bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Admixture, Population Structure and F-statistics

Benjamin M Peter[1,*]

1 Department of Human Genetics, University of Chicago, Chicago IL USA

*** bpeter@uchicago.edu**

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 are 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. Here, we put these statistics in context of phylogenetic and population genetic theory. We show how measures of genetic drift can be interpreted as branch lengths, paths through an admixture graph or in terms of the internal branches in coalescent trees. We show that the admixture tests can be interpreted as testing general properties of phylogenies, allowing us to generalize applications for arbitrary phylogenetic trees. Furthermore, we derive novel expressions for the F-statistics, which enables us to explore the behavior of F-statistic under population structure models. In particular, we show that population substructure may complicate inference.

Author Summary

For the analysis of genetic data from hundreds of populations, a commonly used technique are a set of simple statistics on data from two, three and four populations. These statistics are used to test hypotheses involving the history of populations, in particular whether data is consistent with the history of a set of populations forming a tree.

Here, we provide context to these statistics by deriving novel expressions and by relating them to approaches in comparative phylogenetics. These results are useful because they provide a straightforward interpretation of these statistics under many demographic processes and lead to simplified expressions. However, the result also reveals the limitations of F-statistics, in that population substructure may complicate inference.

Introduction

For humans, whole-genome genotype data is now available for individuals from hundreds of populations [1, 2], opening up the possibility to ask more detailed and complex questions about our history [3], and stimulating the development of new tools for the analysis of the joint history of these populations [4–9]. A simple and intuitive framework for this purpose that has quickly gained in popularity are the F-statistics, introduced by Reich *et al.* [4], and summarized in [5]. In that framework, inference is

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 modern DNA [1, 10–13].

Some care is required with terminology, as the F-statistics *sensu* Reich *et al.* [4] are distinct, but closely related to Wright's fixation indices [4, 14], 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, and typically what we want to estimate using statistics) [15].

In this paper, we will mostly discuss model parameters, and we will therefore refer to them as *drift indices*. The term *F-statistics* will be used when referring to the general framework introduced by Reich *et al.* [4], and Wright's statistics will be 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? [4]
2. (Admixture test): Is a particular present day population descended from multiple ancestral populations? [4]
3. (Admixture ratio): What are the contributions from different population to a focal population [10].
4. (Number of founders): How many founder populations are there for a certain region? [1, 11]
5. (Complex demography): How can mixtures and splits of population explain demography? [5, 7]
6. (Closest relative): What is the closest relative to a contemporary or ancient population [16]

The demographic models under which these questions are addressed, and that motivated the drift indices, 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 genetic drift, 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 further edges that represent population mergers or a significant exchange of migrants.

The three F-statistics proposed by Reich *et al.* [4], labelled F_2, F_3 and F_4, have simple interpretations under a population phylogeny: F_2 corresponds to the path between two samples or vertices in the tree, whereas F_3 and F_4 can be interpreted as external and internal branches of the phylogeny, respectively (Figure 1A, [4]). In an admixture graph, there is no longer a single branch, and interpretations are more complex. However, F-statistics can be thought of as the proportion of genetic drift shared between populations [4].

The fundamental idea exploited in addressing all six questions outlined above is that under a tree model, branch lengths, and thus the drift indices, must satisfy some constraints [4, 17, 18]. 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

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

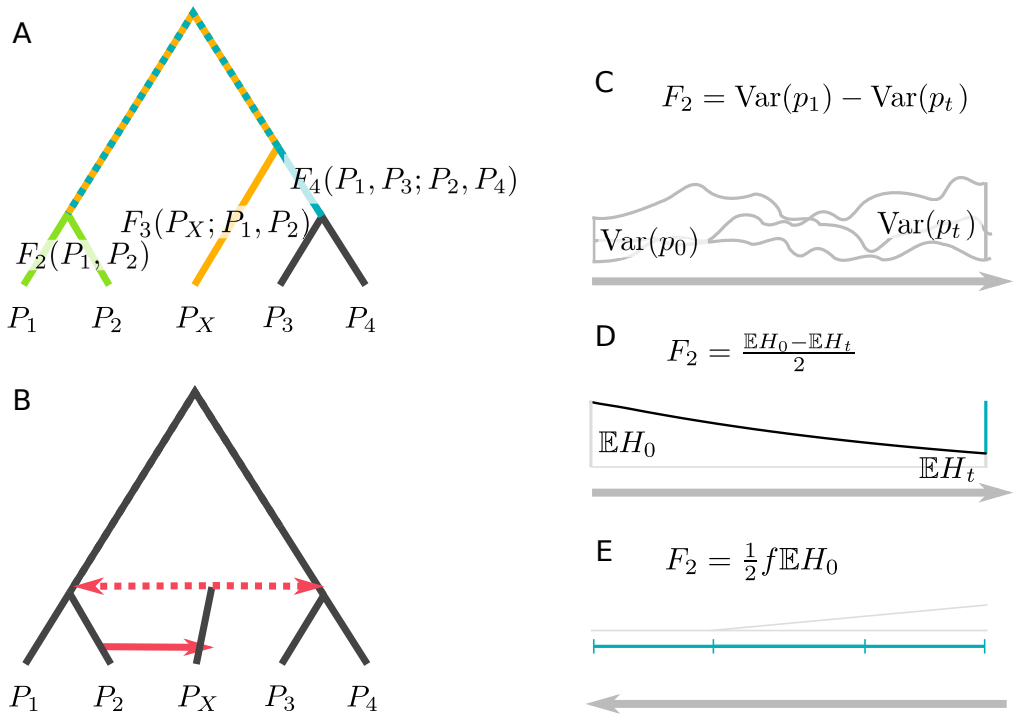


Figure 1. Schematics of gene genalogies, admixture graphs and measures of genetic drift. A: A population phylogeny with branches corresponding to F_2 (green), F_3 (yellow) and F_4 (blue). The dotted branch is part of both F_3 and F_4. B: An Admixture graph, extends population phylogenies by allowing gene flow (red, full line) and admixture events (red, dotted). C-E: Interpretations of F_2 in terms of allele frequency variances (C), heterozygosityies (D) and f, which can be interpreted as probability of coalescence of two lineages, or the probability that they are identical by descent.

with four leaves, there is at most one internal branch (tested using the F_4-admixture test).

The goal of this paper 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. We then propose an alternative definition of F_2 that allows us to simplify some applications of F-statistics, and study them under a wide range of population structure models. We then review some basic properties of distance-based phylogentic trees, show how the admixture tests are interpreted in this context and evaluate their behavior. Many of the results we highlight here are implicit in classical [14, 19–25] and more recent work [5–7], but often not explicitly stated, or given in a different context.

Results & Discussion

In the next sections we will discuss the F-statistics, develop different interpretations and derive some useful expressions. Longer derivations are deferred to the Methods section. A graphical summary of the three interpretations of the statistics is given in Figure 2, and the main formulas are given in Table 1.

Throughout this paper, we label populations as $P_1, P_2, \ldots P_i \ldots$. Often, we will denote a potentially admixed population with P_X, and an ancestral population with

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	$F_2(P_1, P_2)$	$F_3(P_X; P_1, P_2)$	$F_4(P_1, P_2, P_3, P_4)$
Definition	$\mathbb{E}[(p_1 - p_2)^2]$	$\mathbb{E}(p_X - p_1)(p_X - p_2)$	$\mathbb{E}(p_1 - p_2)(p_3 - p_4)$
F_2	-	$\frac{1}{2}\big(F_2(P_1, P_X) + F_2(P_2, P_X) - F_2(P_1, P_2)\big)$	$\frac{1}{2}\big(F_2(P_1, P_4) + F_2(P_2, P_3) - F_2(P_1, P_3) - F_2(P_2, P_4)\big)$
Coalescent times	$2\mathbb{E}T_{12} - \mathbb{E}T_{11} - \mathbb{E}T_{22}$	$\mathbb{E}T_{1X} + \mathbb{E}T_{2X} - \mathbb{E}T_{12} - \mathbb{E}T_{XX}$	$\mathbb{E}T_{14} + \mathbb{E}T_{23} - \mathbb{E}T_{13} - \mathbb{E}T_{24}$
Variance	$\mathrm{Var}(p_1 - p_2)$	$\mathrm{Var}(p_X) + \mathrm{Cov}(p_1, p_2) - \mathrm{Cov}(p_1, p_X) - \mathrm{Cov}(p_2, p_X)$	$\mathrm{Cov}((p_1 - p_2), (p_3 - p_4))$
Branch length	$2\mathcal{B}_c - \mathcal{B}_d$	$2\mathcal{B}_c - \mathcal{B}_d$	$\mathcal{B}_c - \mathcal{B}_d$ or as admixture test: $\mathcal{B}'_d - \mathcal{B}_d$

Table 1. F-statistics in terms of F_2 or tree metrics, coalescent times and allele frequency variances. A constant of proportionality is omitted for coalescence times and branch lengths. Derivations for F_2 are given in the main text, F_3 and F_4 are a simple result of combining Equations 20, 5 with 10b and 14b

.

P_0. 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 paper we will assume that all individuals are haploid. However, all results hold if instead of haploid individuals, we use a random allele of a diploid individual. If necessary, t_i denotes the time when population P_i is sampled. We focus on genetic drift only, and ignore the effects of mutation, selection and other evolutionary forces.

Measuring genetic drift – F_2

The first F-statistic we introduce is F_2, whose purpose is simply to measure genetic dissimilarity or how much genetic drift occurred between two populations. For populations P_1 and P_2, F_2 is defined as [4]

$$F_2(P_1, P_2) = F_2(p_1, p_2) = \mathbb{E}(p_1 - p_2)^2. \tag{1}$$

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 [5], which are assumed to be non-independent replicates of the evolutionary history of the populations.

Why is F_2 a useful measure of genetic drift? As it is generally infeasible to observe the changes in allele frequency directly, we assess the effect of drift indirectly, through its impact on genetic diversity. In general, 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).

