Admixture, Population Structure, and F-Statistics

Benjamin M. Peter
Department of Human Genetics, University of Chicago, Chicago, Illinois 60637
ORCID ID: 0000-0003-2526-8081 (B.M.P.)

ABSTRACT

Many questions about human genetic history can be addressed by examining the patterns of shared genetic variation between sets of populations. A useful methodological framework for this purpose is F-statistics that measure shared genetic drift between sets of two, three, and four populations and can be used to test simple and complex hypotheses about admixture between populations. This article provides context from phylogenetic and population genetic theory. I review how F-statistics can be interpreted as branch lengths or paths and derive new interpretations, using coalescent theory. I further show that the admixture tests can be interpreted as testing general properties of phylogenies, allowing extension of some ideas applications to arbitrary phylogenetic trees. The new results are used to investigate the behavior of the statistics under different models of population structure and show how population substructure complicates inference. The results lead to simplified estimators in many cases, and I recommend to replace F_3 with the average number of pairwise differences for estimating population divergence.

KEYWORDS
admixture; gene flow; phylogenetics; population genetics; phylogenetic network

For humans, whole-genome genotype data are now available for individuals from hundreds of populations (Lazaridis et al. 2014; Yunusbayev et al. 2015), opening up the possibility to ask more detailed and complex questions about our history (Pickrell and Reich 2014; Schraiber and Akey 2015) and stimulating the development of new tools for the analysis of the joint history of these populations (Reich et al. 2009; Patterson et al. 2012; Pickrell and Pritchard 2012; Lipson et al. 2013; Ralph and Coop 2013; Hellenthal et al. 2013). A simple and intuitive approach that has quickly gained in popularity are the F-statistics, introduced by Reich et al. (2009) and summarized in Patterson et al. (2012). In that framework, inference is based on ‘shared genetic drift’ between sets of populations, under the premise that shared drift implies a shared evolutionary history. Tools based on this framework have quickly become widely used in the study of human genetic history, both for ancient and for modern DNA (Green et al. 2010; Reich et al. 2012; Lazaridis et al. 2014; Allentoft et al. 2015; Haak et al. 2015).

Some care is required with terminology, as the F-statistics sensu Reich et al. (2009) are distinct, but closely related to Wright’s fixation indexes (Wright 1931; Reich et al. 2009), which are also often referred to as F-statistics. Furthermore, it is necessary to distinguish between statistics (quantities calculated from data) and the underlying parameters (which are part of the model) (Weir and Cockerham 1984).

In this article, I mostly discuss model parameters, and I therefore refer to them as drift indexes. The term F-statistics is used when referring to the general framework introduced by Reich et al. (2009), and Wright’s statistics are referred to as F_ ST or f.

Most applications of the F-statistic framework can be phrased in terms of the following six questions:

1. Treeness test: Are populations related in a tree-like fashion (Reich et al. 2009)?
2. Admixture test: Is a particular population descended from multiple ancestral populations (Reich et al. 2009)?
3. Admixture proportions: What are the contributions from different populations to a focal population (Green et al. 2010; Haak et al. 2015)?
4. Number of founders: How many founder populations are there for a certain region (Reich et al. 2012; Lazaridis et al. 2014)?
5. Complex demography: How can mixtures and splits of population explain demography (Patterson et al. 2012; Lipson et al. 2013)?
6. Closest relative: What is the closest relative to a contemporary or ancient population (Raghavan et al. 2014)?

The demographic models under which these questions are addressed, and that motivated the drift indexes, are called population phylogenies and admixture graphs. The population phylogeny (or population tree) is a model where populations are related in a tree-like fashion (Figure 1A), and it frequently serves as the null model for admixture tests. The branch lengths in the population phylogeny correspond to how much genetic drift occurred, so that a branch that is subtended by two different populations can be interpreted as the “shared” genetic drift between these populations. The alternative model is an admixture graph (Figure 1B), which extends the population phylogeny by allowing edges that represent population mergers or a significant exchange of migrants.

Under a population phylogeny, the three \(F\)-statistics proposed by Reich et al. (2009), labeled \(F_2\), \(F_3\), and \(F_4\), have interpretations as branch lengths (Figure 1A) between two, three, and four taxa, respectively. Assume populations are labeled as \(P_1, P_2, \ldots\). Then

\[
F_2(P_1, P_2) \text{ corresponds to the path on the phylogeny from } P_1 \text{ to } P_2.
\]

\[
F_3(P_X; P_1, P_2) \text{ represents the length of the external branch from } P_X \text{ to the (unique) internal vertex connecting all three populations. Thus, the first parameter of } F_3 \text{ has a unique role, whereas the other two can be switched arbitrarily.}
\]

\[
F_4^{(B)}(P_1, P_2; P_3, P_4) \text{ represents the internal branch from the internal vertex connecting } P_1 \text{ and } P_2 \text{ to the vertex connecting } P_3 \text{ and } P_4 \text{ (Figure 1A, blue).}
\]

If the arguments are permuted, some \(F\)-statistics will have no corresponding internal branch. In particular, it can be shown that in a population phylogeny, one \(F_4\) index will be zero, implying that the corresponding internal branch is missing. This is the property that is used in the admixture test. For clarity, I add the superscript \(F_4^{(B)}\) if I need to emphasize the interpretation of \(F_4\) as a branch length and \(F_4^{(T)}\) to emphasize the interpretation as a test statistic. For details, see the \(F_4\) subsection in Methods and Results.

In an admixture graph, there is no longer a single branch length corresponding to each \(F\)-statistic, and interpretations are more complex. However, \(F\)-statistics can still be thought of as the proportion of genetic drift shared between populations (Reich et al. 2009). The basic idea exploited in addressing all six questions outlined above is that under a tree model, branch lengths, and thus the drift indexes, must satisfy some constraints (Buneman 1971; Semple and Steel 2003; Reich et al. 2009). The two most relevant constraints are that (i) in a tree, all branches have positive lengths (tested using the \(F_3\)-admixture test) and (ii) in a tree with four leaves, there is at most one internal branch (tested using the \(F_4\)-admixture test).

The goal of this article is to give a broad overview on the theory, ideas, and applications of \(F\)-statistics. Our starting point is a brief review on how genetic drift is quantified in general and how it is measured using \(F_2\). I then propose an alternative definition of \(F_2\) that allows us to simplify applications and study them under a wide range of population structure models. I then review some basic properties of distance-based phylogenetic trees, show how the admixture tests are interpreted in this context, and evaluate their behavior. Many of the results that are highlighted here are implicit in classical (Wahlund 1928; Wright 1931; Cavalli-Sforza and Edwards 1967; Felsenstein 1973, 1981; Cavalli-Sforza and Piazza 1975; Slatkin 1991; Excoffier et al. 1992) and more recent work (Patterson et al. 2012; Pickrell and Pritchard 2012; Lipson et al. 2013), but often not explicitly stated or given in a different context.

**Methods and Results**

The next sections discuss the \(F\)-statistics, introducing different interpretations and giving derivations for some useful expressions. A graphical summary of the three interpretations of the statistics is given in Figure 2, and the main formulas are summarized in Table 1.

Throughout this article, populations are labeled as \(P_1, P_2, \ldots, P_n, \ldots\). Often, \(P_X\) will denote an admixed population. The allele frequency \(p_i\) is defined as the proportion of individuals in \(P_i\) that carry a particular allele at a biallelic locus, and throughout this article I assume that all individuals are diploid. However, all results hold if instead of haploid individuals, an arbitrary allele of a diploid individual is used. I focus on genetic drift only and ignore the effects of mutation, selection, and other evolutionary forces.
Measuring genetic drift—$F_2$

The purpose of $F_2$ is simply to measure how much genetic drift occurred between two populations, i.e., to measure genetic dissimilarity. For populations $P_1$ and $P_2$, $F_2$ is defined as

$$F_2(P_1, P_2) = F_2(p_1, p_2) = E(p_1 - p_2)^2$$  \hspace{1cm} (1)  

(Reich et al. 2009). The expectation is with respect to the evolutionary process, but in practice $F_2$ is estimated from hundreds of thousands of loci across the genome (Patterson et al. 2012), which are assumed to be nonindependent replicates of the evolutionary history of the populations.

Why is $F_2$ a useful measure of genetic drift? As it is infeasible to observe changes in allele frequency directly, the effect of drift is assessed indirectly, through its impact on genetic diversity. Most commonly, genetic drift is quantified in terms of (i) the variance in allele frequency, (ii) heterozygosity, (iii) probability of identity by descent, (iv) correlation (or covariance) between individuals, and (v) the probability of coalescence (two lineages having a common ancestor). In the next sections I show how $F_2$ relates to these quantities in the cases of a single population changing through time and a pair of populations that are partially isolated.

