping subpopulations [2–6]. Many F_{ST} estimators further assume an "island model", in which subpopulations evolved independently from the last common ancestral population [4–6] (Fig. 1A, Fig. 2A). However, populations such as humans are not necessarily naturally subdivided; thus, arbitrarily imposed subdivisions may yield correlated subpopulations [7] (Fig. 1B, Fig. 2B). In this work, we build a generalized F_{ST} definition applicable to arbitrary population structures, including arbitrary evolutionary dependencies.

Natural populations are often structured due to evolutionary forces, population size differences and the constraints of distance and geography [10]. The human genetic population structure, in particular, has been shaped by geography [11], population bottlenecks [12], and numerous admixture events [9, 13, 14]. Notably, human populations display genetic similarity that decays smoothly with geographic distance, rather than with discrete jumps as would be expected for island models [7] (Fig. 1B). Current F_{ST} definitions do not apply to these complex population structures.

Population structure can be quantified by the inbreeding and kinship coefficients, which measure how individuals are related. The inbreeding coefficient f is the probability that the two alleles of an individual, at a random locus, were inherited from a single ancestor, also called "identical by descent" (IBD) [15]. The mean f is positive in a structured population [15], and it also increases slowly over time in finite panmictic populations, an effect known as genetic drift [16]. The kinship coefficient φ is the probability that two random alleles, one from each individual, at a random locus are IBD [2]. Both f and φ combine relatedness due to the population structure with recent or "local" relatedness, such as that of family members [17]. The values of f, φ are relative to an ancestral population, where relationships that predate this population are treated as random [18]. Thus, f and φ increase if the reference ancestral population is an earlier rather than a more recent population.

Given an unstructured subpopulation S, Malécot defined F_{ST} as the mean f in S relative to an ancestral population T [2]. When S is itself structured, Wright defined three coefficients that connect T, S and individuals I in S [3]: F_{IT} ("total f") is the mean f in I relative to T; F_{IS} ("local f") is the mean f in I relative to S, which Wright did not considered to be part of the population structure; lastly, F_{ST} ("structural f") is the mean f relative to T that would result if individuals in S mated randomly. Wright distinguished these quantities in cattle, where F_{IS} can be excessive [15]; however, F_{IS} ought to be small in large, natural populations. The special case $F_{\mathrm{IS}} = 0$ gives

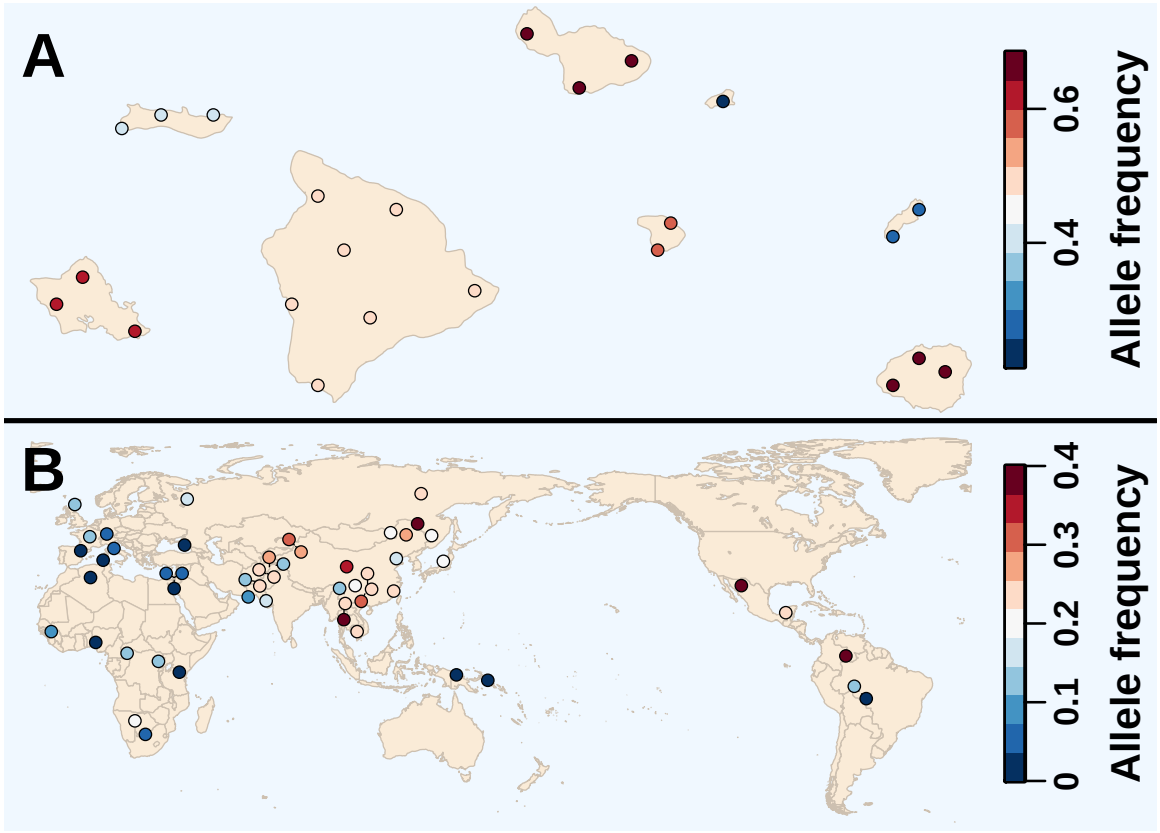

Figure 1: **Illustration of the island model, Human SNP with median differentiation.** In these maps, circles correspond to populations (moved to prevent overlaps in HGDP), and colors are allele frequencies (AFs). **A.** A simulated SNP from the island model (illustration). Individuals from the same island share AFs, while individuals from different islands evolve independently. AFs are drawn from the Balding Nichols distribution [8] with ancestral AF $p = 0.5$ shared by the islands, but F varied by island size: $F = 0.01, 0.1, 0.3$ for the large island, the four medium islands, and the three smallest islands, respectively. **B.** Sample AFs of SNP rs11692531 in the Human Genome Diversity Project (HGDP) [9], illustrates typical differentiation in human populations. HGDP was filtered to remove close relatives resulting in $n = 940$ individuals and $m = 431,345$ SNPs. This SNP had the median F_{ST} estimate (≈ 0.081) in HGDP, using the Weir-Cockerham estimator [4] and the $K = 53$ subpopulations shown in the map. (Note that to improve the dynamic range of the color map, the outlier population "Colombian" (AF ≈ 0.57, z-score ≈ 3.2) is displayed as the next largest AF (0.4).) AFs span a wide range and display strong geographical correlation, so the human population is structured but does not fit the island model.