Single population To make these measures of drift explicit, we assume a single population, measured at two time points ($t_0 \leq t_t$), and label the two samples P_0 and P_t. Then $F_2(P_0, P_t)$ can be interpreted in terms of the variance of allele frequencies (Figure 1C)

$$F_2(P_0, P_t) = \mathrm{Var}(p_t) - \mathrm{Var}(p_0) = \mathrm{Var}(p_t - p_0), \tag{2a}$$

the expected decrease in heterozygosity H_t, between the two sample times (Figure 1D):

$$F_2(P_0, P_t) = \frac{\mathbb{E}H_0 - \mathbb{E}H_t}{2}, \tag{2b}$$

and in terms of the inbreeding coefficient f, which can be interpreted as the probability of two individuals in P_t descend from the same ancestor in P_0, or,

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

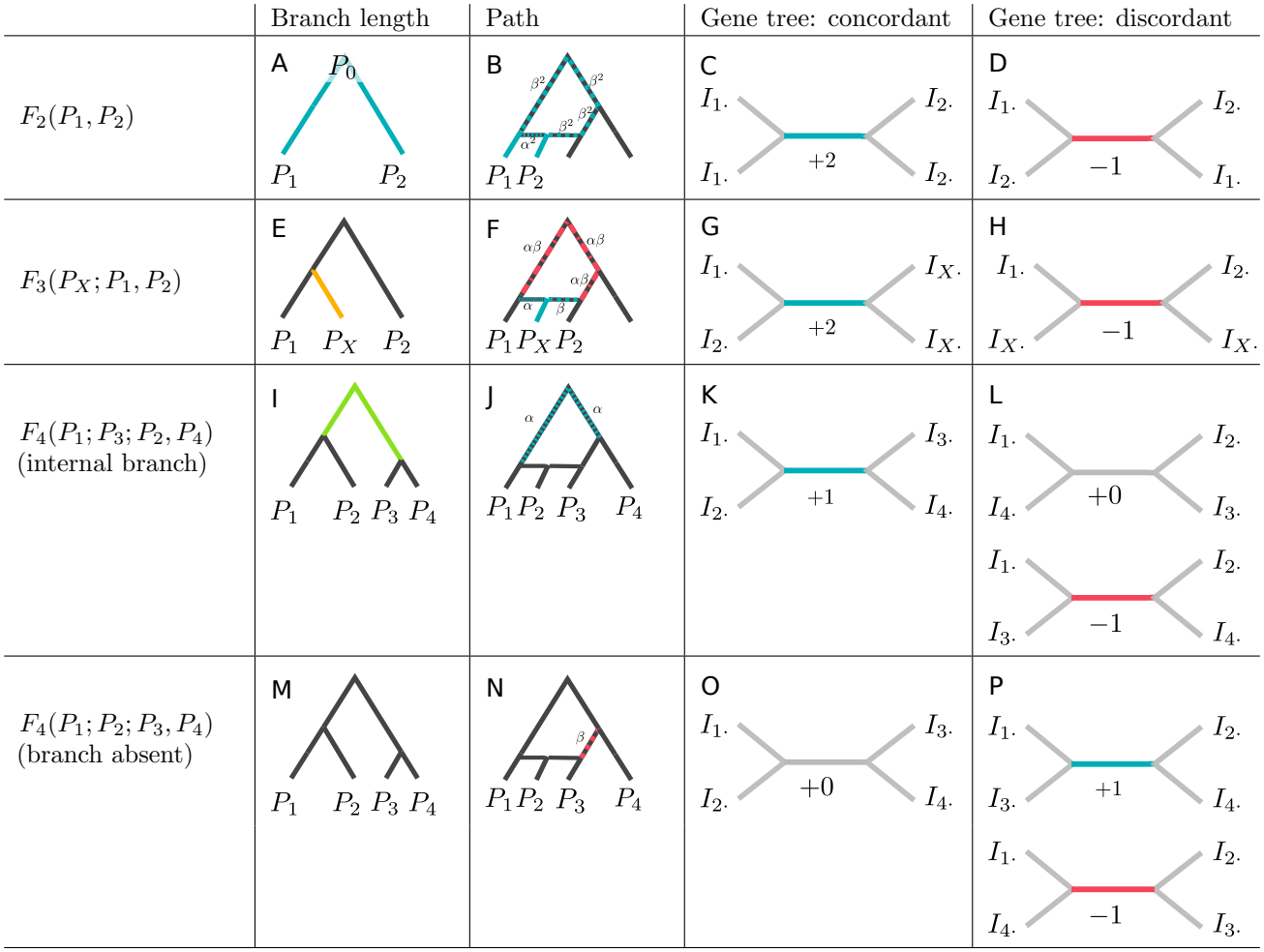


Figure 2. Interpretation of F-statistics. We can interpret the F-statistics i) as branch lengths in a population phylogeny (Panels A,E,I,M), the overlap of paths in an admixture graph (Panels B,F,J,N, see also Figure S1), and in terms of the internal branches of gene-genealogies (see Figures 3, S2 and S3) . For gene trees consistent with the population tree, the internal branch contributes positively (Panels C,G,K), and for discordant branches, internal branches contribute negatively (Panels D,H) or zero (Panel L). For the admixture test, the two possible gene trees contribute to the statistic with different sign, highlighting the similarity to the D-statistic [10] and its expectation of zero in a symmetric model.

equivalently, the probability that two samples from P_t coalesce before t_0. (Figure 1E, [26]):

$$F_2(P_0, P_t) = \frac{1}{2} f \mathbb{E}H_t, \tag{2c}$$

Rearranging Equation 2b, we find that $2F_2$ simply measures the absolute decrease of heterozyosity through time

$$\mathbb{E}H_t = \mathbb{E}H_0 - 2F_2(P_0, P_t). \tag{3a}$$

If we assume that we know p_0 and therefore $\mathrm{Var}(p_0)$ is zero, we can combine 2a and 2b and obtain

$$\mathbb{E}H_t = \mathbb{E}H_0 - 2\mathrm{Var}(p_t). \tag{3b}$$

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Similarly, using equations 2b and 2c we obtain an expression in terms of f.

$$\mathbb{E}H_t = \mathbb{E}H_0(1-f) \tag{3c}$$

Pairs of populations Equations 3b and 3c describing the decay of heterozygosity are – of course – well known by population geneticists, having been established by Wright [14]. In structured populations, very similar relationships exist when we compare the number of heterozygotes expected from the overall allele frequency, H_{obs} with the number of heterozygotes present due to differences in allele frequencies between populations H_{exp} [14, 19].

In fact, already Wahlund showed that for a population made up 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 is easy to see that

$$\begin{aligned} F_2(P_1, P_2) &= \frac{\mathbb{E}H_{exp} - \mathbb{E}H_{obs}}{2} && \text{(4a)} \\ &= \mathrm{Var}(p_1 - p_2) && \text{(4b)} \\ &= \frac{1}{2} F_{ST} \mathbb{E}H_{obs}. && \text{(4c)} \end{aligned}$$

This last equation served as the original motivation of F_2 [4], which was first defined as a numerator of F_{ST}.

Justification for F_2 Our preceding arguments show how the usage of F_2 for both single and structured populations can be justified by the similar effects on heterozygosity and allele frequency variance F_2 measures. However, what is the benefit of using F_2 instead of the established inbreeding coefficient f and fixation index F_{ST}? A conceptual way to approach this is by recalling that Wright motivated f and F_{ST} as *correlation coefficients* between alleles [14, 27]. This has 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. [25]).

F_2 as branch length Reich et al. [4, 5] proposed to partition "drift" (as we established, characterized by 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 & Edwards [20] and Felsenstein [21] 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 [5].

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 independent conditional on p_0, and therefore $\mathrm{Cov}(p_1, p_2) = \mathrm{Var}(p_0)$. Then, using 2a and 4b,

$$\begin{aligned} F_2(P_1, P_2) = \mathrm{Var}(p_1 - p_2) &= \mathrm{Var}(p_1) + \mathrm{Var}(p_2) - 2\mathrm{Cov}(p_1, p_2) \\ &= \mathrm{Var}(p_1) + \mathrm{Var}(p_2) - 2\mathrm{Var}(p_0) \\ &= F_2(P_1, P_0) + F_2(P_2, P_0). \end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Alternative proofs of this statement and more detailed reasoning behind the additivity assumption can be found in [4, 5, 20, 21].

For an admixture graph, we cannot use this approach as lineages are not independent. Reich *et al.* [4] 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 [4, 5]. An example of this representation for F_2 in a simple admixture graph is given in Figure S1, with the result summarized in Figure 2B. Detailed motivation behind this visualization approach is given in Appendix 2 of [5]. In brief, the reasoning is as follows: We write $F_2(P_1, P_2) = \mathbb{E}(p_1 - p_2)(p_1 - p_2)$, and interpret the two terms in parentheses as two paths between P_1 and P_2, and F_2 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_2 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 α they will go left, and with probability $\beta = 1 - \alpha$ they go right. Thus, F_2 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_2 is made up 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, we propose to redefine F_2 using coalescence theory [28]. Instead of allele frequencies on a fixed admixture graph, we track 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 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 paper, Slatkin [24] showed how F_{ST} can be interpreted in terms of the expected coalescence times of gene trees:

$$F_{ST} = \frac{\mathbb{E}T_B - \mathbb{E}T_W}{\mathbb{E}T_B},$$

where $\mathbb{E}T_B$ and $\mathbb{E}T_W$ are the expected coalescence times of two lineages sampled in two different and the same population, respectively.

Unsurprisingly, given the close relationship between F_2 and F_{ST}, we may obtain a similar expression for $F_2(P_1, P_2)$:

$$F_2(P_1, P_2) = \theta \left(\mathbb{E}T_{12} - \frac{\mathbb{E}T_{11} - \mathbb{E}T_{22}}{2} \right), \tag{5}$$

where θ is a scaled mutation parameter, T_{12} is the expected coalescence time for one lineage each sampled from populations P_1 and P_2, and T_{11}, T_{22} are the expected coalescence times for two samples from the P_1 and P_2, respectively. Unlike F_{ST}, the mutation parameter θ does not cancel. However, for most applications, the absolute magnitude of F_2 is of little interest, as we are only interested if a sum of F_2-values is significantly different from zero, significantly negative, or we are comparing

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

F-statistics with the same θ [4]. For this purpose, we may regard θ as a constant of proportionality and largely ignore its effect.

For estimation, the average number of pairwise differences π_{ij} is a commonly used estimator for θT_{ij} [29]. Thus, we can write the estimator for F_2 as

$$\hat{F}_2(P_1, P_2) = \pi_{12} - \frac{\pi_{11} - \pi_{22}}{2}. \tag{6}$$

This estimator of F_2 is numerically equivalent to the unbiased estimator proposed by [4] in terms of the sample allele frequency $\hat{p}_i$ and the sample size n_i (Equation 10 in the Appendix of [4]):

$$\hat{F}_2(P_1, P_2) = (\hat{p}_1 - \hat{p}_2)^2 - \frac{\hat{p}_1(1-\hat{p}_1)}{n_1 - 1} - \frac{\hat{p}_2(1-\hat{p}_2)}{n_2 - 1}. \tag{7}$$

However, the modelling assumptions are different: The original definition only considered loci that were segregating in the the ancestral population; loci not segregating there were discarded. Since ancestral populations are usually unsampled, this is often replaced by ascertainment in an outgroup [5,7]. In contrast, Equation 6 assumes that all markers are used, which is more convenient for sequence data.