Single population: I assume a single population, measured at two time points ($t_0$ and $t$), and label the two samples $P_0$ and $P_t$. Then $F_2(P_0, P_t)$ can be interpreted in terms of the variances of allele frequencies:

$$F_2(P_t, P_0) = E[(p_t - p_0)^2]$$

$$= \text{Var}(p_t) + \text{Var}(p_0) - 2\text{COV}(p_0, p_t)$$

$$= \text{Var}(p_t) + \text{Var}(p_0) - 2\text{COV}(p_0, p_0 + (p_t - p_0))$$

$$= \text{Var}(p_t) - \text{Var}(p_0).$$  \hspace{1cm} (2)

Here, I used $E[p_t - p_0] = \text{COV}(p_0, p_t - p_0) = 0$ to obtain lines three and five. It is worth noting that this result holds for any model of genetic drift where the expected allele frequency is
the current allele frequency and increments are independent. For example, this interpretation of $F_2$ holds also if genetic drift is modeled as a Brownian motion (Cavalli-Sforza and Edwards 1967).

An elegant way to introduce the use of $F_2$ in terms of expected heterozygosities $H_t$ (Figure 3B) and identity by descent (Figure 3C) is the duality

$$E[p_1^n | p_0, n_t] = E[p_0^n | p_0, n_t].$$

(3)

This equation is due to Tavaré (1984), who also provided the following intuition: Given $n_t$ individuals are sampled at time $t$, let $E$ denote the event that all individuals carry allele $x$, conditional on allele $x$ having frequency $p_0$ at time $t_0$. There are two components to this: First, the frequency will change between $t_0$ and $t$, and then all $n_t$ sampled individuals need to carry $x$.

In a diffusion framework,

$$P(E) = \int_0^1 y^{n_t} P(p_t = y | p_0, n_t) dy = E[p_1^n | p_0, n_t].$$

(4)

On the other hand, one may argue using the coalescent: For $E$ to occur, all $n_t$ samples need to carry the $x$ allele. At time $t_0$, they had $n_0$ ancestral lineages, who all carry $x$ with probability $p_0$. Therefore,

$$P(E) = \sum_{i=1}^{n_0} p_0^i = E[p_0^n | p_0, n_t].$$

(5)

Equating (4) and (5) yields Equation 3.

In the present case, the only relevant cases are $n_t = 1, 2$, since

$$E[p_1^n | p_0; n_t = 1] = p_0$$

$$E[p_1^n | p_0; n_t = 2] = p_0 f + p_0^2 (1 - f),$$

where $f$ is the probability that two lineage sampled at time $t$ coalesce before time $t_0$.

This yields an expression for $F_2$ by conditioning on the allele frequency $p_0$.

$$E[(p_0 - p_1)^2 | p_0] = E[p_0^2 | p_0] - E[2p_0 p_1 | p_0] + E[p_1^2 | p_0]$$

$$= p_0^2 - 2p_0^2 + p_0 f + p_0^2 (1 - f)$$

$$= f p_0 (1 - p_0)$$

$$= \frac{1}{2} H_0,$$

where $H_0 = 2p_0 (1 - p_0)$ is the heterozygosity. Integrating over $P(p_0)$ yields

$$F_2(P_0, P_t) = \frac{1}{2} f \ E H_0$$

(6)

and it can be seen that $F_2$ increases as a function of $f$ (Figure 3C). This equation can also be interpreted in terms of probabilities of identity by descent: $f$ is the probability that two individuals are identical by descent in $P_t$, given their ancestors were not identical by descent in $P_0$ (Wright 1931), and $E H_0$ is the probability two individuals are not identical in $P_0$.

Furthermore, $E H_t = (1 - f) E H_0$ (equation 3.4 in Wakeley 2009) and therefore

$$E H_0 - E H_t = E H_0 (1 - (1 - f)) = 2 F_2(P_t, P_0),$$

(7)

which shows that $F_2$ measures the decay of heterozygosity (Figure 3A). A similar argument was used by Lipson et al. (2013) to estimate ancestral heterozygosities and to linearize $F_2$.

These equations can be rearranged to make the connection between other measures of genetic drift and $F_2$ more explicit:

$$E H_t = E H_0 - 2 F_2(P_0, P_t)$$

(8a)

$$= E H_0 - 2 (Var(p_t) - Var(p_0))$$

(8b)

$$= E H_0 (1 - f).$$

(8c)

**Pairs of populations:** Equations 8b and 8c describing the decay of heterozygosity appear–of course–well known by population geneticists, having been established by Wright (1931). In structured populations, very similar relationships exist when the number of heterozygotes expected from the overall allele frequency, $H_{obs}$, is compared with the number of heterozygotes present due to differences in allele frequencies between populations $H_{exp}$ (Wahlund 1928; Wright 1931).

In fact, already Wahlund showed by considering the genotypes of all possible matings in two subpopulations (table 3 in Wahlund 1928) that for a population made of two subpopulations with equal proportions, the proportion of heterozygotes is reduced by

$$H_{obs} = H_{exp} - 2(p_1 - p_2)^2$$

from which it follows that
Furthermore, Variance; (B) the decrease in heterozygosity; and (C) the interpretation as probability of coalescence of two lineages or the probability that they are identical by descent.

$$F_2(P_1, P_2) = \frac{\mathbb{E}H_{\text{exp}} - \mathbb{E}H_{\text{obs}}}{2}.$$  

Furthermore, \(\text{Var}(p_1 - p_2) = \mathbb{E}(p_1 - p_2)^2 - [\mathbb{E}(p_1 - p_2)]^2\), but \(\mathbb{E}(p_1 - p_2) = 0\) and therefore

$$F_2(P_1, P_2) = \text{Var}(p_1 - p_2).$$

Finally, the original definition of \(F_2\) was as the numerator of \(F_{ST}\) (Reich et al. 2009), but \(F_{ST}\) can be written as

$$F_{ST} = 2(p_1 - p_2)^2 / \mathbb{E}H_{\text{exp}},$$

from which follows

$$F_2(P_1, P_2) = \frac{1}{2} F_{ST} \mathbb{E}H_{\text{exp}}.$$  

**Covariance interpretation:** To see how \(F_2\) can be interpreted as a covariance, define \(X_i\) and \(X_j\) as indicator variables that two individuals from the same population sample have the A allele, which has frequency \(p_1\) in one and \(p_2\) in the other population. If individuals are equally likely to be sampled from either population,

$$\mathbb{E}X_i = \mathbb{E}X_j = \frac{1}{2} p_1 + \frac{1}{2} p_2$$

$$\mathbb{E}X_i X_j = \frac{1}{2} p_1^2 + \frac{1}{2} p_2^2$$

$$\text{COV}(X_i, X_j) = \mathbb{E}X_i X_j - \mathbb{E}X_i \mathbb{E}X_j$$

$$= \frac{1}{4} (p_1 - p_2)^2 = \frac{1}{4} F_2(P_1, P_2).$$

**Justification for \(F_2\):** The preceding arguments show how the usage of \(F_2\) for both single and structured populations can be justified by the similar effects of \(F_2\) on different measures of genetic drift. However, what is the benefit of using \(F_2\) instead of the established inbreeding coefficient \(f\) and fixation index \(F_{ST}\)? Recall that Wright motivated \(f\) and \(F_{ST}\) as correlation coefficients between alleles (Wright 1921, 1931). Correlation coefficients have the advantage that they are easy to interpret, as, e.g., \(F_{ST} = 0\) implies panmixia and \(F_{ST} = 1\) implies complete divergence between subpopulations. In contrast, \(F_2\) depends on allele frequencies and is highest for intermediate-frequency alleles. However, \(F_2\) has an interpretation as a covariance, making it simpler and mathematically more convenient to work with. In particular, variances and covariances are frequently partitioned into components due to different effects, using techniques such as analysis of variance and analysis of covariance (e.g., Excoffier et al. 1992).

**\(F_2\) as branch length:** Reich et al. (2009) and Patterson et al. (2012) proposed to partition "drift" (as previously established, measured by covariance, allele frequency variance, or decrease in heterozygosity) between different populations into contribution on the different branches of a population phylogeny. This model has been studied by Cavalli-Sforza and Edwards (1967) and Felsenstein (1973) in the context of a Brownian motion process. In this model, drift on independent branches is assumed to be independent, meaning that the variances can simply be added. This is what is referred to as the additivity property of \(F_2\) (Patterson et al. 2012).

To illustrate the additivity property, consider two populations \(P_1\) and \(P_2\) that split recently from a common ancestral population \(P_0\) (Figure 2A). In this case, \(p_1\) and \(p_2\) are assumed to be independent conditional on \(p_0\), and therefore \(\text{COV}(p_1, p_2) = \text{Var}(p_0)\). Then, using (2) and (10),

$$F_2(P_1, P_2) = \text{Var}(p_1 - p_2) = \text{Var}(p_1) + \text{Var}(p_2) - 2\text{COV}(p_1, p_2)$$

$$= \text{Var}(p_1) + \text{Var}(p_2) - 2\text{Var}(p_0)$$

$$= F_2(P_1, P_0) + F_2(P_2, P_0).$$

Alternative proofs of this statement and more detailed reasoning behind the additivity assumption can be found in Cavalli-Sforza and Edwards (1967), Felsenstein (1973), Reich et al. (2009), and Patterson et al. (2012).

