# The Relationship Between Haplotype-Based $F_{ST}$ and Haplotype Length

**Rohan S. Mehta,*,[1] Alison F. Feder,*,[†] Simina M. Boca,[‡] and Noah A. Rosenberg***
*Department of Biology, Stanford University, Stanford, California 94305, [†]Department of Integrative Biology, University of California, Berkeley, California 94720, and [‡]Innovation Center for Biomedical Informatics, Georgetown University, Washington, DC 20007
ORCID IDs: 0000-0002-6244-9968 (R.S.M.); 0000-0003-2915-089X (A.F.F.); 0000-0002-1400-3398 (S.M.B.)

**ABSTRACT** The population-genetic statistic $F_{ST}$ is used widely to describe allele frequency distributions in subdivided populations. The increasing availability of DNA sequence data has recently enabled computations of $F_{ST}$ from sequence-based "haplotype loci." At the same time, theoretical work has revealed that $F_{ST}$ has a strong dependence on the underlying genetic diversity of a locus from which it is computed, with high diversity constraining values of $F_{ST}$ to be low. In the case of haplotype loci, for which two haplotypes that are distinct over a specified length along a chromosome are treated as distinct alleles, genetic diversity is influenced by haplotype length: longer haplotype loci have the potential for greater genetic diversity. Here, we study the dependence of $F_{ST}$ on haplotype length. Using a model in which a haplotype locus is sequentially incremented by one biallelic locus at a time, we show that increasing the length of the haplotype locus can either increase or decrease the value of $F_{ST}$, and usually decreases it. We compute $F_{ST}$ on haplotype loci in human populations, finding a close correspondence between the observed values and our theoretical predictions. We conclude that effects of haplotype length are valuable to consider when interpreting $F_{ST}$ calculated on haplotypic data.

**KEYWORDS** haplotypes; linkage disequilibrium; population structure; SNPs

THE quantity $F_{ST}$ has seen broad usage in studies of population structure and divergence (Holsinger and Weir 2009). Wright (1951) originally formulated $F_{ST}$ for a biallelic locus; subsequent perspectives that accommodate more than two alleles (Nei 1973) have enabled its computation on multiallelic loci such as microsatellites and haplotype loci.

Calculations of $F_{ST}$ from haplotypic data have provided insight into a variety of questions, especially following the development of a widely used haplotype-based statistical test for population subdivision (Hudson *et al.* 1992). Haplotypic computations of $F_{ST}$ have been useful for studying patterns of population structure, species divergence, and gene flow in numerous organisms (Hanson *et al.* 1996; Clark *et al.* 1998; Rocha *et al.* 2005; Jakobsson *et al.* 2008).


doi: https://doi.org/10.1534/genetics.119.302430
Manuscript received March 6, 2019; accepted for publication June 29, 2019; published Early Online July 8, 2019.
Supplemental material available at FigShare: https://doi.org/10.25386/genetics.8792594.
[1]Corresponding author: Department of Physics, Emory University, 201 Dowman Drive, Atlanta, GA 30322. E-mail: rohan.sushrut.mehta@emory.edu

$F_{ST}$ can be computed from haplotypic data in multiple ways. One method computes sequence differences for pairs of sequences from the same population and from different populations, and relies on a connection between $F_{ST}$, pairwise sequence differences, and coalescence times (Slatkin 1991; Hudson *et al.* 1992). Both this approach and the related analysis of molecular variance framework of Excoffier *et al.* (1992) rely on comparisons of sequences. A fundamentally different method employs a clustering technique to place distinct haplotypes into a set of haplotype clusters, regards the clusters of a sequence at a specified location as alleles, and computes $F_{ST}$ from cluster membership frequencies (Jakobsson *et al.* 2008; San Lucas *et al.* 2012). A third method treats a specific segment of the genome as a "haplotype locus," so that distinct haplotypes over that genomic segment represent distinct "haplotype alleles," and computes $F_{ST}$ from the haplotype alleles (Clark *et al.* 1998; Oleksyk *et al.* 2010).