Gene tree branch lengths An important feature of Equation 5 is that it only depends on the coalescence times between pairs of lineages. Thus, we may fully characterize F_2 by considering a sample of size four, with two random individuals taken from each population, as this allows us to study the joint distribution of T_{12}, T_{11} and T_{22}. For a sample of size four 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, $\mathcal{T}_c^{(2)}$) and one where lineages from different populations coalesce first (which we will refer to as *discordanat* topology $\mathcal{T}_d^{(2)}$). The superscripts refers to the topologies being for F_2, and we will discard them in cases where no ambiguity arises.

Thus, we can condition on the topology, and ask how F_2 depends on the topology:

$$F_2(P_1, P_2) = \mathbb{E}\big[F_2(P_1, P_2)|\mathcal{T}\big].$$

One way to do that is for each topology, consider each of the pairwise differences in Equation 5 separately, and then add the branches (see Figure 3 for a graphical representation).

We see that in both topologies, only the internal branch has a non-zero impact on F_2, and the contribution of the external branches cancels out. The external branch leading to a sample from P_1, for example, is included with 50% probability in T_{12}, but will always be included in T_{11}, so these two terms negate the effect of that branch. The internal branch of $\mathcal{T}_c$ will contribute with a factor of $a_c = 2$ to F_2, since the internal branch is added twice in Figure 3B. In contrast, the length of the internal branch of $\mathcal{T}_d$ is subtracted from F_2, with coefficient $a_d = -1$. Thinking of F_2 as a distance between population that is supposed to be large when the populations are very different from each other, this makes intuitive sense: if the populations are closely related we expect to see $\mathcal{T}_d$ relatively frequently, and F_2 will be low. However, if the populations are more distantly related, then $\mathcal{T}_c$ will be most common, and F_2 will be large.

An interesting way to represent F_2 is therefore in terms of the internal branches over all possible gene genealogies. Let us denote the unconditional average length of the internal branch of T_c as $\mathcal{B}_c$. Similarly, we denote the average length of the internal branch in $\mathcal{T}_d$ as $\mathcal{B}_d$.

$$F_2(P_1, P_2) = \theta\big(2\mathcal{B}_c - 1\mathcal{B}_d\big), \tag{8}$$

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

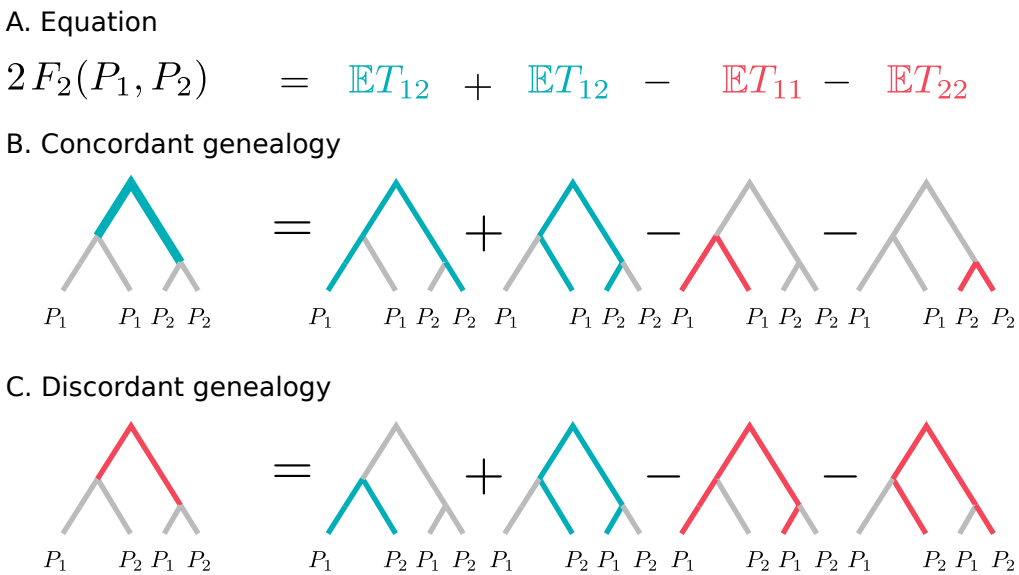


Figure 3. Schematic explanation how F_2 behaves conditioned on gene tree. Blue terms and branches correspond to positive contributions, whereas red branches and terms are subtracted. Labels represent individuals randomly sampled from that population. We see that external branches cancel out, so only the internal branches have non-zero contribution to F_2. 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.

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with the coefficients $a_c = 2, a_d = 1$. A graphical summary of this is given in Figure 2C-D. As a brief sanity check, we can consider the case of a population without structure. In this case, the branch length is independent of the topology and $\mathcal{T}_d$ is twice as likely as $\mathcal{T}_c$. In this case, we see immediately that F_2 will be zero, as expected when there is no difference between topologies.

Testing treeness

In practical cases, we often have dozens or even hundreds of populations [2, 5, 12], and we want 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, we calculate all pairwise F_2 indices between populations, combine them into a *dissimilarity matrix*, and ask if 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 [18, 30–32], and what properties a general dissimilarity matrix needs to have in order to be consistent with a tree [17, 22], in which case the matrix is also called a *tree metric* [18].

There are two central properties for a dissimilarity matrix to be consistent with a tree: The first property is that all edges in a tree have positive length. This is strictly not necessary for phylogenetic trees, and some algorithms may return negative branch lengths [31]; however, since in our case branches have an interpretation of genetic drift, it is clear that negative genetic drift is biologically meaningless, and therefore negative branches should be interpreted as a violation of the modelling assumptions and hence treeness.

The second property of a tree metric that we require 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_i, P_j, P_k and P_l,

$$F_2(P_i, P_j) + F_2(P_k, P_l) \leq \max(F_2(P_i, P_k) + F_2(P_j, P_l), F_2(P_i, P_l) + F_2(P_j, P_k)) \quad (9)$$

that is, if we compare the sums of all possible pairs of distances between populations, then two of these sums will be the same, and no smaller than the third. This theorem, due to Buneman [17, 33] is called the four-point condition or sometimes, more modestly, the "fundamental theorem of phylogenetics". A proof can be found in Chapter 7 of [18].

In terms of a tree, this statement can be understood by noticing 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, or the same length for a topology with no internal branch. Thus, the four-point condition can be informally rephrased as "for any four taxa, a 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_3-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, we can interpret the working of the two test statistics in terms of fundamental properties of phylogenetic trees, with the immediate consequence that they can be applied as treeness-tests for arbitrary dissimilarity matrices.

An early test of treeness, based on a likelihood ratio, was proposed by Cavalli-Sforza & Piazza [22]: They compare the likelihood of the observed F_2-matrix 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,

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

especially for large trees [32] 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.* [4] was that rejection of treeness for subtrees of size three (for F_3) and four (for F_4) is sufficient to reject treeness for the entire tree [4]. Furthermore, tests on these subsets also pinpoint the populations involved in the non-tree-like history.

F_3

In the previous section, we showed how F_2 can be interpreted as a branch length, an overlap of paths or in terms of gene trees (Figure 2). Furthermore, we derived expressions in terms of coalescent times, allele frequency variances and internal branch lengths of gene trees. We now derive analogous results for F_3.

Reich *et al.* [4] defined F_3 as:

$$F_3(P_X; P_1, P_2) = \mathbb{E}(p_X - p_1)(p_X - p_2) \tag{10a}$$

with the goal to test whether P_X is admixed. Recalling the path interpretation detailed in [5], 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 [4]:

$$F_3(p_1; p_2, p_3) = \frac{1}{2}\Big(F_2(p_1, p_2) + F_2(p_1, p_3) - F_2(p_2, p_3)\Big). \tag{10b}$$

If we replace F_2 in Equation 10b with an arbitrary tree metric, Equation 10b is known as the Gromov product [18] in phylogenetics. The Gromov product is a commonly used operation in classical phylogenetic algorithms to calculate the length of the portion of a branch shared bewtween P_1 and P_2 [21, 30, 31]: consistent with the notion that F_3 is the length of an external branch in a phylogeny.

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

Combining Equations 5 and 10b, we find that F_3 can be written in terms of expected coalescence times as

$$F_3(P_X; P_1, P_2)\theta^{-1} = \mathbb{E}T_{1X} + \mathbb{E}T_{2X} - \mathbb{E}T_{12} - \mathbb{E}T_{XX} \tag{10c}$$

Similarly, we may obtain an expression for the variance by combining Equation 20 with 10b, and find that

$$F_3(P_X; P_1, P_2) = \mathrm{Var}(p_X) + \mathrm{Cov}(p_1, p_2) - \mathrm{Cov}(p_1, p_X) - \mathrm{Cov}(p_2, p_X). \tag{10d}$$

This result can also be found in [6].

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 [16]. For an unknown population P_U, 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 $F_3(P_O, P_U, P_i)$, where P_O is an outgroup population that was assumed widely diverged from P_U and all populations in the panel. This measures the shared drift (or shared branch) of P_U with the populations from the panel, and high F_3-values imply close relatedness.

However, using Equation 10c, we see that the outgroup-F_3-statistic is

$$F_3(P_O; P_U, P_i) \propto \mathbb{E}T_{UO} + \mathbb{E}T_{iO} - \mathbb{E}T_{Ui} - \mathbb{E}T_{OO}. \tag{11}$$

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

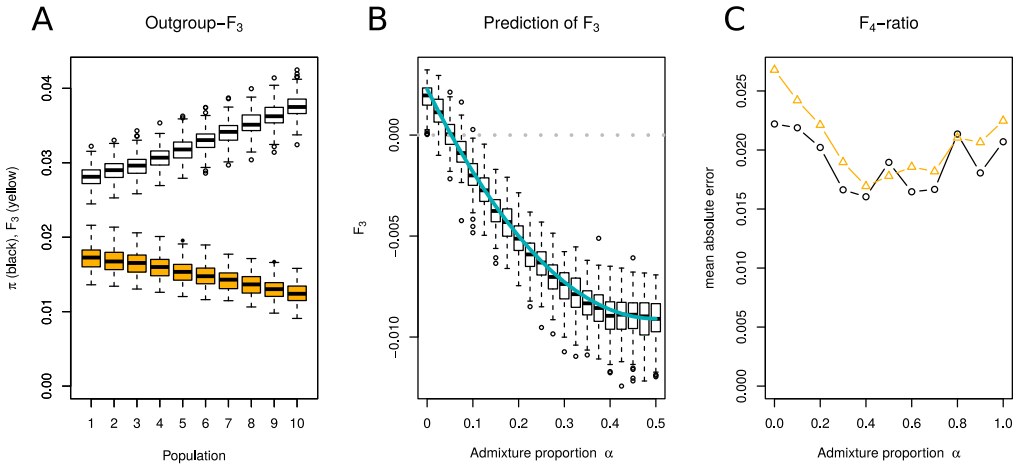


Figure 4. Simulation results. A: Outgroup-F_3 statistics (yellow) and π_{iU}(white) for a panel of populations with linearly increasing divergence time. B: Simulated (boxplots) and predicted(blue) F_3-statistics under a simple admixture model (main text). C: Comparison of F_4-ratio (yellow, Equation 17) and ratio of differences (Equation 19 black)

.