Lineages are not independent in an admixture graph, and so this approach cannot be used. Reich et al. (2009) approached this by conditioning on the possible population trees that are consistent with an admixture scenario. In particular, they proposed a framework of counting the possible paths through the graph (Reich et al. 2009; Patterson et al. 2012). An example of this representation for \(F_2\) in a simple admixture graph is given in Supplemental Material, Figure S1, with the result summarized in Figure 2B. Detailed motivation behind this visualization approach is given in Appendix 2 of Patterson et al. (2012). In brief, the reasoning is as follows: Recall that \(F_2(P_1, P_2) = \text{E}(p_1 - p_2)(p_1 - p_2)\) and interpret the two terms in parentheses as two paths between \(P_1\) and \(P_2\),

\[ F_2 = \frac{1}{2} \mathbb{E}H_0 \]
and $F_s$ as the overlap of these two paths. In a population phylogeny, there is only one possible path, and the two paths are always the same; therefore $F_s$ is the sum of the length of all the branches connecting the two populations. However, if there is admixture, as in Figure 2B, both paths choose independently which admixture edge they follow. With probability $\alpha$ they will go left, and with probability $\beta = 1 - \alpha$ they go right. Thus, $F_s$ can be interpreted by enumerating all possible choices for the two paths, resulting in three possible combinations of paths on the trees (Figure S1), and the branches included will differ, depending on which path is chosen, so that the final $F_s$ is made of an average of the path overlap in the topologies, weighted by the probabilities of the topologies.

However, one drawback of this approach is that it scales quadratically with the number of admixture events, making calculations cumbersome when the number of admixture events is large. More importantly, this approach is restricted to panmictic subpopulations and cannot be used when the population model cannot be represented as a weighted average of trees.

**Gene tree Interpretation:** For this reason, I propose to redefine $F_s$, using coalescent theory (Wakeley 2009). Instead of allele frequencies on a fixed admixture graph, coalescent theory tracks the ancestors of a sample of individuals, tracing their history back to their most recent common ancestor. The resulting tree is called a gene tree (or coalescent tree). Gene trees vary between loci and will often have a different topology from that of the population phylogeny, but they are nevertheless highly informative about a population’s history. Moreover, expected coalescence times and expected branch lengths are easily calculated under a wide array of neutral demographic models.

In a seminal article, Slatkin (1991) showed how $F_{ST}$ can be interpreted in terms of the expected coalescence times of gene trees,$$
F_{ST} = \frac{ET_B - ET_W}{ET_B},
$$
where $ET_B$ and $ET_W$ are the expected coalescence times of two lineages sampled in two different populations and the same population, respectively.

Unsurprisingly, given the close relationship between $F_s$ and $F_{ST}$, an analogous expression exists for $F_s (P_1, P_2)$: The derivation starts by considering $F_s$ for two samples of size $1$. I then express $F_s$ for arbitrary sample sizes in terms of individual-level $F_s$ and obtain a sample-size independent expression by letting the sample size $n$ go to infinity.

In this framework, I assume that mutation is rare such that there is at most one mutation at any locus. In a sample of size 2, let $I_i$ be an indicator random variable that individual $i$ has a particular allele. For two individuals, $F_s (I_1, I_2) = 1$ implies $I_1 = I_2$, whereas $F_s (I_1, I_2) = 0$ implies $I_1 \neq I_2$. Thus, $F_s (I_1, I_2)$ is another indicator random variable with the parameter equal to the probability that a mutation happened on the tree branch between $I_1$ and $I_2$.

Now, instead of a single individual $I_1$, consider a sample of $n_1$ individuals: $P_1 = \{I_{1,1}, I_{1,2}, \ldots, I_{1,n_1}\}$. The sample allele frequency is $\bar{p}_1 = n_1^{-1} \sum I_{1,i}$. And the sample $F_s$ is

$$
F_s (\bar{p}_1, I_2) = F_s \left( \frac{1}{n_1} \sum_{i=1}^{n_1} I_{1,i}, I_2 \right) = \mathbb{E} \left( \frac{1}{n_1} \sum_{i=1}^{n_1} I_{1,i} - I_2 \right)^2
$$

$$
= \mathbb{E} \left[ \frac{1}{n_1^2} \sum I_{1,i}^2 + \frac{2}{n_1} \sum I_{1,i} I_{1,j} - \frac{2}{n_1} \sum I_{1,i} I_2 + \frac{1}{n_1^2} \sum I_2^2 \right]
$$

$$
= \mathbb{E} \left[ \frac{1}{n_1^2} \sum I_{1,i}^2 - \frac{2}{n_1} \sum I_{1,i} I_2 + \frac{n_1^2}{n_1} I_2^2 \right]
$$

$$
+ \mathbb{E} \left[ \frac{2}{n_1^2} \sum I_{1,i} I_{1,j} - \frac{n_1^2 - 1}{n_1^2} \sum I_{1,i}^2 \right].
$$

The first three terms can be grouped into $n_1$ terms of the form $F_s (I_{1,i}, I_2)$, and the last two terms can be grouped into $\binom{n_1}{2}$ terms of the form $F_s (I_{1,i}, I_{1,j})$, one for each possible pair of samples in $P_1$.

Therefore,

$$
F_s (\bar{p}_1, I_2) = \frac{1}{n_1} \sum_{i=1}^{n_1} F_s (I_{1,i}, I_2) - \frac{1}{n_1^2} \sum_{i<j} F_s (I_{1,i}, I_{1,j}),
$$

where the second sum is over all pairs in $P_1$. This equation is equivalent to equation 22 in Felsenstein (1973).

As $F_s (p_1, p_2) = F_s (\bar{p}_1, I_2)$, I can switch the labels and obtain the same expression for a second population $P_2 = \{I_{2,i} : i = 0, \ldots, n_2\}$. Taking the average over all $I_{2,i}$ yields

$$
F_s (\bar{p}_1, \bar{p}_2) = \frac{1}{n_1} \sum_i F_s (I_{1,i}, I_{2,i})
$$

$$
- \frac{1}{n_1^2} \sum_{i<j} F_s (I_{1,i}, I_{1,j}) - \frac{1}{n_2^2} \sum_{i<j} F_s (I_{2,i}, I_{2,j}).
$$

Thus, I can write $F_s$ between the two populations as the average number of differences between individuals from different populations, minus some terms including differences within each sample.

Equation 13 is quite general, making no assumptions on where samples are placed on a tree. In a coalescence framework, it is useful to make the assumptions that all individuals from the same population have the same branch length distribution; i.e., $F_s (I_{x,i}, I_{y,i}) = F_s (I_{x,i}, I_{y,j})$ for all pairs of samples $\{x_1, x_2\}$ and $\{y_1, y_2\}$ from populations $P_1$ and $P_2$. Second, I assume that all samples correspond to the leaves of the tree, so that I can estimate branch lengths in terms of the time to a common ancestor $T_{ij}$. Finally, I assume that mutations occur at a constant rate of $\theta/2$ on each branch. Taken together, these assumptions imply that $F_s (I_{1,i}, I_{1,j}) = \theta ET_{ij}$. For all individuals from populations $P_1$, $P_2$, this simplifies (13) to

$$
F_s (\bar{p}_1, \bar{p}_2) = \theta \left( -\frac{1}{2} \mathbb{E} T_{11} - \frac{1}{2} \mathbb{E} T_{12} - \frac{1}{2} \mathbb{E} T_{22} \right).
$$

1490  B. M. Peter
which, for the cases of \( n = 1, 2 \), was also derived by Petkova et al. (2014). In some applications, \( F_2 \) might be calculated only for segregating sites in a large sample. As the expected number of segregating sites is \( \theta/2T_{\text{tot}} \) (with \( T_{\text{tot}} \) denoting the total tree length), taking the limit where \( \theta \to 0 \) is meaningful (Slatkin 1991; Petkova et al. 2014):

\[
F_2(\hat{p}_1, \hat{p}_2) = \frac{2}{T_{\text{tot}}} \times \left( E_{T_{12}} - \frac{1}{2} \left( 1 - \frac{1}{n_1} \right) E_{T_{11}} - \frac{1}{2} \left( 1 - \frac{1}{n_2} \right) E_{T_{22}} \right). \tag{15}
\]

In either of these equations, \( 2/T_{\text{tot}} \) or \( \theta \) acts as a constant of proportionality and will not change the statistical properties, as a larger \( \theta \) implies more mutations and hence more data.

Estimator for \( F_2 \): An estimator for \( F_2 \) can be derived using the average number of pairwise differences \( \pi_{ij} \) as an estimator for \( \theta T_{ij} \) (Tajima 1983). Thus, a natural estimator for \( F_2 \) is

\[
\hat{F}_2(P_1, P_2) = \pi_{12} - \frac{\pi_{11} + \pi_{22}}{2}. \tag{17}
\]

Strikingly, the estimator in Equation 17 is equivalent to that given by Reich et al. (2009) in terms of the sample allele frequency \( \hat{p} \) and sample size \( n \):

\[
F_2(P_1, P_2) = \pi_{12} - \pi_{11}/2 - \pi_{22}/2 = [\hat{p}_1(1 - \hat{p}_2) + \hat{p}_2(1 - \hat{p}_1)] - \hat{p}_1(1 - \hat{p}_1) n_1/n_1 - 1 - \hat{p}_2(1 - \hat{p}_2) n_2/n_2 - 1
\]

\[
= \hat{p}_1 \left( 1 - \frac{1}{n_1} \right) + \hat{p}_2 \left( 1 - \frac{1}{n_2} \right) - 2\hat{p}_1\hat{p}_2 + \hat{p}_1^2 \left( 1 - \frac{1}{n_1} \right) - \hat{p}_2^2 \left( 1 - \frac{1}{n_2} \right) = (\hat{p}_1 - \hat{p}_2) \left( \frac{\hat{p}_1(1 - \hat{p}_1)}{n_1 - 1} - \frac{\hat{p}_2(1 - q_2)}{n_2 - 1} \right) - \hat{p}_1 \hat{p}_2.
\]

The last line is the same as equation 10 in the appendix of Reich et al. (2009).