$F_{ST} = F_{IT}$ [3]. The F_{ST} definition has been extended to a set of disjoint populations, where it is the average F_{ST} of each population from the last common ancestral population [5, 6].

F_{ST} is known by many names (for example, fixation index [3], coancestry coefficient [5, 19]), and alternative definitions (in terms of the variance of subpopulation allele frequencies [3], variance components [20], correlations [4], and genetic distance [19]). Our generalized F_{ST}, like Wright's F_{ST}, is defined using inbreeding coefficients. There is also a diversity of measures of differentiation that are specialized for a single multiallelic locus, such as G_{ST}, G'_{ST}, and D, which are functions of observed allele frequencies, and which relate to F_{ST} under certain conditions [21–25]. We consider F_{ST} as a genome-wide measure of genetic drift given by the relatedness of individuals, which does not depend on allele frequencies or other locus-specific features.

In our work, we generalize F_{ST} in terms of individual inbreeding coefficients, and exclude local inbreeding on an individual basis. Our F_{ST} applies to arbitrary population structures, generalizing previous F_{ST} definitions restricted to subdivided populations. We also generalize the "pairwise F_{ST}", a quantity often estimated between pairs of populations [6, 11, 26–30], now defined for arbitrary pairs of individuals.

We also define a coancestry model that parametrizes the correlations of "individual-specific allele frequencies" (IAFs) [31, 32], a recent tool that also accommodates arbitrary relationships between individuals. Our model is related to previous models between populations [5, 33]. We prove that our coancestry parameters correspond to kinship coefficients, thereby preserving their probabilistic interpretations, and we relate these parameters to F_{ST}.

We demonstrate our framework by providing a novel F_{ST} analysis in terms of our coancestry model of the widely used Pritchard-Stephens-Donnelly (PSD) admixture model, in which individuals derive their ancestry from intermediate populations given individual-specific admixture proportions [34–36]. We analyze an extension of the PSD model [31, 37–40] that generates intermediate allele frequencies from the Balding-Nichols distribution [8], and propose a more complete coancestry model for the intermediate populations. We derive equations relating F_{ST} to the model parameters of PSD and its extensions.

Our generalized definitions permit the analysis of F_{ST} and kinship estimators under arbitrary population structures, and pave the way forward to new approaches, which are the focus of our following work in this series [41, 42].

2 Generalized definitions in terms of individuals

2.1 Overview of data and model parameters

Let x_{ij} be observed biallelic genotypes for SNP $i \in \{1, ..., m\}$ and diploid individual $j \in \{1, ..., n\}$. Biallelic SNPs are the most common genetic variation in humans; the multiallelic model follows in analogy to the work of [5]. Given a chosen reference allele at each SNP, genotypes are encoded as the

number of reference alleles: $x_{ij} = 2$ is homozygous for the reference allele, $x_{ij} = 0$ is homozygous for the alternative allele, and $x_{ij} = 1$ is heterozygous. Our models assume that the genotype distribution is parametrized solely by the population structure, evolving by genetic drift in the absence of new mutations and selection.

We assume the existence of a panmictic ancestral population T. Relationships that precede T in time are considered random and do not count as IBD, while relationships since T count toward IBD probabilities. Every SNP i is assumed to have been polymorphic in T, with an ancestral reference allele frequency $p_i^T \in (0, 1)$ in T, and no new mutations have occurred since then.

The inbreeding coefficient of individual j relative to T, $f_j^T \in [0, 1]$, is defined as the probability that the two alleles of any random SNP of j are IBD [15]. Therefore, f_j^T measures the amount of relatedness within an individual, or the extent of dependence between its alleles at each SNP. Similarly, the kinship coefficient of individuals j and k relative to T, $\varphi_{jk}^T \in [0, 1]$, is defined as the probability that two alleles at any random SNP, each picked at random from each of the two individuals, are IBD [2]. φ_{jk}^T measures the amount of relatedness between individuals, or the extent of dependence across their alleles at each SNP.

For a panmictic population S that evolved from T, the inbreeding coefficient $f_S^T \in [0, 1]$ of S relative to T equals f_j^T shared by every individual j in S. Thus, f_S^T is equivalent to Wright's F_{ST} for a subdivided population. The random drift in allele frequencies across SNPs from T to S is parametrized by f_S^T alone, combining the contribution of time and sample size history into a single value [16].

2.2 Local populations

Our generalized F_{ST} definition depends on the notion of a local population. Our formulation includes as special cases island models and admixture models, and its generality is in line with recent efforts to model population structure on a fine scale [43, 44], through continuous spatial models [7, 45–47], or in a manner that makes minimal assumptions [32]. We define the *local population* L_j of an individual j as the most recent ancestral population of j. In the simplest case, if j's parents belong to the same population, then that population is L_j and j belongs to it too. However, if j's parents belong to different populations, then L_j is an admixed population (see example below). More broadly, L_j is the most recent population from which the inbreeding coefficient of j can be meaningfully defined. We define the "local" inbreeding coefficient of j to be $f_j^{L_j}$, and j is said to be *locally outbred* if $f_j^{L_j} = 0$.