Out of these four terms, $\mathbb{E}T_{UO}$ and $\mathbb{E}T_{OO}$ do not depend on the panel. Furthermore, if P_O 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 only coalesce once they are in the joint ancestral population. Therefore, only the term $\mathbb{E}T_{Ui}$ is expected to vary between different panel populations, suggesting that using the number of pairwise differences, π_{Ui}, is largely equivalent to using $F_3(P_O; P_U, P_i)$. We confirm this in Figure 4A, where we calculate outgroup-F_3 and π_{iU} for a set of increasingly divergent populations. Linear regression confirms the visual picture that π_{iU} has a higher correlation with divergence time ($R^2 = 0.75$) than F_3 ($R^2 = 0.49$).

F_3 admixture test However, the main motivation of defining F_3 has been as an admixture test [4]. In this context, the null hypothesis is that F_3 is non-negative, i.e. we are testing if the data is consistent with a phylogenetic tree that has positive edge lengths. If this is not the case, we reject the tree model for the more complex admixture graph. From Figure 2F, we see that drift on the path on the internal branches (red) contribute 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 bewteen P_1 and P_2, Reich et al. [4] provided a condition when this is the case (Equation 20 in Supplement 2 of [4]). However, since this condition involves F-statistics with internal, unobserved populations, it is not easily applicable. We can obtain a more useful condition using gene trees:

In the simplest admixture model, an ancestral population splits into P_1 and P_2 and time t_r. At time t_1, the populations mix to form P_X, such that with probability α, individuals in P_X descend from individuals from P_1, and with probability $(1-\alpha)$, they descend from P_2. In this case, $F_3(P_X; P_1, P_2)$ is negative if

$$\frac{1}{(1-c_x)}\frac{t_1}{t_r} < 2\alpha(1-\alpha), \tag{12}$$

where c_x is the probability two individuals sampled in P_X have a common ancestor before t_1. For a constant sized population of size 1, $c_x = 1 - e^{-t_1}$. We see that power

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of F_3 to detect admixture increases the closer they get to fifty percent, and that it only depends on the ratio between the original split and the secondary contact, and coalescence events that happen in P_X.

We obtain a more general condition for negativity of F_3 by considering the internal branches of the possible gene tree topologies, as we did for F_2. Note that Equation 10c includes $\mathbb{E}T_{XX}$, implying that we need two individuals from P_X, but only one each from P_1 and P_2 to study the joint distribution of all terms in (10c). The minimal case is therefore contains again just four samples (Figure S2).

Furthermore, P_1 and P_2 are exchangeable, and thus we can again consider just two unrooted genealogies, a concordant one $\mathcal{T}_c^{(3)}$ where the two lineages from P_X are most closely related, and a discordant genealogy $\mathcal{T}_d^{(3)}$ where the lineages from P_X merge first with the other two lineages. A similar argument as that for F_2 shows (presented in Figure S2) that F_3 can be written as a function of just the internal branches in the topologies:

$$F_3(P_X; P_1, P_2) = \theta\big(2\mathcal{B}_c - \mathcal{B}_d\big), \tag{13}$$

where $\mathcal{B}_c$ and $\mathcal{B}_d$ are the lengths of the internal branches in $\mathcal{T}_c$ and $\mathcal{T}_d$, respectively, and similar to F_2, they have coefficients $a_c = 2$ and $a_d = -1$. Again, if we do the sanity check of all samples coming from a single, randomly mating population, then $\mathcal{T}_d$ is again twice as likely as $\mathcal{T}_c$, and all branches are expected to have the same length. Thus F_3 is zero, as expected. However, for F_3 to be negative, we see that $\mathcal{B}_d$ needs to be more than two times longer than $\mathcal{B}_c$. Thus, F_3 can be seen as a test whether mutations that agree with the population tree are more common than mutations that disagree with it.

We performed a small simulation study to test the accuracy of Equation 12. Parameters were chosen such that F_3 has a negative expectation for $\alpha > 0.05$ (grey dotted line in Figure 4B), so simulations on the left of that line have positive expectation, and samples on the right are true positives. We find that our predicted F_3 fits very well with the simulations (Figure 4B).

F_4

The second admixture statistic, F_4, is defined as [4]

$$F_4(p_1, p_2; p_3, p_4) = \mathbb{E}[(p_1 - p_2)(p_3 - p_4)]. \tag{14a}$$

Similarly to F_3, F_4 can be written as a linear combination of F_2:

$$F_4(p_1p_2; p_3, p_4) = \frac{1}{2}\Big(F_2(p_1, p_4) + F_2(p_2, p_3) - F_2(p_1, p_3) - F_2(p_2, p_4)\Big). \tag{14b}$$

Equations giving F_4 in terms of pairwise coalescence times and as a covariance are given in Table 1.

As four populations are involved, there are $4! = 24$ possible ways of arranging the arguments in Equation 14a. 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, leaving only three unique absolute values, which correspond to the tree possible tree topologies. Thus, we may always find a way of writing F_4 such that the statistic is non-negative (i. e. $F_4(P_1, P_2; P_3, P_4) = -F_4(P_1, P_2; P_4, P_3)$). Out of these three, one F_4 can be written as the sum of the other two, leaving us with just two independent possibilities:

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

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As we did for F_3, we can generalize Equation 14b by replacing F_2 with an arbitrary tree metric. In this case, Equation 14b is known as a tree split [17], as it measures the length of the overlap of the branch lengths between the two pairs (P_1, P_2) and (P_3, P_4). Tree splits have the property that if there exists a branch "splitting" the populations such that the first and third argument are on one side of the branch, and the second and fourth are on the other side (Figure 6I), then it corresponds to the length of that branch. If no such branch exists, then F_4 will be zero.

This can be summarized by the four-point condition [17, 33], or, informally, by noting that any four populations will have at most one internal branch, and thus one of the three F_4-values will be zero, and the other two will have the same value. Therefore, one F_4-index has an interpretation as the internal branch in a genealogy, and the other can be used to test if the data corresponds to a tree. In Figure 2, the third row (Panels I-L) correspond to the internal branch, and the last row (Panels M-P) to the "zero"-branch.

Thus, in the context of testing for admixture, by testing that F_4 is zero we check whether there is in fact only a single internal branch, and if that is not the case, we reject a population phylogeny for an admixture graph.

Evaluating F_4 in terms of gene trees and their internal branches, we have to consider the three different possible gene tree topologies, and depending on if we want to estimate a branch length or do an admixture test, they are interpreted differently.

For the branch length, we see that 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 we can write

$$F_4^{(B)} = \theta(\mathcal{B}_c - \mathcal{B}_d). \tag{15}$$

This equation is slightly different than those for F_2 and F_3, where the coefficient for the discordant genealogy was half that for the concordant genealogy. Note, however, that we have two discordant genealogies, and F_4 only measures one of them. Under a tree, both discordant genealogies are equally likely [34], and thus the expectation of F_4 will be the same.

In contrast, for the admixture test statistic, the contribution of the concordant genealogy will be zero, and the discordant genealogies will contribute with coefficients -1 and $+1$, respectively. Under the population phylogeny, these two gene trees will be equally likely [28], and thus the expectation of F_4 as a test statistic

$$F_4^{(T)} = \theta(\mathcal{B}_d - \mathcal{B}'_d) \tag{16}$$

is zero under the null hypothesis. Furthermore, we see that the statistic is closely related to the ABBA-BABA or D-statistic also used to test for admixture [10, 34], which includes a normalization term, and in our notation is defined as,

$$D = \frac{\mathcal{B}'_d - \mathcal{B}_d}{\mathcal{B}'_d + \mathcal{B}_d}$$

but otherwise tests the exact same hypothesis.

F_4-as a branch

Rank test Two major applications of F_4 use its interpretation as a branch length. First, we can use the rank of a matrix of all F_4-statistics to obtain a lower bound on the number of admixture events required to explain data [11]. The principal idea of this approach is that the number of internal branches in a genealogy is bounded to be

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

at most $n-3$ in an unrooted tree. Since each F_4 corresponds to a sum of internal branches, all F_4-indices 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 [35]) is at most $n-3$, if the data is consistent with a tree. However, admixture events may increase the rank of the matrix, as they add additional internal branches [11]. 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 we are only interested in admixture events in a certain part of the phylogeny. To estimate the number of admixture events on a particular branch of the phylogeny, Reich *et al.* [11], 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 5A). Assuming a phylogeny, all $F_4(S_1, R_1; S_2, R_2)$ will measure the length of the branch absent from Figure 5A, und should be zero, and the rank of the matrix of all F_4 of that form reveals the number of branches of that form.

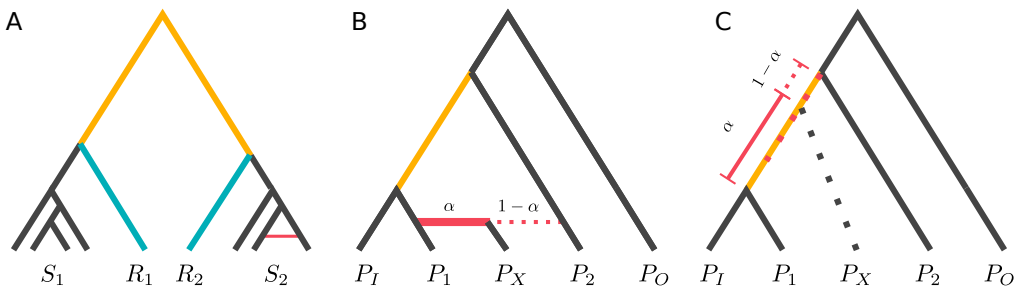


Figure 5. Applications of F_4: A: Visualization of rank test to estimate the number of admixture events. $F_4(S_1, R_1, S_2, R_2)$ measures a branch absent from the phylogeny and should be zero for all populations from S_1 and S_2. B: Model underlying admixture ratio estimate [10]. P_X splits up, and the mean coalescence time of P_X with P_I gives the admixture proportion. C: If the model is violated, α_X measures where on the internal branch in the underlying genealogy P_X (on average) merges

Admixture proportion The second application is by comparing branches between closely related populations to obtain an estimate of mixture proportion, or how much two focal populations correspond to an admixed population. [10]:

$$\alpha_X = \frac{F_4(P_O, P_I; P_X, P_1)}{F_4(P_O, P_I; P_2, P_1)} \tag{17}$$

Here, P_X, is the population whose admixture proportion we are estimating, P_1 and P_2 are the potential contributors, where we assume that they contribute with proportions α_X and $1-\alpha_X$, respectively. and P_O, P_I are reference populations with no direct contribution to P_X (see Figure 5B). P_I has to be more closely related to one of P_1 or P_2 than the other, and P_O is an outgroup.

