F_{ST} AND THE TRIANGLE INEQUALITY FOR BIALLELIC MARKERS

Ilana M. Arbisser[1*], Noah A. Rosenberg[1]

[1]Department of Biology, Stanford University, Stanford, CA 94305 USA

*Correspondence. ilanama@stanford.edu.

November 16, 2018

Keywords: F_{ST}, genetic distance, inequalities, population structure

Abstract

The population differentiation statistic F_{ST}, introduced by Sewall Wright, is often treated as a pairwise distance measure between populations. As was known to Wright, however, F_{ST} is not a true metric because allele frequencies exist for which it does not satisfy the triangle inequality. We prove that a stronger result holds: for biallelic markers whose allele frequencies differ across three populations, F_{ST} *never* satisfies the triangle inequality. We study the deviation from the triangle inequality as a function of the allele frequencies of three populations, identifying frequency vectors at which the deviation is maximal. We also examine the implications of the failure of the triangle inequality for the four-point condition for groups of four populations. Next, we examine the extent to which F_{ST} fails to satisfy the triangle inequality in genome-wide data from human populations, finding that some loci have frequencies that produce deviations near the maximum. We discuss the consequences of the theoretical results for various types of data analysis, including multidimensional scaling and inference of neighbor-joining trees from pairwise F_{ST} matrices.

1. Introduction

Introduced by Wright (1951), F_{ST}, which provides a measure of population structure for population-genetic data, is one of the most commonly used statistics in population genetics (Holsinger and Weir 2009). Pairwise F_{ST} computed between two populations is often viewed as a measure of genetic "distance" between the populations (e.g. Jorde 1985, Rosenberg et al. 2005). Indeed, F_{ST} is frequently treated as a distance in many types of analysis for representing relationships between multiple populations, such as in the distance matrices used for spatially depicting genetic variation and in inference of population trees (e.g. Pérez-Lezaun et al. 1997, Li et al. 2008).

In the formulation of Nei (1973), F_{ST} can be written

$$F_{ST} = \frac{J_S - J_T}{1 - J_T}, \tag{1}$$

where J_S is the mean homozygosity across a set of subpopulations and J_T is the homozygosity of a population formed by pooling the subpopulations, assuming they have equal representation.

For the case of two subpopulations, using p_{ki} for the frequency of allele i in subpopulation k, $J_S = \frac{1}{2}[(\sum_{i=1}^{I} p_{1i}^2) + (\sum_{i=1}^{I} p_{2i}^2)]$ and $J_T = \sum_{i=1}^{I} (\frac{p_{1i}+p_{2i}}{2})^2$, where I is the total number of alleles at a locus of interest. Hence, eq. (1) reduces in this case to

$$F_{ST} = \frac{J_1 + J_2 - 2D_{12}}{4 - J_1 - J_2 - 2D_{12}}, \tag{2}$$

where $J_1 = \sum_{i=1}^{I} p_{1i}^2$, $J_2 = \sum_{i=1}^{I} p_{2i}^2$, and $D_{12} = \sum_{i=1}^{I} p_{1i}p_{2i}$.

F_{ST}, eq. (2), has some of the properties required by a mathematical measure of distance: it is symmetric with respect to a change in the population labels, it is nonnegative, and it is equal to 0 if and only if two populations have the same allele frequencies ($p_{1i} = p_{2i}$ for all i). Yet, F_{ST} is not a true distance metric because it does not satisfy the triangle inequality: with three populations, the sum of two of the distances can be smaller than the third distance. In fact, Sewall Wright (1978, p. 89) was aware of this fact, offering a counterexample of three populations whose allele frequencies result in values of F_{ST} that do not satisfy the inequality: a biallelic locus that is monomorphic for one allele in population 1 and monomorphic for the other allele in population 2, and has equal frequencies for the two alleles in population 3.

Here, we generalize beyond Wright's counterexample to show that not only is it possible for F_{ST} to violate the triangle inequality, for a biallelic locus with distinct allele frequencies in three populations, F_{ST} *never* satisfies the triangle inequality. We explore the extent to which F_{ST} fails to satisfy the triangle inequality over the space of possible allele frequencies, finding that the maximal deviation from the condition specified by the triangle inequality occurs precisely in Wright's counterexample. We also show that the failure to satisfy the triangle inequality has as a consequence a failure of the four-point condition associated with construction of evolutionary trees. To consider the context of our theoretical results in data analysis, we examine the extent to which F_{ST} fails to satisfy the triangle inequality in data from three human populations. We also examine the impact of the mathematical results in multidimensional scaling analysis and on inference of population trees by neighbor-joining.

2. The triangle inequality never holds for biallelic markers with distinct allele frequencies

We consider a biallelic locus in three populations. We choose one allele and label its frequencies in populations 1, 2, and 3, by p_1, p_2, and p_3, respectively. Without loss of generality, we assume $0 \leqslant p_1 \leqslant p_2 \leqslant p_3 \leqslant 1$. We can define $F(p_i, p_j)$ as the value of F_{ST} measured between two populations i and j, in which the frequencies of the chosen allele are p_i and p_j, respectively.

Simplifying the expression for F_{ST} from eq. (2) by noting that the other allele of the biallelic locus has frequency $q_i = 1 - p_i$ and $q_j = 1 - p_j$ in populations i and j, respectively, $F_{ST}(p_i, p_j)$, or F_{ij} for short, can be written

$$F_{ij} = \frac{\frac{1}{4}(p_i - p_j)^2}{(\frac{p_i+p_j}{2})(1 - \frac{p_i+p_j}{2})} = \frac{(p_i - p_j)^2}{(p_i + p_j)(2 - p_i - p_j)}. \tag{3}$$

At $p_i = p_j = 0$ and at $p_i = p_j = 1$, we define F_{ij} to be 0. We disregard the cases of $p_1 = p_2 = p_3 = 0$ and $p_1 = p_2 = p_3 = 1$, as these cases do not represent polymorphic loci.

The triangle inequality holds for F_{ST} in three populations if and only if all three of the following inequalities hold:

$$F_{12} + F_{13} \geqslant F_{23} \tag{4}$$

$$F_{13} + F_{23} \geqslant F_{12} \tag{5}$$

$$F_{12} + F_{23} \geqslant F_{13}. \tag{6}$$

We show that eqs. (4) and (5) always hold, as these statements place the largest of the three F_{ST} values, F_{13}, on the larger side of the inequality. We also show that when $p_1 = p_2 \leqslant p_3$ or $p_1 \leqslant p_2 = p_3$, eq. (6) also holds, so that the triangle inequality is satisfied. However, we show that when $p_1 < p_2 < p_3$, the triangle inequality fails: while eqs. (4) and (5) do hold, eq. (6) does not.

2.1. **At least two of the three of the inequalities always hold.** To show that eqs. (4) and (5) hold, it suffices to show that

$$F_{12} \leqslant F_{13} \tag{7}$$

$$F_{23} \leqslant F_{13}. \tag{8}$$

We need only show eq. (7) to also prove eq. (8). In particular, because $p_i = 1 - q_i$ and $p_1 \leqslant p_2 \leqslant p_3$, $q_3 \leqslant q_2 \leqslant q_1$ and $F_{ST}(p_i, p_j) = F_{ST}(q_i, q_j)$. Hence, by switching the population labels for populations 1 and 3 in eq. (8) and using frequencies q_i in place of p_i, eqs. (7) and (8) are equivalent.

To prove eqs. (4) and eq. (5), it remains to prove eq. (7). By eq. (3), we wish to show:

$$\frac{(p_2 - p_1)^2}{(p_1 + p_2)(2 - p_1 - p_2)} \leqslant \frac{(p_3 - p_1)^2}{(p_1 + p_3)(2 - p_1 - p_3)}.$$

For convenience, we define

$$\sigma_{12} = p_2 + p_1$$
$$\delta_{12} = p_2 - p_1$$
$$\delta_{23} = p_3 - p_2.$$

By these definitions,

$$0 \leqslant \delta_{12} \leqslant \min\{\sigma_{12}, 1\}$$
$$0 \leqslant \sigma_{12} \leqslant 2$$
$$0 \leqslant \delta_{23} \leqslant 1.$$

Therefore, we seek to show:

$$\frac{\delta_{12}^2}{\sigma_{12}(2 - \sigma_{12})} \leqslant \frac{(\delta_{12} + \delta_{23})^2}{(\sigma_{12} + \delta_{23})(2 - \sigma_{12} - \delta_{23})}. \tag{9}$$

For the cases in which $\sigma_{12} = 0$ or $\sigma_{12} = 2$, we have $p_1 = p_2 = 0$ and $p_1 = p_2 = 1$, respectively, which we defined in eq. (3) to have $F_{12} = 0$. If $F_{12} = 0$, then $F_{12} \leqslant F_{13}$ trivially because $0 \leqslant F_{13}$. Hence, noting that $\delta_{12} = 0$ if $p_1 = p_2$, inequality (9) always holds for $p_1 = p_2 = p_3$.

Similarly, if $\sigma_{12} + \delta_{23} = 0$, then $p_1 = p_3 = 0$, and if $\sigma_{12} + \delta_{23} = 2$, then $p_1 = p_3 = 1$. It follows that $p_1 = p_3 = p_2$, so the locus is not polymorphic. Consequently, the cases of $\sigma_{12} + \delta_{23} = 0$ and $\sigma_{12} + \delta_{23} = 0$ are excluded by our assumption that the locus is polymorphic.

Having dealt with the cases in which denominators in eq. (9) are zero, we rearrange eq. (9) and find that eq. (9) holds if

$$(\delta_{12} + \delta_{23})^2(\sigma_{12})(2 - \sigma_{12}) - \delta_{12}^2(\sigma_{12} + \delta_{23})(2 - \sigma_{12} - \delta_{23}) \geqslant 0. \tag{10}$$

Inequality (10) is equivalent to:

$$\delta_{23}\Big[2\delta_{12}[\sigma_{12}(2 - \sigma_{12}) - \delta_{12}(1 - \sigma_{12})] + \delta_{23}\sigma_{12}(2 - \sigma_{12}) + \delta_{12}^2\delta_{23}\Big] \geqslant 0. \tag{11}$$

Because $\delta_{23} \geqslant 0$, δ_{23} does not change the sign of the expression in eq. (11). Hence, eq. (11) holds if $\delta_{23} = 0$, or if $2\delta_{12}[\sigma_{12}(2 - \sigma_{12}) - \delta_{12}(1 - \sigma_{12})] + \delta_{23}\sigma_{12}(2 - \sigma_{12}) + \delta_{12}^2\delta_{23} \geqslant 0$. The latter inequality always holds, as $\delta_{23}\sigma_{12}(2 - \sigma_{12}) \geqslant 0$, $\delta_{12}^2\delta_{23} \geqslant 0$, and $\sigma_{12}(2 - \sigma_{12}) - \delta_{12}(1 - \sigma_{12}) \geqslant 0$, noting that $\sigma_{12} \geqslant \delta_{12}$, $2 - \sigma_{12} \geqslant 0$, and $2 - \sigma_{12} > 1 - \sigma_{12}$. Therefore, eq. (11) holds, eq. (7) follows, and eqs. (4) and (5) are both true.

2.2. **If two of the three populations have identical frequencies, then eq. (6) always holds.** We show that eq. (6) always holds when $p_1 = p_2 \leqslant p_3$ or $p_1 \leqslant p_2 = p_3$. If $p_1 = p_2$, then $F_{23} = F_{13}$ and $F_{12} = 0$. Eq. (6) holds trivially as $0 + F_{23} \geqslant F_{23}$. Similarly, if $p_2 = p_3$, then $F_{12} = F_{13}$ and $F_{23} = 0$. Eq. (6) holds trivially, as $F_{12} + 0 \geqslant F_{12}$. Because eqs. (4)-(6) are all satisfied, the triangle inequality is satisifed for three populations if either $p_1 = p_2 \leqslant p_3$ or $p_1 \leqslant p_2 = p_3$.