For any population T ancestral to L_j, the parameter trio $(f_j^T,\ f_j^{L_j},\ f_{L_j}^T)$ are individual-level analogs of Wright's trio $(F_{\mathrm{IT}},\ F_{\mathrm{IS}},\ F_{\mathrm{ST}})$ defined for a subdivided population [3]. Moreover, just like Wright's coefficients satisfy

$$(1 - F_{\mathrm{IT}}) = (1 - F_{\mathrm{IS}})\,(1 - F_{\mathrm{ST}})\,, \tag{1}$$

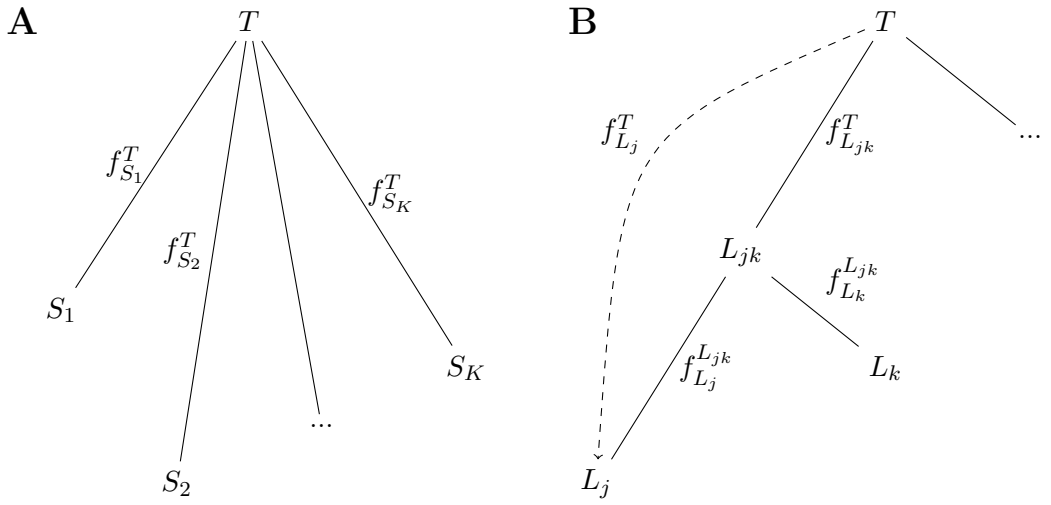

Figure 2: **Comparison between island model and arbitrary population structures.** These trees illustrate relationships (edges) between populations (nodes). Edge lengths are proportional to inbreeding coefficients. **A.** In the island model, K populations $S_1, ..., S_K$ evolved independently from an ancestral population T. $f^T_{S_u}$ is the inbreeding coefficient of population S_u relative to T. **B.** In an arbitrary population structure, each individual j has its own local population L_j, and every pair of individuals (j, k) have a jointly local population L_{jk} from which L_j and L_k evolved. Note that we do not assume a bifurcating tree process; the case for three or more individuals cannot generally be visualized as a tree. The coefficients $f^T_{L_j}$ and $f^T_{L_{jk}}$ are relative to T, while $f^{L_{jk}}_{L_j}$, $f^{L_{jk}}_{L_k}$ are relative to L_{jk}.

our individual-level parameters satisfy

$$\left(1 - f^T_j\right) = \left(1 - f^{L_j}_j\right)\left(1 - f^T_{L_j}\right), \tag{2}$$

since the absence of IBD of j relative to T requires independent absence of IBD at two levels: of j relative to L_j, and of L_j relative to T. Note that an individual j is locally outbred ($f^{L_j}_j = 0$) if and only if $f^T_{L_j} = f^T_j$.

Similarly, we define the *jointly local population* L_{jk} of the pair of individuals j and k as the most recent ancestral population of j and k. Hence, L_{jk} is ancestral to both L_j and L_k (Fig. 2B). We define the "local" kinship coefficient to be $\varphi^{L_{jk}}_{jk}$, and j and k are said to be *locally unrelated* if $\varphi^{L_{jk}}_{jk} = 0$. Since the inbreeding coefficient of an individual is the kinship of its parents [2], it follows that a locally-outbred individual has locally-unrelated parents.

Consider an individual j in an admixture model, deriving alleles from two populations A and B with proportions q_{jA} and $q_{jB} = 1 - q_{jA}$. Then L_j has allele frequencies $\pi_{ij} = q_{jA}p^A_i + q_{jB}p^B_i$ at each SNP i, where p^A_i and p^B_i are the allele frequencies in A and B, respectively. Considering a pair of individuals (j, k), the jointly local population L_{jk} at one extreme equals $L_j = L_k$ if $q_{jA} = q_{kA}$;

at the other extreme L_{jk} is the last common ancestral population T of A and B if $q_{jA} = 1$ and $q_{kA} = 0$ or vice versa (i.e., individuals are not admixed and belong to separate populations).

2.3 The generalized $F_{\mathbf{ST}}$ for arbitrary structures

Recall the individual-level analog of Wright's F_{ST} is $f^T_{L_j}$, which measures the inbreeding coefficient of individual j relative to T due exclusively to the population structure (Fig. 2B), as discussed in the last section. We generalize F_{ST} for a set of individuals as

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f^T_{L_j}, \tag{3}$$

where T is the most recent ancestral population common to all individuals under consideration, and $w_j > 0, \sum_{j=1}^{n} w_j = 1$ are fixed weights for individuals. The simplest weights are $w_j = \frac{1}{n}$ for all j. However, we allow for flexibility in the weights so that one may assign them to reflect how individuals were sampled, such as a skewed or uneven sampling scheme.

This generalized F_{ST} definition summarizes the population structure with a single value, intuitively measuring the average distance of our individuals from T. Moreover, our definition contains the previous F_{ST} definition as a special case, as discussed shortly. For simplicity, we kept Wright's traditional F_{ST} notation [3] rather than using something that resembles our f^T_S notation. A more consistent notation could be $F^T_{\{L_j\}}(\{w_j\})$, which more clearly denotes the weighted average of $f^T_{L_j}$ across individuals. Our definition is more general because the traditional S population is replaced by a set of local populations $\{L_j\}$, which may differ for every individual.