The canonical way [5] to interpret this ratio is as follows : the denominator is the branch length from the common ancestor population from P_I and P_1 to the common ancestor of P_I with P_2. (Figure 5C, yellow line), The numerator has a similar interpretation as an internal branch (red dotted line). In an admixture scenarios, (Figure 5B, this is not unique, and is replaced by a linear combination of lineages merging at the common ancestor of P_I and P_1 (with probability α_X), and lineages merging at the common ancestor of P_I with P_2 (with probability $1-\alpha_X$).

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Thus, a more general interpretation is that α_X measures how much closer the common ancestor of P_X and P_I is to the common ancestor of P_I and P_1 and the common ancestor of P_I and P_2, indicated by the gray dotted line in Figure 5B. This quantity is defined also when the assumptions underlying the admixture test are violated, and if the assumptions are not carefully checked, might lead to misinterpretations of the data. In particular, α_X is well-defined in cases where no admixture occurred, or in cases where either of P_1 and P_2 did not experience any admixture.

Furthermore, it is evident from Figure 5 that if all populations are sampled at the same time, $\mathbb{E}T_{OX} = \mathbb{E}T_{O1} = \mathbb{E}T_{O2} = \mathbb{E}T_{OI}$, and therefore,

$$\alpha_X = \frac{\mathbb{E}T_{I1} - \mathbb{E}T_{IX}}{\mathbb{E}T_{I1} - \mathbb{E}T_{I2}}. \tag{18}$$

Thus,

$$\alpha_X = \frac{\pi_{I1} - \pi_{IX}}{\pi_{I1} - \pi_{I2}} \tag{19}$$

is another estimator for α_X that can be used even if no outgroup is available. We compared Equations 17 and 19 for varying admixture proportions in Figure 4C using the mean absolute error in the admixture proportion. Both estimators perform very well, but we find that (19) performs slightly better in cases where the admixture proportion is low. However, in most cases this minor improvement possibly does not negate the drawback that Equation 19 is only applicable when populations are sampled at the same time.

Structure models

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 well-approximated by a tree? Equation 5 allows us to derive expectations for F_3 and F_4 under a wide variety of models of population structure (Figure 6). 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_4 is also zero, and F_3 is inversely proportional to the migration rate. Results are similar under a hierarchical island model, except that the number of demes has a small effect. This corresponds to a population phylogeny that is star-like and has no internal branches, which is explained by the strong symmetry of the island model. Thus, looking at different F_3 and F_4-statistics may be a simple heuristic to see if data is broadly consistent with an island model; if F_3-values vary a lot between populations, or if F_4 is substantially different from zero, an island model might be a poor choice. When looking at a finite stepping stone model, we find that F_3 and F_4 are both non-zero, highlighting that F_4 (and the ABBA-BABA-D-statistic) is susceptible to migration between any pair of populations. Thus, for applications, F_4 should only be used if there is good evidence that gene flow between some pairs of the populations was severely restricted. A hierarchical stepping stone model, where demes are combined into populations, is the only case 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: we can think of admixture as a (temporary) reduction in gene flow between individuals from the same population. Finally, for a simple serial founder model without migration, we find that F_3 measures the time between subsequent founder events.

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

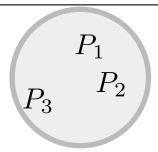
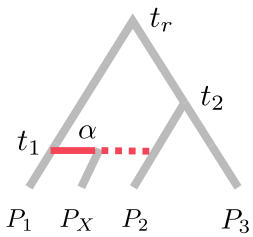
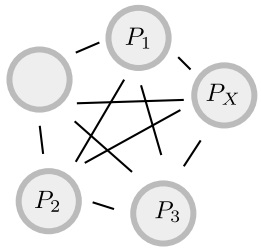
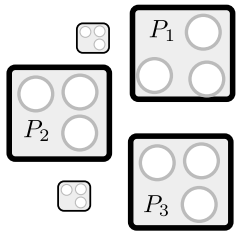
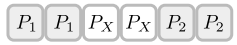
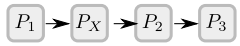
Model		$F_3(P_X; P_1, P_2)$	$F_4(P_1; P_X; P_2, P_3)$	Parameters
Panmictic		0	0	
Admixture Graph		$t_1 - 2\alpha(1-\alpha) \times (1-c_x)t_r$	$(1-\alpha)(t_2 - t_1)$	α: admixture ratio; t_1; admixture time; t_2 merging time of P_2 and P_3; t_r global ancestor
Island Model		$\frac{1}{M}$	0	M: Migration rate
Hierarchical Island Model		$\frac{n(d-1)}{M}$	0	M: Migration rate n: # of island d: # of demes per island
Stepping stone		$\frac{2}{7M}$	$-\frac{8}{7M}$	M: Migration rate between adjacient demes
Hierarchical stepping stone		$-\frac{\mathbf{0.06}}{\mathbf{M}}$	$\frac{14}{55M}$	M: Migration rate between adjacient demes
Serial founder model		t_x	0	t_x: time when P_X is first colonized

Figure 6. Expectations for F_3 and F_4 under select models.

Conclusions

We showed that there are three main ways to interpret F-statistics: First, we can think of them as the branches in a population phylogeny. Second, we can think of them as the shared drift, or paths in an admixture graph. And third, we can think of them in terms of coalescence times and the lengths of the internal branches of gene genealogies. This last interpretation allows us to make the connection to the ABBA-BABA-statistic explicit, and allows us to investigate the behavior of the F-statistics under arbitrary demographic models.

If we have indices for two, three and four populations, should there be corresponding quantities for five or more populations(e.g. [36])? 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, we can write all F-statistics in terms of four-individual trees, but this is not possible for

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

five samples. This seems to suggest that there may not exist a five-population statistic as general as the F-statistics we discussed here, but they will still be valid for questions pertaining to a very specific demographic model [36].

A well-known drawback of F_3 is that it may have a positive expectation under some admixture scenarios [5]. Here, we 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 may frequently very correlated. Thus, negative F_3-values indicate that individuals from the admixed population are more likely to coalesce with individuals from the two other populations, than with other individuals from the same population!

Overall, when F_3 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 F_3-values are commonly found [4, 5], 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 and F_4-statistics [10]. Here, we show that non-symmetric population structure such as in stepping stone models can lead to non-zero F_4-values, showing that both ancestral and persisting population structure may result in false-positives when the statistics are applied in an incorrect setting.

Furthermore, we showed that the F-statistics can be seen as a special case of a tree-metric, and that both F_3 and F_4 can be interpreted, for arbitrary tree metrics, as tests for properties of phylogenetic trees.

From this perspective, it is worth re-raising the issue pointed out by Felsenstein [21], how and when allele-frequency data should be transformed for within-species phylogenetic inference.While F_2 has become a *de facto* standard, which, as we have shown, leads to useful interpretations, the F_3 and F_4-tests can be used for arbitrary tree metrics, and different transformations of allele frequencies might be useful in some cases.

But it is clear that, when we are applying F-statistics, we are implicitly using phylogenetic theory to test hypotheses about simple phylogenetic networks [37].

This close relationship provides ample opportunities for interaction between these currently diverged fields: Theory [37, 38] and algorithms for finding phylogenetic networks such as Neighbor-Net [39] may provide a useful alternative to tools specifically developed for allele frequencies and F-statistics [5–7], 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.

Acknowledgements

I would like to thank Heejung Shim, Choongwon Jeong, Evan Koch, Lauren Blake, Joel Smith and John Novembre for helpful comments and discussions.

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Equivalence of drift interpretations

First, we show that F_2 can be interpreted as the difference in variance of allele frequencies (Figure 1C):

As in the Results section, let P_i denote a population with allele frequency, sample size and sampling time with p_i, n_i and t_i, respectively. Then, for $t_0 < t_t$:

$$\begin{aligned} F_2(p_t, p_0) = \mathbb{E}[(p_t - p_0)^2] &= \mathbb{E}\left[\mathbb{E}[(p_t - p_0)^2]|p_0\right] \\ &= \mathbb{E}\left[\mathbb{E}[(p_t - \mathbb{E}(p_t))^2|p_0]\right] \\ &= \mathbb{E}\left[\mathrm{Var}(p_t|p_0)\right] \\ &= \mathrm{Var}(p_t) - \mathrm{Var}\left(\mathbb{E}[p_t|p_0]\right) \\ &= \mathrm{Var}(p_t) - \mathrm{Var}(p_0) \end{aligned} \tag{20}$$

Here, we used $\mathbb{E}[p_t|p_0] = p_0$ on lines two and five (which holds if there is no mutation, no selection and P_t is a descendant of P_0). The fourth line is obtained using the law of total variance. It is worth noting that this result holds for any model of genetic drift where the expected allele frequency is the current allele frequency (the process describing the allele frequency is a martingale). For example, this this interpretation of F_2 holds also if we model genetic drift as a Brownian motion.

A heterozygosity model The interpretation of F_2 in terms of the decay in heterozygosity and identity by descend can be derived elegantly using duality between the diffusion process and the coalescent: Let again $t_0 < t_t$ Furthermore, let f be the probability that two individuals sampled at time t_t have coalesced at time t_0.

Then,

$$\mathbb{E}_{p_t}\left[p_t^{n_t}|p_0\right] = \mathbb{E}_{n_0}\left[p_0^{n_0}|n_t\right]. \tag{21}$$

This equation is due to Tavaré [40], who also provided the following intuition: Given we sample n_t individuals at time t_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_t, and then we need all n_t sampled individuals to carry x.