Note that the triangle inequality is satisfied in the case that two of three points are the same for any function, f on some set X, which is symmetric and has the identity of indiscernibles, i.e. $f(x, x) = 0$ for all $x \in X$. With the added condition of nonnegativity, such a function is sometimes called a distance *function* or distance *measure* (to distinguish it from a true distance *metric* that also satisfies the triangle inequality).

2.3. **If the three populations have distinct frequencies, then eq. (6) never holds.** To show that eq. (6) does not hold when $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$, we consider a function

$$\psi(p_1, p_2, p_3) = F_{ST}(p_1, p_2) + F_{ST}(p_2, p_3) - F_{ST}(p_1, p_3). \tag{12}$$

Eq. (6) holds and hence the triangle inequality holds if and only if $\psi(p_1, p_2, p_3) \geqslant 0$. We proceed in several steps to show that $\psi(p_1, p_2, p_3) < 0$ when $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$.

(i) We write $\psi(p_1, p_2, p_3)$ as a fraction with positive denominator and relate $\psi(p_1, p_2, p_3)$ to its numerator $\omega(p_1, p_2, p_3)$. Because the denominator of $\psi(p_1, p_2, p_3)$ is always positive, $\psi(p_1, p_2, p_3) < 0$ if and only if $\omega(p_1, p_2, p_3) < 0$.

(ii) Next, we show that as a function of any one of its variables, $\omega(p_1, p_2, p_3)$ is a quartic function. Considering $\omega(p_1, p_2, p_3)$ as a function of p_2, two of its roots lie at $p_2 = p_1$ and $p_2 = p_3$. Therefore,

we can define a quadratic function in p_2, $\phi(p_1, p_2, p_3)$:

$$\omega(p_1, p_2, p_3) = (p_1 - p_2)(p_2 - p_3)\phi(p_1, p_2, p_3),$$

Because we assume $p_1 < p_2$ and $p_2 < p_3$, $\omega(p_1, p_2, p_3) < 0$ if and only if $\phi(p_1, p_2, p_3) < 0$.

(iii) We then consider the roots of $\phi(p_1, p_2, p_3)$ as functions of p_2, $r_1(p_1, p_3)$ and $r_2(p_1, p_3)$. We show that $r_1(p_1, p_3) \leqslant p_1$, $r_2(p_1, p_3) \geqslant p_3$, and $\phi(p_1, p_2, p_3) < 0$ for $p_2 \in (r_1(p_1, p_3), r_2(p_1, p_3))$.

(iv) We conclude that because $\phi(p_1, p_2, p_3) < 0$, $\omega(p_1, p_2, p_3) < 0$, which implies $\psi(p_1, p_2, p_3) < 0$. Hence when $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$, eq. (6) does not hold, and the triangle inequality does not hold for a biallelic marker with distinct allele frequencies in three populations.

2.3.1. *$\psi(p_1, p_2, p_3) < 0$ if and only if $\omega(p_1, p_2, p_3) < 0$ for $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$.* We simplify eq. (12) for $\psi(p_1, p_2, p_3)$ by using eq. (3).

Define $\omega(x, y, z)$ as:

$$\begin{aligned}\omega(x, y, z) = &(x - y)^2(x + z)(2 - x - z)(z + y)(2 - z - y)\\ &+ (z - y)^2(x + y)(2 - x - y)(x + z)(2 - x - z)\\ &- (z - x)^2(x + y)(2 - x - y)(z + y)(2 - z - y).\end{aligned}$$

If $x = p_1$, $y = p_2$, and $z = p_3$, then

$$\psi(p_1, p_2, p_3) = \frac{\omega(p_1, p_2, p_3)}{(p_1 + p_2)(2 - p_1 - p_2)(p_2 + p_3)(2 - p_2 - p_3)(p_1 + p_3)(2 - p_1 - p_3)}.$$

The denominator is always nonnegative when $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$, so $\psi(p_1, p_2, p_3) < 0$ if and only if $\omega(p_1, p_2, p_3) < 0$.

2.3.2. *$\omega(p_1, p_2, p_3) < 0$ if and only if $\phi(p_1, p_2, p_3) < 0$ for $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$.* Consider $\omega(x, y, z)$ with $x = p_1$ and $z = p_3$ fixed, so that $\omega(p_1, y, p_3)$ is only a function of y. Because ω is quartic in y, it has at most four distinct roots, each of which can be expressed as a function of p_1 and p_3.

It is trivial to show that $y = p_1$ and $y = p_3$ are both roots of $\omega(p_1, y, p_3)$. Consequently, $\omega(p_1, y, p_3)$ has at most two other roots for y. Define a quadratic function $\phi(p_1, y, p_3)$ such that:

$$\omega(p_1, y, p_3) = (p_1 - y)(y - p_3)\phi(p_1, y, p_3).$$

Performing polynomial division, we can write

$$\phi(p_1, y, p_3) = y^2\alpha(p_1, p_3) + y\beta(p_1, p_3) + \gamma(p_1, p_3),$$

where:

$$\alpha(p_1, p_3) = 4p_1 - p_1^2 + 4p_3 - 6p_1p_3 - p_3^2 \tag{13}$$

$$\beta(p_1, p_3) = -8p_1 + 4p_1^2 + p_1^3 - 8p_3 + 24p_1p_3 - 9p_1^2p_3 + 4p_3^2 - 9p_1p_3^2 + p_3^3 \tag{14}$$

$$\gamma(p_1, p_3) = -16p_1p_3 + 12p_1^2p_3 - p_1^3p_3 + 12p_1p_3^2 - 6p_1^2p_3^2 - p_1p_3^3. \tag{15}$$

For $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$, $p_1 - p_2 < 0$ and $p_2 - p_3 < 0$, so $\omega(p_1, p_2, p_3) < 0$ if and only if $\phi(p_1, p_2, p_3) < 0$.

2.3.3. $\phi(p_1, p_2, p_3) < 0$ *when* $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$. We know that $\phi(p_1, y, p_3)$ has at most two roots, $r_1(p_1, p_3)$ and $r_2(p_1, p_3)$. If these two roots are distinct, then $\phi(p_1, y, p_3) > 0$ or $\phi(p_1, y, p_3) < 0$ for values of y between the roots. Without loss of generality, assume $r_1(p_1, p_3) \leqslant r_2(p_1, p_3)$. To show that $\phi(p_1, p_2, p_3) < 0$ for $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$, we need to show all of the following:

(1) If $r_1(p_1, p_3) < y < r_2(p_1, p_3)$, then $\phi(p_1, y, p_3) < 0$,
(2) $r_1(p_1, p_3) \leqslant p_1$,
(3) $r_2(p_1, p_3) \geqslant p_3$.

Note that because $p_1 < p_3$, demonstrating (2) and (3) suffices to show that the two roots $r_1(p_1, p_3)$ and $r_2(p_1, p_3)$ are distinct.

2.3.3.1. *If* $r_1(p_1, p_3) < y < r_2(p_1, p_3)$, *then* $\phi(p_1, y, p_3) < 0$. $\phi(p_1, y, p_3)$ is quadratic in y with leading coefficient $\alpha(p_1, p_3)$. If $\alpha(p_1, p_3) > 0$ and the two roots $r_1(p_1, p_3)$ and $r_3(p_1, p_3)$ are distinct, then $\phi(p_1, y, p_3) < 0$ between the roots of ϕ.

To show that $\alpha(p_1, p_3) > 0$, rewrite $\alpha(p_1, p_3)$ as follows:

$$\alpha(p_1, p_3) = p_1(1 - p_1) + p_3(1 - p_3) + 3p_1 - 6p_1p_3 + 3p_3.$$

Because $p_1 \geqslant p_1^2$ and $p_3 \geqslant p_3^2$, we then have

$$\alpha(p_1, p_3) \geqslant p_1(1 - p_1) + p_3(1 - p_3) + 3(p_1 - p_3)^2,$$

from which we conclude $\alpha(p_1, p_3) > 0$ because $p_1 \neq p_3$.

2.3.3.2. $r_1(p_1, p_3) \leqslant p_1$. Note that $r_1(p_1, p_3) \leqslant 0$ implies that $r_1(p_1, p_3) \leqslant p_1$ because $0 \leqslant p_1$. We can solve the quadratic equation $\phi(p_1, y, p_3) = 0$ for the value of y as a function of p_1 and p_3, taking the smaller root to be $r_1(p_1, p_3)$:

$$r_1(p_1, p_3) = \frac{-\beta(p_1, p_3) - \sqrt{\beta(p_1, p_3)^2 - 4\alpha(p_1, p_3)\gamma(p_1, p_3)}}{2\alpha(p_1, p_3)}.$$

To show $r_1(p_1, p_3) \leqslant 0 \leqslant p_1$, because we have demonstrated that $\alpha(p_1, p_3) > 0$, we must show

$$-\beta(p_1, p_3) - \sqrt{\beta(p_1, p_3)^2 - 4\alpha(p_1, p_3)\gamma(p_1, p_3)} \leqslant 0.$$

It suffices to show that $\gamma(p_1, p_3) \leqslant 0$.

Write $\gamma(p_1, p_3) = p_1p_3\tilde{\gamma}(p_1, p_3)$, where

$$\tilde{\gamma}(p_1, p_3) = -16 + 12p_1 - p_1^2 + 12p_3 - 6p_1p_3 - p_3^2.$$

The partial derivatives of $\tilde{\gamma}(p_1, p_3)$ are positive for $0 \leqslant p_1 < p_3 \leqslant 1$:

$$\frac{\partial\tilde{\gamma}(p_1, p_3)}{\partial p_1} = 12 - 2p_1 - 6p_3 > 4$$
$$\frac{\partial\tilde{\gamma}(p_1, p_3)}{\partial p_3} = 12 - 2p_3 - 6p_1 > 4.$$

Hence, $\tilde{\gamma}(p_1, p_3)$ is increasing in $p_1 \in [0, 1]$ and $p_3 \in [p_1, 1]$ and is maximized at $(p_1, p_3) = (1, 1)$. Because $\tilde{\gamma}(1, 1) = 0$, it follows that $\tilde{\gamma}(p_1, p_3) < 0$ for all (p_1, p_3) with $0 \leqslant p_1 < 1$ and $p_1 \leqslant p_3 < 1$.

We conclude $\gamma(p_1, p_3) \leqslant 0$ and therefore $r_1(p_1, p_3) \leqslant p_1$.

2.3.3.3. $r_2(p_1, p_3) \geqslant p_3$. It suffices to show $r_2(p_1, p_3) \geqslant 1 \geqslant p_3$.

Taking the positive root of the quadratic equation $\phi(p_1, p_2, p_3) = 0$,

$$r_2(p_1, p_3) = \frac{-\beta(p_1, p_3) + \sqrt{\beta(p_1, p_3)^2 - 4\alpha(p_1, p_3)\gamma(p_1, p_3)}}{2\alpha(p_1, p_3)}$$

Because $\alpha(p_1, p_3) > 0$ and $\gamma(p_1, p_3) \leqslant 0$, and leaving off the arguments, it suffices to show $\alpha + \beta + \gamma \leqslant 0$. If $\alpha + \beta + \gamma \leqslant 0$ then $4\alpha^2 + 4\alpha\beta + 4\alpha\gamma \leqslant 0$, $(2\alpha + \beta)^2 \leqslant \beta^2 - 4\alpha\gamma$, and, thus, $1 \leqslant \frac{1}{2\alpha}(-\beta + \sqrt{\beta^2 - 4\alpha\gamma})$.

Define $g(p_1, p_3) = \alpha + \beta + \gamma$. We can simplify the condition $g(p_1, p_3) \leqslant 0$, by using eqs. (13)-(15):

$$g(p_1, p_3) = -(1 - p_1)(1 - p_3)(4p_1 + p_1^2 + 4p_3 + 6p_1p_3 + p_3^2).$$

If $0 \leqslant p_1 < p_3 \leqslant 1$, then $g(p_1, p_3) \leqslant 0$, as all of the factors in parentheses are nonnegative. Because $g(p_1, p_3) \leqslant 0$, it follows that $r_2(p_1, p_3) \geqslant p_3$.