This last approach, treating each distinct haplotype as its own distinct allele, provides a theoretical framework for understanding an observed dependence of $F_{ST}$ on haplotype length. Studies that have computed $F_{ST}$ using both individual single-nucleotide polymorphisms (SNPs) and haplotypes in the same data set have consistently observed that haplotype

$F_{ST}$ tends to be smaller than SNP $F_{ST}$ [Clark *et al.* 1998; Jakobsson *et al.* 2008 (Figure S29); Oleksyk *et al.* 2010; Sjöstrand *et al.* 2014 (Figure 2)]. An explanation for this basic pattern is suggested by the dependence of $F_{ST}$ on the frequency of the most frequent allelic type (Jakobsson *et al.* 2013; Edge and Rosenberg 2014; Alcala and Rosenberg 2017). A lower frequency for the most frequent type at a locus generally results in lower values of $F_{ST}$, and the most frequent haplotype at a particular haplotype locus is necessarily no more frequent than the most frequent SNP allele that it contains. We would then expect that because longer haplotype loci are likely to have a lower frequency for the most frequent haplotype, such loci would generate lower $F_{ST}$ values.

Here, we examine the effect of haplotype length on $F_{ST}$. We derive the value of $F_{ST}$ upon the addition of a biallelic SNP locus to an existing haplotype locus. Using this result, we predict the effect of haplotype length on values of $F_{ST}$, assuming for mathematical convenience that added SNPs are in linkage equilibrium with existing haplotype loci. Comparing values of $F_{ST}$ for haplotype loci in human genomic data to those obtained by our theoretical predictions, we find that our predictions largely match the observed values, despite the presence of linkage disequilibrium (LD) between the added SNPs and the existing haplotype loci in the data but not in the theory. In addition, we find that haplotype-based $F_{ST}$ computations are likely to reduce $F_{ST}$ compared to single-SNP $F_{ST}$ computations. We propose that a variety of haplotype lengths be used when computing $F_{ST}$ from haplotype loci and that the length of the haplotype locus be considered when interpreting the resulting $F_{ST}$ values.

## Model

### Definitions

We compute $F_{ST}$ on a multiallelic locus in a pair of populations, 1 and 2, of equal size. Denote by $p_{ki}$ the frequency of allele $i$ in population $k$, with $p_{ki} \geqslant 0$ for all $(k, i)$. For each $k$, $\sum_{i=1}^{I} p_{ki} = 1$, where $I$ is the total number of distinct alleles at the locus. We use Nei's (1973) formulation of $F_{ST}$,

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

where

$$J_S = \frac{1}{2}\sum_{k=1}^{2}\sum_{i=1}^{I} p_{ki}^2 \tag{2}$$

is the mean of the two population homozygosities, and

$$J_T = \sum_{i=1}^{I}\left[\frac{1}{2}\sum_{k=1}^{2} p_{ki}\right]^2 \tag{3}$$

is the homozygosity of the population obtained by pooling populations 1 and 2 together.

For $k = 1$ and $k = 2$, we define the population homozygosities by

$$J_k = \sum_{i=1}^{I} p_{ki}^2. \tag{4}$$

We define the dot product between the two population allele frequency vectors by

$$D_{12} = \sum_{i=1}^{I} p_{1i}p_{2i}. \tag{5}$$

Using Equations 4 and 5, we rewrite $F_{ST}$ (Equation 1) in the form that we use for our analysis:

$$F_{ST} = \frac{J_1 + J_2 - 2D_{12}}{4 - J_1 - J_2 - 2D_{12}}. \tag{6}$$

Note that a constraint exists on $D_{12}$ given $J_1$ and $J_2$:

$$0 \leqslant D_{12} \leqslant \sqrt{J_1 J_2}, \tag{7}$$

with equality in the upper bound if, and only if, each allele has the same frequency in both populations. Note that the upper bound is only achievable if $J_1 = J_2$ (see further discussion in Appendix A). The lower bound can be obtained by making each distinct allele unique to one of the two populations.

### Adding a SNP to a haplotype locus

We are concerned with the scenario in which the multiallelic locus is a "haplotype locus," a genomic region of specified length for which each distinct haplotype is regarded as a distinct "allele." We add a biallelic locus to our multiallelic locus, corresponding to a scenario in which the "haplotype locus" is augmented by one SNP. We refer to the multiallelic locus as a "haplotype locus," to each of its alleles as a "haplotype," and to the biallelic locus as a SNP. However, our results can apply to any kind of multiallelic locus augmented by a biallelic locus. We refer to the haplotype locus augmented by a SNP as an "extended haplotype locus."