2.3.1 Mean heterozygosity in a structured population

Our generalized F_{ST} is connected to the mean heterozygosity in a structured population, and illustrates its properties. Here we will assume locally outbred individuals, for which $f^T_{L_j} = f^T_j$. The expected proportion of heterozygotes H_{ij} of an individual with inbreeding coefficient f^T_j at SNP i with an ancestral allele frequency p^T_i is given by [15]

$$H_{ij} = \Pr(x_{ij} = 1|T) = 2p^T_i \left(1 - p^T_i\right)\left(1 - f^T_j\right).$$

The weighted mean of these expected proportion of heterozygotes across individuals, $\bar{H}_i$, is given by our generalized F_{ST}:

$$\bar{H}_i = \sum_j w_j H_{ij} = 2p^T_i \left(1 - p^T_i\right)\left(1 - F_{\mathrm{ST}}\right).$$

Hence, individuals have Hardy-Weinberg proportions ($\bar{H}_i = 2p^T_i \left(1 - p^T_i\right)$) if and only if $F_{\mathrm{ST}} = 0$, which in turn happens if and only if $f^T_j = 0$ for each j. In the other extreme, individuals have fully-fixated alleles ($\bar{H}_i = 0$), if and only if $F_{\mathrm{ST}} = 1$, which in turn happens if and only if $f^T_j = 1$ for each j.

2.3.2 F_{ST} under the island model

Here we show that our generalized F_{ST} contains as a special case the currently used F_{ST} definition for a subdivided population. As discussed above, F_{ST} estimators often assume what we call the "island model," in which the population is subdivided into K non-overlapping subpopulations that evolved independently from their last common ancestral population T [4–6]. For simplicity, individuals are often further assumed to be locally outbred and locally unrelated. These assumptions result in the following block structure for our parameters,

$$f_j^T = f_{S_u}^T \quad \text{for} \quad j \in S_u,$$
$$\varphi_{jk}^T = \begin{cases} f_{S_u}^T & j \in S_u, k \in S_u, j \neq k, \\ 0 & j \in S_u, k \in S_{u'}, u \neq u', \end{cases}$$

where $S_u, S_{u'}$ are disjoint subpopulations treated as sets containing individuals. This population structure is illustrated as a tree in Fig. 2A.

In the notation of our generalized F_{ST}, we have under the island model assumptions that

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f_j^T = \sum_{u=1}^{K} \frac{1}{K} f_{S_u}^T,$$

where the weights w_j are such that $\sum_{j \in S_u} w_j = \frac{1}{K}$. Note also that the S_u here act as the K unique local populations, where $L_j = S_u$ whenever $j \in S_u$.

2.4 The individual-level pairwise F_{ST}

An important special case of F_{ST} is the "pairwise" F_{ST}, which is the F_{ST} of two subpopulations. When the assumption holds that individuals belong to one of the two unstructured populations, this pairwise F_{ST} can be estimated consistently [6], and is used frequently in the literature [11, 26–30]. Here we generalize this parameter to be between two individuals, and clarify its relationship to inbreeding coefficients measured relative to ancestral population T.

Let L_{jk} denote the last common ancestral population of the pair of individuals j and k, which we defined above as their jointly local population (Fig. 2B). We define the "individual-level pairwise F_{ST}" to be

$$F_{jk} = \frac{f_{L_j}^{L_{jk}} + f_{L_k}^{L_{jk}}}{2},$$

which is the special case of our generalized F_{ST} for two populations, $T = L_{jk}$, and equal weights $w_j = w_k = \frac{1}{2}$. Note that L_j and L_k being independent relative to L_{jk} enables consistent estimation of F_{jk} [6, 41]; the same is not generally possible for three or more individuals relative to their most recent ancestral population T.

Given $f^T_{L_j}$ and $f^T_{L_{jk}}$ relative to some earlier ancestral population $T \neq L_{jk}$ (Fig. 2B), the desired parameters $f^{L_{jk}}_{L_j}$ are given by,

$$\left(1 - f^T_{L_j}\right) = \left(1 - f^{L_{jk}}_{L_j}\right)\left(1 - f^T_{L_{jk}}\right), \tag{4}$$

which follows analogously to Eqs. (1) and (2). Solving for $f^{L_{jk}}_{L_j}$, repeating for $f^{L_{jk}}_{L_k}$, and replacing them into our individual-level pairwise F_{ST}, we obtain an equation for arbitrary T:

$$F_{jk} = \frac{\frac{f^T_{L_j} + f^T_{L_k}}{2} - f^T_{L_{jk}}}{1 - f^T_{L_{jk}}}. \tag{5}$$

When there is no local relatedness, $f^T_{L_j} = f^T_j$ is the usual inbreeding coefficient and $f^T_{L_{jk}} = \varphi^T_{jk}$ is the usual kinship coefficient, both measuring population structure only and yielding

$$F_{jk} = \frac{\frac{f^T_j + f^T_k}{2} - \varphi^T_{jk}}{1 - \varphi^T_{jk}}.$$

Note that the mean individual-level pairwise F_{ST} for $n > 2$, given by

$$\bar{F} = \sum_{j=1}^{n}\sum_{k=1}^{n} w_j w_k F_{jk} = \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} w_j w_k \left(f^{L_{jk}}_{L_j} + f^{L_{jk}}_{L_k}\right)$$

gives a lower bound for the "global" $F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f^T_{L_j}$, since $f^{L_{jk}}_{L_j} \leq f^T_{L_j}$. Thus, $\bar{F} = F_{\mathrm{ST}}$ for $n > 2$ if and only if all individuals are independent, where $f^T_{L_{jk}} = 0$ for all $j \neq k$.

2.5 Shifting IBD probabilities for change of reference ancestral population

In developing the generalized F_{ST} and the individual-level pairwise F_{ST}, we have made use of equations that relate IBD probabilities in a hierarchy. Here we present more general forms of these equations, which allow for transformations of probabilities under a change of reference ancestral population. Our relationships are straightforward generalizations of Wright's equation relating F_{IT}, F_{IS}, and F_{ST} in Eq. (1), now more generally applicable.

Let A be a population ancestral to population B, which is in turn ancestral to population C. The inbreeding coefficients relating every pair of populations in $\{A, B, C\}$ satisfy

$$\left(1 - f^A_C\right) = \left(1 - f^B_C\right)\left(1 - f^A_B\right),$$

which generalizes Eq. (4). A similar form applies for individual inbreeding and kinship coefficients given relative to populations A and B, respectively,

$$\left(1 - f^A_j\right) = \left(1 - f^B_j\right)\left(1 - f^A_B\right),$$
$$\left(1 - \varphi^A_{jk}\right) = \left(1 - \varphi^B_{jk}\right)\left(1 - f^A_B\right),$$

which generalizes Eq. (2). All of these cases follow since the absence of IBD of C (or j, or j, k) relative to A requires independent absence of IBD at two levels: of C (or j, or j, k) relative to B, and of B relative to A.