We conclude that $\phi(p_1, p_2, p_3) < 0$ if $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$. We have $r_1(p_1, p_3) \leqslant p_1$ and $r_1(p_1, p_3) \geqslant p_3$, the roots of $\phi(p_1, y, p_3) = 0$ are distinct, and $\phi(p_1, y, p_3) < 0$ between the roots, $r_1(p_1, p_3)$ and $r_2(p_1, p_3)$.

2.3.4. *Concluding the proof.* Because we have shown that for $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$, $\phi(p_1, p_2, p_3) < 0$ and $\phi(p_1, p_2, p_3) < 0$ implies $\omega(p_1, p_2, p_3) < 0$, in turn implying $\psi(p_1, p_2, p_3) < 0$ for $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$, eq. (6) is never satisfied, and the triangle inequality is never satsified for biallelic markers with $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$.

3. The maximal deviation from the triangle inequality occurs at Sewall Wright's counterexample

3.1. **Visualization of $\psi(p_1, p_2, p_3)$.** As shown in Section 2.3, $\psi(p_1, p_2, p_3)$, measuring the extent to which the triangle inequality fails to be satisfied, is always less than or equal to 0 for $0 \leqslant p_1 \leqslant p_2 \leqslant p_3 \leqslant 1$. We illustrate the value of $\psi(p_1, p_2, p_3)$ over the space of possible allele frequencies (p_1, p_2, p_3) in Figure 1, holding p_2 constant at each of several values and plotting $\psi(p_1, p_2, p_3)$ as a function of (p_1, p_3) over the permissible domain $[0, p_2] \times [p_2, 1]$.

In each plot at a fixed p_2, the value of ψ appears to decrease monotonically from 0 along lines of constant p_1 and along lines of constant p_3 to a minimum at $(p_1, p_3) = (0, 0)$. Moreover, considering all plots at different values of p_2, the minimum at $(p_1, p_3) = (0, 1)$ appears lowest in the case that $p_2 = \frac{1}{2}$. The plots suggest that the point at which $\psi(p_1, p_2, p_3)$ is the most negative—where F_{ST} fails the triangle inequality by the largest amount—is where p_1, p_2, and p_3 are furthest apart. They suggest that the minimum of $\psi(p_1, p_2, p_3)$ lies at $(0, \frac{1}{2}, 1)$, exactly the triplet Sewall Wright (1978, p. 89) offered in his counterexample. We next prove this to be the case.

3.2. **The minimum of $\psi(p_1, p_2, p_3)$ is $-\frac{1}{3}$ and occurs at $(p_1, p_2, p_3) = (0, \frac{1}{2}, 1)$.** We seek to find the minimum of $\psi(p_1, p_2, p_3)$, as described in eq. (12), considering all possible (p_1, p_2, p_3) with $0 \leqslant p_1 \leqslant p_2 \leqslant p_3 \leqslant 1$. Note that we can assume $0 \leqslant p_1 < p_2 < p_3 \leqslant 1$, because if $p_1 = p_2 \leqslant p_3$ or $p_1 \leqslant p_2 = p_3$, then $\psi(p_1, p_2, p_3) = 0$. As we showed previously, $\psi(p_1, p_2, p_3) = 0$ is the maximal value of ψ, so in finding the minimum, we can assume p_1, p_2, and p_3 are distinct.

We show that the minimum of $\psi(p_1, p_2, p_3)$ occurs at $\psi(0, \frac{1}{2}, 1) = -\frac{1}{3}$. The proof proceeds in three steps:

(1) For fixed p_2, p_3, we show $\psi(0, p_2, p_3) < \psi(p_1, p_2, p_3)$, for all p_1 with $0 < p_1 < 1$.
(2) For fixed p_2, we show $\psi(0, p_2, 1) < \psi(0, p_2, p_3)$ for all p_3 with $0 < p_3 < 1$.
(3) We show $\psi(0, \frac{1}{2}, 1) < \psi(0, p_2, 1)$ for all p_2 with $0 < p_2 < \frac{1}{2}$ or $\frac{1}{2} < p_2 < 1$.

Showing (1), (2), and (3) suffices to show that the minimum is $\psi(0, \frac{1}{2}, 1)$, as Step 1 shows that the minimum has $p_1 = 0$, Step 2 shows that $p_3 = 1$, and Step 3 shows that $p_2 = \frac{1}{2}$.

3.2.1. $\psi(0, p_2, p_3) < \psi(p_1, p_2, p_3)$. To show that the minimum of $\psi(0, p_2, p_3)$ over $0 \leqslant p_1 \leqslant 1$ at fixed (p_2, p_3) occurs at $p_1 = 0$, we seek to show that there is no minimum of $\psi(0, p_2, p_3)$ for $0 < p_1 < 1$, so that the minimum must occur on the boundary of the unit interval. If $\partial\psi/\partial p_1 > 0$ for $0 < p_1 < 1$, then a minimum occurs at the lower bound of p_1: $p_1 = 0$. To show that $\psi(0, p_2, p_3) < \psi(p_1, p_2, p_3)$ for all p_1, we show that $\partial\psi/\partial p_1 > 0$ everywhere in $0 < p_1 < 1$.

Note that

$$\begin{aligned}\frac{\partial\psi}{\partial p_1} &= \frac{\partial F_{ST}(p_1, p_2)}{\partial p_1} + \frac{\partial F_{ST}(p_2, p_3)}{\partial p_1} - \frac{\partial F_{ST}(p_1, p_3)}{\partial p_1} \\ &= \frac{\partial F_{ST}(p_1, p_2)}{\partial p_1} - \frac{\partial F_{ST}(p_1, p_3)}{\partial p_1}.\end{aligned}$$

To show $\partial F_{ST}(p_1, p_2)/\partial p_1 - \partial F_{ST}(p_1, p_3)/\partial p_1 > 0$, it suffices to show $\partial F_{ST}(p_1, p_2)/\partial p_1 > \partial F_{ST}(p_1, p_3)/\partial p_1$.

Define a function $f(p_1, \rho) = \partial F_{ST}(p_1, \rho)/\partial p_1$. Note that showing $f(p_1, p_2) > f(p_1, p_3)$ where $p_2 < p_3$ is the same as showing that $\partial f(p_1, \rho)/\partial\rho < 0$, for $0 < \rho < 1$ (ρ must be strictly in the bounds of its domain because $\rho = 0$ would imply $p_1 = p_2$ and $\rho = 1$ would imply $p_2 = p_3$). Showing that $\partial^2 F_{ST}(p_1, \rho)/\partial p_1 \partial\rho < 0$ implies that $\partial\psi(p_1, p_2, p_3)/p_1 > 0$.

Taking the partial derivative of $F_{ST}(p_1, \rho)$ with respect to p_1 gives us

$$\frac{\partial F_{ST}(p_1, \rho)}{\partial p_1} = -\frac{(\rho - p_1)^2}{(2 - \rho - p_1)(\rho + p_1)^2} - \frac{2(\rho - p_1)}{(2 - \rho - p_1)(\rho + p_1)} + \frac{(\rho - p_1)^2}{(2 - \rho - p_1)^2(\rho + p_1)}.$$

Taking the partial derivative again with respect to ρ yields

(16)
$$\frac{\partial^2 F_{ST}}{\partial p_1 \partial\rho} = \frac{2(\rho - p_1)^2}{(2 - \rho - p_1)(\rho + p_1)^3} - \frac{2(\rho - p_1)^2}{(2 - \rho - p_1)^2(\rho + p_1)^2} + \frac{2}{(2 - \rho - p_1)(\rho + p_1)} + \frac{2(\rho - p_1)^2}{(2 - \rho - p_1)^3(\rho + p_1)}.$$

We seek to show that eq. (16) is strictly less than 0. By rearranging terms, it is equivalent to show that

(17)
$$\frac{(\rho - p_1)^2}{(\rho + p_1)^2} - \frac{(\rho - p_1)^2}{(2 - \rho - p_1)(\rho + p_1)} + \frac{(\rho - p_1)^2}{(2 - \rho - p_1)^2} < 1.$$

First, consider that $p_1 - p_1^2 > 0$ because $p_1 < 1$ and $\rho - \rho^2 > 0$ because $\rho < 1$. Thus,

$$\begin{aligned}-2(p_1 - p_1^2) + 2(\rho - \rho^2) &< 2(p_1 - p_1^2) + 2(\rho - \rho^2) \\ 2(p_1 - p_1^2) - 2(\rho - \rho^2) &< 2(p_1 - p_1^2) + 2(\rho - \rho^2).\end{aligned}$$

Hence,

$$|2(p_1 - p_1^2) - 2(\rho - \rho^2)| < 2(p_1 - p_1^2) + 2(\rho - \rho^2).$$

Because $\rho - p_1$, $\rho + p_1$, and $2 - \rho - p_1$ are positive, we have

$$\left| \frac{-2p_1 + 2p_1^2 + 2\rho - 2\rho^2}{(\rho + p_1)(2 - \rho - p_1)} \right| < \frac{2p_1 - 2p_1^2 + 2\rho - 2\rho^2}{(2 - \rho - p_1)(\rho + p_1)},$$

which can be rearranged to show

$$\left| \frac{\rho - p_1}{\rho + p_1} - \frac{\rho - p_1}{2 - \rho - p_1} \right| < 1 - \frac{(\rho - p_1)^2}{(2 - \rho - p_1)(\rho + p_1)}. \tag{18}$$

By noting that $(\rho - p_1)^2/[(2 - \rho - p_1)(\rho + p_1)]$ is the expression for F_{ST} (eq. (3)), which is non-negative and less than or equal to 1, we find that both sides of eq. (18) are bounded in $[0, 1]$:

$$0 \leqslant \left| \frac{\rho - p_1}{\rho + p_1} - \frac{\rho - p_1}{2 - \rho - p_1} \right| < 1 - \frac{(\rho - p_1)^2}{(2 - \rho - p_1)(\rho + p_1)} \leqslant 1.$$

Therefore,

$$\left(\frac{\rho - p_1}{\rho + p_1} - \frac{\rho - p_1}{2 - \rho - p_1} \right)^2 < \left| \frac{\rho - p_1}{\rho + p_1} - \frac{\rho - p_1}{2 - \rho - p_1} \right| < 1 - \frac{(\rho - p_1)^2}{(2 - \rho - p_1)(\rho + p_1)}.$$

By adding $(\rho - p_1)^2/[(2 - \rho - p_1)(\rho + p_1)]$ to both sides of

$$\left(\frac{\rho - p_1}{\rho + p_1} - \frac{\rho - p_1}{2 - \rho - p_1} \right)^2 < 1 - \frac{(\rho - p_1)^2}{(2 - \rho - p_1)(\rho + p_1)},$$

we have eq. (17).

Because eq. (17) holds, we have completed our proof that eq. (16) is less than 0 for $0 < p_1 < 1$ and $0 < \rho < 1$. Therefore, $\partial\psi/\partial p_1 > 0$ for $0 < p_1 < 1$ at fixed p_2 and p_3, and the minimum of ψ occurs at $p_1 = 0$. We conclude $\psi(0, p_2, p_3) < \psi(p_1, p_2, p_3)$.