In a diffusion framework, we can write

$$\mathbb{P}(E) = \int_0^1 y^{n_t}\mathbb{P}(p_t = y|p_0)dy = \mathbb{E}[p_0^{n_0}|p_t]. \tag{22}$$

On the other hand, we may argue using the coalescent: For E to occur, all n_1 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,

$$\mathbb{P}(E) = \sum_{i=1}^{n_0} p_0^i \mathbb{P}\left(n_0 = i|n_t\right) = \mathbb{E}\left[p_0^{n_0}|n_t\right]. \tag{23}$$

Equating (22) and (23) yields Equation 21.

In the present case, we are most interested in the cases of $n_t = 1, 2$, since:

$$\mathbb{E}\left[p_t^2|p_0\right] = p_0 f + p_0^2(1-f) \tag{24a}$$

$$\mathbb{E}\left[p_t^1|p_0\right] = p_0 \tag{24b}$$

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To derive an expression for F_2, we start by conditioning on the allele frequency p_0,

$$\begin{aligned}\mathbb{E}\left[(p_0-p_t)^2|p_0\right] &= \mathbb{E}\left[p_0^2|p_A\right] - \mathbb{E}\left[2p_tp_0|p_t\right] + \mathbb{E}\left[p_t^2|p_0\right] \\ &= p_0^2 - 2p_0^2 + p_0 f + p_0^2(1-f) \\ &= f p_0(1-p_0) \\ &= \frac{1}{2} f H_0.\end{aligned}$$

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

$$F_2(p_0, p_t) = \frac{1}{2} f \mathbb{E} H_0 \tag{25}$$

and we see that F_2 increases as a function of f (Figure 1E). 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, and $\mathbb{E}H_0$ is the probability two individuals are not identical by descent in P_0. Thus, F_2 is half the probability of the event that two individuals in P_t are identical by descent, and they were not in P_0.

Furthermore, $\mathbb{E}H_t = (1-f)\mathbb{E}H_0$ (Equation 3.4 in [28]) and therefore

$$\mathbb{E}H_0 - \mathbb{E}H_t = \mathbb{E}H_0\big(1-(1-f\big) = 2F_2(p_t, p_0). \tag{26}$$

which shows that F_2 measures the decay of heterozygosity (Figure 1C). A similar argument was used by in [7] to estimate ancestral heterozygosities using F_2 and to linearize F_2.

Two populations F_2 in terms of the difference in expected and observed heterozygosity follows directly from the result from [19], which was obtained by considering the genotypes of all possible matings in the two subpopulations, and the variance case follows directly because $\mathrm{Var}(p_1-p_2) = \mathbb{E}(p_1-p_2)^2 - \left[\mathbb{E}(p_1-p_2)\right]^2$, but $\mathbb{E}(p_1-p_2) = 0$. Lastly, we relate F_2 to F_{ST} by using the definition of F_2 as a variance in the definition of F_{ST}: $F_{ST} = \frac{2(p_1-p_2)^2}{\mathbb{E}H_{exp.}}$

Covariance interpretation To see how F_2 can be interpreted as a covariance between two individuals from the same population, 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 and p_2 in the other population. If we are equally likely to pick from either population,

$$\begin{aligned}\mathbb{E}X_i = \mathbb{E}X_j &= \frac{1}{2}p_1 + \frac{1}{2}p_2 \\ \mathbb{E}X_iX_j &= \frac{1}{2}p_1^2 + \frac{1}{2}p_2^2 \\ \mathrm{Cov}(X_i, X_j) &= \mathbb{E}X_iX_j - \mathbb{E}X_i\mathbb{E}X_j \\ &= \frac{1}{4}(p_1-p_2)^2 = \frac{1}{4}F_2(P_1, P_2)\end{aligned} \tag{27}$$

The expectations can be interpreted that we pick a population, and then with probability equal to the allele frequency an individual will have the A allele. The joint expectation is similar, except we need two individuals.

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Derivation of F_2 for gene trees

To derive equation (5), we start by considering F_2 for two samples of size one, express F_2 for arbitrary sample sizes in terms of individual-level F_2, and obtain a sample-size independent expression by letting the sample size n go to infinity.

In this framework, we assume that mutation is rare such that there is at most one mutation at any locus. In a sample of size two, let the genotypes of the two haploid individuals be denoted as I_1, I_2. $I_i \in \{0, 1\}$ and $F_2(I_1, I_2) = 1$ implies $I_1 = I_2$, whereas $F_2(I_1, I_2) = 0$ implies $I_1 \neq I_2$. We can think of $F_2(I_1, I_2)$ as an indicator random variable with parameter equal to the branch length between I_1 and I_2, times the probability that a mutation occurs on that branch.

Now, replace I_1 with a sample $P_1 = \{I_{1,1} I_{1,2}, \ldots, I_{1,n_1}\}$. The sample allele frequency is $\hat{p}_1 = n_1^{-1} \sum_i I_{1,i}$. And the sample-$F_2$ is

$$\begin{aligned} F_2\left(\hat{p}_1, I_2\right) &= F_2\left(\frac{1}{n_1}\sum_{i=1}^{n_1} I_{1,i}, I_2\right) = \left(\frac{1}{n_1}\sum_{i=1}^{n_1} I_{1,i} - I_2\right)^2 \\ &= \frac{1}{n_1^2}\sum I_{1,i}^2 + \frac{2}{n_1^2}\sum I_{1,i} I_{1,j} - \frac{2}{n_1}\sum I_{1,i} I_2 + I_2^2 \\ &= \frac{1}{n_1}\sum I_{1,i}^2 - \frac{2}{n_1}\sum I_{1,i} I_2 + \frac{n_1}{n_1} I_2^2 + \frac{2}{n_1^2}\sum I_{1,i} I_{1,j} - \frac{n_1 - 1}{n_1^2}\sum I_{1,i}^2 \end{aligned}$$

The first three terms can be grouped into n_1 terms of the form $F_2(I_{1,i}, I_2)$, and the last two terms can be grouped into $\binom{n_1}{2}$ terms of the form $F_2(I_{1,i}, I_{1,j})$, one for each possible pair of samples in P_1.

Therefore,

$$F_2\left(\hat{p}_1, I_2\right) = \frac{1}{n_1}\sum_i F_2(I_{1,i}, I_2) - \frac{1}{n_1^2}\sum_{i<j} F_2(I_{1,i}, I_{1,j}) \tag{28}$$

where the second sum is over all pairs in P_1.

As $F_2(\hat{p}_1, \hat{p}_2) = F_2(\hat{p}_2, \hat{p}_1)$, we can switch the labels, and obtain the same expression for population $P_2 = \{I_{2,i}, i = 0, \ldots, n_2\}$ Taking the average over all $I_{2,j}$ yields

$$F_2\left(\hat{p}_1, \hat{p}_2\right) = \frac{1}{n_1}\sum_i F_2(I_{1,i}, I_{2,j}) - \frac{1}{n_1^2}\sum_{i<j} F_2(I_{1,i}, I_{1,j}) - \frac{1}{n_2^2}\sum_{i<j} F_2(I_{2,i}, I_{2,j}). \tag{29}$$

Thus, we can write F_2 between the two populations as the average number of differences we see between individuals from different populations, minus some terms including differences *within* each sample.

Equation 29 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. $F_2(I_{x_1,i} = I_{y_1,j}) = F_2(I_{x_2,i} = I_{y_2,j}) =$ for all pairs of samples (x_1, x_2) and (y_1, y_2) from populations P_i and P_j. Secondly, we assume that all samples correspond to the leaves of the tree, so that we can estimate branch lengths in terms of the time to a common ancestor T_{ij}. Finally, we assume that mutations occur at a constant rate of $\theta/2$ on each branch. Taken together, these assumptions imply that $F_2(I_{i,k}, I_{j,l}) = \theta \mathbb{E} T_{ij}$ for all individuals from populations P_i, P_j, this simplifies to

$$F_2(\hat{p}_1, \hat{p}_2) = \theta \times \left(\mathbb{E}T_{12} - \frac{1}{2}\left(1 - \frac{1}{n_1}\right)\mathbb{E}T_{11} - \frac{1}{2}\left(1 - \frac{1}{n_2}\right)\mathbb{E}T_{22}\right) \tag{30}$$

which, for the cases of $n = 1, 2$ was also derived by Petkova [41]. In most applications, we wish to calculate F_2 per segregating site in a large sample. As the expected number

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of segregating sites is $\frac{\theta}{2}T_{tot}$, we can follow [24, 41] and take the limit where $\theta \to 0$:

$$F_2(\hat{p}_1, \hat{p}_2) = \frac{2}{T_{tot}} \times \left(\mathbb{E}T_{12} - \frac{1}{2}\left(1 - \frac{1}{n_1}\right)\mathbb{E}T_{11} - \frac{1}{2}\left(1 - \frac{1}{n_2}\right)\mathbb{E}T_{22} \right) \tag{31}$$

to obtain an expression independent of the mutation rate. In either of these equations, we can see $\frac{2}{T_{tot}}$ or θ as a constant of proportionality that is the same for all statistics calculated from the same data. Since we are either interested in the relative magnitude of F_2, or whether a sum of F_2-values is different from zero, this constant has no impact on inference.

Furthermore, we can obtain a population-level statistic by taking the limit when the number of individuals per sample n_1 and n_2 go to infinity:

$$F_2(p_1, p_2) = \lim_{n_1, n_2 \to \infty} F_2(\hat{p}_1, \hat{p}_2). \tag{32}$$

This yields Equation 5. Using θ as the constant of proportionality, we find that

$$\mathbb{E}[2\pi_{12} - \pi_{11} - \pi_{22}] = F_2(P_1, P_2), \tag{33}$$

leading to the estimator given in 6.

It is straightforward to check that this estimator is equivalent to that given by Reich *et al.* [4]:

$$\begin{aligned}
F_2(P_1, P_2) &= \pi_{12} - \pi_{11}/2 - \pi_{22}/2 \\
&= \hat{p}_1(1-\hat{p}_2) + \hat{p}_2(1-\hat{p}_1) - \hat{p}_1(1-\hat{p}_1)\frac{n_1}{n_1-1} - \hat{p}_2(1-\hat{p}_2)\frac{n_2}{n_2-1} \\
&= \hat{p}_1\left(1 - 1 - \frac{1}{n_1}\right) + \hat{p}_2\left(1 - 1 - \frac{1}{n_2}\right) - 2\hat{p}_1\hat{p}_2 \\
&+ \hat{p}_1\left(1 - \frac{1}{n_1-1}\right) - \hat{p}_2\left(1 - \frac{1}{n_2-1}\right) \\
&= (\hat{p}_1 - \hat{p}_2)^2 - \frac{\hat{p}_1(1-\hat{p}_1)}{n_1-1} - \frac{\hat{p}_2(1-q2)}{n_2-1}
\end{aligned}$$

which is Equation 10 in the Appendix of [4].