2.6 Genotype moments under the kinship model

In the kinship model, genotypes x_{ij} are random variables with first and second moments given by

$$\mathrm{E}[x_{ij}|T] = 2p_i^T, \tag{6}$$

$$\mathrm{Var}(x_{ij}|T) = 2p_i^T\left(1-p_i^T\right)\left(1+f_j^T\right), \tag{7}$$

$$\mathrm{Cov}(x_{ij}, x_{ik}|T) = 4p_i^T\left(1-p_i^T\right)\varphi_{jk}^T. \tag{8}$$

Eq. (6) is a consequence of assuming no selection or new mutations, leaving random drift as the only evolutionary force acting on genotypes [15]. Eq. (7) shows how inbreeding modulates the genotype variance: an outbred individual relative to T ($f_j^T = 0$) has the Binomial variance of $2p_i^T\left(1-p_i^T\right)$ that corresponds to independently-drawn alleles; a fully inbred individual ($f_j^T = 1$) has a scaled Bernoulli variance of $4p_i^T\left(1-p_i^T\right)$ that corresponds to maximally correlated alleles [3]. Lastly, Eq. (8) shows how kinship modulates the correlations between individuals: unrelated individuals relative to T ($\varphi_{jk}^T = 0$) have uncorrelated genotypes, while $\varphi_{jk}^T = 1$ holds for the extreme of identical and fully inbred twins, which have maximally correlated genotypes [2, 48]. Hence, f_j^T and φ_{jk}^T parametrize the frequency of non-independent allele draws within and between individuals. The "self kinship", arising from comparing Eq. (7) to the $j = k$ case in Eq. (8), implies $\varphi_{jj}^T = \frac{1}{2}\left(1+f_j^T\right)$, which is a rescaled inbreeding coefficient resulting from comparing an individual with itself or its identical twin.

3 The coancestry model for individual allele frequencies

F_{ST} and its estimators are most often studied in terms of population allele frequencies [4–6, 33]. Here we introduce a coancestry model for individuals, which is based on *individual-specific allele frequencies* (IAFs) [31, 32] that accomodate arbitrary population-level relationships between individuals. Some authors use the terms "coancestry" and "kinship" exchangeably [5, 49, 50]; in our framework, kinship coefficients are general IBD probabilities (following [17]), and we reserve coancestry coefficients for the IAFs covariance parameters (in analogy to the work of [5]).

In this section we introduce two parameters. First, $\pi_{ij} \in [0, 1]$ is the IAF of individual j at SNP i. Individual j draws its alleles independently according to probability π_{ij}. Allowing every SNP-individual pair to have a potentially unique allele frequency allows for arbitrary forms of population structure at the level of allele frequencies [32]. Second, $\theta_{jk}^T \in [0, 1]$ is the coancestry coefficient of individuals j and k relative to an ancestral population T, which modulate the covariance of π_{ij} and π_{ik} as shown below.

3.1 The coancestry model

In our coancestry model, the IAFs π_{ij} have the following first and second moments,

$$\mathrm{E}[\pi_{ij}|T] = p_i^T, \tag{9}$$

$$\mathrm{Cov}(\pi_{ij}, \pi_{ik}|T) = p_i^T \left(1 - p_i^T\right) \theta_{jk}^T, \tag{10}$$

$$x_{ij}|\pi_{ij} \sim \mathrm{Binomial}(2, \pi_{ij}). \tag{11}$$

Eq. (9) implies that random drift is the only force acting on the IAFs, and is analogous to Eq. (6) in the kinship model. Eq. (10) is analogous to Eqs. (7) and (8) in the kinship model, with individual coancestry coefficients (θ_{jk}^T) playing the role of the kinship and inbreeding coefficients (for $j = k$), a relationship elaborated in the next section. Lastly, Eq. (11) draws the two alleles of a genotype independently from the IAF, which models locally outbred ($f_j^{L_j} = 0$) and locally unrelated ($\varphi_{jk}^{L_{jk}} = 0$) individuals [5]. Hence, the coancestry model excludes local relationships, so it is more restrictive than the kinship model.

Our coancestry model between individuals is closely related to previous models between populations [5, 33]. However, previous models allowed $\theta_{jk}^T < 0$ [5]. We require that $\theta_{jk}^T \in [0, 1]$ for two reasons: (1) covariance is non-negative in latent structure models [51], such as population structure, and (2) it is necessary in order to relate θ_{jk}^T to IBD probabilities as shown next.

3.2 Relationship between coancestry and kinship coefficients

Here we show that the coancestry coefficients for IAFs, θ_{jk}, defined above can be written in terms of the kinship and inbreeding coefficients utilized in our more general model. We do so by relating our coancestry coefficients to general kinship coefficients by matching moments. Conditional on the IAFs, genotypes in the coancestry model have a Binomial distribution, so

$$\mathrm{E}[x_{ij}|\pi_{ij}] = 2\pi_{ij},$$

$$\mathrm{Cov}(x_{ij}, x_{ik}|\pi_{ij}, \pi_{ik}) = \begin{cases} 2\pi_{ij}(1 - \pi_{ij}) & j = k \\ 0 & j \neq k \end{cases}.$$

We calculate total moments by marginalizing the IAFs. The total expectation is

$$\mathrm{E}[x_{ij}|T] = \mathrm{E}[\mathrm{E}[x_{ij}|\pi_{ij}]|T] = \mathrm{E}[2\pi_{ij}|T] = 2p_i^T,$$

which agrees with Eq. (6) of the kinship model. The total covariance is calculated using

$$\mathrm{Cov}(x_{ij}, x_{ik}|T) = \mathrm{E}\left[\mathrm{Cov}(x_{ij}, x_{ik}|\pi_{ij}, \pi_{ik})|T\right] + \mathrm{Cov}\left(\mathrm{E}[x_{ij}|\pi_{ij}], \mathrm{E}[x_{ik}|\pi_{ik}]|T\right).$$

