Theoretical Population Biology 80 (2011) 208–216

Contents lists available at SciVerse ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Mathematical properties of F_{st} between admixed populations and their parental source populations

Simina M. Boca [a],*, Noah A. Rosenberg [b]

[a] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205, USA

[b] Department of Human Genetics, Center for Computational Medicine and Bioinformatics, and the Life Sciences Institute, University of Michigan, 100 Washtenaw Ave., Ann Arbor, MI 48109, USA

ARTICLE INFO

Article history:
Received 13 February 2011
Available online 25 May 2011

Keywords:
Admixture
Allele frequencies
F_{st}

ABSTRACT

We consider the properties of the F_{st} measure of genetic divergence between an admixed population and its parental source populations. Among all possible populations admixed among an arbitrary set of parental populations, we show that the value of F_{st} between an admixed population and a specific source population is maximized when the admixed population is simply the most distant of the other source populations. For the case with only two parental populations, as a function of the admixture fraction, we further demonstrate that this F_{st} value is monotonic and convex, so that F_{st} is informative about the admixture fraction. We illustrate our results using example human population-genetic data, showing how they provide a framework in which to interpret the features of F_{st} in admixed populations.

1. Introduction

The well-known "fixation index" F_{st} quantifies the extent to which a polymorphic population is subdivided into subpopulations (Wright, 1951; Excoffier, 2001; Rousset, 2001; Balding, 2003; Holsinger and Weir, 2009). In a definition due to Nei (1973, 1987), F_{st} is defined in terms of the expected heterozygosity of the overall population and the mean expected heterozygosity across the subpopulations.

Definition. The *expected heterozygosity* in a population for a given locus with I distinct alleles is defined as $H = 1 - \sum_{i=1}^{I} p_i^2$, where p_i is the frequency of allele i.

Definition. At a given locus, the *fixation index*, F_{st}, is defined as $F_{st} = (H_t - H_s)/H_t$, where H_t is the expected heterozygosity of the overall population, and H_s is the mean expected heterozygosity across subpopulations.

Assuming that the subpopulations have equal contribution to the total population, H_t is computed by pooling the various subpopulations in equal proportion, and H_s is calculated by weighting the various subpopulations equally. Recall that F_{st} always lies between 0 and 1, and that F_{st} is 0 if and only if $H_t = H_s$, meaning that the pooled population is unstructured. F_{st} increases as the genetic differentiation between the various subpopulations increases, and the theoretical maximum of 1 is reached if and only if each subpopulation is entirely monomorphic (and homozygous).

We consider the fixation index in the context of admixture, where one population arises from the amalgamation of multiple populations, typically after a long period of relative isolation for the founding groups. Admixture scenarios have been abundant over the course of human history, and for admixture events that have occurred recently, the admixture process has left a detectable signature in the genomes of admixed individuals. For example, the contributions of European, Native American, and African populations to the genetic history of African American (Parra et al., 1998, 2001; Salas et al., 2005; Tang et al., 2006b; Tishkoff et al., 2009; Zakharia et al., 2009; Bryc et al., 2010a) and Hispanic and Mestizo (Bonilla et al., 2004; Seldin et al., 2007; Tang et al., 2007; Wang et al., 2008; Risch et al., 2009; Silva-Zolezzi et al., 2009; Bryc et al., 2010b) populations have been the focus of much investigation. Genetic studies of admixture have further been of interest not only for what they reveal about population history, but also because admixed populations can be used in locating disease-associated genomic regions through methods that search in admixed individuals for regions of the genome with excess ancestry from the ancestral population in which a disease is more prevalent (reviewed by McKeigue, 2005; Reich and Patterson, 2005; Smith and O'Brien, 2005; Seldin, 2007; Buerkle and Lexer, 2008; Zhu et al., 2008; Winkler et al., 2010).

In an admixture setting, we provide a theoretical framework for explaining the properties of F_{st} between admixed populations and their parental populations. We examine the values of F_{st}

* Corresponding author.
E-mail addresses: sboca@jhsph.edu (S.M. Boca), rnoah@umich.edu (N.A. Rosenberg).

0040-5809/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2011.05.003

between pairs of populations, in which one member of the pair is an admixed population and the other is one of its parental source populations. After introducing notation and an example dataset in Section 2, in Section 3, we prove our main theorem, which concerns the value of F_{st} between a population formed by admixture of K founding populations and a specific one of the K founding populations. We show that for any $K \geq 2$, considering all admixture combinations for a given set of founding populations, this F_{st} expression is maximized when the admixed population is in fact one of the other founding populations. In Section 4, we then consider the special case of $K = 2$ founding populations, proving in Section 4.1 that F_{st} is monotonic and convex in the admixture coefficient. Section 4.2 uses microsatellite genotype data on Mestizo populations to demonstrate that F_{st} values predicted using our theoretical results closely match observed F_{st} values. Section 4.3 then suggests an estimator of admixture on the basis of F_{st}. Finally, in Section 5, we summarize the main results and discuss their broader implications.

2. Notation and data

We use a simplified form of the F_{st} value between two populations that have the same contribution to the total population and that when pooled together produce a polymorphic population. Denote by p_{1i} the frequency of allele i in population 1 and by p_{2i} the frequency of allele i in population 2. We then have

$$H_t = 1 - \sum_{i=1}^{I}\left[\frac{1}{2}(p_{1i}+p_{2i})\right]^2 = 1 - \frac{1}{4}\sum_{i=1}^{I}(p_{1i}+p_{2i})^2$$

$$\begin{aligned} H_s &= \frac{1}{2}\left[\left(1-\sum_{i=1}^{I}p_{1i}^2\right)+\left(1-\sum_{i=1}^{I}p_{2i}^2\right)\right] \\ &= 1 - \frac{1}{2}\sum_{i=1}^{I}(p_{1i}^2+p_{2i}^2) \end{aligned}$$

$$\begin{aligned} H_t - H_s &= \frac{1}{4}\sum_{i=1}^{I}[2p_{1i}^2+2p_{2i}^2-(p_{1i}+p_{2i})^2] \\ &= \frac{1}{4}\sum_{i=1}^{I}(p_{1i}-p_{2i})^2 \end{aligned}$$

$$\begin{aligned} F_{st} &= \frac{H_t-H_s}{H_t} = \frac{\frac{1}{4}\sum_{i=1}^{I}(p_{1i}-p_{2i})^2}{1-\frac{1}{4}\sum_{i=1}^{I}(p_{1i}+p_{2i})^2} \\ &= \frac{\sum_{i=1}^{I}(p_{1i}-p_{2i})^2}{4-\sum_{i=1}^{I}(p_{1i}+p_{2i})^2}. \qquad (1) \end{aligned}$$

Table 1 summarizes the notation for the scenario in which $K \geq 2$ founding populations give rise to an admixed population. Denote by r_i the frequency of allele i in the admixed population. Let γ_k represent the admixture fraction corresponding to population k, so that fraction γ_k of the ancestry of the admixed population derives from source population k. The admixed population then has

$$r_i = \sum_{k=1}^{K}\gamma_k p_{ki},$$

where $\gamma_k \geq 0$ for $1 \leq k \leq K$ and $\sum_{k=1}^{K}\gamma_k = 1$. We assume throughout the paper that there exists at least one pair of founding populations, k and ℓ, for which $\underline{p_k} \neq \underline{p_\ell}$. This assumption corresponds to an assumption that the founding populations do not all have identical allele frequencies.

Table 1
Notation.