3.2.2. $\psi(0, p_2, 1) < \psi(0, p_2, p_3)$. To show $\psi(0, p_2, 1) < \psi(0, p_2, p_3)$, we first comment that $\psi(p_1, p_2, p_3)$ symmetric with respect to the choice of allele, so that

$$\psi(p_1, p_2, p_3) = \psi(1 - p_3, 1 - p_2, 1 - p_1). \tag{19}$$

F_{ST} is symmetric with respect to an exchange of populations: $F_{ST}(p_i, p_j) = F_{ST}(p_j, p_i)$. It is also symmetric in the choice of allele used for the computation, so that $F_{ST}(p_i, p_j) = F_{ST}(1-p_i, 1-p_j)$. Thus, we have

$$\begin{aligned}
\psi(p_1, p_2, p_3) &= F_{ST}(p_1, p_2) + F_{ST}(p_2, p_3) - F_{ST}(p_1, p_3) \\
&= F_{ST}(1 - p_1, 1 - p_2) + F_{ST}(1 - p_2, 1 - p_3) - F_{ST}(1 - p_1, 1 - p_3) \\
&= F_{ST}(1 - p_3, 1 - p_2) + F_{ST}(1 - p_2, 1 - p_1) - F_{ST}(1 - p_3, 1 - p_1) \\
&= \psi(1 - p_3, 1 - p_2, 1 - p_1). \qquad (20)
\end{aligned}$$

From Section 3.2.1, we have $\psi(0, p_2, p_3) < \psi(p_1, p_2, p_3)$. By the symmetry in eq. (20), we have $\psi(1-p_3, 1-p_2, 1-0) < \psi(1-p_1, 1-p_2, 1-p_3)$. Defining $q_1 = 1-p_3$, $q_2 = 1-p_2$, and $q_3 = 1-p_1$, where $0 \leqslant q_1 < q_2 < q_3 \leqslant 1$, we can also express this inequality as $\psi(q_1, q_2, 1) < \psi(q_1, q_2, q_3)$, for all $0 \leqslant q_1 < q_2 < q_3 \leqslant 1$. Therefore $p_1 = 0$ minimizes ψ for all values of (p_2, p_3) with $0 < p_2 < 1$ and $0 < p_3 \leqslant 1$, and $p_3 = 1$ minimizes ψ for all values of (p_1, p_2) with $0 \leqslant p_1 < 1$ and $0 < p_2 < 1$. We then have $\psi(0, p_2, 1) < \psi(p_1, p_2, p_3)$, which concludes the proof of the claim.

3.2.3. $\psi(0, \frac{1}{2}, 1) < \psi(0, p_2, 1)$. Given that we know that $p_1 = 0$ and $p_3 = 1$ minimize $\psi(p_1, p_2, p_3)$ at fixed p_2, we have reduced this last step to a single variable problem to determine what value of p_2 minimizes $\psi(p_1, p_2, p_3)$. Consider $\psi(0, p_2, 1)$:

$$\begin{aligned}\psi(0, p_2, 1) &= F_{ST}(0, p_2) + F_{ST}(p_2, 1) - F_{ST}(0, 1) \\ &= \frac{p_2^2}{p_2(2 - p_2)} + \frac{(1 - p_2)^2}{(1 + p_1)(1 - p_1)} - 1 \\ &= \frac{3(1 - p_2 + p_2^2)}{(-2 + p_2)(1 + p_2)}. \end{aligned} \tag{21}$$

We can take the derivative of eq. (21) with respect to p_2:

$$\frac{\partial \psi}{\partial p_2} = \frac{6(-1 + 2p_2)}{(-2 + p_2)^2(1 + p_2)^2}. \tag{22}$$

Eq. (22) is only equal to 0 when $p_2 = \frac{1}{2}$. Therefore, p_2 is a critical point for ψ in the domain $0 \leqslant p_2 \leqslant 1$ and specifically, $\psi(0, \frac{1}{2}, 1) = -\frac{1}{3}$ is a minimum because ψ is greater when $p_2 = 0$ or $p_2 = 1$: $\psi(0, 0, 1) = \psi(0, 1, 1) = 0$.

4. The four-point condition never holds for biallelic markers with distinct allele frequencies

The failure of F_{ST} to satisfy the triangle inequality for distinct allele frequencies (Section 2.3) raises the issue of the status of F_{ST} with respect to the four-point condition of Buneman (1974). The four-point condition is satisfied for a function d on a set X if and only if for all choices of four points $x_1, x_2, x_3, x_4 \in X$, not necessarily distinct, all of the following hold:

$$d(x_1, x_2) + d(x_3, x_4) \leqslant \max\{d(x_1, x_3) + d(x_2, x_4), d(x_1, x_4) + d(x_2, x_3)\} \tag{23}$$

$$d(x_1, x_3) + d(x_2, x_4) \leqslant \max\{d(x_1, x_2) + d(x_3, x_4), d(x_1, x_4) + d(x_2, x_3)\} \tag{24}$$

$$d(x_1, x_4) + d(x_2, x_3) \leqslant \max\{d(x_1, x_2) + d(x_3, x_4), d(x_1, x_3) + d(x_2, x_4)\}. \tag{25}$$

Equivalently to eqs. (23)-(25), two of the quantities $d(x_1, x_2) + d(x_3, x_4)$, $d(x_1, x_3) + d(x_2, x_4)$, and $d(x_1, x_4) + d(x_2, x_3)$ are equal and greater than or equal to the third.

The four-point condition can be satisfied for some set of four points $x_1, x_2, x_3, x_4 \in X$ without necessarily holding for all sets of four points in X. For a specific set of four points, if and only if the four-point condition is satisfied, those points can be placed as the leaves of an unrooted tree whose edges are associated with lengths in such a manner that the pairwise distances between points computed along the tree accord with the function d (Buneman 1974, Steel 2016, p. 112).

For a function d that fails to satisfy the triangle inequality for all distinct points x_1, x_2, x_4 in a set X, the four-point condition sometimes fails; supposing $d(x_1, x_2) + d(x_2, x_4) < d(x_1, x_4)$, we simply take $x_3 = x_2$. Noting that $d(x_2, x_3) = 0$, eq. (25) does not hold. However, the four-point condition can be satisfied for x_1, x_2, x_3, x_4 even if the triangle inequality is not satisfied for any three of the points, as is the case if $(d(x_1, x_2), d(x_1, x_3), d(x_1, x_4), d(x_2, x_3), d(x_2, x_4), d(x_3, x_4)) = (2, 5, 8, 2, 5, 2)$.

We now demonstrate that for F_{ST}, not only does the triangle inequality fail for all sets of three points that correspond to the allele frequencies of a biallelic marker with distinct frequencies in three

populations, as shown in Section 2.3, the four-point condition also fails for all sets of four points that correspond to frequencies of a biallelic marker with distinct frequencies in four populations. This result has the consequence that sets of four populations cannot be placed on an unrooted tree in such a way that pairwise distances, as computed along the tree, accord with F_{ST}.

Consider four populations whose frequencies of a particular allele at a biallelic marker satisfy $0 \leqslant p_1 < p_2 < p_3 < p_4 \leqslant 1$. We show that

$$F_{ST}(p_1, p_3) + F_{ST}(p_2, p_4) > \max\{F_{ST}(p_1, p_2) + F_{ST}(p_3, p_4), F_{ST}(p_1, p_4) + F_{ST}(p_2, p_3)\},$$

so that eq. (24) fails to be satisfied with F_{ST} in the role of d.

Define $s(p_i, p_j, p_k, p_\ell) = F_{ST}(p_i, p_j) + F_{ST}(p_k, p_\ell)$. For $p_i \leqslant p_j \leqslant p_k$, recall from eq. (12) that $\psi(p_i, p_j, p_k) = F_{ST}(p_i, p_j) + F_{ST}(p_j, p_k) - F_{ST}(p_i, p_k)$. Applying the result of Section 2.3, $\psi(p_i, p_j, p_k) < 0$ for $0 \leqslant p_i < p_j < p_k \leqslant 1$.

We can then use eq. (12) to write

(26)
$$s(p_1, p_3, p_2, p_4) = F_{ST}(p_1, p_2) + F_{ST}(p_2, p_3) - \psi(p_1, p_2, p_3) + F_{ST}(p_2, p_3) + F_{ST}(p_3, p_4) - \psi(p_2, p_3, p_4)$$

(27)
$$s(p_1, p_4, p_2, p_3) = F_{ST}(p_1, p_2) + F_{ST}(p_2, p_3) - \psi(p_1, p_2, p_3) + F_{ST}(p_3, p_4) - \psi(p_1, p_3, p_4) + F_{ST}(p_2, p_3).$$

Noting that $\psi(p_1, p_2, p_3)$, $\psi(p_1, p_3, p_4)$, and $\psi(p_2, p_3, p_4)$ are all bounded above by 0 owing to the failure of the triangle inequality for F_{ST}, we can cancel equal terms and use the positivity of F_{ST} in eq. (3) for distinct allele frequencies to obtain $s(p_1, p_2, p_3, p_4) < s(p_1, p_3, p_2, p_4)$ and $s(p_1, p_2, p_3, p_4) < s(p_1, p_4, p_2, p_3)$. We then have

$$s(p_1, p_2, p_3, p_4) < \max\{s(p_1, p_3, p_2, p_4), s(p_1, p_4, p_2, p_3)\}.$$

To show that the four-point condition does not hold, we must show $s(p_1, p_3, p_2, p_4) \neq s(p_1, p_4, p_2, p_3)$, so that with F_{ST} in the role of d and p_i in the role of x_i, eqs. (24) and (25) cannot both hold simultaneously. Examining eqs. (26) and (27), we have

$$s(p_1, p_3, p_2, p_4) = s(p_1, p_2, p_3, p_4) - \psi(p_1, p_2, p_3) - \psi(p_2, p_3, p_4) \tag{28}$$

$$s(p_1, p_4, p_2, p_3) = s(p_1, p_2, p_3, p_4) - \psi(p_1, p_2, p_3) - \psi(p_1, p_3, p_4). \tag{29}$$

Thus, this problem reduces to showing that

$$\psi(p_1, p_3, p_4) \neq \psi(p_2, p_3, p_4).$$

We have already shown in Section 3.2.1 that for $0 \leqslant p_i < p_j < p_k < 1$, $\partial\psi(p_i, p_j, p_k)/\partial p_i > 0$. Because $p_1 < p_2$, we can therefore conclude

$$\psi(p_1, p_3, p_4) < \psi(p_2, p_3, p_4),$$

and thus, $s(p_1, p_3, p_2, p_4) < s(p_1, p_4, p_2, p_3)$. Hence, eq. (25) fails, so that the four-point condition does not hold for distinct allele frequencies p_1, p_2, p_3, p_4. Therefore, beyond the failure of the four-point condition that results quickly when $p_3 = p_2$ from the triangle inequality not holding for $0 \leqslant p_1 < p_2 < p_4 \leqslant 1$, F_{ST} never satisfies the four-point condition when $0 \leqslant p_1 < p_2 < p_3 < p_4 \leqslant 1$.

5. Distribution of allele frequencies in the parameter space

Next, in the context of our F_{ST} results, we consider the placement of loci from human populations in the space of possible allele frequencies. For this analysis, we examined 590,461 single-nucleotide polymorphisms (SNPs) taken from the HapMap (International HapMap 3 Consortium 2010), as used by Verdu et al. (2014) and Kang et al. (2016). We considered three populations, CEU with sample size 112 individuals, CHB with 137 individuals, and YRI with 140 individuals.

We identified ordered triples (p_1, p_2, p_3) of frequencies, with p_1 representing an allele frequency in CHB, p_2 representing the frequency of the same allele in CEU, and p_3 in YRI, and with $p_1 \leqslant p_2 \leqslant p_3$. The SNPs can be divided into three groups based on which of the three populations has allele frequencies that lie between those of the other two populations. For the 265,517 SNPs with CEU in the intermediate position, we relabeled alleles such that $p_2 \leqslant \frac{1}{2}$. Note that we placed CEU in the intermediate position in case of ties. At nonzero allele frequencies, we observed 380 two-way ties with CHB, 621 two-way ties with YRI, and 7 three-way ties.