The first term is zero for $j \neq k$, and for $j = k$ it is

$$\begin{aligned} \mathrm{E}\left[\mathrm{Var}(x_{ij}|\pi_{ij})|T\right] &= \mathrm{E}\left[2\pi_{ij}(1 - \pi_{ij})|T\right] \\ &= 2\left(\mathrm{E}[\pi_{ij}] - \mathrm{Var}(\pi_{ij}|T) - \mathrm{E}[\pi_{ij}]^2\right) \\ &= 2p_i^T(1 - p_i^T)\left(1 - \theta_{jj}^T\right) \end{aligned}$$

The second term equals $4\,\mathrm{Cov}\left(\pi_{ij}, \pi_{ik}|T\right)$ for all (j,k) cases, which is given by Eq. (10). All together,

$$\mathrm{Cov}(x_{ij}, x_{ik}|T) = \begin{cases} 2p_i^T(1-p_i^T)\left(1+\theta_{jj}^T\right) & j=k, \\ 4p_i^T\left(1-p_i^T\right)\theta_{jk}^T & j\neq k. \end{cases}$$

Comparing the above to Eqs. (7) and (8), we find that

$$\theta_{jk}^T = \begin{cases} f_j^T & \text{if} \quad j=k, \\ \varphi_{jk}^T & \text{if} \quad j\neq k. \end{cases} \tag{12}$$

Therefore, our coancestry coefficients are equal to kinship coefficients, except that self coancestries are equal to inbreeding coefficients.

Since individuals in our IAF coancestry model are locally outbred and unrelated, we also have $f_{L_j}^T = \theta_{jj}^T$ and $f_{L_{jk}}^T = \theta_{jk}^T$ for $j \neq k$. Replacing these quantities in Eqs. (3) and (5), we obtain the generalized F_{ST} and pairwise F_{ST} in terms of coancestry coefficients.

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j \theta_{jj}^T, \tag{13}$$

$$F_{jk} = \frac{\frac{\theta_{jj}^T + \theta_{kk}^T}{2} - \theta_{jk}^T}{1-\theta_{jk}^T}. \tag{14}$$

4 Coancestry and F_{ST} in admixture models

The Pritchard-Stephens-Donnelly (PSD) admixture model [34] is a well-established, tractable model of structure that is more complex than island models. There are several algorithms available to estimate the PSD model parameters [34–36, 40, 52]. This model assumes the existence of "intermediate" populations, from which individuals draw alleles according to their admixture proportions. However, the PSD model was not developed with F_{ST} in mind; we will present a modified model that is compatible with our coancestry model.

The PSD model is a special case of our coancestry model with the following additional parameters. The number of intermediate populations is denoted by K. Let $p_i^{S_u} \in [0,1]$ be the reference allele frequency at SNP i and intermediate population S_u. Lastly, $q_{ju} \in [0,1]$ is the admixture proportion of individual j for intermediate population S_u. These proportions satisfy $\sum_{u=1}^{K} q_{ju} = 1$ for each j.

4.1 The PSD model with Balding-Nichols allele frequencies

The original algorithm for fitting the PSD model [34] utilizes prior distributions for intermediate population allele frequencies and admixture proportions according to

$$(q_{ju})_{u=1}^{K} \sim \mathrm{Dirichlet}\left(\alpha, ..., \alpha\right), \tag{15}$$

$$p_i^{S_u} \sim \mathrm{Beta}(1,1). \tag{16}$$

It has been shown [32, 36] that their model is then equivalent to forming IAFs

$$\pi_{ij} = \sum_{u=1}^{K} p_i^{S_u} q_{ju} \tag{17}$$

where genotypes are then drawn independently according to $x_{ij} \sim \text{Binomial}(2, \pi_{ij})$.

Here we consider an extension of this, which we call the "BN-PSD" model, by replacing Eq. (16) with the Balding-Nichols (BN) distribution [8] to generate the intermediate allele frequencies $p_i^{S_u}$. This combined model has been used to simulate structured genotypes [31, 38, 39], and is the target of some inference algorithms [37, 40]. The BN distribution is the following reparametrized Beta distribution,

$$p^* \sim \text{BN}(p, F) = \text{Beta}\left(p\left(\frac{1}{F} - 1\right), (1-p)\left(\frac{1}{F} - 1\right)\right),$$

where p is the ancestral allele frequency and F is the inbreeding coefficient [8]. The resulting allele frequencies p^* fit into our coancestry model, since $\text{E}[p^*] = p$ and $\text{Var}(p^*) = p(1-p)F$ hold.

In BN-PSD, the allele frequencies $p_i^{S_u}$ are generated independently from

$$p_i^{S_u} | T \sim \text{BN}\left(p_i^T, f_{S_u}^T\right),$$

resulting in an island model structure for the intermediate populations S_u.

We calculate the coancestry parameters of this model by matching moments conditional on the admixture proportions $\mathbf{Q} = (q_{ju})$. We calculate the expectation as

$$\text{E}[\pi_{ij} | \mathbf{Q}, T] = \sum_{u=1}^{K} q_{ju} \,\text{E}\left[p_i^{S_u} \middle| T\right] = \sum_{u=1}^{K} q_{ju} p_i^T = p_i^T.$$

and the IAF covariance is

$$\text{Cov}(\pi_{ij}, \pi_{ik} | \mathbf{Q}, T) = \sum_{u=1}^{K} q_{ju} q_{ku} \,\text{Var}\left(p_i^{S_u} \middle| T\right) = p_i^T (1 - p_i^T) \sum_{u=1}^{K} q_{ju} q_{ku} f_{S_u}^T.$$

By matching these to Eq. (10), we arrive at coancestry coefficients and F_{ST} of

$$\theta_{jk}^T = \sum_{u=1}^{K} q_{ju} q_{ku} f_{S_u}^T,$$
$$F_{\text{ST}} = \sum_{j=1}^{n} \sum_{u=1}^{K} w_j q_{ju}^2 f_{S_u}^T.$$