Type of quantity	Symbol	Description
Indices	$i = 1, \ldots, I$	Index over alleles
	$k = 1, \ldots, K$	Index over populations
Allele frequencies	p_{ki}	Frequency of allele i in population k
	$\underline{p_k}$	Vector of allele frequencies for population k
	r_i	Frequency of allele i in the admixed population
Admixture fractions	γ_k	Admixture fraction for population k
	$\underline{\gamma}$	Vector of admixture fractions

The example scenarios that we consider use a subset of the data from Wang et al. (2008), consisting of genotypes of 249 Mestizos, 160 Europeans, 463 Native Americans, and 123 Africans at 678 autosomal microsatellite loci. The Mestizo samples provide an example of admixture primarily between European and Native American founding populations, and to a lesser extent, African populations. In our analyses, we focus on the European and Native American contributions, treating the full Mestizo sample as an admixed population and the European and Native American samples as its founding populations. In each of the population samples, except where otherwise specified (in particular, in Section 4.2), we treat the sample allele frequencies from Wang et al. (2008) as the parametric allele frequencies.

3. General case: *K* founding populations

Our goal in this section is to examine F_{st} to a specific founding population over the space of admixture vectors possible for a population admixed among a given collection of K founding populations. We begin by providing an expression for F_{st} between the admixed population and a specific founding population. Without loss of generality, we investigate F_{st} between population 1 and the admixed population. Using Eq. (1) and viewing F_{st} as a function of two allele frequency vectors, we have

$$F_{st}(\underline{p_1}, \underline{r}) = \frac{\sum_{i=1}^{I}\left(p_{1i}-\sum_{k=1}^{K}\gamma_k p_{ki}\right)^2}{4-\sum_{i=1}^{I}\left(p_{1i}+\sum_{k=1}^{K}\gamma_k p_{ki}\right)^2}. \qquad (2)$$

In Theorem 2, we obtain a result concerning the maximum over admixed populations of F_{st} between the admixed population and an arbitrarily chosen founding population. We first prove a preliminary result involving F_{st} between one population and a population formed by admixture of two other populations.

Lemma 1 (*Three-Population Lemma*)**.** *Denote by $\underline{p_1}, \underline{p_2}, \underline{p_3}$ the vectors corresponding to the allele frequencies of three populations. Consider a population formed by admixture between populations 2 and 3, where $\gamma \in [0, 1]$ represents the admixture fraction for population 2. Then*

$$\max_{\gamma\in[0,1]} F_{st}[\underline{p_1}, \gamma\underline{p_2}+(1-\gamma)\underline{p_3}] = \max\{F_{st}(\underline{p_1}, \underline{p_2}), F_{st}(\underline{p_1}, \underline{p_3})\}.$$

Proof. We start by applying Eq. (1) to calculate F_{st} between population 1 and the population formed by admixture of populations 2 and 3. We also introduce some additional variables, $\delta_{13i} = p_{1i}-p_{3i}$, $\delta_{23i} = p_{2i}-p_{3i}$, and $\tau_{13i} = p_{1i}+p_{3i}$, for $1 \leq i \leq I$, to simplify the notation. Then

$$F_{st}[\underline{p_1}, \gamma\underline{p_2}+(1-\gamma)\underline{p_3}] = \frac{\sum_{i=1}^{I}[p_{1i}-\gamma p_{2i}-(1-\gamma)p_{3i}]^2}{4-\sum_{i=1}^{I}[p_{1i}+\gamma p_{2i}+(1-\gamma)p_{3i}]^2}$$

$$= \frac{\sum_{i=1}^{I}(\delta_{13i} - \gamma\delta_{23i})^2}{4 - \sum_{i=1}^{I}(\tau_{13i} + \gamma\delta_{23i})^2}. \qquad (3)$$

We denote the function $F_{st}[\underline{p_1}, \gamma\underline{p_2} + (1-\gamma)\underline{p_3}]$ by $G(\gamma)$, to emphasize the fact that we aim to maximize this quantity with respect to γ. To calculate the derivative of $G(\gamma)$, we first do some preliminary calculations:

$$\frac{d}{d\gamma}\sum_{i=1}^{I}(\delta_{13i} - \gamma\delta_{23i})^2 = -2\sum_{i=1}^{I}\delta_{23i}(\delta_{13i} - \gamma\delta_{23i})$$

$$\frac{d}{d\gamma}\left[4 - \sum_{i=1}^{I}(\tau_{13i} + \gamma\delta_{23i})^2\right] = -2\sum_{i=1}^{I}\delta_{23i}(\tau_{13i} + \gamma\delta_{23i}).$$

Putting these results together, we obtain the equation given in Box I. The denominator in Box I is always positive, so we focus on the numerator to see where it is greater or less than 0. We denote half the numerator by $E(\gamma)$, and note that $E(\gamma)$ is a polynomial in γ, of degree at most 2. We denote the coefficients of γ^2, γ, and 1 in $E(\gamma)$, by a, b, and c, respectively, and calculate them individually:

$$a = -2\left(\sum_{i=1}^{I}\delta_{23i}^2\right)\left(\sum_{i=1}^{I}p_{1i}\delta_{23i}\right)$$

$$b = 4\left(\sum_{i=1}^{I}\delta_{23i}^2\right)\left(1 - \sum_{i=1}^{I}p_{1i}p_{3i}\right)$$

$$c = -4\left(\sum_{i=1}^{I}\delta_{13i}\delta_{23i}\right) + \left(\sum_{i=1}^{I}\delta_{13i}\delta_{23i}\right)\left(\sum_{i=1}^{I}\tau_{13i}^2\right) + \left(\sum_{i=1}^{I}\tau_{13i}\delta_{23i}\right)\left(\sum_{i=1}^{I}\delta_{13i}^2\right). \qquad (4)$$

We next show that $E(\gamma)$ is increasing or decreasing on [0, 1]. If $a = 0$, then $E(\gamma)$ is linear, and the claim is trivial (a and b cannot both be zero, because $a = b = 0$ implies that populations 2 and 3 have identical allele frequencies). Suppose $a \neq 0$. We show that the position of the vertex of the parabola $E(\gamma) = a\gamma^2 + b\gamma + c$, or $-b/(2a)$, is not in (0, 1). We first note that $1 = \sum_{i=1}^{I} p_{1i} \geq \sum_{i=1}^{I} p_{1i}p_{3i}$, so that $b \geq 0$. Consequently, if $-b/(2a) > 0$, then $a < 0$ and $\sum_{i=1}^{I} p_{1i}p_{2i} > \sum_{i=1}^{I} p_{1i}p_{3i}$. If $0 < -b/(2a) < 1$, then we also have $1 - \sum_{i=1}^{I} p_{1i}p_{3i} < \sum_{i=1}^{I} p_{1i}p_{2i} - \sum_{i=1}^{I} p_{1i}p_{3i}$, which means that $1 < \sum_{i=1}^{I} p_{1i}p_{2i}$. This inequality clearly does not hold, because $1 = \sum_{i=1}^{I} p_{1i} \geq \sum_{i=1}^{I} p_{1i}p_{2i}$. As a result, $-b/(2a) \notin (0, 1)$, so that the extrema of $E(\gamma)$ on [0, 1] must occur at $\gamma = 0$ and $\gamma = 1$. It follows that $E(\gamma)$ is either increasing or decreasing for $\gamma \in (0, 1)$.

To demonstrate that $E(\gamma)$ is increasing, we show that $E(1) > E(0)$:

$$E(1) - E(0) = -2\left(\sum_{i=1}^{I}\delta_{23i}^2\right)\left(\sum_{i=1}^{I}p_{1i}\delta_{23i}\right) + 4\left(\sum_{i=1}^{I}\delta_{23i}^2\right)\left(1 - \sum_{i=1}^{I}p_{1i}p_{3i}\right)$$
$$= 2\left(\sum_{i=1}^{I}\delta_{23i}^2\right)\left(2 - \sum_{i=1}^{I}p_{1i}p_{2i} - \sum_{i=1}^{I}p_{1i}p_{3i}\right).$$

We note that $\sum_{i=1}^{I}\delta_{23i}^2 > 0$ by the assumption that populations 2 and 3 do not have identical allele frequencies, and $2 = \sum_{i=1}^{I} p_{2i} + \sum_{i=1}^{I} p_{3i} > \sum_{i=1}^{I} p_{1i}p_{2i} + \sum_{i=1}^{I} p_{1i}p_{3i}$. This inequality is strict by this same assumption, as population 1 cannot have identical allele frequencies to both populations 2 and 3. Consequently, $E(1) > E(0)$, and $E(\gamma)$ is increasing on $\gamma \in [0, 1]$. Note that if populations 2 and 3 were switched, so that we instead considered the value of $F_{st}[\underline{p_1}, (1-\gamma)\underline{p_2} + \gamma\underline{p_3}]$ rather than $F_{st}[\underline{p_1}, \gamma\underline{p_2} + (1-\gamma)\underline{p_3}]$, then $E(1) - E(0)$ would not change, and we would reach the same conclusion that $E(1) > E(0)$ and $E(\gamma)$ is increasing on the interval.