Our goal is to compute $F_{ST}$ over the extended haplotype locus defined by adding the SNP to the haplotype locus, given the population frequencies of the alleles of the haplotype locus and the SNP. The SNP has two alleles, a major allele—with frequency greater than or equal to $\frac{1}{2}$—and a minor allele. We identify these alleles by examining the mean allele frequency between the two populations, so that the minor allele has mean frequency $\frac{1}{2}$ or less, even if it is the more common allele in one but not the other of the two populations.

The alleles of the extended haplotype locus are cooccurrences of the alleles of the SNP with the haplotypes of the haplotype locus. Each of the $I$ distinct haplotypes can cooccur with either the major or the minor allele of the SNP. Therefore, $2I$ alleles are possible for the extended haplotype locus as a result of combining the haplotype locus with the SNP. For each $i$ from 1 to $I$, we index the allele formed by cooccurrence of the $i$th haplotype with the SNP minor allele by $2i - 1$, and the allele

**Table 1 Haplotype and SNP allele frequency notation**

| | | Allele at the haplotype locus, population 1 | | | | Allele at the haplotype locus, population 2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | *I* | Total | 1 | 2 | *I* | Total |
| SNP allele | Minor | $p_{11}q_{11}$ | $p_{12}q_{12}$ | $p_{1I}q_{1I}$ | $q_1$ | $p_{21}q_{21}$ | $p_{22}q_{22}$ | $p_{2I}q_{2I}$ | $q_2$ |
| | Major | $p_{11}(1-q_{11})$ | $p_{12}(1-q_{12})$ | $p_{1I}(1-q_{1I})$ | $1-q_1$ | $p_{21}(1-q_{21})$ | $p_{22}(1-q_{22})$ | $p_{2I}(1-q_{2I})$ | $1-q_2$ |
| | Total | $p_{11}$ | $p_{12}$ | $p_{1I}$ | 1 | $p_{21}$ | $p_{22}$ | $p_{2I}$ | 1 |

Table entries represent allele frequencies of an extended haplotype locus (Equations 8 and 9). Columns for alleles 3, 4, ..., *I*-1 are omitted from the table.

formed by cooccurrence of the *i*th haplotype with the SNP major allele by 2*i*. Denote by $q_{ki}$ the frequency of the minor allele of the SNP on the *i*th haplotype in population *k*. In other words, $q_{ki}$ is the probability that haplotype *i* contains the minor allele of the SNP when augmented by the SNP. By a slight abuse of notation, using $p_{ki}$ for the frequency of allele *i* of the haplotype locus in population *k*, for each *i* from 1 to *I*, the allele frequencies of the extended haplotype locus in population *k* are

$$p_{k,2i-1} = p_{ki}q_{ki} \tag{8}$$

$$p_{k,2i} = p_{ki}(1-q_{ki}). \tag{9}$$

For convenience, we drop the comma in subscripts when possible.

Written with conditional probability, if *A* is the event that the SNP minor allele is observed and *B* is the event that haplotype *i* is observed, then cooccurrence of *A* and *B* has probability $P(A\cap B) = P(A|B)P(B)$. Equation 8 merely encodes this result, with $P(B) = p_{ki}$, $P(A\cap B) = p_{k,2i-1}$, and $P(A|B) = q_{ki}$. If $\overline{A}$ is the event that the major allele of the SNP is observed, then Equation 9 can be obtained by noting that $P(\overline{A}\cap B) = P(\overline{A}|B)P(B)$ and $P(\overline{A}\cap B) + P(A\cap B) = P(B)$, so that $P(\overline{A}|B) = P(\overline{A}\cap B)/P(B) = 1 - P(A\cap B)/P(B) = 1 - P(A|B) = 1 - q_{ki}$.

Note that $q_{ki}$ is not necessarily equal to the overall frequency of the SNP minor allele in population *k*, or $q_k$. The notation in Equations 8 and 9 allows us to write $q_k$ as

$$q_k = \sum_{i=1}^{I} p_{k,2i-1} = \sum_{i=1}^{I} p_{ki}q_{ki} \tag{10}$$

and the minor allele frequency of the SNP across all populations, *q*, as

$$q = \frac{1}{2}\sum_{k=1}^{2} q_k = \frac{1}{2}\sum_{k=1}^{2}\sum_{i=1}^{I} p_{ki}q_{ki}. \tag{11}$$

Table 1 summarizes our allele frequency notation and Figure 1 provides a schematic of the process of adding a SNP to a set of haplotypes.