4.2 The BN-PSD model with full coancestry

The BN-PSD contains a restriction that the K intermediate populations are independent. Suppose instead that the intermediate population allele frequencies $p_i^{S_u}$ satisfy our more general coancestry

model:

$$\mathrm{E}\left[p_i^{S_u}\middle|T\right] = p_i^T,$$
$$\mathrm{Cov}\left(p_i^{S_u}, p_i^{S_v}\middle|T\right) = p_i^T\left(1-p_i^T\right)\vartheta_{uv}^T,$$

where ϑ_{uv}^T is the coancestry of the intermediate populations S_u and S_v. Note that the previous BN-PSD model satisfies $\vartheta_{uu}^T = f_{S_u}^T$ and $\vartheta_{uv}^T = 0$ for $u \neq v$. Repeating our calculations assuming our full coancestry setting, individual coancestry coefficients and F_{ST} are given by

$$\theta_{jk}^T = \sum_{u=1}^{K}\sum_{v=1}^{K} q_{ju}q_{kv}\vartheta_{uv}^T,$$
$$F_{\mathrm{ST}} = \sum_{j=1}^{n}\sum_{u=1}^{K}\sum_{v=1}^{K} w_j q_{ju}q_{jv}\vartheta_{uv}^T.$$

Therefore, all coancestry coefficients of the intermediate populations influence the coancestry coefficients between individuals and the overall F_{ST}. The form for θ_{jk}^T above has a simple probabilistic interpretation: the probability of IBD at random SNPs between individuals j and k corresponds to the sum for each pair of ancestries u and v of the probability of the pairing ($q_{ju}q_{kv}$) times the probability of IBD between these populations (ϑ_{uv}^T).

5 Discussion

We presented a generalized F_{ST} definition corresponding to a weighted mean of individual-specific inbreeding coefficients. Compared to previous F_{ST} definitions, ours is applicable to arbitrary population structures, and in particular does not require the existence of discrete subpopulations. A special case of our generalized F_{ST} is the pairwise F_{ST} of two individuals, which generalizes the pairwise F_{ST} between two populations that is part of many modern analyses [6, 11, 26–30].

We considered two closely-related population structure models with individual-level resolution: the kinship model for genotypes, and our new coancestry model for IAFs (individual-specific allele frequencies). The kinship model is the most general, applicable to the genotypes in arbitrary sets of individuals. Our IAF model requires a local form of Hardy-Weinberg equilibrium to hold, and it does not model locally related or locally inbred individuals. Nevertheless, IAFs arise in many applications, including admixture models [35], estimation of local kinship [31], genome-wide association studies [53], and the logistic factor analysis [32]. We prove that kinship coefficients, which control genotype covariance, also control IAF covariance under our coancestry model.

We also calculated F_{ST} for admixture models. To achieve this, we framed the PSD (Pritchard-Stephens-Donnelly) admixture model as a special case of our IAF coancestry model, and studied extensions where the intermediate populations are more structured. F_{ST} was previously studied in an admixture model under Nei's F_{ST} definition for one locus, where F_{ST} in the admixed population

is given by a ratio involving admixture proportions and intermediate population allele frequencies [54]. On the other hand, our F_{ST} is an IBD probability shared by all loci and independent of allele frequencies. Under our framework, the F_{ST} of an admixed individual is a sum of products, which is quadratic in the admixture proportions and linear in the coancestry coefficients of the intermediate populations. In the future, inference algorithms for our admixture model with fully correlated intermediate populations could yield improved results, including coancestry and F_{ST} estimates.

Our probabilistic model reconnects F_{ST} [5, 6] to inbreeding and kinship coefficients [17, 50, 55], all quantities of great interest in population genetics, but which are studied in increasing isolation. The main reason for this isolation is that F_{ST} estimation assumes the island model, in which kinship coefficients are uninteresting. However, study of the generalized F_{ST} in arbitrary population structures requires the consideration of arbitrary kinship coefficients [17]. Our work lays the foundation necessary to study estimation of the generalized F_{ST}, which is the focus of our next publications in this series [41, 42].

References

[1] S. Wright. "Coefficients of Inbreeding and Relationship". *The American Naturalist* 56(645) (1922), pp. 330–338.

[2] G. Malécot. *Mathématiques de l'hérédité*. Masson et Cie, 1948.

[3] S. Wright. "The Genetical Structure of Populations". *Annals of Eugenics* 15(1) (1951), pp. 323–354.

[4] B. S. Weir and C. C. Cockerham. "Estimating F-Statistics for the Analysis of Population Structure". *Evolution* 38(6) (1984), pp. 1358–1370.

[5] B. S. Weir and W. G. Hill. "Estimating F-Statistics". *Annual Review of Genetics* 36(1) (2002), pp. 721–750.

[6] G. Bhatia et al. "Estimating and interpreting FST: The impact of rare variants". *Genome Res.* 23(9) (2013), pp. 1514–1521.

[7] J. Novembre et al. "Genes mirror geography within Europe". *Nature* 456(7218) (2008), pp. 98–101.

[8] D. J. Balding and R. A. Nichols. "A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity". *Genetica* 96(1-2) (1995), pp. 3–12.

[9] N. A. Rosenberg et al. "Genetic Structure of Human Populations". *Science* 298(5602) (2002), pp. 2381–2385.

[10] S. Wright. "Isolation by Distance". *Genetics* 28(2) (1943), pp. 114–138.

[11] G. Coop et al. "The Role of Geography in Human Adaptation". *PLoS Genet* 5(6) (2009), e1000500.

[12] J. Z. Li et al. "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation". *Science* 319(5866) (2008), pp. 1100–1104.

[13] J. K. Pickrell and J. K. Pritchard. "Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data". *PLoS Genet* 8(11) (2012), e1002967.

[14] N. Patterson et al. "Ancient Admixture in Human History". *Genetics* 192(3) (2012), pp. 1065–1093.

[15] S. Wright. "Systems of Mating. V. General Considerations". *Genetics* 6(2) (1921), pp. 167–178.

[16] S. Wright. "Evolution in Mendelian Populations". *Genetics* 16(2) (1931), pp. 97–159.