However, while the estimators are identical, the underlying modeling assumptions are different: The original definition considered only loci that were segregating in an ancestral population; loci not segregating there were discarded. Since ancestral populations are usually unsampled, this is often replaced by ascertainment in an outgroup (Patterson et al. 2012; Lipson et al. 2013). In contrast, Equation 17 assumes that all markers are used, which is the more natural interpretation for sequence data.

Gene tree branch lengths: An important feature of Equation 16 is that it depends only on the expected coalescence times between pairs of lineages. Thus, the behavior of \( F_2 \) can be fully characterized by considering a sample of size \( 4 \), with two random individuals taken from each population. This is all that is needed to study the joint distribution of \( T_{12}, T_{11}, \) and \( T_{22} \) and hence \( F_2 \). By linearity of expectation, larger samples can be accommodated by summing the expectations over all possible quartets.

For a sample of size \( 4 \) with two pairs, there are only two possible unrooted tree topologies: one, where the lineages from the same population are more closely related to each other [called concordant topology, \( T_c(2) \)] and one where lineages from different populations coalesce first [which I refer to as discordant topology, \( T_d(2) \)]. The superscripts refers to the topologies being for \( F_2 \), and I discard them in cases where no ambiguity arises.

Conditioning on the topology yields

\[
F_2(P_1, P_2) = \mathbb{E}[F_2(P_1, P_2)|T] = \mathbb{P}(T_c|T)\mathbb{E}[F_2(P_1, P_2)|T_c] + \mathbb{P}(T_d|T)\mathbb{E}[F_2(P_1, P_2)|T_d].
\]

Figure 4 contains graphical representations for \( \mathbb{E}[F_2(P_1, P_2)|T_c] \) (Figure 4B) and \( \mathbb{E}[F_2(P_1, P_2)|T_d] \) (Figure 4C), respectively.

In this representation, \( T_{12} \) corresponds to a path from a random individual from \( P_1 \) to a random individual from \( P_2 \), and \( T_{11} \) represents the path between the two samples from \( P_1 \). For \( T_c \), the internal branch is always included in \( T_{12} \), but never in \( T_{11} \) or \( T_{22} \). External branches, on the other hand, are included with 50% probability in \( T_{12} \) on any path through the tree. \( T_{11} \) and \( T_{22} \), on the other hand, consist only of external branches, and the lengths of the external branches cancel.

On the other hand, for \( T_d \), the internal branch is always included in \( T_{11} \) and \( T_{22} \), but only half the time in \( T_{12} \). Thus, they contribute negatively to \( F_2 \), but only with half the magnitude of \( T_c \). As for \( T_d \), each \( T \) contains exactly two external branches, cancelling the external branches from \( T_{12} \).

An interesting way to represent \( F_2 \) is therefore in terms of the internal branches over all possible gene genealogies. Denote the unconditional average length of the internal branch of \( T_c \) as \( B_c \) and the average length of the internal branch in \( T_d \) as \( B_d \). Then, \( F_2 \) can be written in terms of these branch lengths as
resulting in the representation given in Figure 2, C and D.

As a brief sanity check, consider the case of a population without structure. In this case, the branch length is independent of the topology and $T_d$ is twice as likely as $T_c$ and hence $B_d = 2B_c$, from which it follows that $F_2$ will be zero, as expected in a randomly mating population.

This argument can be transformed from branch lengths to observed mutations by recalling that mutations occur on a branch at a rate proportional to its length. $F_2$ is increased by doubletons that support the assignment of populations (i.e., the two lineages from the same population have the same allele), but reduced by doubletons shared by individuals from different populations. All other mutations have a contribution of zero.

### Testing treeness

Many applications consider tens or even hundreds of populations simultaneously (Patterson et al. 2012; Pickrell and Pritchard 2012; Haak et al. 2015; Yunusbayev et al. 2015), with the goal to infer where and between which populations admixture occurred. Using $F$-statistics, the approach is to interpret $F_2(P_1, P_2)$ as a measure of dissimilarity between $P_1$ and $P_2$, as a large $F_2$ value implies that populations are highly diverged. Thus, the strategy is to calculate all pairwise $F_2$ indexes between populations, combine them into a dissimilarity matrix, and ask whether that matrix is consistent with a tree.

One way to approach this question is by using phylogenetic theory: Many classical algorithms have been proposed that use a measure of dissimilarity to generate a tree (Fitch et al. 1967; Saitou and Nei 1987; Semple and Steel 2003; Felsenstein 2004) and what properties a general dissimilarity matrix needs to have to be consistent with a tree (Buneman 1971; Cavalli-Sforza and Piazza 1975), in which case the matrix is also called a tree metric (Semple and Steel 2003). Thus, testing for admixture can be thought of as testing treeness.

For a dissimilarity matrix to be consistent with a tree, there are two central properties it needs to satisfy: First, the length of all branches has to be positive. This is strictly not necessary for phylogenetic trees, and some algorithms may return negative branch lengths (e.g. Saitou and Nei 1987); however, since in our case branches have an interpretation of genetic drift, negative genetic drift is biologically nonsensical, and therefore negative branches should be interpreted as a violation of the modeling assumptions and hence of treeness.

The second property of a tree metric important in the present context is a bit more involved: A dissimilarity matrix (written in terms of $F_2$) is consistent with a tree if for any four populations $P_1$, $P_2$, $P_k$, and $P_l$,

$$F_2(P_1, P_j) + F_2(P_k, P_l) \leq \max(F_2(P_1, P_k) + F_2(P_j, P_l), F_2(P_1, P_l) + F_2(P_j, P_k)),$$

that is, if the sums of pairs of distances are compared, two of these sums will be the same, and no smaller than the third one. This theorem, due to Buneman (1971, 1974), is called the four-point condition or sometimes, more modestly, the “fundamental theorem of phylogenetics.” A proof can be found in Semple and Steel (2003, Chap. 7).

Informally, this statement can be understood by noting that on a tree, two of the pairs of distances will include the internal branch, whereas the third one will not and therefore be shorter. Thus, the four-point condition can be colloquially rephrased as “any four-taxa tree has at most one internal branch.”

Why are these properties useful? It turns out that the admixture tests based on $F$-statistics can be interpreted as tests of these properties: The $F_2$ test can be interpreted as a test for the positivity of a branch and the $F_4$ as a test of the four-point condition. Thus, the working of the two test statistics can be interpreted in terms of fundamental properties of phylogenetic trees, with the immediate consequence that they may be applied as treeness tests for arbitrary dissimilarity matrices.

An early test of treeness, based on a likelihood ratio, was proposed by Cavalli-Sforza and Piazza (1975): They compared the likelihood of the observed $F_2$ matrix to that induced
by the best-fitting tree (assuming Brownian motion), rejecting the null hypothesis if the tree likelihood is much lower than that of the empirical matrix. In practice, however, finding the best-fitting tree is a challenging problem, especially for large trees (Felsenstein 2004), and so the likelihood test proved to be difficult to apply. From that perspective, the $F_3$ and $F_4$ tests provide a convenient alternative: Since treeness implies that all subsets of taxa are also trees, the ingenious idea of Reich et al. (2009) was that rejection of treeness for subtrees of sizes 3 (for $F_3$) and 4 (for $F_4$) is sufficient to reject treeness for the entire tree. Furthermore, tests on these subtrees also pinpoint the populations involved in the non-tree-like history.

**$F_3$: Three population statistic**

In the previous section, I showed how $F_2$ can be interpreted as a branch length, as an overlap of paths, or in terms of gene trees (Figure 2F). Furthermore, I gave expressions in terms of coalescence times, allele frequency variances, and internal branch lengths of gene trees. In this section, I give analogous results for $F_3$.

Reich et al. (2009) defined $F_3$ as

$$F_3(P_X; P_1, P_2) = F_3(p_X; p_1, p_2) = \mathbb{E}(p_X - p_1)(p_X - p_2)$$

(20a)

with the goal to test whether $P_X$ is admixed. Recalling the path interpretation detailed in Patterson et al. (2012), $F_3$ can be interpreted as the shared portion of the paths from $P_X$ to $P_1$ with the path from $P_X$ to $P_1$. In a population phylogeny (Figure 2E) this corresponds to the branch between $P_X$ and the internal node. Equivalently, $F_3$ can also be written in terms of $F_2$ (Reich et al. 2009):

$$F_3(P_X; P_1, P_2) = \frac{1}{2} \left( F_2(P_X, P_1) + F_2(P_X, P_2) - F_2(P_1, P_2) \right).$$

(20b)

If $F_2$ in Equation 20b is generalized to an arbitrary tree metric, Equation 20b is known as the Gromov product in phylogenetics (Semple and Steel 2003). The Gromov product is a commonly used operation in classical phylogenetic algorithms to calculate the length of the portion of a branch shared between $P_1$ and $P_2$ (Fitch et al. 1967; Felsenstein 1973; Saitou and Nei 1987), consistent with the notion that $F_3$ is the length of an external branch in a population phylogeny.