We plotted the values of (p_1, p_2, p_3) over the permissible domain (Figure 2). Owing to the general similarity of allele frequencies among human populations, most points tend to have p_1 only slightly less than p_2 and p_3 only slightly greater than p_2. In regions with similar frequencies for p_1, p_2, and p_3, $\psi(p_1, p_2, p_3)$ is only slightly less than zero. However, nontrivial numbers of points are placed in the upper left corner of the plots, where the deviation from 0 is greatest. Therefore, some SNPs in the three populations do indeed produce substantial deviations from the triangle inequality.

6. Discussion

In this paper, we have expanded on the observation of Sewall Wright (1978) that F_{ST} does not always satisfy the triangle inequality. In particular, we found that F_{ST} *never* satisfies the triangle inequality for biallelic markers with distinct allele frequencies. Interestingly, Wright's case—arguably the simplest counterexample owing to its use of 0, $\frac{1}{2}$, and 1 rather than more obscure frequency values—is the triplet that fails the triangle inequality by the largest amount.

6.1. **Consequences for statistical methods.**

We have found that failure to satisfy the triangle inequality for all triplets of distinct allele frequencies implies failure to satisfy the four-point condition for all sets of four distinct allele frequencies. These failures to satisfy the triangle inequality and the four-point condition for all 3-tuples and 4-tuples of distinct allele frequencies for biallelic markers have implications for various forms of data analysis using F_{ST}.

6.1.1. *Multidimensional scaling (MDS).*

Matrices of pairwise dissimilarity among a set of populations are commonly used as a basis for visually depicting similarities of the populations in two or three dimensions by mulitdimensional scaling analysis (Jombart et al. 2009, Wang et al. 2010). These depictions find a representation of the matrix in a two- or three-dimensional space that has the property that Euclidean distances between points in the space approximate the matrix entries. Because F_{ST} does not satisfy the triangle inequality for biallelic markers, however, three distinct populations considered for a biallelic marker in an F_{ST} matrix cannot be represented as points in Euclidean space in such a way that Euclidean distances in the triangle connecting them correspond

to the entries in the F_{ST} matrix. This imperfection of the spatial representation applies for any subset of three distinct points in a larger collection; thus, Euclidean distances between points in an MDS representation of an F_{ST} matrix necessarily only approximate the matrix entries.

Although MDS cannot always perfectly recapitulate the dissimilarities in the input matrix, MDS is frequently performed on distance matrices that are not Euclidean (Mardia et al. 1979, Cox and Cox 2001). Typical metric MDS finds a best-fit of distances between points in Euclidean space to dissimilarities in the non-Euclidean matrix. The matrix entries can also be transformed so that sets of three points necessarily satisfy the triangle inequality. One adjustment adds a constant c to each matrix entry (Cailliez 1983). For a dissimilarity d, after a large enough constant is added to obtain a new dissimilarity $d' = d + c$, the transformed distances satisfy the triangle inequality: if $d(x_1, x_2) + d(x_2, x_3) < d(x_1, x_3)$, then a choice $c > d(x_1, x_3) - d(x_1, x_2) - d(x_2, x_3)$ leads to $d'(x_1, x_2) + d'(x_2, x_3) > d'(x_1, x_3)$. Alternatively, taking the square root of values in $[0, 1]$ yields larger values still in $[0, 1]$, so that the sum of any two of three transformed values is more likely to exceed the third one (Legendre and Legendre 1998, p. 433). In Figure 3, we apply this transformation, finding that $\sqrt{F_{ST}}$ satisfies the triangle inequality for all triplets plotted.

We note, however, that the choice of transformation does affect the resulting MDS representation. In Figure 4, we compare the output of the Caillez transformation and the square root transformation on F_{ST} dissimilarity matrices with the same five allele frequencies chosen independently at random from a uniform distribution. The MDS plots differ and, in some cases, two points that are close together in one plot are distant in the other. Considering the distances between the output in the plots, neither distance matrix results in the the same matrix as the original unmodified F_{ST} dissimilarities, because F_{ST} cannot be represented as distances in Euclidean space. The choice of transform ultimately affect the results, and is relevant to report for interpretation of MDS results.

6.1.2. *Neighbor-joining inference of evolutionary trees.* A second form of analysis affected by the failure of the triangle inequality is tree reconstruction from matrices of F_{ST} values computed from allele frequencies (e.g. Takezaki and Nei 1996, Pérez-Lezaun et al. 1997, Bosch et al. 2000). Here, we consider the behavior of the neighbor-joining (NJ) algorithm applied to F_{ST} dissimilarity matrices for biallelic markers with distinct frequencies.

If a dissimilarity matrix is generated exactly from a population tree by calculating path lengths between population pairs on the tree, then NJ recovers the generating tree (Saitou and Nei 1987, Studier and Keppler 1988, Atteson 1999). Because of the failure of the four-point condition, population trees constructed from F_{ST} matrices for biallelic markers do not perfectly represent those matrices. Moreover, for any four leaves of the population tree, the minimal path connecting the leaves does not faithfully represent the F_{ST} matrix entries associated with those leaves.

Nevertheless, the inferred tree might still place more genetically similar populations close together on the tree. Using results from Mihaescu et al. (2009), we determine the tree inferred by neighbor-joining from F_{ST} matrices for 4, 5, 6, and 7 taxa. In particular, we prove the following proposition.

Proposition 1. *Consider a biallelic marker in n populations, with allele frequencies $0 \leqslant p_1 < p_2 < \ldots < p_n \leqslant 1$ for a specified allele. For these populations, neighbor-joining applied to an F_{ST} dissimilarity matrix, d, produces the tree topologies in Figure 5 for $4 \leqslant n \leqslant 7$.*

The proposition states that with the populations ordered by allele frequency, neighboring populations in the sequence are placed in adjacent positions on the neighbor-joining tree. We begin with a lemma that addresses the case of $n = 4$.

Lemma 1. *Consider a biallelic marker in 4 populations, with allele frequencies* $0 \leqslant p_1 < p_2 < p_3 < p_4 \leqslant 1$. *For these populations, neighbor-joining applied to an* F_{ST} *dissimilarity matrix,* d, *produces the quartet ((1,2),(3,4)).*

Proof. Proposition 6 of Mihaescu et al. (2009) demonstrates that neighbor-joining applied to a dissimilarity d in four populations i_1, i_2, i_3, i_4 returns the quartet $((i_1, i_2), (i_3, i_4))$ if

$$d(i_1, i_2) + d(i_3, i_4) < \min\{d(i_1, i_3) + d(i_2, i_4), d(i_1, i_4) + d(i_2, i_3)\}.$$

For dissimilarity measure F_{ST}, we have already demonstrated in Section 4 that, using our notation, $s(p_1, p_2, p_3, p_4) < s(p_1, p_3, p_2, p_4)$ and $s(p_1, p_2, p_3, p_4) < s(p_1, p_4, p_2, p_3)$, so that

$$s(p_1, p_2, p_3, p_4) < \min\{s(p_1, p_3, p_2, p_4), s(p_1, p_4, p_2, p_3)\}.$$

Using the definition $s(p_i, p_j, p_k, p_\ell) = F_{ST}(p_i, p_j) + F_{ST}(p_k, p_\ell)$, the condition in Proposition 6 of Mihaescu et al. (2009) is obtained. □

To prove Proposition 1, we rely on the concept of *quartet consistency*, which Mihaescu et al. (2009) introduced for assessing the output topology T of neighbor-joining in contexts in which no tree T exactly captures entries in the dissimilarity matrix. By Definition 8 of Mihaescu et al. (2009), a dissimilarity map d for n populations is *quartet consistent* with a tree T if for every quartet $((i, j), (k, \ell)) \in T$, $w_d(ij : k\ell) > \max[w_d(ik : j\ell), w_d(i\ell : jk)]$, where

$$w_d(xy : wz) = \frac{1}{2}[d(x, w) + d(x, z) + d(y, w) + d(y, z)] - d(x, y) - d(w, z).$$

Theorem 9 of Mihaescu et al. (2009) states that for $4 \leqslant n \leqslant 7$, if there exists a tree T that is *quartet consistent* with a dissimilarity map $d : X \times X \to \mathbb{R}$, then NJ outputs a tree with the same topology as T. This theorem provides a method of determining the NJ tree from a dissimilarity map on n taxa, $4 \leqslant n \leqslant 7$, without proceeding through the steps of the NJ algorithm: it suffices to exhibit T with which d is quartet consistent.

Lemma 2. *Consider a biallelic marker in* n *populations, with allele frequencies* $0 \leqslant p_1 < p_2 < \ldots < p_n \leqslant 1$ *for a specified allele. For any* i_1, i_2, i_3, i_4 *with* $i_1 < i_2 < i_3 < i_4$, *using* F_{ST} *for the dissimilarity* d, $w_d(i_1 i_2 : i_3 i_4) > \max[w_d(i_1 i_3 : i_2 i_4), w_d(i_1 i_4 : i_2 i_3)]$.

Proof. By definition of d, we have

$$w_d(i_1 i_2 : i_3 i_4) = \frac{1}{2}[s(1, 3, 2, 4) + s(1, 4, 2, 3)] - s(1, 2, 3, 4)$$
$$w_d(i_1 i_3 : i_2 i_4) = \frac{1}{2}[s(1, 2, 3, 4) + s(1, 4, 2, 3)] - s(1, 3, 2, 4)$$
$$w_d(i_1 i_4 : i_2 i_3) = \frac{1}{2}[s(1, 3, 2, 4) + s(1, 2, 3, 4)] - s(1, 4, 2, 3).$$

Recalling eqs. (28) and (29) and the result that $\psi(p_{i_1}, p_{i_2}, p_{i_3}) < 0$ for $0 \leqslant p_{i_1} < p_{i_2} < p_{i_3} \leqslant 1$,

$$s(1,3,2,4) = s(1,2,3,4) + |\psi(p_1,p_2,p_3)| + |\psi(p_2,p_3,p_4)|$$
$$s(1,4,2,3) = s(1,2,3,4) + |\psi(p_1,p_2,p_3)| + |\psi(p_1,p_3,p_4)|.$$

We then have:

$$w_d(i_1i_2 : i_3i_4) = |\psi(p_1,p_2,p_3)| + \frac{1}{2}|\psi(p_2,p_3,p_4)| + \frac{1}{2}|\psi(p_1,p_3,p_4)|$$
$$w_d(i_1i_3 : i_2i_4) = -\frac{1}{2}|\psi(p_1,p_2,p_3)| - |\psi(p_2,p_3,p_4)| + \frac{1}{2}|\psi(p_1,p_3,p_4)|$$
$$w_d(i_1i_4 : i_2i_3) = -\frac{1}{2}|\psi(p_1,p_2,p_3)| + \frac{1}{2}|\psi(p_2,p_3,p_4)| - |\psi(p_1,p_3,p_4)|.$$

We conclude $w_d(i_1i_2 : i_3i_4) > \max[w_d(i_1i_3 : i_2i_4), w_d(i_1i_4 : i_2i_3)]$. □

Proof of Proposition 1. We consider n populations labeled $1, 2, \ldots, n$ such that the frequency of a specific allele has $0 \leqslant p_1 \leqslant p_2 \leqslant \ldots \leqslant p_n \leqslant 1$. Consider a tree topology T in which populations 1 and 2 are in a cherry, $n-1$ and n are in a cherry, and $3, \ldots, n-2$ are arranged in numerical sequence incidental to the internal branch of the quartet $((1,2),(n-1,n))$ (Figure 5). It suffices to show that the F_{ST} dissimilarity matrix d is quartet consistent with such a topology.

Consider an arbitrary subset of four populations $\{i_1, i_2, i_3, i_4\}$, where $p_{i_1} < p_{i_2} < p_{i_3} < p_{i_4}$ but i_1, i_2, i_3, i_4 are not necessarily consecutive. To show that F_{ST} is quartet consistent with T, it suffices to show that the quartet displayed by T for populations $\{i_1, i_2, i_3, i_4\}$ has a larger value of w than either of the alternative quartets possible for the four populations.