Four-point-condition and F_4

We prove the statement that for any tree, two of the three possible F_4 values will be equal, and the last will be zero. First, notice that permuting one of the two pairs only changes the sign of the statistic, i.e.

$$F_4(p_1, p_2; p_3, p_4) = -F_4(p_2, p_1; p_3, p_4) \tag{34}$$

Using F_2 as a tree-metric, the four-point condition [17] can be written as

$$F_2(p_1, p_2) + F_2(p_3, p_4) \leq \min\Big[F_2(p_1, p_3) + F_2(p_2, p_4), F_2(p_1, p_4) + F_2(p_2, p_3)\Big] \tag{35}$$

which holds for any permutations of the samples.

Applying this to the first two and last two terms on the right-hand-side in equation 14b yields

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\begin{aligned}2F_4(p_1,p_2;p_3,p_4) =& F_2(p_1,p_3)+F_2(p_2,p_4)-(F_2(p_1,p_4)+F_2(p_2,p_3))\\ \leq& \min\big(F_2(p_1,p_2)+F_2(p_3,p_4),F_2(p_1,p_4)+F_2(p_2,p_3)\big)\\ &-\min\big(F_2(p_1,p_2)+F_2(p_3,p_4),F_2(p_1,p_3)+F_2(p_2,p_4)\big)\end{aligned} \tag{36}$$

The four-point condition states that two of the sums of disjoint F_2 statistics need to be identical, and the third one should be less or equal than that. This gives us four cases to evaluate (36) under:

1. If $F_2(p_1,p_2)+F_2(p_3,p_4)$ is smallest: (36) is zero
2. If $F_2(p_1,p_3)+F_2(p_2,p_4)$ is smallest: (36) is $F_4(p_1,p_4;p_2,p_3)>0$
3. If $F_2(p_1,p_4)+F_2(p_2,p_3)$ is smallest: (36) is $-F_4(p_1,p_4;p_2,p_3)<0$
4. All sums of F_2 are equal: (36) is zero

If the F_2 are not all equal, then for each F_4 with distinct pairs, one of conditions 2-4 is true, and we see that indeed one will be zero and the other two will have the same absolute value.

Derivation of F under select models

Here, we use Equation 5 together with Equations 10b and 14b to derive expectations for F_3 and F_4 under some simple models.

Panmixia Under panmixia with arbitrary population size changes, P_1 and P_2 are taken from the same pool of individuals and therefore $T_{12}=T_{11}=T_{22}$, $\mathbb{E}F_2=\mathbb{E}F_3=\mathbb{E}F_4=0$.

Island models A (finite) island model has D subpopulations of size 1 each. Migration occurs at rate M between subpopulations. It can be shown [42] that $\mathbb{E}T_{11}=\mathbb{E}T_{22}=D$. $\mathbb{E}T_{12}$ satisfies the recursion

$$\mathbb{E}T_{12}=\frac{1}{(D-1)M}+\frac{D-2}{D-1}\mathbb{E}T_{12}+\frac{1}{D-1}\mathbb{E}T_{11}. \tag{37}$$

with solution $\mathbb{E}T_{12}=1+M^{-1}$. This results in the equation in figure 6. The derivations for the hierarchical island models is marginally more complicated, but similar. It is given in [43].

Admixture models These are the model for which the F-statistics were originally developed. Many details, applications, and the origin of the path representation are found in [5]. For simplicity, we look at the simplest possible tree of size four, where P_X is admixed from P_1 and P_2 with contributions α and $\beta=(1-\alpha)$, respectively. We assume that all populations have the same size, and that this size is one. Then,

$$\begin{aligned}F_3(P_X;P_1,P_2) =& \mathbb{E}T_{1X}+\mathbb{E}T_{2X}-\mathbb{E}T_{12}-\mathbb{E}T_{XX}\\ =&(\alpha t_1+\beta t_r+1)+(\alpha t_r+\beta t_1+1)-t_r-1\\ &-\alpha^2 1\ -(1-\alpha)^2 1-2\alpha(1-\alpha)[(1-c_x)t_r+1]\\ =& t_1-2\alpha(1-\alpha)(1-c_x)t_r\end{aligned} \tag{38}$$

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

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Thus, we find that F_3 is negative if

$$\frac{t_1}{(1-c_x)t_r} < 2\alpha(1-\alpha), \tag{39}$$

which is more likely if α is large, the admixture is recent and the overall coalescent is far in the past.

For F_4, we have, omitting the within-population coalescence time of 1:

$$\begin{aligned} F_4(P_1P_X;P_2,P_3) =& \mathbb{E}T_{12} + \mathbb{E}T_{3X} - \mathbb{E}T_{13} - \mathbb{E}T_{2X} \\ =& t_r + \alpha t_r + \beta t_{23} - t_r - \alpha t_r - \beta t_{2X} \\ =& \beta(t_2 - t_1) \end{aligned}$$

Stepping-stone models For the stepping stone models, we 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, we 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, lineages in demes 1 and 4 and v) lineages in demes 2 and 3. Note that the symmetry of this system allows us to collapse some states. The transition matrix for this system is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 2m & 1-3m & m & 0 & 0 \\ 0 & m & 1-3m & m & m \\ 0 & 0 & 2m & 1-2m & 0 \\ 2m & 0 & 2m & 0 & 1-4m \end{pmatrix} \tag{40}$$

We can end the system once lineages are in the same deme, as the time to coalescence time is independent of the deme in isotropic migration models [42], and cancels from the F-statistics.

Therefore, we can find the vector v of the expected time until two lineages using standard Markov Chain theory by solving $v = (\mathbf{I} - \mathbf{T})^{-1})\mathbf{1}$, where $\mathbf{T}$ is the transition matrix involving only the transitive states in the Markov chain (all but the first state), and $\mathbf{1}$ is a vector of ones.

Finding the expected coalescent time involves solving a system of 5 equations. The terms involved in calculating the F-statistics (Table 1) are the entries in v corresponding to these states.

The hierarchical case is similar, except there are 6 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).

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2m & 1-3m & m & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & m & 1-3m & m & 0 & 0 & m & 0 & 0 & 0 \\ 0 & 0 & m & 1-3m & m & 0 & 0 & m & 0 & 0 \\ 0 & 0 & 0 & m & 1-3m & m & 0 & 0 & m & 0 \\ 0 & 0 & 0 & 0 & 2m & 1-2m & 0 & 0 & 0 & 0 \\ 2m & 0 & m & 0 & 0 & 0 & 1-4m & m & 0 & 0 \\ 0 & 0 & 0 & m & 0 & 0 & m & 1-4m & m & m \\ 0 & 0 & 0 & 0 & 2m & 0 & 0 & 2m & 1-4m & 0 \\ 2m & 0 & 0 & 0 & 0 & 0 & 0 & 2m & 0 & 1-4m \end{pmatrix}$$

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

And we can solve the same equation as in the non-hierarchical case to get all pairwise coalescence times. Then, all we have to do is average the coalescence times over all possibilities. E.g.

$$\mathbb{E}T_{1X} = \frac{v_2 + v_3 + v_6 + v_7}{4}. \tag{41}$$

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

We average the two left demes to P_1, the two right demes to P_2 and the two middle demes to P_X. The

Range expansion model We use a range expansion model with no migration [44]. Under that model, we 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 ids even further away from the expansion origin. Then $\mathbb{E}T_{12} = t_1 + \mathbb{E}T_{11}$, where $\mathbb{E}t_1$ is the time required for a lineage sampled further away in the expansion to end up in D_1. (Note that t_1 only depends on the deme that is closer to the origin). Thus, for three demes,

$$\begin{aligned} F_3(p_2; p_1, p_3) &\propto \mathbb{E}T_{12} - \mathbb{E}T_{13} + \mathbb{E}T_{23} - \mathbb{E}T_{22} \\ &\propto \mathbb{E}T_{11} + t_1 - \mathbb{E}T_{11} - t_1 + \mathbb{E}T_{22} + t_2 - \mathbb{E}T_{22} \\ &\propto t_2. \end{aligned}$$

and

$$\begin{aligned} F_4(p_1, p_2; p_3, p_4) &\propto \mathbb{E}T_{13} - \mathbb{E}T_{14} + \mathbb{E}T_{24} - \mathbb{E}T_{23} \\ &\propto \mathbb{E}T_{11} + t_1 - \mathbb{E}T_{11} - t_1 + \mathbb{E}T_{22} + t_2 - \mathbb{E}T_{22} - t_2 \\ &= 0. \end{aligned}$$

More interesting is

$$\begin{aligned} F_4(p_1, p_3; p_2, p_4) &\propto \mathbb{E}T_{12} - \mathbb{E}T_{14} + \mathbb{E}T_{34} - \mathbb{E}T_{23} \\ &\propto \mathbb{E}T_{11} + t_1 - \mathbb{E}T_{11} - t_1 + \mathbb{E}T_{33} + t_3 - \mathbb{E}T_{22} - t_2 \\ &\propto \mathbb{E}T_{33} + t_3 - \mathbb{E}T_{22} - t_2. \end{aligned}$$

Simulations

Simulations were performed using `ms` [45]. Specific commands used are

```
ms 1201 100 -t 10 -I 13 100 100 100 100 100 100 100 100 100 100 100
100 1 -ej 0.01 2 1 -ej 0.02 3 1 -ej 0.04 4 1 -ej 0.06 5 1 -ej 0.08 6 1
-ej 0.10 7 1 -ej 0.12 8 1 -ej 0.14 9 1 -ej 0.16 10 1 -ej 0.16 11 1 -ej
0.3 12 1 -ej 0.31 13 1
```

for the outgroup-F_3-statistic (Figure 4A),

```
ms 301 100 -t 10 -I 4 100 100 100 1 -es 0.001 2 $ALPHA  -ej 0.03 2 1
-ej 0.03 5 3 -ej 0.3 3 1 -ej 0.31 4 1
```

for Figure 4B, where the admixture proportion `$ALPHA` was varied in increments of 0.025 from 0 to 0.5, with 200 data sets generated per `$ALPHA`.