In an admixture graph, there is no longer a single external branch; instead all possible trees have to be considered, and $F_3$ is the (weighted) average of paths through the admixture graph (Figure 2F).

Combining Equations 16 and 20b gives an expression of $F_3$ in terms of expected coalescence times:

$$F_3(P_X; P_1, P_2) = \frac{\theta}{2} \left( \mathbb{E}T_{1X} + \mathbb{E}T_{2X} - \mathbb{E}T_{12} - \mathbb{E}T_{XX} \right).$$

(20c)

Similarly, an expression in terms of variances is obtained by combining Equation 2 with Equation 20b,

$$F_3(P_X; P_1, P_2) = \text{Var}(p_X) + \text{COV}(p_1, p_2) - \text{COV}(p_1, p_X) - \text{COV}(p_2, p_X).$$

(20d)

which was also noted by Pickrell and Pritchard (2012).

**Outgroup $F_3$ statistics:**

A simple application of the interpretation of $F_3$ as a shared branch length are the “outgroup” $F_3$ statistics proposed by Raghavan et al. (2014). For an unknown population $P_3$, they wanted to find the most closely related population from a panel of $k$ extant populations $\{P_i, i = 1, 2, \ldots, k\}$ They did this by calculating $F_3(P_3; P_0, P_i)$, where $P_0$ is an outgroup population that was assumed widely diverged from $P_3$ and all populations in the panel. This measures the shared drift (or shared branch) of $P_3$ with the populations from the panel, and high $F_3$ values imply close relatedness.

However, using Equation 20c, the outgroup $F_3$ statistic can be written as

$$F_3(P_3; P_0; P_i, P_i) = \mathbb{E}T_{1O} + \mathbb{E}T_{IO} - \mathbb{E}T_{U1} - \mathbb{E}T_{OO}.$$ (21)

Of these four terms, $\mathbb{E}T_{1O}$ and $\mathbb{E}T_{OO}$ do not depend on $P_i$. Furthermore, if $P_0$ is truly an outgroup, then all $\mathbb{E}T_{IO}$ should be the same, as pairs of individuals from the panel population and the outgroup can coalesce only once they are in the joint ancestral population. Therefore, only the term $\mathbb{E}T_{1O}$ is expected to vary between different panel populations, suggesting that using the number of pairwise differences, $\pi_{1i}$, is largely equivalent to using $F_3$ ($F_3(P_0; P_i, P_i)$). I confirm this in Figure 5A by calculating outgroup $F_3$ and $\pi_{1i}$ for a set of increasingly divergent populations, with each population having its own size, sample size, and sequencing error probability. Linear regression confirms the visual picture that $\pi_{1i}$ has a higher correlation with divergence time ($R^2 = 0.90$) than $F_3$ ($R^2 = 0.73$). Hence, the number of pairwise differences may be a better metric for population divergence than $F_3$.

**$F_3$ admixture test:**

However, $F_3$ is motivated and primarily used as an admixture test (Reich et al. 2009). In this context, the null hypothesis is that $F_3$ is nonnegative; i.e., the null hypothesis is that the data are generated from a phylogenetic tree that has positive edge lengths. If this is not the case, the null hypothesis is rejected in favor of the more complex admixture graph. From Figure 2F it may be seen that drift on the path on the internal branches (red) contributes negatively to $F_3$. If these branches are long enough compared to the branch after the admixture event (blue), then $F_3$ will be negative. For the simplest scenario where $P_X$ is admixed between $P_1$ and $P_2$, Reich et al. (2009) provided a condition when this is the case (equation 20 in supplement 2 of Reich et al. 2009). However, since this condition involves $F$-statistics with internal, unobserved populations, it cannot be used in practical applications. A more useful condition is obtained using Equation 20c.
In the simplest admixture model, an ancestral population splits into $P_1$ and $P_2$ at time $t_r$. At time $t_1$, the populations mix to form $P_X$, such that with probability $\alpha$, individuals in $P_X$ descend from individuals from $P_1$, and with probability $(1 - \alpha)$, they descend from $P_2$ (see Figure 7 for an illustration). In this case, $F_3$ ($P_X; P_1, P_2$) is negative if

$$\frac{t_1}{(1 - c_x)} t_r < 2\alpha(1 - \alpha), \quad (22)$$

where $c_x$ is the probability two individuals sampled in $P_X$ have a common ancestor before $t_1$. For a randomly mating population with changing size $N(t)$,

$$c_x = 1 - \exp\left(-\int_0^{t_1} \frac{1}{N(s)} ds\right).$$

Thus, the power of $F_3$ to detect admixture is large (1) if the admixture proportion $\alpha$ is close to 50%; (2) if the ratio between the times of the original split and the time of secondary contact is large; and (3) if the probability of coalescence before the admixture event in $P_X$ is small, i.e., the size of $P_X$ is large.

A more general condition for negativity of $F_3$ is obtained by considering the internal branches of the possible gene tree topologies, analogously to that given for $F_2$ in the Gene tree branch lengths section. Since Equation 20c includes $\mathbb{E}T_{XX}$, only two individuals from $P_X$ are needed and one each from $P_1$ and $P_2$ to study the joint distribution of all terms in (20c). The minimal case therefore contains again just four samples (Figure S2).

Furthermore, $P_1$ and $P_2$ are exchangeable, and thus there are again just two distinct genealogies, a concordant one $T^{(3)}_c$ where the two lineages from $P_X$ are most closely related and a discordant genealogy $T^{(3)}_d$ where the lineages from $P_X$ merge first with the other two lineages. A similar argument to that for $F_2$ shows (presented in Figure S2) that $F_4$ can be written as a function of just the internal branches in the topologies,

$$F_3(P_X; P_1, P_2) = \theta(2B_c - B_d), \quad (23)$$

where $B_c$ and $B_d$ are the lengths of the internal branches in $T^{(3)}_c$ and $T^{(3)}_d$, respectively, and similar to $F_2$, concordant branches have twice the weight of discordant ones. Again, the case of all individuals coming from a single population serves as a sanity check: In this case $T_d$ is twice as likely as $T_c$, and all branches are expected to have the same length, resulting in $F_3$ being zero. However, for $F_3$ to be negative, note that $B_d$ needs to be more than twice longer than $B_c$. Since mutations are proportional to $B_d$ and $B_c$, $F_3$ can be interpreted as a test whether mutations that agree with the population tree are more than twice as common as mutations that disagree with it.

I performed a small simulation study to test the accuracy of Equation 22. Parameters were chosen such that $F_3$ has a negative expectation for $\alpha > 0.05$, and I find that the predicted $F_3$ fitted very well with the simulations (Figure 5B).

### $F_4$: Four population study

The second admixture statistic, $F_4$, is defined as

$$F_4(P_1, P_2; P_3, P_4) = F_4(p_1, p_2; p_3, p_4) = \mathbb{E}\left[(p_1 - p_2)(p_3 - p_4)\right] \quad (24a)$$

(Reich et al. 2009). Similarly to $F_3$, $F_4$ can be written as a linear combination of $F_2$,

$$F_4(P_1, P_2; P_3, P_4) = \frac{1}{2} \left(F_2(P_1, P_2) + F_2(P_2, P_3) - F_2(P_1, P_3) - F_2(P_2, P_4)\right), \quad (24b)$$

which leads to

$$F_4(P_1, P_2; P_3, P_4) = \frac{\theta}{2} \left(\mathbb{E}T_{14} + \mathbb{E}T_{23} - \mathbb{E}T_{13} - \mathbb{E}T_{24}\right). \quad (24c)$$

As four populations are involved, there are $4! = 24$ possible ways of arranging the populations in Equation 24a. However, there are four possible permutations of arguments that will lead to identical values, leaving only six unique $F_4$ values for any four populations. Furthermore, these six values come in pairs that have the same absolute value and a different sign [i.e., $F_4(P_1, P_2; P_3, P_4) = -F_4(P_1, P_2; P_4, P_3)$], leaving only three unique absolute values, which correspond to the three possible tree topologies. Of these three, one $F_4$ can be written as the sum of the other two, leaving just two independent possibilities:

$$F_4(P_1, P_2; P_3, P_4) + F_4(P_1, P_3; P_2, P_4) = F_4(P_1, P_4; P_2, P_3).$$

As for $F_3$, Equation 24b can be generalized by replacing $F_2$ with an arbitrary tree metric. In this case, Equation 24b is known as a tree split (Buneman 1971), as it measures the length of the overlap of the branch lengths between the two pairs. As there are two independent $F_3$ indexes for a fixed tree, there are two different interpretations for the $F_4$ indexes. Consider the tree from Figure 1A: $F_4(P_1, P_2; P_3, P_4)$ can be interpreted as the overlap between the paths from $P_1$ to $P_2$ and from $P_3$ to $P_4$. However, these paths do not overlap in Figure 1A, and therefore $F_4 = 0$. This is how $F_4$ is used as a test statistic. On the other hand, $F_4(P_1, P_3; P_2, P_4)$ measures the overlap between the paths from $P_1$ to $P_3$ and from $P_2$ to $P_4$, which is the internal branch in Figure 1A, and will be positive.