By construction, T restricted to $\{i_1, i_2, i_3, i_4\}$ gives quartet $((i_1, i_2), (i_3, i_4))$. By Lemma 2, given $\{i_1, i_2, i_3, i_4\}$ with $i_1 < i_2 < i_3 < i_4$, using F_{ST} for the dissimilarity d, $w_d(i_1i_2 : i_3i_4) > \max[w_d(i_1i_3 : i_2i_4), w_d(i_1i_4 : i_2i_3)]$. Hence, F_{ST} is quartet consistent with T, and by Theorem 9 of Mihaescu et al. (2009), NJ applied to the F_{ST} dissimilarity matrix produces tree T. □

By the proposition, despite the fact that F_{ST} matrices cannot be perfectly represented on a tree, for $4 \leqslant n \leqslant 7$ populations, NJ applied to F_{ST} places populations with neighboring allele frequencies in adjacent positions on the tree. The proof requires $n \leqslant 7$ in applying Theorem 9 of Mihaescu et al. (2009). According to that proposition, for $4 \leqslant n \leqslant 7$, exhibiting T with which a dissimilarity d is quartet consistent suffices to determine the NJ output topology. However, for $n > 7$, quartet consistency does not suffice: it is possible to identify a tree T with which the dissimilarity matrix d is quartet consistent but that has a different topology than the tree produced by NJ.

6.1.3. *Estimators of F_{ST}.* Our results thus far have examined values of F_{ST} assuming that they are computed from true population allele frequencies. We can also examine the relationship of F_{ST} to the triangle inequality for an estimator $\hat{F}_{ST}$.

For a biallelic locus in K populations with equal sample of size n diploid individuals, the Weir–Cockerham estimator (Weir and Cockerham 1984) is computed according to

$$\hat{\theta} = \frac{s^2 - \frac{1}{2n-1}\left[\hat{M}(1-\hat{M}) - \frac{K-1}{K}s^2)\right]}{\hat{M}(1-\hat{M}) + \frac{s^2}{K}},$$

where $\hat{M} = \frac{1}{K}\sum_{k=1}^{K}\hat{p}_k$ and $s^2 = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{p}_k - \hat{M})^2$ (Weir 1996, p. 173). For pairwise F_{ST}, with $K = 2$, this expression simplifies to:

$$\hat{\theta} = \frac{2(\hat{p}_1 - \hat{p}_2)^2 - \frac{1}{2n-1}\left[(\hat{p}_1 + \hat{p}_2)(2 - \hat{p}_1 - \hat{p}_2) - (\hat{p}_1 - \hat{p}_2)^2\right]}{(\hat{p}_1 + \hat{p}_2)(2 - \hat{p}_1 - \hat{p}_2) + (\hat{p}_1 - \hat{p}_2)^2}.$$

As $n \to \infty$, we have

$$\hat{\theta} \to \frac{2(\hat{p}_1 - \hat{p}_2)^2}{(\hat{p}_1 + \hat{p}_2)(2 - \hat{p}_1 - \hat{p}_2) + (\hat{p}_1 - \hat{p}_2)^2}. \tag{30}$$

By applying the formula for F_{ST} from eq. (3), we can rewrite this limit

$$\hat{\theta} = \frac{2F_{ST}(\hat{p}_1, \hat{p}_2)}{F_{ST}(\hat{p}_1, \hat{p}_2) + 1}. \tag{31}$$

To examine whether the large-sample limit of the estimator in eq. (30) satisfies the triangle inequality, we can consider the function

$$\hat{\psi}(\hat{p}_1, \hat{p}_2, \hat{p}_3) = \hat{\theta}(\hat{p}_1, \hat{p}_2) + \hat{\theta}(\hat{p}_2, \hat{p}_3) - \hat{\theta}(\hat{p}_1, \hat{p}_3). \tag{32}$$

Note that $2x/(x+1)$ is a monotonically increasing function for x in $[0, 1]$. Hence, using eq. (31), because $F_{ST}(p_1, p_3) > F_{ST}(p_1, p_2)$ and $F_{ST}(p_1, p_3) > F_{ST}(p_2, p_3)$ from eqs. (7) and (8), we have $\hat{\theta}(\hat{p}_1, \hat{p}_3) > \hat{\theta}(\hat{p}_1, \hat{p}_2)$ and $\hat{\theta}(\hat{p}_1, \hat{p}_3) > \hat{\theta}(\hat{p}_2, \hat{p}_3)$. It follows that $\hat{\theta}(\hat{p}_1, \hat{p}_2) + \hat{\theta}(\hat{p}_1, \hat{p}_3) > \hat{\theta}(\hat{p}_2, \hat{p}_3)$ and $\hat{\theta}(\hat{p}_1, \hat{p}_3) + \hat{\theta}(\hat{p}_2, \hat{p}_3) > \hat{\theta}(\hat{p}_1, \hat{p}_3)$. Consequently, $\hat{\theta}$ satisfies the triangle inequality for $(\hat{p}_1, \hat{p}_2, \hat{p}_3)$ if and only if $\hat{\psi}(\hat{p}_1, \hat{p}_2, \hat{p}_3) \geqslant 0$

Figure 6 shows that for most of the values for $(\hat{p}_1, \hat{p}_2, \hat{p}_3)$ shown, $\hat{\psi}$ lies below zero, and hence $\hat{\theta}$ fails the triangle inequality. We did not find any values with $\hat{\psi}(\hat{p}_1, \hat{p}_2, \hat{p}_3) > 0$. However, we did observe $\hat{\psi} = 0$ for some values with mutually distinct $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_3$, meaning that at some triples of distinct allele frequencies, $\hat{\psi}$ satisfies the triangle inequality. In particular, in Wright's counterexample, $\hat{\psi}(0, \frac{1}{2}, 1) = 0$. This case illustrates that the minimum of $\hat{\psi}$ does not occur in the same place as the minimum for ψ. The minimum for $\hat{\psi}$ is not as far below zero as the corresponding minimum for ψ. From this large-sample analysis, we can conclude that the deviation from the triangle inequality is potentially not as great for the Weir–Cockerham estimator as it is for parametric F_{ST}.

6.2. **Conclusions.** We have seen that in allele frequencies from three human populations, the frequencies sometimes lie in parts of the allele frequency space in which the deviation is fairly large. Although we have considered sets of only small numbers of populations, relationships in a larger set of populations are constrained by features of relationships in smaller subsets, so that the results based on 3 and 4 populations that MDS and NJ representations do not perfectly represent F_{ST} matrices apply to larger sets. In the case of neighbor-joining, because demonstrating quartet consistency of a dissimilarity matrix with a tree is not sufficient to obtain the inferred tree for $n > 7$ taxa, it remains to assess the precise relationship of F_{ST} matrices and NJ.

The failure of F_{ST} to satisfy the triangle inequality can often be mitigated in data analysis. First, in the human data, the loci at which the failures are most severe correspond to points with relatively large allele frequency differences and are relatively sparse in the genome. Second, in MDS analysis, transformations can be applied to data matrices to produce spatial representations that

more closely accord with the input matrix. Another solution is to use non-metric MDS, which is designed for input dissimilarity measures that are not necessarily metric; because the MDS visualization is affected by choices made in the analysis—both the version of the multidimensional scaling algorithm chosen and any associated transformations applied—it is desirable for these choices to be documented as part of the analysis (Jombart et al. 2009). Third, in NJ inference, although F_{ST} dissimilarity matrices cannot be perfectly recapitulated by a tree, we have found that NJ inference from F_{ST} produces predictable and intuitively sensible topologies for $n \leqslant 7$ taxa.

We note that we have only considered biallelic markers. Recall that the triangle inequality is satisfied for a distance between three populations if the sum of the distance between any two is greater than the third distance. Consider a modified version of ψ from eq. (12) for allele frequency vectors $\mathbf{p}$, $\mathbf{q}$, and $\mathbf{r}$ of a multiallelic marker:

$$\psi_{\text{multi}}(\mathbf{p},\mathbf{q},\mathbf{r}) = \min\{F_{ST}(\mathbf{p},\mathbf{q}) + F_{ST}(\mathbf{q},\mathbf{r}) - F_{ST}(\mathbf{p},\mathbf{r}), F_{ST}(\mathbf{p},\mathbf{r}) + F_{ST}(\mathbf{q},\mathbf{r}) - F_{ST}(\mathbf{p},\mathbf{q}),$$
$$F_{ST}(\mathbf{p},\mathbf{q}) + F_{ST}(\mathbf{p},\mathbf{r}) - F_{ST}(\mathbf{q},\mathbf{r})\},$$

where F_{ST} follows the general eq. (1) or (2). If $\psi_{\text{multi}} \geqslant 0$, then the triangle inequality is satisfied and if $\psi_{\text{multi}} < 0$, then it fails. Taking two examples of $(\mathbf{p},\mathbf{q},\mathbf{r})$, we have $\psi_{\text{multi}}((0,0,1),(1,0,0),(0,1,0)) = 1$ and $\psi_{\text{multi}}((0,0,1),(1,0,0),(\frac{1}{4},\frac{1}{2},\frac{1}{4})) = -\frac{3}{17}$. Thus, for multiallelic markers, F_{ST} sometimes satisfies the triangle inequality and sometimes does not. The multiallelic case does not have a result as simple as the biallelic result that the triangle inequality is never satisfied for distinct allele frequency vectors, and merits a more detailed analysis.

Sewall Wright's use of a counterexample to demonstrate the failure of the triangle inequality for F_{ST} has suggested a broader investigation of the nature of F_{ST} dissimilarity matrices. The results illustrate that even fundamental statistics such as F_{ST} and simple properties such as the triangle inequality continue to permit rich mathematical analysis.

Acknowledgments. We thank Jonathan Kang for assistance with the SNP genotypes. Support was provided by NIH grants R01 GM117590, R01 GM131404, and R01 HG005855.

References

Atteson, K. (1999). The performance of neighbor-joining methods for phylogenetic reconstruction. *Algorithmica 25*, 251–278.

Bosch, E., F. Calafell, A. Pérez-Lezaun, J. Clarimón, D. Comas, E. Mateu, R. Martínez-Arias, B. Morera, Z. Brakez, O. Akhayat, A. Sefiani, G. Hariti, A. Cambon-Thomsen, and J. Bertranpetit (2000). Genetic structure of north-west africa revealed by STR analysis. *European Journal of Human Genetics 8*, 360–366.

Buneman, P. (1974). A note on the metric properties of trees. *Journal of Combinatorial Theory B 17*, 48–50.

Cailliez, F. (1983). The analytical solution of the additive constant problem. *Psychometrika 48*, 305–308.

Cox, T. F. and M. A. A. Cox (2001). *Multidimensional Scaling*. Boca Raton: Chapman & Hall/CRC.

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

International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature 467*, 52–58.

Jombart, T., D. Pontier, and A.-B. Dufour (2009). Genetic markers in the playground of multivariate analysis. *Heredity 102*, 330–341.

Jorde, L. B. (1985). Human genetic distance studies: present status and future prospects. *Annual Review of Anthropology 14*, 343–373.

Kang, J. T. L., A. Goldberg, M. D. Edge, D. M. Behar, and N. A. Rosenberg (2016). Consanguinity rates predict long runs of homozygosity in Jewish populations. *Human Heredity 82*, 87–102.

Legendre, P. and L. Legendre (1998). *Numerical Ecology* (2nd ed.). Amsterdam: Elsevier.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science 319*, 1100–1104.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Amsterdam: Academic Press.

Mihaescu, R., D. Levy, and L. Pachter (2009). Why neighbor-joining works. *Algorithmica 54*, 1–24.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences USA 70*, 3321–3323.

Pérez-Lezaun, A., F. Calafell, E. Mateu, D. Comas, R. Ruiz-Pacheco, and J. Bertranpetit (1997). Microsatellite variation and the differentiation of modern humans. *Human Genetics 99*, 1–7.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics 1*, 660–671.