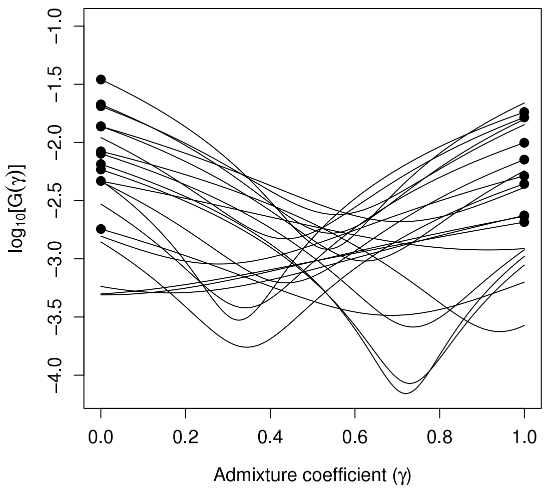


Fig. 1. F_{st} between a population and a hypothetical second population that is admixed between two other populations. $\log_{10}[G(\gamma)]$, seen as a function of γ, is plotted against γ, where $G(\gamma)$ is $F_{st}[\underline{p_1}, \gamma\underline{p_2} + (1-\gamma)\underline{p_3}]$ (Eq. (3)). Populations 1, 2, and 3 represent populations of Mestizo, European, and Native American descent, respectively, and $\underline{p_1}$, $\underline{p_2}$, and $\underline{p_3}$ are based on allele frequencies estimated from Wang et al. (2008). Twenty randomly selected loci are considered, with each curve representing a different locus. The dots indicate the maxima for individual curves. In accordance with Lemma 1, for $\gamma \in [0, 1]$, the maximal value of $\log_{10}[G(\gamma)]$, and therefore of $G(\gamma)$, is always at $\gamma = 0$ or $\gamma = 1$.

Three possibilities exist for the location of 0: $E(1) > E(0) \geq 0$, $E(1) > 0 > E(0)$, or $0 \geq E(1) > E(0)$. Because $E(\gamma)$ only differs from $\frac{d}{d\gamma}G(\gamma)$ by a positive factor, these three possibilities correspond to three possibilities for the shape of $G(\gamma)$ on $\gamma \in [0, 1]$: $G(\gamma)$ is increasing, $G(\gamma)$ is decreasing up to a point, then increasing, or $G(\gamma)$ is decreasing. In each of these three cases, it follows that

$$\max_{\gamma\in[0,1]} F_{st}[\underline{p_1}, \gamma\underline{p_2} + (1-\gamma)\underline{p_3}] = \max_{\gamma\in[0,1]} G(\gamma)$$
$$= \max\{G(0), G(1)\} = \max\{F_{st}(\underline{p_1}, \underline{p_2}), F_{st}(\underline{p_1}, \underline{p_3})\}. \quad \square \qquad (5)$$

The result of this lemma is illustrated on a data example in Fig. 1, in which $\log_{10}[G(\gamma)]$ is plotted against γ. In this case, population 1 represents a Mestizo admixed population, and populations 2 and 3 represent European and Native American populations, respectively. For each of twenty loci considered, as in the lemma, the maximum of the function is located either at $\gamma = 0$ or at $\gamma = 1$.

We now use the three-population lemma to prove that for any K, F_{st} between a population formed by admixture of K founding populations and a specific one of those founding populations is maximized when the admixed population is in fact one of the other $K - 1$ founding populations.

Theorem 2. *Denote by $\underline{p_1}, \ldots, \underline{p_K}$ the vectors corresponding to the allele frequencies of K populations. Consider a population formed by admixture between the K populations, where $\gamma_k \in [0, 1]$ represents the admixture fraction for population k, for $1 \leq k \leq K$, such that $\sum_{k=1}^{K}\gamma_k = 1$. Then*

$$\max_{\underline{\gamma}} F_{st}[\underline{p_1}, \gamma_1\underline{p_1} + \cdots + \gamma_K\underline{p_K}]$$
$$= \max\{F_{st}(\underline{p_1}, \underline{p_2}), \ldots, F_{st}(\underline{p_1}, \underline{p_K})\}. \qquad (6)$$

Proof. We prove this result by induction on K.

$$\frac{d}{d\gamma}G(\gamma)=\frac{-2\left[\sum_{i=1}^{I}\delta_{23i}(\delta_{13i}-\gamma\delta_{23i})\right]\left[4-\sum_{i=1}^{I}(\tau_{13i}+\gamma\delta_{23i})^2\right]+2\left[\sum_{i=1}^{I}(\delta_{13i}-\gamma\delta_{23i})^2\right]\left[\sum_{i=1}^{I}\delta_{23i}(\tau_{13i}+\gamma\delta_{23i})\right]}{\left[4-\sum_{i=1}^{I}(\tau_{13i}+\gamma\delta_{23i})^2\right]^2}.$$

Box I.

Step 1. $K=2$. Taking $\underline{p_1}$ and $\underline{p_2}$ in place of $\underline{p_2}$ and $\underline{p_3}$, respectively, Lemma 1 already demonstrates the result in the case that $K=2$:

$$\begin{aligned}&\max_{\gamma\in[0,1]} F_{st}[\underline{p_1},\gamma\underline{p_1}+(1-\gamma)\underline{p_2}]\\&\quad=\max\{F_{st}(\underline{p_1},\underline{p_1}),F_{st}(\underline{p_1},\underline{p_2})\}=F_{st}(\underline{p_1},\underline{p_2}),\end{aligned}$$

where, for simplicity, we replace γ_1 by γ and γ_2 by $1-\gamma$.

Step 2. $K\to K+1$. We now show that if the result in Eq. (6) holds for K populations, then it also holds for $K+1$ populations (with $\gamma_{K+1}>0$):

$$\begin{aligned}&F_{st}[\underline{p_1},\gamma_1\underline{p_1}+\cdots+\gamma_{K-1}\underline{p_{K-1}}+\gamma_K\underline{p_K}+\gamma_{K+1}\underline{p_{K+1}}]\\&\quad=F_{st}\left[\underline{p_1},\gamma_1\underline{p_1}+\cdots+\gamma_{K-1}\underline{p_{K-1}}+(\gamma_K+\gamma_{K+1})\right.\\&\qquad\left.\times\left(\frac{\gamma_K}{\gamma_K+\gamma_{K+1}}\underline{p_K}+\frac{\gamma_{K+1}}{\gamma_K+\gamma_{K+1}}\underline{p_{K+1}}\right)\right]\\&\quad\le\max\left\{F_{st}(\underline{p_1},\underline{p_2}),\ldots,F_{st}(\underline{p_1},\underline{p_{K-1}}),\right.\\&\qquad\left.F_{st}\left(\underline{p_1},\frac{\gamma_K}{\gamma_K+\gamma_{K+1}}\underline{p_K}+\frac{\gamma_{K+1}}{\gamma_K+\gamma_{K+1}}\underline{p_{K+1}}\right)\right\},\end{aligned}$$