It is cumbersome that the interpretation of $F_4$ depends on the ordering of its arguments. To make the intention clear, instead of switching the arguments around for the two interpretations, I introduce the superscripts (T) (for test) and (B) (for branch length):

$$F_4^{(T)}(P_1, P_2; P_3, P_4) = F_4(P_1, P_2; P_3, P_4) \quad (25a)$$

$$F_4^{(B)}(P_1, P_2; P_3, P_4) = F_4(P_1, P_3; P_2, P_4). \quad (25b)$$
Four-point condition and \( F_4 \): Tree splits, and hence \( F_4 \), are closely related to the four-point condition (Buneman 1971, 1974), which, informally, states that a (sub)tree with four populations will have at most one internal branch. Thus, if data are consistent with a tree, \( F_4^{(B)} \) will be the length of that branch, and \( F_4^{(T)} \) will be zero. Figure 2, I–L, corresponds to the internal branch and Figure 2, M–P, to the “zero” branch.

Thus, in the context of testing for admixture, testing that \( F_4 \) is zero is equivalent to checking whether there is in fact only a single internal branch. If that is not the case, the population phylogeny is rejected. This statement can be generalized to arbitrary tree metrics: The four-point condition (Buneman 1971) can be written as

\[
F_2 (P_1, P_2) + F_2 (P_3, P_4) \leq \min \{ F_2 (P_1, P_3) + F_2 (P_2, P_4), F_2 (P_1, P_4) + F_2 (P_2, P_3) \} \tag{26}
\]

for any permutations of the samples. This implies that two of the sums need to be the same and larger than the third one. The claim is that if the four-point condition holds, at least one of the \( F_4 \) values will be zero, and the others will have the same absolute value.

Without loss of generality, assume that

\[
F_2 (P_1, P_2) + F_2 (P_3, P_4) = F_2 (P_1, P_3) + F_2 (P_2, P_4) \leq F_2 (P_1, P_4) + F_2 (P_2, P_3).
\]

Simply plugging this into the three possible \( F_4 \) equations yields

\[
\begin{align*}
F_4 (P_1, P_2; P_3, P_4) &= 0 \\
F_4 (P_1, P_3; P_2, P_4) &= k \\
F_4 (P_1, P_4; P_2, P_3) &= -k,
\end{align*}
\]

where \( k = F_2 (P_1, P_3) + F_2 (P_2, P_4) - F_2 (P_1, P_2) + F_2 (P_3, P_4). \)

It is worth noting that the converse is false. If

\[
F_2 (P_1, P_2) + F_2 (P_3, P_4) > F_2 (P_1, P_3) + F_2 (P_2, P_4) \\
F_2 (P_1, P_3) + F_2 (P_2, P_4) = F_2 (P_1, P_4) + F_2 (P_2, P_3),
\]

the four-point condition is violated, but \( F_4 (P_1, P_2; P_3, P_4) \) is still zero, and the other two \( F_4 \) values have the same magnitude.

Gene trees: Evaluating \( F_4 \) in terms of gene trees and their internal branches, there are three different gene tree topologies that have to be considered, whose interpretation depends on whether the branch length or test-statistic interpretation is considered.

For the branch length \( [F_4^{(B)}] \), the gene tree corresponding to the population tree has a positive contribution to \( F_4 \), and the other two possible trees have a zero and negative contribution, respectively (Figure S3). Since the gene tree corresponding to the population tree is expected to be most frequent, \( F_4 \) will be positive and can be written as

\[
F_4^{(B)} = \theta(B_c - B_d). \tag{27}
\]

This equation is slightly different from those for \( F_2 \) and \( F_3 \), where the coefficient for the discordant genealogy was half that for the concordant genealogy. Note, however, that \( F_4 \) includes only one of the two discordant genealogies. Under a tree, both discordant genealogies are equally likely (Durand et al. 2011), and thus the expectation of \( F_4 \) will be the same.

In contrast, for the admixture test statistic \( [F_4^{(T)}] \), the contribution of the discordant genealogy will be zero, and the discordant genealogies will contribute with coefficients \(-1\) and \(+1\), respectively and thus the expectation of \( F_4 \) as a test statistic

\[
F_4^{(T)} = \theta(B_c - B_d) \tag{28}
\]
is zero under the null hypothesis. Furthermore, the statistic is
closely related to the ABBA-BABA or D-statistic also used to
test for admixture (Green et al. 2010; Durand et al. 2011),
which includes a normalization term and conditions on al-
leles being derived. In our notation the expectation of D is
\[
E[D] = \frac{B_d - B_d}{B_d + B_d}
\]
and thus, \( F^{(T)}_4 \) and D are different test statistics for the same
null hypothesis.

Rank test: Two major applications of \( F_4 \) use its interpretation
as a branch length. First, the rank of a matrix of all \( F_4 \) statistics is
used to obtain a lower bound on the number of admixture
events required to explain data (Reich et al. 2012). The principal
idea of this approach is that the number of internal branches in a
genealogy is bounded to be at most \( n - 3 \) in an unrooted tree.
Since each \( F_4 \) is a sum of the length of tree branches, all \( F_4 \)
indexes should be sums of \( n - 3 \) branches or \( n - 3 \) independent
components. This implies that the rank of the matrix (see, e.g.,
section 4 in McCullagh 2009) is at most \( n - 3 \), if the data are
consistent with a tree. However, admixture events may increase
the rank of the matrix, as they add additional internal branches
(Reich et al. 2012). Therefore, if the rank of the matrix is \( r \), the
number of admixture events is at least \( r - n + 3 \).

One issue is that the full \( F_4 \) matrix has size \( \binom{n}{2} \times \binom{n}{2} \) and
may thus become rather large. Furthermore, in many cases only admixture
events in a certain part of the phylogeny are of interest. To estimate the minimum number of admixture events on a particular branch of the phylogeny, Reich et al. (2012)
proposed to find two sets of test populations \( S_1 \) and \( S_2 \) and
two reference populations for each set \( R_1 \) and \( R_2 \) that are
presumed unadmixed (see Figure 6A). Assuming a phylogeny, all
\( F_4^{(B)} (S_1, R_1; S_2, R_2) \) will measure the length of the same branch,
and all \( F_4^{(T)} (S_1, R_1; S_2, R_2) \) should be zero. Since each admixture
event introduces at most one additional branch, the rank of the
resulting matrix will increase by at most one, and the rank of
either the matrix of all \( F_4^{(T)} \) or the matrix of all \( F_4^{(B)} \) may reveal
the number of branches of that form.

Admixture proportion: The second application is by compar-
ring branches between closely related populations to obtain an
estimate of mixture proportion or how much two focal populations
correspond to an admixed population (Green et al. 2010):
\[
\alpha = \frac{F_4 (P_0, P_1; P_2, P_1)}{F_4 (P_0, P_1; P_2, P_1)}
\]
(29)
Here, \( P_X \) is the population whose admixture proportion is esti-
mated; \( P_1 \) and \( P_2 \) are the potential contributors, where I assume
that they contribute with proportions \( \alpha \) and \( 1 - \alpha \), respectively;
and \( P_0 \), \( P_1 \) are reference populations with no direct contribution
to \( P_X \) (see Figure 6B). \( P_1 \) has to be more closely related to one of
\( P_1 \) or \( P_2 \) than the other, and \( P_0 \) is an outgroup.

The canonical way (Patterson et al. 2012) to interpret this
ratio is as follows: The denominator is the branch length from
the common ancestor population from \( P_1 \) and \( P_1 \) to the com-
mon ancestor of \( P_1 \) with \( P_2 \) (Figure 6C, yellow line). The
numerator has a similar interpretation as an internal branch
(Figure 6C, red dotted line). In an admixture scenario (Figure
6B), this is not unique and is replaced by a linear combination
of lineages merging at the common ancestor of \( P_1 \) and \( P_1 \)
(with probability \( \alpha \) and lineages merging at the common
ancestor of \( P_1 \) with \( P_2 \) (with probability \( 1 - \alpha \)).

Thus, a more general interpretation is that \( \alpha \) measures how
much closer the common ancestor of \( P_X \) and \( P_1 \) is to the
common ancestor of \( P_1 \) and \( P_1 \), and \( P_1 \) and \( P_1 \) indicated by the red dotted line in Figure 6C. This quantity
is defined also when the assumptions underlying the admix-
ture test are violated and, if the assumptions are not carefully
checked, might lead to misinterpretations of the data. In par-

cular, \( \alpha \) is well defined in cases where no admixture oc-
curred or in cases where either one of \( P_1 \) and \( P_2 \) did not
experience any admixture.