## Results

### *General formula: arbitrary LD between haplotype locus and SNP*

We seek to evaluate $F_{ST}$ on the set of $2I$ alleles of the extended haplotype locus. We call this quantity $F_{ST}^{+}$. To compute $F_{ST}^{+}$ using Equation 6, we use Equations 8 and 9 to obtain the values of the component quantities $J_1^{+}$, $J_2^{+}$, and $D_{12}^{+}$ (Equations 4 and 5) for the extended haplotype locus:

$$\begin{aligned} J_k^{+} &= \sum_{i=1}^{I} p_{k,2i-1}^2 + p_{k,2i}^2 \\ &= \sum_{i=1}^{I} p_{ki}^2 q_{ki}^2 + p_{ki}^2(1-q_{ki})^2 \\ &= J_k - 2\sum_{i=1}^{I} p_{ki}^2 q_{ki}(1-q_{ki}) \end{aligned} \tag{12}$$

$$\begin{aligned} D_{12}^{+} &= \sum_{i=1}^{I} p_{1,2i-1}p_{2,2i-1} + p_{1,2i}p_{2,2i} \\ &= \sum_{i=1}^{I} (p_{1i}q_{1i})(p_{2i}q_{2i}) + [p_{1i}(1-q_{1i})][p_{2i}(1-q_{2i})] \\ &= D_{12} - \sum_{i=1}^{I} p_{1i}p_{2i}(q_{1i} + q_{2i} - 2q_{1i}q_{2i}). \end{aligned} \tag{13}$$

Addition of the SNP splits each haplotype into two new alleles, so homozygosity (Equation 12) cannot increase: $J_k^{+} \leq J_k$. For a fixed set of $p_{ki}$ for the haplotype locus in population *k*, equality can occur if and only if for all *i*, $q_{ki}$ is either 0 or 1. This condition is obtained if and only if each haplotype is associated with only a single SNP allele. Otherwise, adding a SNP always decreases homozygosity at the extended haplotype locus compared to the haplotype locus itself. Figure 2, A and B, provides geometric intuition for this result.

The dot product (Equation 13) also cannot increase, as $q_{1i} + q_{2i} - 2q_{1i}q_{2i} = q_{1i}(1-q_{2i}) + q_{2i}(1-q_{1i}) \geq 0$. Equality occurs if and only if: (1) for all *i*, $p_{ki} = 0$ for some *k*, or (2) for each *i*, $q_{1i}$ and $q_{2i}$ are both 0 or both 1. In the former case, the alleles of the haplotype locus are each private to a single population. In the latter case, the SNP is partitioned so that each haplotype is associated with a single SNP allele, the same one in both populations. Otherwise, adding the SNP decreases the dot product at the extended haplotype locus. Figure 2, C and D, provides geometric intuition for this result.

Note that if $q = 0$, so that $q_1 = q_2 = 0$, then $q_{1i} = q_{2i} = 0$ for all *i*. We then have $J_1^{+} = J_1, J_2^{+} = J_2$, and $D_{12}^{+} = D_{12}$. In this case, $F_{ST}^{+}$ is equal to the $F_{ST}$ for the initial haplotype locus (Equation 6). Thus, addition of a monomorphic locus does not change $F_{ST}$.

Because $F_{ST}$ (Equation 6) monotonically increases with $J_1 + J_2$, decreasing homozygosity decreases $F_{ST}$. In contrast, $F_{ST}$ monotonically decreases with $D_{12}$, so decreasing $D_{12}$