where the last step follows by the inductive hypothesis, Eq. (6), for the case with K populations. The expression $[\gamma_K/(\gamma_K+\gamma_{K+1})]\underline{p_K}+[\gamma_{K+1}/(\gamma_K+\gamma_{K+1})]\underline{p_{K+1}}$ has the form $\gamma\underline{p_K}+(1-\gamma)\underline{p_{K+1}}$, with $\gamma_K/(\gamma_K+\gamma_{K+1})$ taking on the role of γ. Consequently, using Lemma 1,

$$\begin{aligned}&F_{st}\left[\underline{p_1},\frac{\gamma_K}{\gamma_K+\gamma_{K+1}}\underline{p_K}+\frac{\gamma_{K+1}}{\gamma_K+\gamma_{K+1}}\underline{p_{K+1}}\right]\\&\quad\le\max\{F_{st}(\underline{p_1},\underline{p_K}),F_{st}(\underline{p_1},\underline{p_{K+1}})\},\end{aligned}$$

so that

$$\begin{aligned}&F_{st}[\underline{p_1},\gamma_1\underline{p_1}+\cdots+\gamma_{K-1}\underline{p_{K-1}}+\gamma_K\underline{p_K}+\gamma_{K+1}\underline{p_{K+1}}]\\&\quad\le\max\{F_{st}(\underline{p_1},\underline{p_2}),\ldots,F_{st}(\underline{p_1},\underline{p_{K-1}}),F_{st}(\underline{p_1},\underline{p_K}),F_{st}(\underline{p_1},\underline{p_{K+1}})\}.\end{aligned}$$

Thus, the induction is complete, and we have shown that the result in Eq. (6) holds for arbitrary K. □

The theorem is sensible, in that considering all possible populations admixed among a given collection of source populations, the most "distant" populations from source population 1 are combinations that do not include ancestry from population 1. The theorem demonstrates that the most distant admixed population according to F_{st} is precisely one of the remaining source populations; it is not a nontrivial mixture of those source populations, either with each other or with population 1.

An interesting corollary of Theorem 2 is that given a set of K founding populations, considering all admixed populations that can be constructed from those founding populations, the value of F_{st} between an admixed population and the founding population from which it is maximally distant, or $\max_{k\in\{1,\ldots,K\}}F_{st}(\underline{p_k},\underline{r})$, is bounded above by the maximal F_{st} among pairs of founding populations, or $\max_{k,\ell\in\{1,\ldots,K\}}F_{st}(\underline{p_k},\underline{p_\ell})$. This result is obtained by simply noting that given k, as a result of the theorem, $F_{st}(\underline{p_k},\underline{r})$ is bounded above by $\max_{\ell\in\{1,\ldots,K\}}F_{st}(\underline{p_k},\underline{p_\ell})$, and by then taking the maximum over k. Thus, according to the F_{st} measure, an admixed population can be no more distant from any of its founding populations than the two most distant among the founding populations are from each other.

Three examples illustrating the theorem with $K=3$ are presented in Fig. 2. Sample allele frequencies for three genetic loci in three populations – European, Native American, and African – are used as $\underline{p_1}$, $\underline{p_2}$, and $\underline{p_3}$, respectively. The triangular region shown in the figure for a given locus represents the space of possible admixture vectors $(\gamma_1,\gamma_2,\gamma_3)$. For each locus, the maximal value of F_{st} between the admixed population and population 1 occurs either at the corner represented by $\gamma_2=1$ or at the corner represented by $\gamma_3=1$, as established in the theorem.

4. Special case: two founding populations

We now consider the special case in which only two founding populations give rise to an admixed population. We can simplify the notation of the previous section, so that $\gamma=\gamma_1$, $1-\gamma=\gamma_2$, $\tau_i=p_{1i}+p_{2i}$, and $\delta_i=p_{1i}-p_{2i}$. Using Eq. (1), the F_{st} value between population 1 and the admixed population can be written as

$$\begin{aligned}F_{st}[\underline{p_1},\gamma\underline{p_1}+(1-\gamma)\underline{p_2}]&=\frac{\sum_{i=1}^{I}[p_{1i}-\gamma p_{1i}-(1-\gamma)p_{2i}]^2}{4-\sum_{i=1}^{I}[p_{1i}+\gamma p_{1i}+(1-\gamma)p_{2i}]^2}\\&=\frac{(1-\gamma)^2\sum_{i=1}^{I}\delta_i^2}{4-\sum_{i=1}^{I}(\tau_i+\gamma\delta_i)^2}.\end{aligned}\qquad(7)$$

Thus, in this case, F_{st} is a function only of the admixture fraction, γ, and the sums and differences of the allele frequencies of the two founding populations. The $K=2$ case has fewer parameters than the general case of arbitrary K, and it is therefore possible to more precisely examine the properties of F_{st} as a function of the single admixture coefficient γ.

We first note that it was shown in the proof of Theorem 2 that the maximal F_{st} between population 1 and the admixed population is obtained when the admixed population is in fact population 2, so that $\gamma=0$:

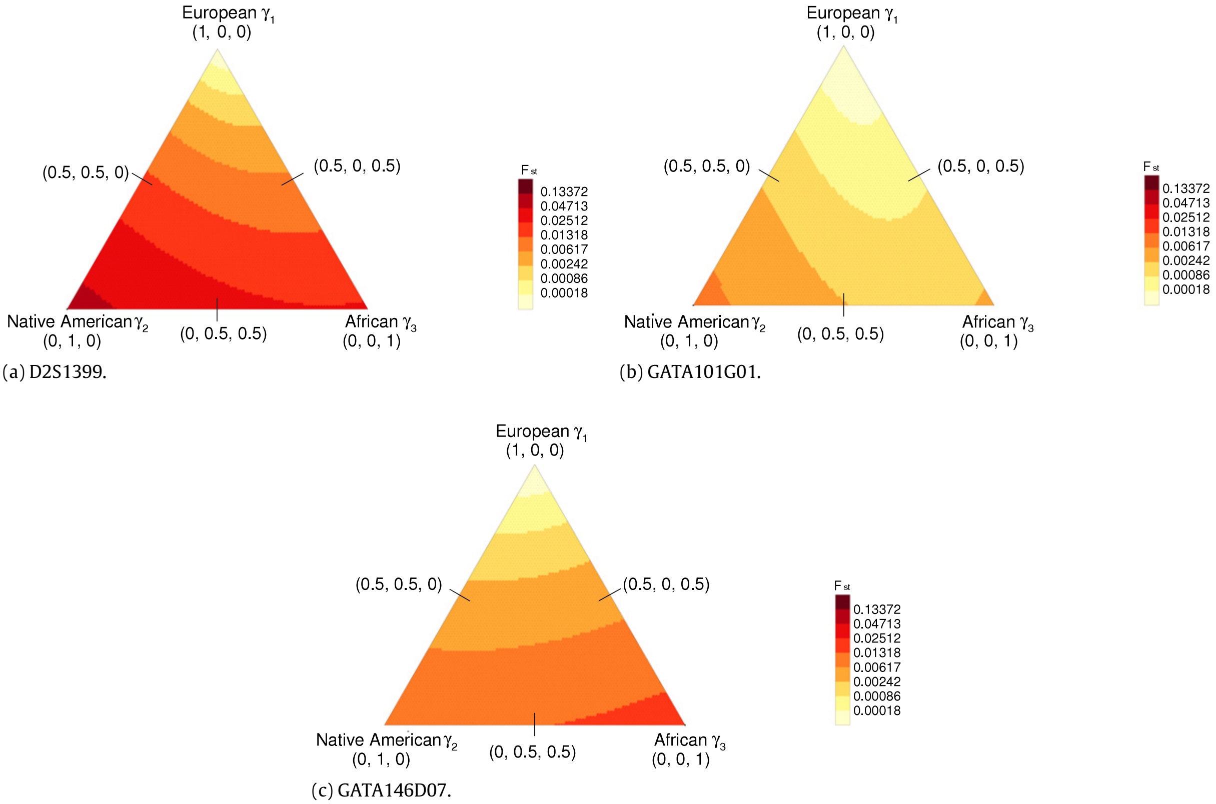
$$\max_{\gamma\in[0,1]}F_{st}[\underline{p_1},\gamma\underline{p_1}+(1-\gamma)\underline{p_2}]=F_{st}(\underline{p_1},\underline{p_2}).$$

4.1. F_{st} is monotonic and convex in the admixture coefficient

For the case with two founding populations, it is of interest to determine whether F_{st} behaves in a predictable way as a function of γ. We now show that F_{st} between a founding population and the admixed population is monotonic in the admixture coefficient.