Furthermore, it is evident from Figure 6 that if all populations
are sampled at the same time, \( \Xi T_{10} = \Xi T_{01} = \Xi T_{02} = \Xi T_{03} \),
and therefore
\[
\alpha = \frac{\Xi T_{11} - \Xi T_{10}}{\Xi T_{11} - \Xi T_{12}}.
\]
(30)
Thus,
\[
\alpha = \frac{\Xi_{11} - \Xi_{10}}{\Xi_{11} - \Xi_{12}}
\]
(31)
is another estimator for \( \alpha \) that can be used even if no outgroup is
available. I compare Equations 29 and 31 for varying admixture
proportions in Figure 5C, using the mean absolute error in the
admixture proportion. Both estimators perform very well, but
(31) performs slightly better in cases where the admixture pro-
portion is low. However, in most cases this minor improvement
possibly does not negate the drawback that Equation 31 is ap-
licable only when populations are sampled at the same time.

An area of recent development is how these estimates can be
extended to more populations. A simple approach is to
assume a fixed series of admixture events, in which case
admixture proportions for each event can be extracted from
a series of \( F_4 \) ratios (Lazaridis et al. 2014, SI 13). A more
sophisticated approach estimates mixture weights using the
rank of the \( F_4 \) matrix, as discussed in the Rank test section
(Haak et al. 2015, SI 10). Then, it is possible to estimate
admixture proportions, using a model similar to that introduced
in the program structure (Pritchard et al. 2000), by obtaining
a low-rank approximation for the \( F_4 \) matrix.

Population structure models
Here, I use Equation 16 together with Equations 20b and 24b to
derive expectations for \( F_3 \) and \( F_4 \) under some simple models.

Panmixia: In a randomly mating population (with arbitrary
population size changes), \( P_1 \) and \( P_2 \) are taken from the same
pool of individuals and therefore \( \mathbb{E}T_{12} = \mathbb{E}T_{11} = \mathbb{E}T_{22}, \mathbb{E}F_2 = \mathbb{E}F_3 = \mathbb{E}F_4 = 0. \)

**Island models:** A (finite) island model has \( D \) subpopulations of size 1 each. Migration occurs at rate \( M \) between subpopulations. It can be shown (Strobeck 1987) that \( \mathbb{E}T_{11} = \mathbb{E}T_{22} = D (\ldots) \), and \( \mathbb{E}T_{12} \) satisfies

\[
\mathbb{E}T_{12} = \frac{1}{D-1} M + \frac{D-2}{D-1} \mathbb{E}T_{12} + \frac{1}{D-1} \mathbb{E}T_{11}
\]

with solution \( \mathbb{E}T_{12} = 1 + M^{-1} \). This results in the equation in Figure 7. The derivation of coalescence times for the hierarchical island models is marginally more complicated, but similar. It is given in Slatkin and Voelm (1991).

**Admixture models:** These are the models for which the \( F \)-statistics were originally developed. Many details, applications, and the origin of the path representation are found in Patterson et al. (2012). For simplicity, I look at the simplest possible tree with four populations, where \( P_X \) is admixed from \( P_1 \) and \( P_2 \) with contributions \( \alpha \) and \( \beta = (1 - \alpha) \), respectively. I assume that all populations have the same size and that this size is 1. Then,

\[
F_3(P_X; P_1, P_2) \propto \mathbb{E}T_{1X} + \mathbb{E}T_{2X} - \mathbb{E}T_{12} - \mathbb{E}T_{XX} = (\alpha t_1 + \beta t_r + 1) + (\alpha t_r + \beta t_1 + 1) - t_r - 1 - \alpha^2 1 - (1-\alpha)2^1 - 2\alpha(1-\alpha)[(1-c_x)t_r + 1] = t_1 - 2\alpha(1-\alpha)(1-c_x)t_r.
\]

Here, \( c_x \) is the probability that the two lineages from \( P_X \) coalesce before the admixture event.

Thus, \( F_3 \) is negative if

\[
\frac{t_1}{(1-c_x)t_r} < 2\alpha(1-\alpha),
\]

which is more likely if \( \alpha \) is large, the admixture is recent, and the overall coalescent is far in the past.

For \( F_4 \), omitting the within-population coalescence time of 1,

\[
F_4(P_1P_XP_2P_3) \propto \mathbb{E}T_{12} + \mathbb{E}T_{3X} - \mathbb{E}T_{13} - \mathbb{E}T_{2X} = t_r + \alpha t_r + \beta t_{23} - t_r - \alpha t_r - \beta t_{2X} = \beta(t_2 - t_1).
\]

**Stepping-stone models:** For the stepping-stone models, I have to solve the recursions of the Markov chains describing the location of all lineages in a sample of size 2. For the standard stepping-stone model, I assumed there were four demes, all of which exchange migrants at rate \( M \). This results in a Markov chain with the following five states: (i) lineages in same deme, (ii) lineages in demes 1 and 2, (iii) lineages in demes 1 and 3, (iv) lineages in demes 1 and 4, and (v) lineages in demes 2 and 3. Note that the symmetry of this system allows collapsing some states. The transition matrix for this system is

\[
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
2M & 1 - 3M & 0 & 0 & 0 \\
0 & M & 1 - 3M & M & M \\
0 & 0 & 2M & 1 - 2M & 0 \\
2M & 0 & 2M & 0 & 1 - 4M
\end{pmatrix}.
\]

Once lineages are in the same deme, the system terminates as the time to coalescence time is independent of the deme in isotropic migration models (Strobeck 1987) and cancels from the \( F \)-statistics.

Therefore, the vector \( \nu \) of the expected time until two lineages are in the same deme is found using standard Markov chain theory by solving \( \nu = (I - T)^{-1}1 \), where \( T \) is the transition matrix involving only the transitive states in the Markov chain (all but the first state), and \( 1 \) is a vector of 1’s.

Finding the expected coalescence time involves solving a system of five equations. The terms involved in calculating the \( F \)-statistics (Table 1) are the entries in \( \nu \) corresponding to these states.

The hierarchical case is similar, except there are six demes and 10 equations. Representing states as lineages being in demes (same), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 3), (2, 4), (2, 5), (3, 4),

As in the nonhierarchical case, solving this system yields all pairwise coalescence times. Then, all I have to do is average the coalescence times over all possibilities; e.g.,

\[
\mathbb{E}T_{1X} = \frac{\nu_2 + \nu_3 + \nu_6 + \nu_7}{4}.
\]

For \( F_4 \), I assume that demes 1 and 2 are in \( P_1 \), demes 3 and 4 are in \( P_X \), and demes 5 and 6 correspond to \( P_2 \) and \( P_3 \), respectively.

**Range expansion model:** I use a serial founder model with no migration (Peter and Slatkin 2015), where I assume that the expansion is recent enough such that the effect of migration
after the expansion finished can be ignored. Under that model, I assume that samples $P_1$ and $P_2$ are taken from demes $D_1$ and $D_2$, with $D_1$ closer to the origin of the expansion and populations with high identification numbers even farther away from the expansion origin. Then $E \tau_{12} = t_1 + ET_{11}$, where $\tau_{11}$ is the time required for a lineage sampled farther away in the expansion to end up in $D_1$. (Note that $t_1$ depends only on the deme that is closer to the origin.) Thus, for three demes,

$$F_3(P_2; P_1, P_2) \propto ET_{12} - ET_{13} + ET_{23} - ET_{22}$$

$$\propto ET_{11} + t_1 - ET_{11} - t_1 + ET_{22} + t_2 - ET_{22}$$

and

$$F_4^{(T)}(P_1, P_2; P_3, P_4) \propto ET_{13} - ET_{14} + ET_{24} - ET_{23}$$

$$\propto ET_{11} + t_1 - ET_{11} - t_1 + ET_{22} + t_2 - ET_{22} - t_2$$

$$= 0.$$
More interesting is

\[ F_4^{(B)}(p_1, p_2; p_3, p_4) \propto ET_{12} - ET_{14} + ET_{34} - ET_{23} \]

\[ \times ET_{11} + t_1 - ET_{11} - t_1 + ET_{33} + t_3 - ET_{22} - t_2 \]

\[ \times ET_{33} + t_3 - ET_{22} - t_2. \]

A hierarchical stepping-stone model, where demes are combined into populations, is the only case I studied (besides the admixture graph) where \( F_3 \) can be negative. This effect indicates that admixture and population structure models may be the two sides of the same coin: Admixture is a (temporary) reduction in gene flow between individuals from the same population. Finally, for a simple serial founder model without migration, I find that \( F_3 \) measures the time between subsequent founder events.

### Simulations

Simulations were performed using ms (Hudson 2002). Specific commands used are

```
ms 466 100 -t 10 -r 10 10 00000 -i 1 1 2 1 2 1 6 61 49 57 33
43 34 40 84 13 24 -en 0 27 2 -en 0 3 2 -en 0 4 4 -en 0 5 2 -en 0 6 4 4 -en 0 7 3 2 -en 0 8 4 8 -en 0 9 0 2
-2 0 1 3 2 -en 0 1 1 0 2 -en 0 1 2 7 -ej 0 0 1 2 1
-ej 0 0 2 3 1 -ej 0 0 4 1 -ej 0 0 6 5 1 -ej 0 0 8 6 1
-ej 0 1 0 7 1 -ej 0 1 2 8 1 -ej 0 1 4 9 1 -ej 0 1 6 1 0
-ej 0 1 8 1 1 -ej 0 3 1 2 1
```

for the outgroup \( F_3 \) statistic (Figure 5A). Sample sizes and population sizes were picked randomly, but kept the same over all 100 replicates. Additionally, I randomly assigned each population an error rate uniformly between 0 and 0.05. Errors were introduced by adding additional singletons and flipping alleles at that rate.