[17] W. Astle and D. J. Balding. "Population Structure and Cryptic Relatedness in Genetic Association Studies". *Statist. Sci.* 24(4) (2009). Mathematical Reviews number (MathSciNet): MR2779337, pp. 451–471.

[18] E. A. Thompson. "Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations". *Genetics* 194(2) (2013), pp. 301–326.

[19] J. Reynolds, B. S. Weir, and C. C. Cockerham. "Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance". *Genetics* 105(3) (1983), pp. 767–779.

[20] C. C. Cockerham. "Variance of Gene Frequencies". *Evolution* 23(1) (1969), pp. 72–84.

[21] M. Nei. "Analysis of Gene Diversity in Subdivided Populations". *PNAS* 70(12) (1973), pp. 3321–3323.

[22] P. W. Hedrick. "A Standardized Genetic Differentiation Measure". *Evolution* 59(8) (2005), pp. 1633–1638.

[23] L. Jost. "GST and its relatives do not measure differentiation". *Molecular Ecology* 17(18) (2008), pp. 4015–4026.

[24] M. C. Whitlock. "Gst' and D do not replace FST". *Molecular Ecology* 20(6) (2011), pp. 1083–1091.

[25] M. Jakobsson, M. D. Edge, and N. A. Rosenberg. "The Relationship Between FST and the Frequency of the Most Frequent Allele". *Genetics* 193(2) (2013), pp. 515–528.

[26] G. Coop et al. "Using Environmental Correlations to Identify Loci Underlying Local Adaptation". *Genetics* 185(4) (2010), pp. 1411–1423.

[27] A. Moreno-Estrada et al. "The genetics of Mexico recapitulates Native American substructure and affects biomedical traits". *Science* 344(6189) (2014), pp. 1280–1285.

[28] S. Leslie et al. "The fine-scale genetic structure of the British population". *Nature* 519(7543) (2015), pp. 309–314.

[29] W. Haak et al. "Massive migration from the steppe was a source for Indo-European languages in Europe". *Nature* 522(7555) (2015), pp. 207–211.

[30] M. E. Allentoft et al. "Population genomics of Bronze Age Eurasia". *Nature* 522(7555) (2015), pp. 167–172.

[31] T. Thornton et al. "Estimating Kinship in Admixed Populations". *The American Journal of Human Genetics* 91(1) (2012), pp. 122–138.

[32] W. Hao, M. Song, and J. D. Storey. "Probabilistic models of genetic variation in structured populations applied to global human studies". *Bioinformatics* 32(5) (2016), pp. 713–721.

[33] G. Nicholson et al. "Assessing population differentiation and isolation from single-nucleotide polymorphism data". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(4) (2002), pp. 695–715.

[34] J. K. Pritchard, M. Stephens, and P. Donnelly. "Inference of Population Structure Using Multilocus Genotype Data". *Genetics* 155(2) (2000), pp. 945–959.

[35] H. Tang et al. "Estimation of individual admixture: Analytical and study design considerations". *Genet. Epidemiol.* 28(4) (2005), pp. 289–301.

[36] D. H. Alexander, J. Novembre, and K. Lange. "Fast model-based estimation of ancestry in unrelated individuals". *Genome Res.* 19(9) (2009), pp. 1655–1664.

[37] D. Falush, M. Stephens, and J. K. Pritchard. "Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies". *Genetics* 164(4) (2003), pp. 1567–1587.

[38] A. L. Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". *Nat Genet* 38(8) (2006), pp. 904–909.

[39] T. Thornton and M. S. McPeek. "ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure". *The American Journal of Human Genetics* 86(2) (2010), pp. 172–184.

[40] A. Raj, M. Stephens, and J. K. Pritchard. "fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets". *Genetics* 197(2) (2014), pp. 573–589.

[41] A. Ochoa and J. D. Storey. "F_{ST} and kinship for arbitrary population structures II: Method of moments estimators". Submitted.

[42] A. Ochoa and J. D. Storey. "F_{ST} and kinship for arbitrary population structures III: A new estimation framework". In preparation.

[43] D. J. Lawson et al. "Inference of Population Structure using Dense Haplotype Data". *PLoS Genet* 8(1) (2012), e1002453.

[44] G. Hellenthal et al. "A Genetic Atlas of Human Admixture History". *Science* 343(6172) (2014), pp. 747–751.

[45] G. Guillot et al. "A Spatial Statistical Model for Landscape Genetics". *Genetics* 170(3) (2005), pp. 1261–1280.

[46] W.-Y. Yang et al. "A model-based approach for analysis of spatial structure in genetic data". *Nat Genet* 44(6) (2012), pp. 725–731.

[47] J. M. Rañola, J. Novembre, and K. Lange. "Fast spatial ancestry via flexible allele frequency surfaces". *Bioinformatics* 30(20) (2014), pp. 2915–2922.

[48] A. Jacquard. *Structures génétiques des populations.* Paris: Masson et Cie, 1970.

[49] B. S. Weir, A. D. Anderson, and A. B. Hepler. "Genetic relatedness analysis: modern data and new challenges". *Nat Rev Genet* 7(10) (2006), pp. 771–780.

[50] D. Speed and D. J. Balding. "Relatedness in the post-genomic era: is it still useful?" *Nat Rev Genet* 16(1) (2015), pp. 33–44.

[51] P. Mccullagh. *Structured covariance matrices in multivariate regression models.* Tech. rep. 2006.

[52] P. Gopalan et al. "Scaling probabilistic models of genetic variation to millions of humans". *bioRxiv* (2014), p. 013227.

[53] M. Song, W. Hao, and J. D. Storey. "Testing for genetic associations in arbitrarily structured populations". *Nat Genet* 47(5) (2015), pp. 550–554.

[54] S. M. Boca and N. A. Rosenberg. "Mathematical properties of between admixed populations and their parental source populations". *Theoretical Population Biology* 80(3) (2011), pp. 208–216.

[55] J. Yang et al. "GCTA: A Tool for Genome-wide Complex Trait Analysis". *The American Journal of Human Genetics* 88(1) (2011), pp. 76–82.