Theorem 3. *As a function of γ, $F_{st}[\underline{p_1},\gamma\underline{p_1}+(1-\gamma)\underline{p_2}]$ is decreasing for $\gamma\in[0,1]$.*

Proof. Let $\alpha=1-\gamma$. We can use portions of the proof of Lemma 1, with $\underline{p_1}$ in the role of $\underline{p_3}$. Following the proof of Lemma 1, for $F_{st}[\underline{p_1},\alpha\underline{p_2}+(1-\alpha)\underline{p_1}]$, $E(\alpha)=a\alpha^2+b\alpha+c$. Plugging in $\alpha=0$,



(a) D2S1399.

(b) GATA101G01.

(c) GATA146D07.

Fig. 2. F_{st} between a population and hypothetical admixtures of that population with two other populations. $F_{st}[\underline{p_1}, \gamma_1\underline{p_1} + \gamma_2\underline{p_2} + \gamma_3\underline{p_3}]$, seen as a function of $\underline{\gamma}$, is plotted for all possible values of $\underline{\gamma}$ (Eq. (2)). Populations 1–3 represent populations of European, Native American, and African descent, respectively, and $\underline{p_1}$, $\underline{p_2}$, and $\underline{p_3}$ are based on allele frequencies estimated from Wang et al. (2008). Three loci are considered. In each triangle, the admixture fractions γ_1, γ_2, and γ_3 vary along the three axes, and darker colors correspond to higher values of F_{st}. In accordance with Theorem 2, considering all possible $\underline{\gamma}$, the maximal value of F_{st} is always at $\gamma_2 = 1$ or $\gamma_3 = 1$. (a) Locus D2S1399. (b) Locus GATA101G01. (c) Locus GATA146D07.

$E(\alpha) = c$. Noting in Eq. (4) that $\delta_{13i} = 0$ when $\underline{p_3} = \underline{p_1}$, we get $E(0) = 0$. We then have $E(1) > E(0) \geq 0$, so that $G(\alpha)$ is increasing on $\alpha \in [0, 1]$. As a result, $G(1-\alpha) = G(\gamma) = F_{st}[\underline{p_1}, \gamma\underline{p_1} + (1-\gamma)\underline{p_2}]$ is decreasing in γ on $\gamma \in [0, 1]$. □

The theorem supports the intuitive perspective that increasing the admixture fraction from source population 2 increases the genetic divergence of an admixed population from source population 1. We can in fact prove a stronger result. Not only is $F_{st}[\underline{p_1}, \gamma\underline{p_1} + (1-\gamma)\underline{p_2}]$ monotonic in γ, we can also show that it is convex as a function of the admixture fraction.

Theorem 4. *As a function of* γ, $F_{st}[\underline{p_1}, \gamma\underline{p_1} + (1-\gamma)\underline{p_2}]$ *is convex for* $\gamma \in [0, 1]$.

Proof. It suffices to demonstrate that the second derivative of $F_{st}[\underline{p_1}, \gamma\underline{p_1} + (1-\gamma)\underline{p_2}]$ as a function of γ is nonnegative. We insert $\underline{p_1}$ and $\underline{p_2}$ in place of $\underline{p_2}$ and $\underline{p_3}$, respectively, in the proof of Lemma 1. Thus, we aim to show that $\frac{d^2G(\gamma)}{d\gamma^2} > 0$. We first obtain the equation in Box II. To verify that the second derivative of $G(\gamma)$ is nonnegative, we need only show that $\frac{1}{2\sum_{i=1}^{I}\delta_i^2}\frac{d^2G(\gamma)}{d\gamma^2}$ is nonnegative. Consider the following expressions:

$$E_1(\gamma) = 4 + \sum_{i=1}^{I}\delta_i^2 - \sum_{i=1}^{I}\tau_i^2 - 2\gamma\sum_{i=1}^{I}\tau_i\delta_i - 2\gamma\sum_{i=1}^{I}\delta_i^2$$

$$E_2(\gamma) = 4 - \sum_{i=1}^{I}(\tau_i + \gamma\delta_i)^2$$

$$E_3(\gamma) = 4 - \sum_{i=1}^{I}\tau_i^2 - \sum_{i=1}^{I}\tau_i\delta_i - \gamma\sum_{i=1}^{I}\tau_i\delta_i - \gamma\sum_{i=1}^{I}\delta_i^2$$

$$E_4(\gamma) = \sum_{i=1}^{I}\delta_i(\tau_i + \gamma\delta_i).$$

We note some relationships between these expressions:

$$E_1(\gamma) = E_2(\gamma) + (1-\gamma)^2\sum_{i=1}^{I}\delta_i^2 \qquad (8)$$

$$E_3(\gamma) = E_2(\gamma) - (1-\gamma)E_4(\gamma). \qquad (9)$$

Using Eq. (9),

$$\frac{1}{2\sum\limits_{i=1}^{I}\delta_i^2}\frac{d}{d\gamma}G(\gamma) = \frac{-(1-\gamma)E_3(\gamma)}{E_2^2(\gamma)}. \qquad (10)$$

Differentiating the individual expressions and then combining them,

$$\begin{aligned}\frac{d}{d\gamma}[-E_3(\gamma)] &= \frac{d}{d\gamma}\left\{-\left[4 - \sum_{i=1}^{I}(\tau_i + \gamma\delta_i)^2\right.\right.\\ &\qquad \left.\left. + (1-\gamma)\sum_{i=1}^{I}\delta_i(\tau_i + \gamma\delta_i)\right\}\right.\\ &= \sum_{i=1}^{I}\tau_i\delta_i + \sum_{i=1}^{I}\delta_i^2\end{aligned}$$

$$\frac{d}{d\gamma}[-(1-\gamma)E_3(\gamma)] = \sum_{i=1}^{I}\tau_i\delta_i + \sum_{i=1}^{I}\delta_i^2$$

$$\frac{d}{d\gamma}G(\gamma) = \frac{-2\left[\sum_{i=1}^{I}\delta_i(\delta_i-\gamma\delta_i)\right]\left[4-\sum_{i=1}^{I}(\tau_i+\gamma\delta_i)^2\right]+\left[\sum_{i=1}^{I}(\delta_i-\gamma\delta_i)^2\right]\left[2\sum_{i=1}^{I}\delta_i(\tau_i+\gamma\delta_i)\right]}{\left[4-\sum_{i=1}^{I}(\tau_i+\gamma\delta_i)^2\right]^2}$$

$$= 2(1-\gamma)\left(\sum_{i=1}^{I}\delta_i^2\right)\left\{\frac{-\left[4-\sum_{i=1}^{I}(\tau_i+\gamma\delta_i)^2\right]+(1-\gamma)\sum_{i=1}^{I}\delta_i(\tau_i+\gamma\delta_i)}{\left[4-\sum_{i=1}^{I}(\tau_i+\gamma\delta_i)^2\right]^2}\right\}.$$

Box II.