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

Steel, M. (2016). *Phylogeny: Discrete and Random Processes in Evolution*. Philadelphia: Society for Industrial and Applied Mathematics.

Studier, J. A. and K. J. Keppler (1988). A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution 5*, 729–731.

Takezaki, N. and M. Nei (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics 144*, 389–399.

Verdu, P., T. J. Pemberton, R. Laurent, B. M. Kemp, A. Gonzalez-Oliver, C. Gorodezky, C. E. Hughes, M. R. Shattuck, B. Petzelt, J. Mitchell, H. Harry, T. William, R. Worl, J. S. Cybulski, N. A. Rosenberg, and R. S. Malhi (2014). Patterns of admixture and population structure in native populations of northwest North America. *PLoS Genetics 10*, e1004530.

Wang, C., Z. A. Szpiech, J. H. Degnan, M. Jakobsson, T. J. Pemberton, J. A. Hardy, A. B. Singleton, and N. A. Rosenberg (2010). Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Statistical Applications in Genetics and Molecular Biology 9*, 13.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sunderland, MA: Sinauer.

Weir, B. S. and C. C. Cockerham (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution 38*, 1358–1370.

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

Wright, S. (1978). *Evolution and the Genetics of Populations Volume 4*. Chicago: University of Chicago Press.

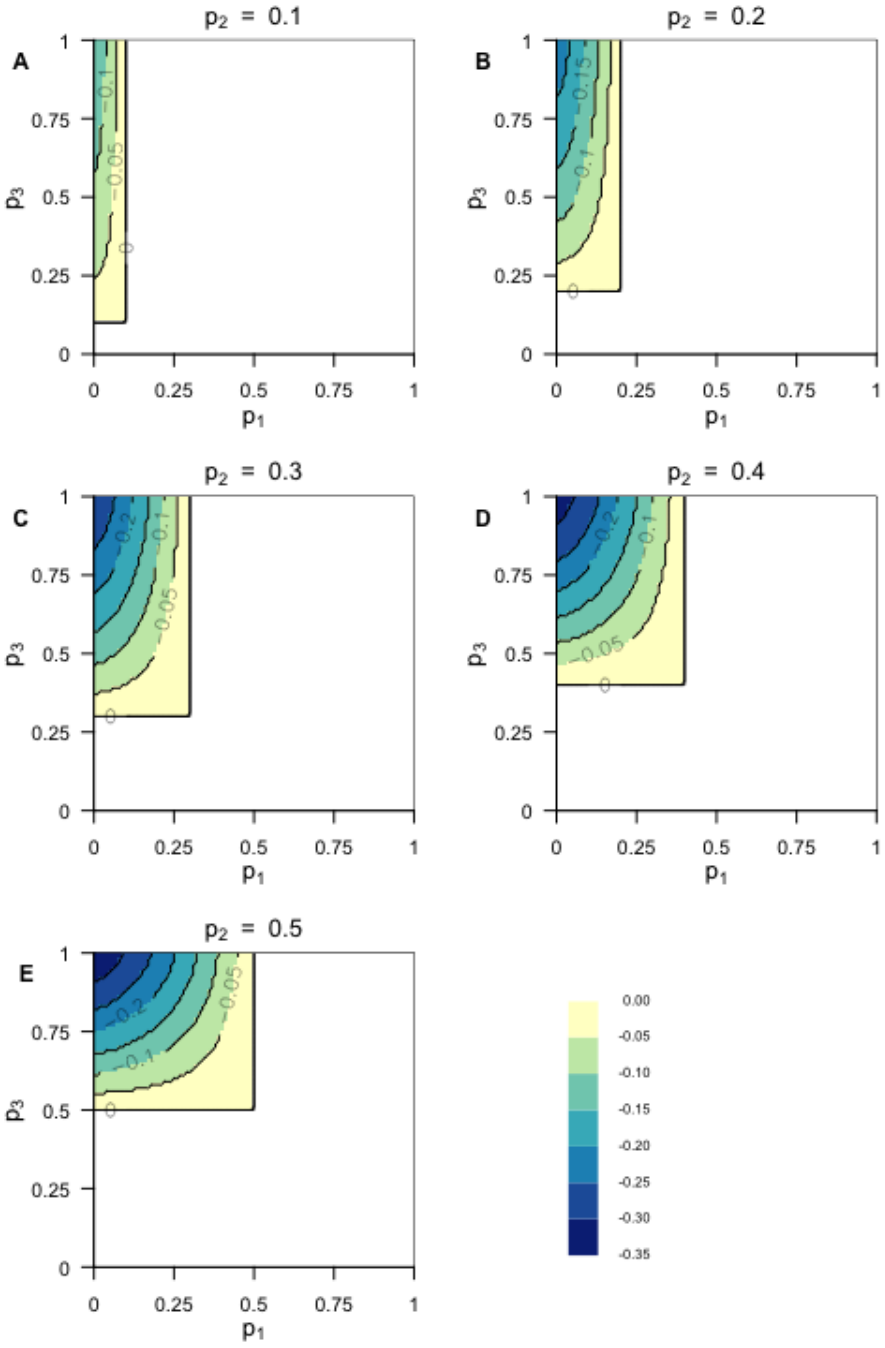

FIGURE 1. The extent to which the triangle inequality is violated, $\psi(p_1, p_2, p_3)$, for constant p_2 (eq. (12)). (A) $p_2 = 0.1$. (B) $p_2 = 0.2$. (C) $p_2 = 0.3$. (D) $p_2 = 0.4$. (E) $p_2 = 0.5$. The color at a point represents the value of ψ at that point. Because we define $p_1 \leqslant p_2 \leqslant p_3$, the bottom and right portions of the graph are empty. Because of the symmetry of ψ with respect to choice of allele, so that $\psi(p_1, p_2, p_3) = \psi(1 - p_3, 1 - p_2, 1 - p_1)$ (eq. (19)), we show plots only for $p_2 \leqslant 0.5$.

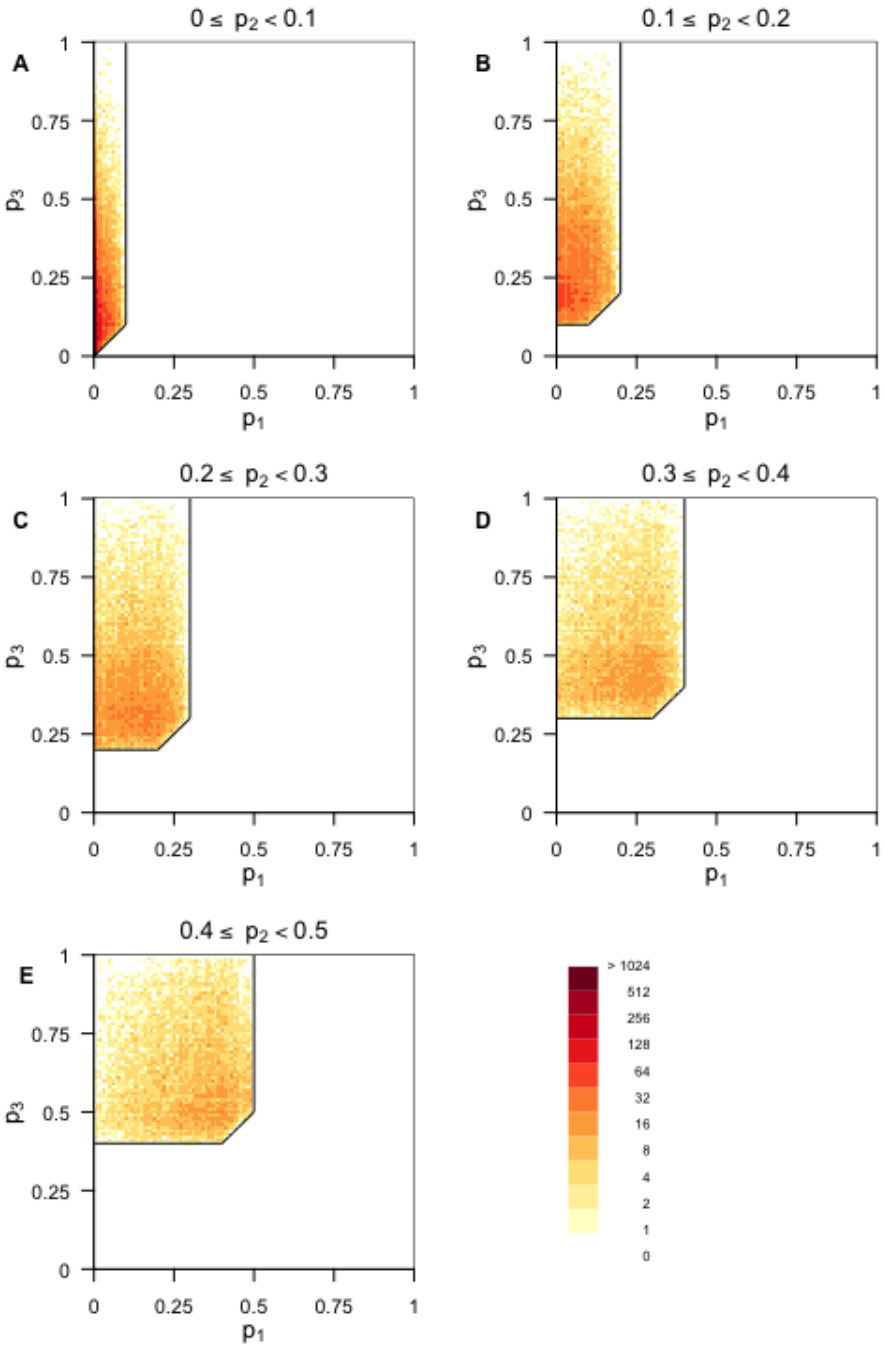

FIGURE 2. Two-dimensional histogram of population allele frequencies for three human populations: CHB (p_1), CEU (p_2), and YRI (p_3). The plots consider 185,522 loci for which $p_1 \leqslant p_2 \leqslant p_3$ when alleles are polarized by the frequency in CEU so that $p_2 \leqslant 0.5$. (A) $0 \leqslant p_2 \leqslant 0.1$, (B) $0.1 < p_2 \leqslant 0.2$, (C) $0.2 < p_2 \leqslant 0.3$, (D) $0.3 < p_2 \leqslant 0.4$, (E) $0.4 < p_2 \leqslant 0.5$. The color of the box represents the number of points present, on a logarithmic scale. Because we define $p_1 \leqslant p_2 \leqslant p_3$, the domain requires $p_1 \leqslant p_2$, $p_2 \leqslant p_3$, and $p_1 \leqslant p_3$. Because $p_2 \leqslant 0.5$ is required, the upper right regions of the graph are empty. The notches in the nearly rectangular domains arise from the requirement that $p_1 \leqslant p_3$.