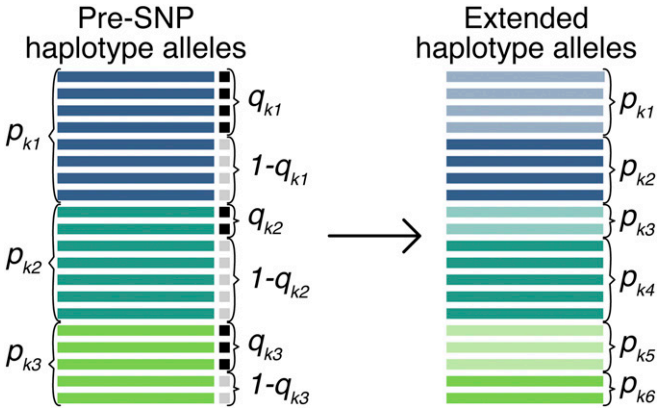


**Figure 1** Schematic of the process of creating an extended haplotype locus by adding a SNP to a set of existing haplotypes in a population $k$. Colors represent different haplotypes ($i = 1, 2, 3$), gray (major) and black (minor) represent the two SNP alleles, and color intensity in the right panel differentiates between the two extended haplotype alleles corresponding to a single haplotype allele prior to the addition of the SNP. Notation is defined in Table 1, updating the meaning of the $p_{ki}$ for the extended haplotype locus.

increases $F_{ST}$. Therefore, it is not immediately evident if modifying $J_1, J_2$, and $D_{12}$ in the manner of Equations 12 and 13 increases or decreases $F_{ST}$. Whether $F_{ST}$ increases or decreases with the addition of a SNP to a haplotype locus depends on whether the decrease in homozygosity (Equation 12) or the decrease in dot product (Equation 13) has a larger effect on Equation 6.

We can investigate the relative impact of the decreases in $J_1, J_2$, and $D_{12}$ on the value of $F_{ST}$ by using Equations 12 and 13 in Equation 6 to compute

$$F_{ST}^{+} = \frac{J_1 + J_2 - 2D_{12} - 2\sum_{i=1}^{I} x_i}{4 - J_1 - J_2 - 2D_{12} + 2\sum_{i=1}^{I} y_i}, \tag{14}$$

where

$$\begin{aligned} x_i &= (p_{1i}q_{1i} - p_{2i}q_{2i})[p_{1i}(1 - q_{1i}) - p_{2i}(1 - q_{2i})] \\ y_i &= (p_{1i}q_{1i} + p_{2i}q_{2i})[p_{1i}(1 - q_{1i}) + p_{2i}(1 - q_{2i})]. \end{aligned} \tag{15}$$

We now proceed to examine Equation 14 in the simplest case, in which the SNP and the haplotype locus are in linkage equilibrium separately in the two populations.

### *Special case: linkage equilibrium between haplotype locus and SNP*

We focus the remainder of our analysis on the situation in which the SNP is in linkage equilibrium with the haplotype locus. Under this condition of independence, the frequency of the minor allele of the SNP on a particular haplotype $i$ in population $k$, $q_{ki}$, is just the population frequency of the minor allele of the SNP in population $k$, $q_k$ (Equation 10).

Plugging $q_{ki} = q_k$ into Equations 12 and 13 yields

$$J_k^{+} = [1 - 2q_k(1 - q_k)]J_k \tag{16}$$

$$D_{12}^{+} = [1 - (q_1 + q_2 - 2q_1q_2)]D_{12}. \tag{17}$$

If we denote the homozygosity of the SNP in population $k$, $1 - 2q_k(1 - q_k), j_k$, and the dot product of the SNP allele frequency vectors in the two populations, $1 - (q_1 + q_2 - 2q_1q_2), d_{12}$, then we can write the quantities in Equations 16 and 17 by

$$J_k^{+} = j_k J_k \tag{18}$$

$$D_{12}^{+} = d_{12}D_{12}. \tag{19}$$

Using $J_k^{+}$ and $D_{12}^{+}$ from Equations 18 and 19 in Equation 6 yields the special case of Equation 14 in which the SNP is in linkage equilibrium with the haplotype locus:

$$F_{ST}^{+} = \frac{j_1J_1 + j_2J_2 - 2d_{12}D_{12}}{4 - j_1J_1 - j_2J_2 - 2d_{12}D_{12}}. \tag{20}$$

Thus, adding an independent SNP to a set of existing haplotypes amounts to multiplying the haplotype homozygosities and dot product by the SNP homozygosities and dot product, respectively, and recomputing $F_{ST}$ (Equation 6) using the resulting products. This result also holds if the appended locus has more than two alleles. The general case appears in Appendix B.

Figure 3 provides a schematic of the special case of adding a SNP to a set of haplotypes where the SNP and the haplotypes are in linkage equilibrium.