$$\begin{aligned}
&-\gamma\sum_{i=1}^{I}\tau_i\delta_i-\gamma\sum_{i=1}^{I}\delta_i^2+4-\sum_{i=1}^{I}\tau_i^2-\sum_{i=1}^{I}\tau_i\delta_i\\
&-\gamma\sum_{i=1}^{I}\tau_i\delta_i-\gamma\sum_{i=1}^{I}\delta_i^2\\
&=4+\sum_{i=1}^{I}\delta_i^2-\sum_{i=1}^{I}\tau_i^2-2\gamma\sum_{i=1}^{I}\tau_i\delta_i-2\gamma\sum_{i=1}^{I}\delta_i^2\\
&=E_1(\gamma)
\end{aligned}$$

$$\begin{aligned}
\frac{d}{d\gamma}[E_2^2(\gamma)] &= \frac{d}{d\gamma}\left[4-\sum_{i=1}^{I}(\tau_i+\gamma\delta_i)^2\right]^2\\
&= -4\left[4-\sum_{i=1}^{I}(\tau_i+\gamma\delta_i)^2\right]\left[\sum_{i=1}^{I}\delta_i(\tau_i+\gamma\delta_i)\right]\\
&= -4E_2(\gamma)E_4(\gamma).
\end{aligned}$$

We now differentiate Eq. (10) and use the expressions above, obtaining:

$$\begin{aligned}
&\frac{1}{2\sum_{i=1}^{I}\delta_i^2}\frac{d^2G(\gamma)}{d\gamma^2}\geq 0\\
&\Leftrightarrow\frac{d}{d\gamma}\left[\frac{-(1-\gamma)E_3(\gamma)}{E_2^2(\gamma)}\right]\geq 0\\
&\Leftrightarrow E_1(\gamma)E_2^2(\gamma)-4(1-\gamma)E_3(\gamma)E_2(\gamma)E_4(\gamma)\geq 0\\
&\Leftrightarrow E_1(\gamma)E_2(\gamma)-4(1-\gamma)E_3(\gamma)E_4(\gamma)\geq 0.
\end{aligned}$$

The last step follows because $E_2(\gamma) \geq 0$, as $E_2(\gamma)$ corresponds to four times the (nonnegative) heterozygosity of the pooled population consisting of population 1 and the admixed population with allele frequency vector $\gamma\underline{p_1}+(1-\gamma)\underline{p_2}$. By applying Eqs. (8) and (9) and simplifying, we obtain:

$$\begin{aligned}
&E_1(\gamma)E_2(\gamma)-4(1-\gamma)E_3(\gamma)E_4(\gamma)\\
&=E_2^2(\gamma)+\left[(1-\gamma)^2\sum_{i=1}^{I}\delta_i^2\right]E_2(\gamma)\\
&\quad-4(1-\gamma)E_2(\gamma)E_4(\gamma)+4(1-\gamma)^2E_4^2(\gamma)\\
&=[E_2(\gamma)-2(1-\gamma)E_4(\gamma)]^2+(1-\gamma)^2\sum_{i=1}^{I}\delta_i^2E_2(\gamma)\geq 0
\end{aligned}$$

because both terms are nonnegative. □

An illustration of Theorems 3 and 4 appears in Fig. 3. The same twenty loci from Fig. 1 are used; populations 1 and 2 are the European and Native American populations, respectively. For each locus, the F_{st} value between population 1 and a population formed by the admixture of populations 1 and 2 can be seen to be decreasing and convex in $\gamma \in [0, 1]$, where γ is the admixture fraction for population 1.

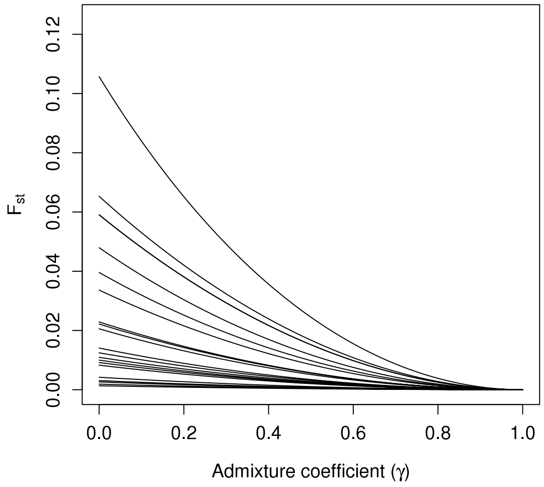


Fig. 3. F_{st} between a population and a hypothetical admixture of that population with a second population. $F_{st}[\underline{p_1}, \gamma\underline{p_1}+(1-\gamma)\underline{p_2}]$, seen as a function of γ, is plotted against γ (Eq. (7)). Populations 1 and 2 represent populations of European and Native American descent, respectively, and $\underline{p_1}$ and $\underline{p_2}$ are based on allele frequencies estimated from Wang et al. (2008). The same twenty randomly selected loci as in Fig. 1 are considered, with each curve representing a different locus. In accordance with Theorems 3 and 4, F_{st} is always decreasing and convex in γ.

4.2. Comparison of predicted F_{st} to observed F_{st}

When allele frequencies are available on both the admixed population and the founding populations, we are able to calculate the *observed* F_{st} value between a specific founding population and a population formed by admixture of multiple founding populations. For example, it is possible to calculate the observed F_{st} between an African American population and a putative African founding population, or between a Mestizo population and a putative European founding population. In practice, the true founding populations are not precisely known, no longer exist, or may not have data available, so that in general, only an approximation is possible.

In such cases, our results provide a way of *predicting* the F_{st} value between a population formed by an admixture of multiple founding populations and a specific founding population, on the basis of measured allele frequencies and admixture coefficients. The predicted F_{st} value can be calculated when the allele frequencies and the admixture coefficients are available or can be estimated for the founding populations. Estimation of the admixture fractions at a given locus for the various founding populations can be achieved via maximum likelihood (Millar, 1987) or other techniques.

For the Wang et al. (2008) data, we estimated the fraction of European ancestry in the Mestizo population at each of the 678 loci, treating the European and Native American populations as founding populations. This approach followed the procedure

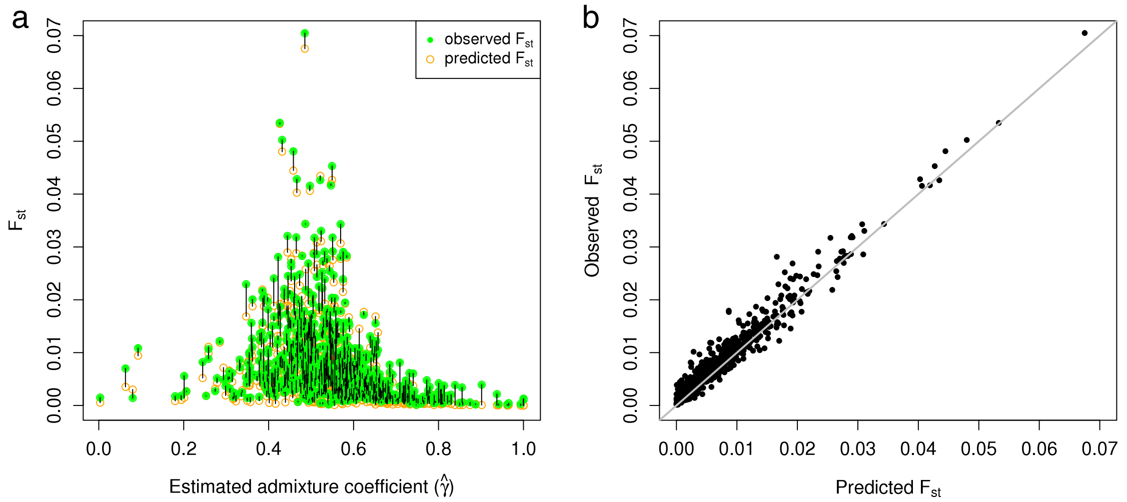


Fig. 4. Predicted and observed F_{st}. (a) The predicted and observed F_{st} values between an admixed Mestizo population and a European founding population are plotted against the European admixture fraction γ in the Mestizo population, estimated by maximum likelihood. The prediction is based on Eq. (7), using the European and Native American allele frequencies estimated from Wang et al. (2008) as $\underline{p_1}$ and $\underline{p_2}$, respectively, together with the maximum likelihood estimate of γ. The observation is based on F_{st} estimated from Eq. (1), inserting estimated allele frequencies from Wang et al. (2008) on European and Mestizo populations. (b) The observed F_{st} value is plotted against the predicted F_{st} value. The identity line is shown in gray. In both panels, each point represents one of the 678 loci used. The correlation coefficient between the predicted and observed F_{st} values is 0.978.