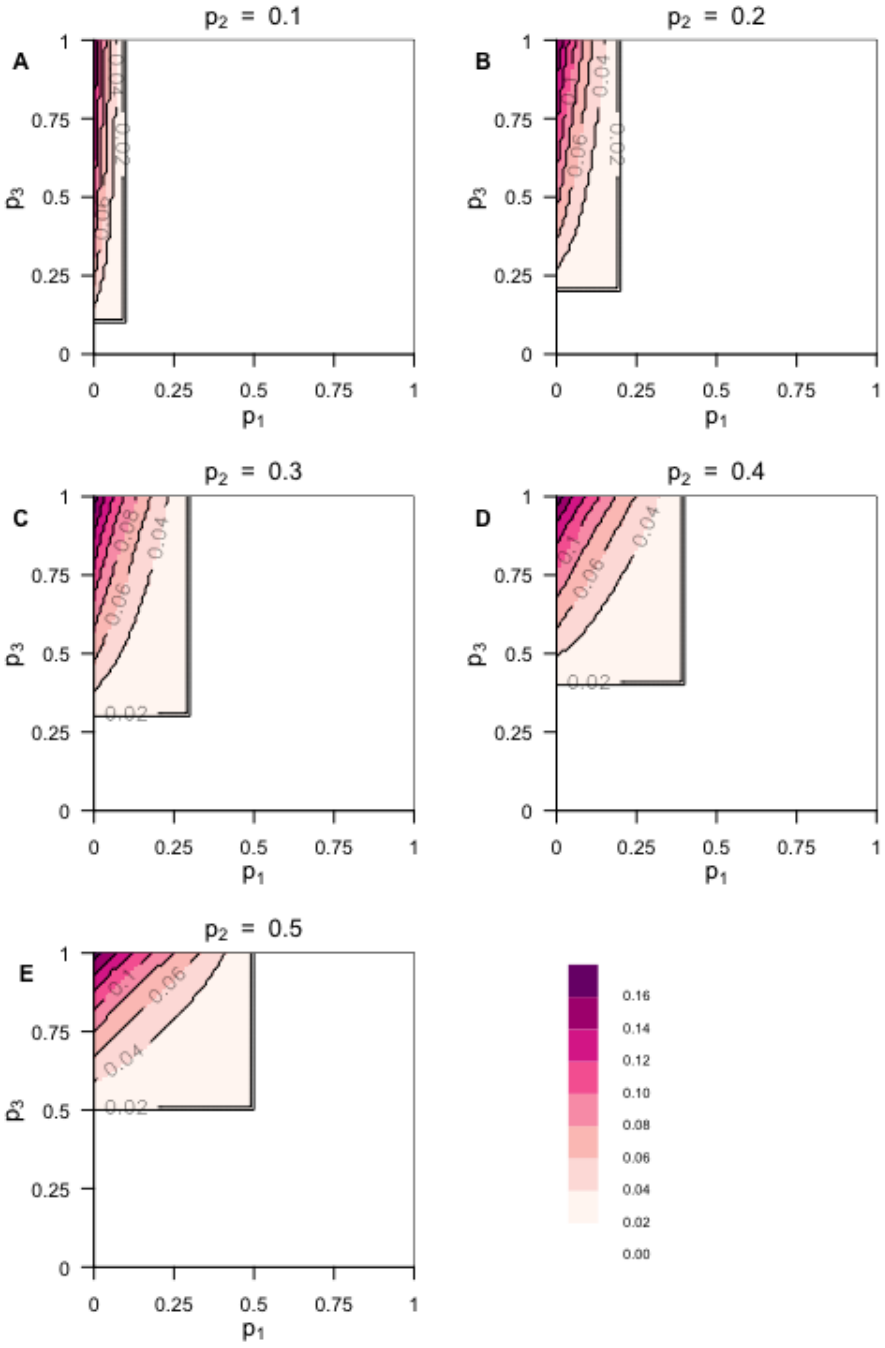

FIGURE 3. Contour plot of the relationship to the triangle inequality of the square root transformation of F_{ST}. The function plotted is $\tilde{\psi}(p_1, p_2, p_3) = \sqrt{F_{ST}(p_1, p_2)} + \sqrt{F_{ST}(p_2, p_3)} - \sqrt{F_{ST}(p_1, p_3)}$. (A) $p_2 = 0.1$. (B) $p_2 = 0.2$. (C) $p_2 = 0.3$. (D) $p_2 = 0.4$. (E) $p_2 = 0.5$. The color at a point represents the value of $\tilde{\psi}$ at that point. Because we define $p_1 \leqslant p_2 \leqslant p_3$, the bottom and right portions of the graph are empty. Because of the symmetry of $\tilde{\psi}$ with respect to choice of allele, so that $\tilde{\psi}(p_1, p_2, p_3) = \tilde{\psi}(1 - p_3, 1 - p_2, 1 - p_1)$ (eq. (19)), we show plots only for $p_2 \leqslant 0.5$. For all plotted values of $0 \leqslant p_1 \leqslant p_2 \leqslant p_3 \leqslant 1$, $\tilde{\psi}(p_1, p_2, p_3) \geqslant 0$.

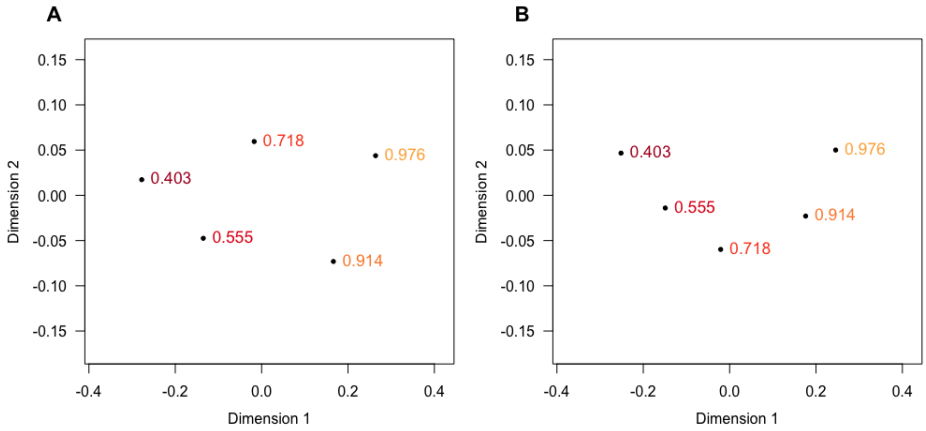

FIGURE 4. Classical multidimensional scaling (MDS) applied to transformed F_{ST} dissimilarity matrices. Five allele frequencies were chosen independently at random from a uniform-$[0, 1]$ distribution and an F_{ST} matrix was calculated. (A) A Cailliez constant is added to all non-diagonal elements of the F_{ST} matrix. (B) The entries in the distance matrix are the square root of the pairwise F_{ST} values. The MDS output was obtained by `cmdscale` in R. Using the `procrustes` function in the R `vegan` package, a Procrustes analysis was used to optimally rotate the MDS output from the square-root-transformed matrix to obtain the best alignment with the output from the Caillez-transformed matrix.

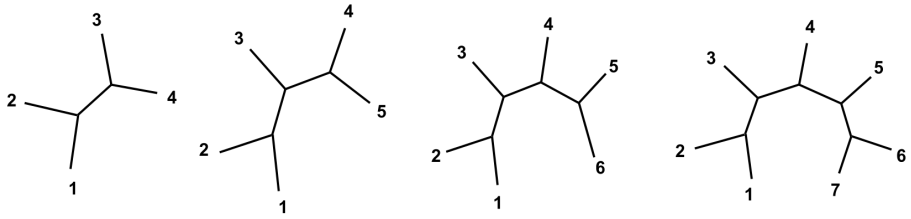

FIGURE 5. Neighbor-joining topologies for F_{ST} dissimilarity matrices with $n = 4$, 5, 6, and 7 taxa. The two cherries are $(1, 2)$ and $(n - 1, n)$, and all other leaves are placed sequentially along the path connecting the cherries.

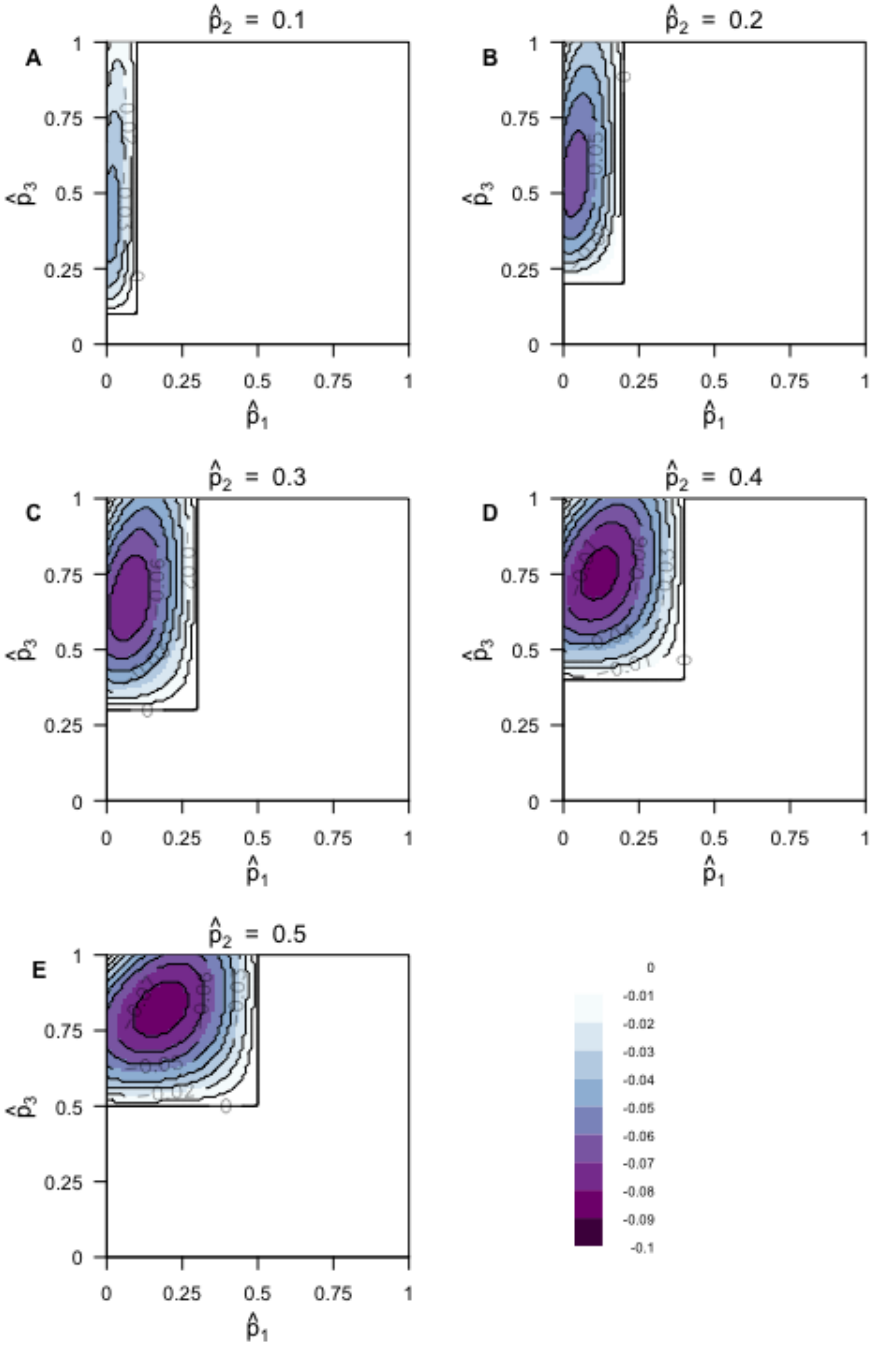

FIGURE 6. The extent to which the triangle inequality is violated for the estimator of F_{ST} as $n \to \infty$ (eq. (31)), $\hat{\psi}(\hat{p}_1, \hat{p}_2, \hat{p}_3)$, for constant $\hat{p}_2$ (eq. (32)). (A) $\hat{p}_2 = 0.1$. (B) $\hat{p}_2 = 0.2$. (C) $\hat{p}_2 = 0.3$. (D) $\hat{p}_2 = 0.4$. (E) $\hat{p}_2 = 0.5$. The color at a point represents the value of $\hat{\psi}$ at that point. Because we define $\hat{p}_1 \leqslant \hat{p}_2 \leqslant \hat{p}_3$, the bottom and right portions of the graph are empty. Because of the symmetry of $\hat{\psi}$ with respect to choice of allele, so that $\hat{\psi}(\hat{p}_1, \hat{p}_2, \hat{p}_3) = \hat{\psi}(1 - \hat{p}_3, 1 - \hat{p}_2, 1 - \hat{p}_1)$ (eq. (19)), we show plots only for $\hat{p}_2 \leqslant 0.5$.