Lastly, data for Figure 4C was simulated using

```
ms 501 100 -t 50 -r 50 10000 -I 6 100 100 100 100 100 1 -es 0.001 3
$ALPHA -ej 0.03 3 2 -ej 0.03 7 4 -ej 0.1 2 1 -ej 0.2 4 1 -ej 0.3 5 1 -ej
0.31 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 [6].

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513(7518):409–413.

2. Yunusbayev B, Metspalu M, Metspalu E, Valeev A, Litvinov S, Valiev R, et al. The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. PLoS Genet. 2015 Apr;11(4):e1005068.

3. Pickrell JK, Reich D. Toward a new history and geography of human genes informed by ancient DNA. Trends in Genetics. 2014 Sep;30(9):377–389.

4. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461(7263):489–494.

5. Patterson NJ, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient Admixture in Human History. Genetics. 2012 Sep;p. genetics.112.145037.

6. Pickrell JK, Pritchard JK. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet. 2012 Nov;8(11):e1002967.

7. Lipson M, Loh PR, Levin A, Reich D, Patterson N, Berger B. Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. Molecular Biology and Evolution. 2013 Aug;30(8):1788–1802.

8. Ralph P, Coop G. The Geography of Recent Genetic Ancestry across Europe. PLoS Biol. 2013 May;11(5):e1001555.

9. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A Genetic Atlas of Human Admixture History. Science. 2014 Feb;343(6172):747–751.

10. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. science. 2010;328(5979):710.

11. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al. Reconstructing Native American population history. Nature. 2012 Aug;488(7411):370–374.

12. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015 Jun;522(7555):207–211.

13. Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of Bronze Age Eurasia. Nature. 2015 Jun;522(7555):167–172.

14. Wright S. Evolution in Mendelian populations. Genetics. 1931;16:97–159.

15. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984 Nov;38(6):1358–1370. ArticleType: research-article / Full publication date: Nov., 1984 / Copyright © 1984 Society for the Study of Evolution.

16. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2014;505(7481):87–91.

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

17. Buneman P. The recovery of trees from measures of dissimilarity. Mathematics in the Archaeological and Historical Sciences. 1971;.

18. Semple C, Steel MA. Phylogenetics. Oxford University Press; 2003.

19. Wahlund S. Zusammensetzung Von Populationen Und Korrelationserscheinungen Vom Standpunkt Der Vererbungslehre Aus Betrachtet. Hereditas. 1928 May;11(1):65–106.

20. Cavalli-Sforza LL, Edwards AWF. Phylogenetic Analysis: Models and Estimation Procedures. Evolution. 1967;21(3):550–570. ArticleType: research-article / Full publication date: Sep., 1967 / Copyright © 1967 Society for the Study of Evolution.

21. Felsenstein J. Maximum-likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics. 1973 Sep;25(5):471–492.

22. Cavalli-Sforza LL, Piazza A. Analysis of evolution: Evolutionary rates, independence and treeness. Theoretical Population Biology. 1975 Oct;8(2):127–165.

23. Felsenstein J. Evolutionary Trees From Gene Frequencies and Quantitative Characters: Finding Maximum Likelihood Estimates. Evolution. 1981 Nov;35(6):1229–1242.

24. Slatkin M. Inbreeding coefficients and coalescence times. Genetic Research. 1991;58:167–175.

25. Excoffier L, Smouse PE, Quattro JM. Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. Genetics. 1992;131:479–491.

26. Malecot G, et al. Mathematics of heredity. Les mathematiques de l'heredite. 1948;.

27. Wright S. Systems of mating. Genetics. 1921;6(2):111–178.

28. Wakeley J. Coalescent theory: an introduction. Roberts & Co. Publishers; 2009.

29. Tajima F. Evolutionary Relationship of Dna Sequences in Finite Populations. Genetics. 1983 Oct;105(2):437–460.

30. Fitch WM, Margoliash E, others. Construction of phylogenetic trees. Science. 1967;155(3760):279–284.

31. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987 Jul;4(4):406–425.

32. Felsenstein J. Inferring phytogenies. Sunderland, Massachusetts: Sinauer Associates. 2004;.

33. Buneman P. A note on the metric properties of trees. Journal of Combinatorial Theory, Series B. 1974;17(1):48–50.

34. Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Molecular Biology and Evolution. 2011;28(8):2239–2252.

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

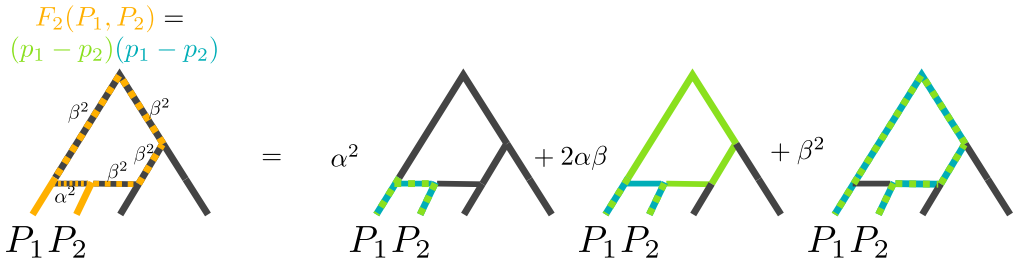


Figure S1. Path interpretation of F_2: F_2 is interpreted as two possible paths from P_1 to P_2, which we color green and blue, respectively. With probability α, a path takes the left admixture edge, and with probability $\beta = 1 - \alpha$, the right one. The dotted lines give the overlap of the two paths, conditional on which admixture edge they take, and the result is summarized as the weighted sum of branches in the left-most tree. For a more detailed explanation, see [5].

35. McCullagh P. Marginal likelihood for distance matrices. Statistica Sinica. 2009;19(2):631.

36. Pease JB, Hahn MW. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. Systematic Biology. 2015 Jul;64(4):651–662.

37. Huson DH, Rupp R, Scornavacca C. Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press; 2010.

38. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Molecular biology and evolution. 2006;23(2):254–267.

39. Bryant D, Moulton V. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. Molecular Biology and Evolution. 2004 Feb;21(2):255–265.

40. Tavaré S. Line-of-descent and genealogical processes, and their applications in population genetics models. Theoretical Population Biology. 1984 Oct;26(2):119–164.

41. Petkova D, Novembre J, Stephens M. Visualizing spatial population structure with estimated effective migration surfaces. bioRxiv. 2014 Nov;p. 011809.

42. Strobeck C. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics. 1987 Sep;117(1):149–153.

43. Slatkin M, Voelm L. FST in a Hierarchical Island Model. Genetics. 1991;127:627–629.

44. Peter BM, Slatkin M. The effective founder effect in a spatially expanding population. Evolution. 2015 Mar;69(3):721–734.

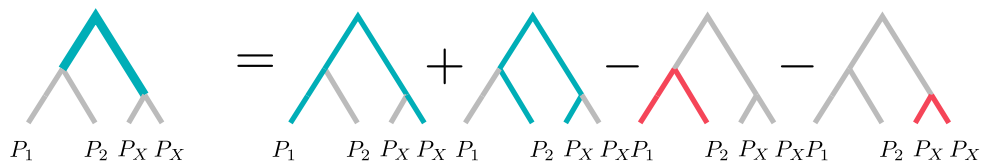
45. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics (Oxford, England). 2002 Feb;18(2):337–338.

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A. Equation

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

B. Concordant genealogy



C. Discordant genealogy

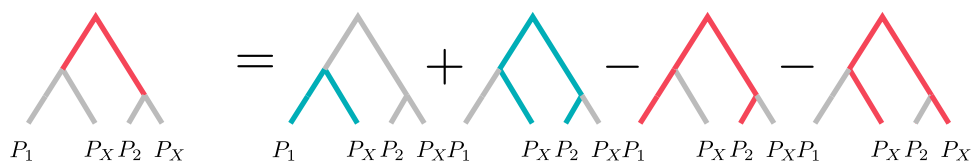


Figure S2. Schematic explanation how F_3 behaves conditioned on gene tree. Blue terms and branches correspond to positive contributions, whereas red branches and terms are subtracted. Labels represent individuals randomly sampled from that population. We see that external branches cancel out, so only 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.

bioRxiv preprint doi: https://doi.org/10.1101/028753; this version posted October 9, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

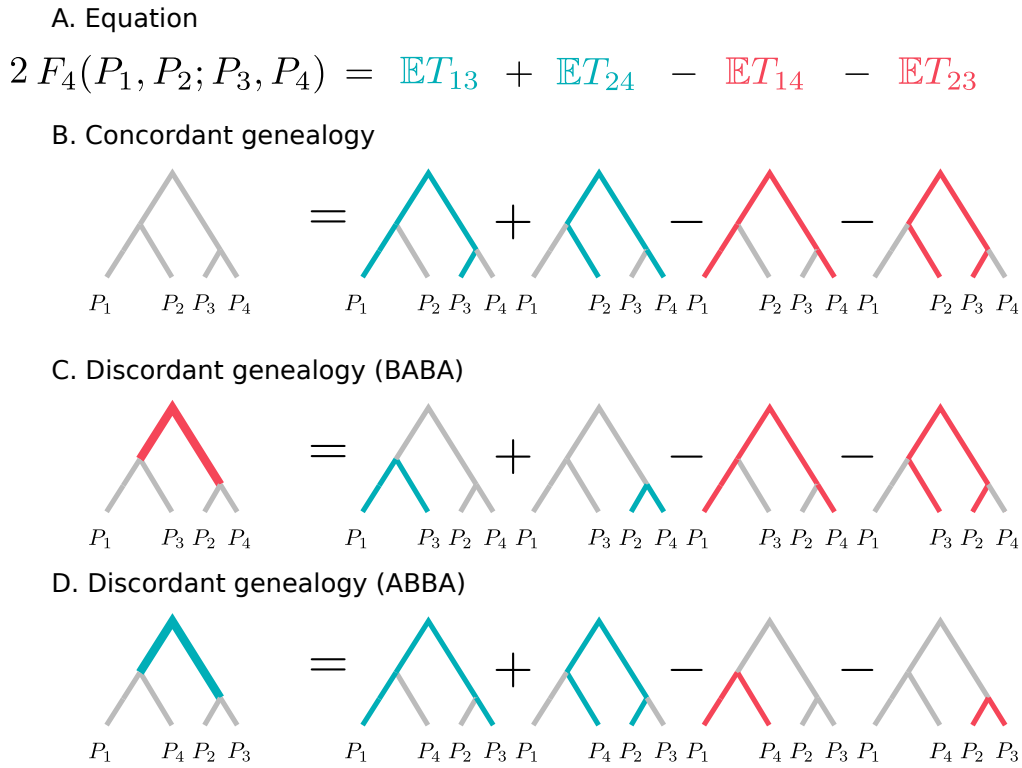


Figure S3. Schematic explanation how F_4 behaves conditioned on gene tree. Blue terms and branches correspond to positive contributions, whereas red branches and terms are subtracted. Labels represent individuals randomly sampled from that population. We see that all branches cancel out in the concordant 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.