of Schroeder et al. (2009), with all of the various subgroups in the Mestizo sample of Wang et al. (2008) pooled together (indeed the admixture estimates are the same as those used in the "Combined admixed sample" analysis in Table 1 of Schroeder et al. (2009)). Following Schroeder et al. (2009), for any allele present in at least one individual in the Mestizo population but not present in both founding populations, and for each founding population that did not possess the allele, a single copy of the allele was artificially added to that ancestral population. Sample allele frequencies that were then obtained for Europeans and Native Americans were treated as true allele frequencies for use in the maximum likelihood inference of the European admixture proportion, assuming Hardy–Weinberg equilibrium in the admixed population. Maximum likelihood estimates were obtained numerically and were used to obtain the predicted F_{st} according to Eq. (7).

The observed and predicted F_{st} values for individual loci are compared in Fig. 4. In general, we find that the observation closely matches the prediction. In most cases (549 of 678 loci), however, the prediction provides an underestimate of the observed value. This systematic underestimation might arise from the use of estimated rather than true values to obtain the prediction; in particular, the prediction relies on both the estimated allele frequencies and the maximum likelihood estimate of γ obtained from the same data used to estimate the allele frequencies.

4.3. An admixture estimator on the basis of F_{st}

As an alternative to use of an estimated admixture coefficient to predict F_{st}, an observed F_{st} value between a population formed by admixture of two founding populations and a specific founding population can be used as a way of estimating the admixture fraction. In the case of two founding populations, the quadratic equation in Eq. (7) can be solved to provide an estimator in the style of the method of moments. This approach is reasonable, as the monotonicity result in Theorem 3 indicates that for fixed allele frequencies, γ is identifiable from F_{st}. The resulting estimator is non-parametric, in that it does not make assumptions on the form of the probability distribution of the allele frequencies at a given locus (see Box III). It can be shown that $\hat{\gamma}_+ \geq 1$ for all possible values of F_{st} and the δ_i and τ_i. Therefore, if $\hat{\gamma}_-$ is between 0 and 1, it is chosen as the estimate. In the case in which $\hat{\gamma}_- < 0$, 0 is chosen as the estimate, and if $\hat{\gamma}_- > 1$, 1 is chosen as the estimate.

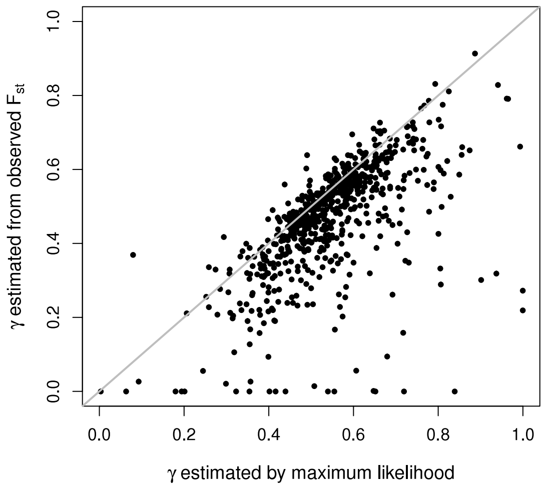


Fig. 5. Admixture estimates obtained from observed F_{st} (Eq. (11)) versus estimates obtained by maximum likelihood. The plot represents a scenario in which a European and a Native American population are the founding populations and a Mestizo population is the admixed population, and allele frequencies estimated from Wang et al. (2008) on all three populations are used to estimate γ, the European admixture fraction in the Mestizo population. The identity line is shown in gray, and each point represents one of the 678 loci used. The correlation coefficient between the two sets of estimates is 0.618.

For the example data, Fig. 5 presents a plot of the estimate of γ from the observed F_{st} versus the maximum likelihood estimate of γ, with Europeans and Native Americans as populations 1 and 2, and with Mestizos as the admixed population. The correlation between the estimates from these two methods is 0.618, and in general, the moment estimator produces smaller estimates than the maximum likelihood method, including several estimates of zero in cases where maximum likelihood obtains a positive value.

5. Discussion

In this paper, we have considered the F_{st} measure in the context of admixed populations. We have explored the F_{st} value

$$\hat{\gamma}_{\pm} = \frac{\left(\sum_{i=1}^{I}\delta_i^2 - F_{st}\sum_{i=1}^{I}\delta_i\tau_i\right) \pm \sqrt{\left(\sum_{i=1}^{I}\delta_i^2 - F_{st}\sum_{i=1}^{I}\delta_i\tau_i\right)^2 - (1+F_{st})\left(\sum_{i=1}^{I}\delta_i^2\right)\left(\sum_{i=1}^{I}\delta_i^2 + F_{st}\sum_{i=1}^{I}\tau_i^2 - 4F_{st}\right)}}{(1+F_{st})\sum_{i=1}^{I}\delta_i^2}. \quad (11)$$

Box III.

between a population formed by the admixture of K founding populations and one of those founding populations. In the general case of arbitrary $K \geq 2$, we have demonstrated that this value is maximized when the admixed population is in fact one of the other founding populations. In the particular case of $K = 2$, this F_{st} value is monotonic and convex in the admixture fraction. We have also provided a formula for predicting F_{st} in an admixed population on the basis of the estimated admixture coefficient and the allele frequencies in the founding populations, producing very similar values to those observed in an empirical example utilizing the data of Wang et al. (2008).

Further, we discussed a non-parametric method of estimating the admixture fraction from the observed F_{st} values, and we compared it to the maximum likelihood method. In general, the non-parametric estimator is useful primarily for the purpose of illustrating the close relationship between F_{st} and the admixture coefficient, and its statistical properties are likely to be poorer than those of modern genome-based approaches to admixture estimation (Falush et al., 2003; Hoggart et al., 2004; Tang et al., 2006a; Sankararaman et al., 2008; Alexander et al., 2009; Price et al., 2009; Engelhardt and Stephens, 2010). However, as a straightforward formula that is calculated from quantities that are easily obtained, it provides a convenient approach when a computationally simple initial estimate is desirable.

Our results can provide a basis for interpreting F_{st} in admixed populations. In particular, in the case in which $K = 2$, the monotonicity and convexity of F_{st} in the admixture coefficient imply that F_{st} is informative about the level of admixture, and vice versa. This relationship can be a useful starting point for measurement of admixture, and a comparison of observed and predicted F_{st} values can be used as an initial check on the extent to which estimates of the admixture fraction obtained by maximum likelihood or other algorithms are sensible.

We note several limitations of our work. First, our admixture model does not involve a mechanistic evolutionary process, considering only the linear combination of allele frequencies that occurs when an admixed population is produced instantaneously from a set of source populations. Second, we examine admixture only at the population level, disregarding variation that might exist in admixture levels across individuals within a population. Third, as in many methods for analysis of admixed populations, we caution that our work presumes that the source populations for a given admixed population have been correctly specified. It is encouraging, however, that in spite of these concerns, the predicted F_{st} values generally agree with the values observed in our empirical example. As illustrated by the results presented here, further analysis of the properties of F_{st} in an admixture setting will continue to facilitate the understanding of population-genetic issues in the context of admixture research.

Acknowledgments

Support for this work was provided by NIH grants R01 GM081441 and T32 GM074906, NSF grant BCS-1024627, the Burroughs Wellcome Foundation, and the Johns Hopkins Sommer Scholar Program. We also thank two anonymous reviewers for comments that significantly improved the manuscript.