For Figure 5B, the command was

```
ms 301 100 -t 10 -r 1 4 10 100 100 1 -es 0 0 0 1 1 2
$ALPHA -ej 0 0 3 2 1 -ej 0 0 3 5 3 -ej 0 3 3 1 -ej 0 3 1 4 1
```

with the admixture proportion \$ALPHA \$ set to increments of 0.025 from 0 to 0.5, with 200 data sets generated per \$ALPHA.

Finally, data for Figure 5C were simulated using

```
ms 501 100 -t 5 0 -r 5 0 1 0 0 0 0 0 1 6 1 0 0 1 1 0 1 00 1 00 100 1 100 100 100 100 100 100 100 100
1 -es 0 0 1 1 3 $ALPHA -ej 0 0 3 2 1 -ej 0 0 3 7 4 -ej 0 1 2 1 -ej 0 2 4 1 -ej 0 3 5 1 -ej 0 3 1 6 1
```

Here, the admixture proportion \$ALPHA \$ was varied in increments of 0.1 from 0 to 1, again with 200 data sets generated per \$ALPHA.

\( F_3 \) and \( F_4 \) statistics were calculated using the implementation from Pickrell and Pritchard (2012).

### Estimation and testing

In this article, I focused almost exclusively on the theoretical properties of the \( F \)-statistics, glancing over the statistical problems of how they are estimated. Many procedures are implemented in the software package ADMIXTOOLS and described in Patterson et al. (2012). Alternatively, the software package treemix (Pickrell and Pritchard 2012) contains lightweight alternatives for calculating \( F_3 \) and \( F_4 \) statistics. Both use a block-jackknife approach to estimate standard errors, taking linkage between markers into account.

### Discussion

There are three main ways to interpret \( F \)-statistics: In the simplest case, they represent branches in a population phylogeny. In the case of an admixture graph, the idea of shared drift in terms of paths is most convenient. Finally, the expressions in terms of coalescence times and the lengths of the internal branches of gene genealogies are useful for more complex scenarios. This last interpretation makes the connection to the ABBA-BABA statistic explicit and allows the investigation of the behavior of the \( F \)-statistics under arbitrary demographic models.

If drift indexes exist for two, three, and four populations, should there be corresponding quantities for five or more populations (e.g., Pease and Hahn 2015)? Two of the interpretations speak against this possibility: First, a population phylogeny can be fully characterized by internal and external branches, and it is not clear how a five-population statistic could be written as a meaningful branch length. Second, all \( F \)-statistics can be written in terms of four-individual trees, but this is not possible for five samples. This seems to suggest that there may not exist a five-population statistic as general as the three \( F \)-statistics I discussed here, but they will still be useful for questions pertaining to a specific demographic model.

A well-known drawback of \( F_3 \) is that it may have a positive expectation under some admixture scenarios (Patterson et al. 2012). Here, I showed that \( F_3 \) is positive if and only if the branch supporting the population tree is longer than the two branches discordant with the population tree. Note that this is (possibly) distinct from the probabilities of tree topologies, although the average branch length of the internal branch in a topology and the probability of that topology are frequently strongly correlated. Thus, negative \( F_3 \) values indicate that individuals from the admixed population are likely to coalesce with individuals from the two other populations, before they coalesce with other individuals from their own population!

For practical purposes, it is useful to know how the admixture tests perform under demographic models different from population phylogenies and admixture graphs and in which cases the assumptions made for the tests are problematic. In other words, under which demographic models is population structure distinguishable from a tree? Equation 16 enables the derivation of expectations for \( F_3 \) and \( F_4 \) under a wide variety of models of population structure (Figure 7). The simplest case is that of a single panmictic population. In that case, all \( F \)-statistics have an expectation of zero, consistent with the assumption that no structure and therefore no population phylogeny exists. Under island models, \( F_3 \) is also zero, and \( F_3 \) is inversely proportional to the migration rate. Results are similar under a hierarchical island model, except that the
many applications to humans, negative such as in stepping-stone models can lead to nonzero 2010). Here, I show that nonsymmetric population structure confounder for the human populations, the majority of markers support a dis-

of a tree metric and that using for within-species phylogenetic inference. While of how and when allele-frequency data should be transformed may result in false positives when assumptions are violated. Furthermore, I showed that F2 can be seen as a special case of a tree metric and that using F-statistics is equivalent to using phylogenetic theory to test hypotheses about simple phyloge-

netic networks (Huson et al. 2010). From this perspective, it is worth raising again the issue pointed out by Felsenstein (1973) for ancient admixture test only if there is good evidence that gene flow between some pairs of the populations was severely restricted.

Overall, when F3 is applicable, it is remarkably robust to population structure, requiring rather strong substructure to yield false positives. Thus, it is a very striking finding that in many applications to humans, negative F3 values are commonly found (Patterson et al. 2012), indicating that for most human populations, the majority of markers support a discordant gene tree, which suggests that population structure and admixture are widespread and that population phylogenies are poorly suited to describe human evolution.

Ancient population structure was proposed as possible confounder for the D-statistic and F3 statistic (Green et al. 2010). Here, I show that nonsymmetric population structure such as in stepping-stone models can lead to nonzero F3 values, showing that both ancestral and persisting population structure may result in false positives when assumptions are violated.

Furthermore, I showed that F3 can be seen as a special case of a tree metric and that using F-statistics is equivalent to using phylogenetic networks. While F2 has become a de facto standard, different transformations of allele frequencies might be useful in some cases, as both F3 and F4 can be interpreted as tests for treeness for arbitrary tree metrics.

This relationship provides ample opportunities for interaction between these currently diverged fields: Theory (Huson and Bryant 2006; Huson et al. 2010) and algorithms for finding phylogenetic networks such as Neighbor-Net (Bryant and Moulton 2004) may provide a useful alternative to tools specifically developed for allele frequencies and F-statistics (Patterson et al. 2012; Pickrell and Pritchard 2012; Lipson et al. 2013), particularly in complex cases. On the other hand, the tests and different interpretations described here may be useful to test for treeness in other phylogenetic applications, and the complex history of humans may provide motivation to further develop the theory of phylogenetic networks and stress its usefulness for within-species demographic analyses.

Acknowledgments

I thank Heejung Shim, Rasmus Nielsen, John Novembre, and all members of the Novembre laboratory for helpful comments and discussions. I am further grateful for comments from Nick Patterson and an anonymous reviewer. B.M.P. is supported by a Swiss National Science Foundation early postdoctoral mobility fellowship. Additional funding for this work was provided by National Institutes of Health grant R01 HG007089 to John Novembre.

Literature Cited


Fitch, W. M., and E. Margoliash, 1967 Construction of phyloge-


Haak, W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick et al., 2015 Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522: 207–211.


Lazaridis, I., N. Patterson, A. Mittnik, G. Renaud, S. Mallick et al., 2014 Ancient human genomes suggest three ancestral popula-

Admixture, Population Structure, and $F$-Statistics

Benjamin M. Peter
Figure S1 Path interpretation of $F_2$: $F_2$ is interpreted as the covariance of two possible paths from $P_1$ to $P_2$, which I color green and blue, respectively. Only branches that are taken by both paths contribute to the covariance. With probability $\alpha$, a path takes the left admixture edge, and with probability $\beta = 1 - \alpha$, the right one. I then condition on which admixture edge the paths follow: In the first tree on the right-hand side, both paths take the right admixture edge (with probability $\alpha^2$, in the second and third tree they take different or the right path, respectively. The result is summarized as the weighted sum of branches in the left-hand side tree. For a more detailed explanation, see Patterson et al. (2012).
A. Equation
\[ 2F_3(P_X, P_1, P_2) = ET_{1X} + ET_{2X} - ET_{12} - ET_{XX} \]

B. Concordant genealogy

C. Discordant genealogy

Figure S2 Schematic explanation how \( F_3 \) behaves conditioned on gene tree. Blue terms and branches correspond to positive contributions, whereas red branches and terms are subtracted. Labels represent individuals randomly sampled from that population. The external branches cancel out, so only the internal branches have non-zero contribution to \( F_3 \). In the concordant genealogy (Panel B), the contribution is positive (with weight 2), and in the discordant genealogy (Panel C), it is negative (with weight 1). The mutation rate as constant of proportionality is omitted.
A. Equation

\[ 2 F_4(P_1, P_2; P_3, P_4) = ET_{13} + ET_{24} - ET_{14} - ET_{23} \]

B. Concordant genealogy

C. Discordant genealogy (BABA)

D. Discordant genealogy (ABBA)

Figure S3 Schematic explanation how \( F_4 \) behaves conditioned on gene tree. Blue terms and branches correspond to positive contributions, whereas red branches and terms are subtracted. Labels represent individuals randomly sampled from that population. All branches cancel out in the concordant genealogy (Panel B), and that the two discordant genealogies contribute with weight +2 and -2, respectively. The mutation rate as constant of proportionality is omitted.