References

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19, 1655–1664.

Balding, D.J., 2003. Likelihood-based inference for genetic correlation coefficients. Theoretical Population Biology 63, 221–230.

Bonilla, C., Parra, E.J., Pfaff, C.L., Dios, S., Marshall, J.A., Hamman, R.F., Ferrell, R.E., Hoggart, C.L., McKeigue, P.M., Shriver, M.D., 2004. Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. Annals of Human Genetics 68, 139–153.

Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S.A., Bustamante, C.D., 2010a. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proceedings of the National Academy of Sciences of the United States of America 107, 786–791.

Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C.D., Ostrer, H., 2010b. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proceedings of the National Academy of Sciences of the United States of America 107, 8954–8961.

Buerkle, C.A., Lexer, C., 2008. Admixture as the basis for genetic mapping. Trends in Ecology and Evolution 23, 686–694.

Engelhardt, B.E., Stephens, M., 2010. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS Genetics 6, e1001117.

Excoffier, L., 2001. Analysis of population subdivision. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, pp. 271–307. (Chapter 10).

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

Hoggart, C.J., Shriver, M.D., Kittles, R.A., Clayton, D.G., McKeigue, P.M., 2004. Design and analysis of admixture mapping studies. American Journal of Human Genetics 74, 965–978.

Holsinger, K.E., Weir, B.S., 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nature Reviews Genetics 10, 639–650.

McKeigue, P.M., 2005. Prospects for admixture mapping of complex traits. American Journal of Human Genetics 76, 1–7.

Millar, R.B., 1987. Maximum likelihood estimation of mixed stock fishery composition. Canadian Journal of Fisheries and Aquatic Sciences 44, 583–590.

Nei, M., 1973. Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences of the United States of America 70, 3321–3323.

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Parra, E.J., Kittles, R.A., Argyropoulos, G., Pfaff, C.L., Hiester, K., Bonilla, C., Sylvester, N., Parrish-Gause, D., Garvey, W.T., Jin, L., McKeigue, P.M., Kamboh, M.I., Ferrell, R.E., Pollitzer, W.S., Shriver, M.D., 2001. Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. American Journal of Physical Anthropology 114, 18–29.

Parra, E.J., Marcini, A., Akey, J., Martinson, J., Batzer, M.A., Cooper, R., Forrester, T., Allison, D.B., Deka, R., Ferrell, R.E., Shriver, M.D., 1998. Estimating African American admixture proportions by use of population-specific alleles. American Journal of Human Genetics 63, 1839–1851.

Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., Myers, S., 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genetics 5, e1000519.

Reich, D., Patterson, N., 2005. Will admixture mapping work to find disease genes? Philophical Transactions of the Royal Society of London B—Biological Sciences 360, 1605–1607.

Risch, N., Choudhry, S., Via, M., Basu, A., Sebro, R., Eng, C., Beckman, K., Thyne, S., Chapela, R., Rodriguez-Santana, J.R., Rodriguez-Cintron, W., Avila, P.C., Ziv, E., Burchard, E.G., 2009. Ancestry-related assortative mating in Latino populations. Genome Biology 10, R132.

Rousset, F., 2001. Inferences from spatial population genetics. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, pp. 239–269. (Chapter 9).

Salas, A., Carracedo, A., Richards, M., Macaulay, V., 2005. Charting the ancestry of African Americans. American Journal of Human Genetics 77, 676–680.

Sankararaman, S., Sridhar, S., Kimmel, G., Halperin, E., 2008. Estimating local ancestry in admixed populations. American Journal of Human Genetics 82, 290–303.

Schroeder, K.B., Jakobsson, M., Crawford, M.H., Schurr, T.G., Boca, S.M., Conrad, D.F., Tito, R.Y., Osipova, L.P., Tarskaia, L.A., Zhadanov, S.I., Wall, J.D., Pritchard, J.K., Malhi, R.S., Smith, D.G., Rosenberg, N.A., 2009. Haplotypic background of a private allele at high frequency in the Americas. Molecular Biology and Evolution 26, 995–1016.

Seldin, M.F., 2007. Admixture mapping as a tool in gene discovery. Current Opinion in Genetics & Development 17, 177–181.

Seldin, M.F., Tian, C., Shigeta, R., Scherbarth, H.R., Silva, G., Belmont, J.W., Kittles, R., Gamron, S., Allevi, A., Palatnik, S.A., Alvarellos, A., Paira, S., Caprarulo, C., Guillerón, C., Catoggio, L.J., Prigione, C., Berbotto, G.A., García, M.A., Perandones, C.E., Pons-Estel, B.A., Alarcon-Riquelme, M.E., 2007. Argentine population genetic structure: large variance in Amerindian contribution. American Journal of Physical Anthropology 132, 455–462.

Silva-Zolezzi, I., Hidalgo-Miranda, A., Estrada-Gil, J., Fernandez-Lopez, J.C., Uribe-Figueroa, L., Contreras, A., Balam-Ortiz, E., del Bosque-Plata, L., Velazquez-Fernandez, D., Lara, C., Goya, R., Hernandez-Lemus, E., Davila, C., Barrientos, E., March, S., Jimenez-Sanchez, G., 2009. Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. Proceedings of the National Academy of Sciences of the United States of America 106, 8611–8616.

Smith, M.W., O'Brien, S.J., 2005. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nature Reviews Genetics 6, 623–632.

Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E.G., Risch, N.J., 2007. Recent genetic selection in the ancestral admixture of Puerto Ricans. American Journal of Human Genetics 81, 626–633.

Tang, H., Coram, M., Wang, P., Zhu, X., Risch, N., 2006a. Reconstructing genetic ancestry blocks in admixed individuals. American Journal of Human Genetics 79, 1–12.

Tang, H., Jorgenson, E., Gadde, M., Kardia, S.L.R., Rao, D.C., Zhu, X., Schork, N.J., Hanis, C.L., Risch, N., 2006b. Racial admixture and its impact on BMI and blood pressure in African and Mexican Americans. Human Genetics 119, 624–633.

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A.T., Kotze, M.J., Lema, G., Moore, J.H., Mortensen, H., Nyambo, T.B., Omar, S.A., Powell, K., Pretorius, G.S., Smith, M.W., Thera, M.A., Wambebe, C., Weber, J.L., Williams, S.M., 2009. The genetic structure and history of Africans and African Americans. Science 324, 1035–1044.

Wang, S., Ray, N., Rojas, W., Parra, M.V., Bedoya, G., Gallo, C., Poletti, G., Mazzotti, G., Hill, K., Hurtado, A.M., Camrena, B., Nicolini, H., Klitz, W., Barrantes, R., Molina, J.A., Freimer, N.B., Bortolini, M.C., Salzano, F.M., Petzl-Erler, M.L., Tsuneto, L.T., Dipierri, J.E., Alfaro, E.L., Bailliet, G., Bianchi, N.O., Llop, E., Rothhammer, F., Excoffier, L., Ruiz-Linares, A., 2008. Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genetics 4, e1000037.

Winkler, C.A., Nelson, G.W., Smith, M.W., 2010. Admixture mapping comes of age. Annual Review of Genomics and Human Genetics 11, 65–89.

Wright, S., 1951. The genetical structure of populations. Annals of Eugenics 15, 323–354.

Zakharia, F., Basu, A., Absher, D., Assimes, T.L., Go, A.S., Hlatky, M.A., Iribarren, C., Knowles, J.W., Li, J., Narasimhan, B., Sidney, S., Southwick, A., Myers, R.M., Quertermous, T., Risch, N., Tang, H., 2009. Characterizing the admixed African ancestry of African Americans. Genome Biology 10, R141.

Zhu, X., Tang, H., Risch, N., 2008. Admixture mapping and the role of population structure for localizing disease genes. Advances in Genetics 60, 547–569.