***Subcase: the SNP has the same minor allele frequency in the two populations:*** We now consider a series of further constraints on the alleles. First, we consider an independent SNP that is not differentiated between the two populations. This procedure is equivalent to taking all haplotypes and labeling them with two different labels in the same proportions in both populations. It might be expected to decrease $F_{ST}$, because within-population diversity increases but haplotypes are not split differently between the two populations.

If the SNP has identical minor allele frequency in the two populations, then $q_1 = q_2 = q$, with $0 \leqslant q \leqslant \frac{1}{2}$. Inserting $q_1 = q_2 = q$ into Equations 16 and 17 and applying Equation 6 yields

$$F_{ST}^{+}(q) = \frac{J_1 + J_2 - 2D_{12}}{\frac{4}{1 - 2q(1 - q)} - J_1 - J_2 - 2D_{12}}. \tag{21}$$

Equation 21 also follows from Equation 20, noting that for this case, $j_1 = j_2 = d_{12} = 1 - 2q(1 - q)$.

The constant 4 in the denominator of Equation 21 is divided by a quantity that is at most 1, with equality only in the monomorphic case of $q = 0$. Hence, the denominator of Equation 21 is always greater than or equal to that of Equation 6. Thus, the addition of a polymorphic SNP with the same minor allele frequency in the two populations always decreases $F_{ST}$.

The function in Equation 21 decreases monotonically with increasing minor allele frequency $q$ (Figure 4).

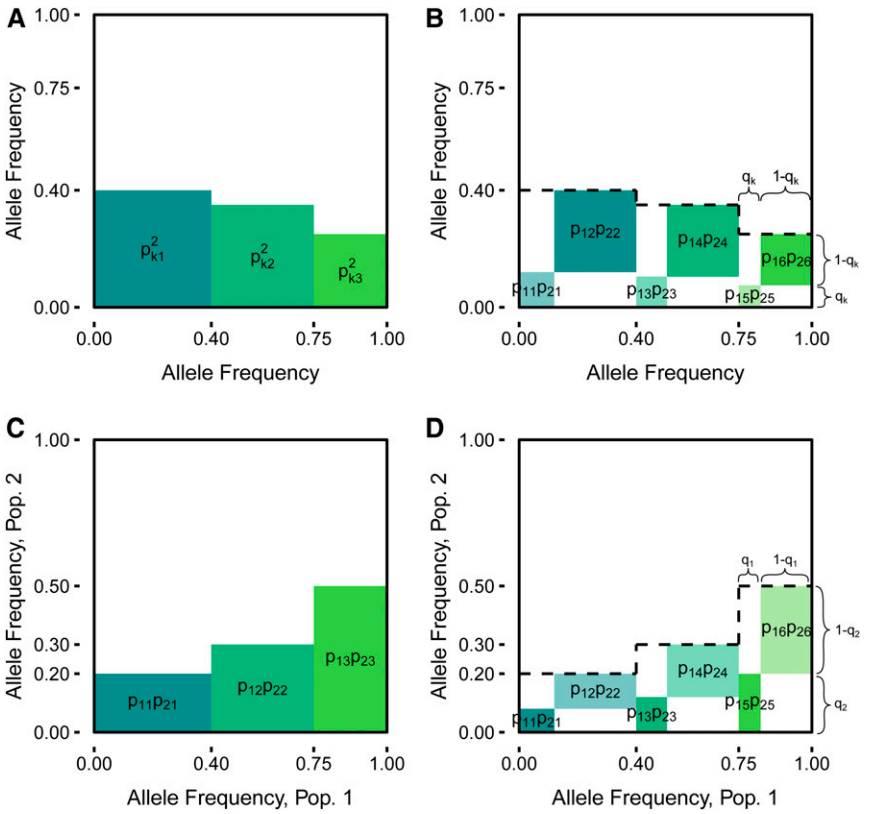


**Figure 2** The components of $F_{ST}$ (Equation 6) all decrease upon the addition of a SNP. (A) Homozygosity $J_k$ of a single population at a haplotype locus whose three haplotypes have frequencies $p_{k1} = 0.4, p_{k2} = 0.35$, and $p_{k3} = 0.25$. Homozygosity is represented geometrically by the total area of squares with side lengths $p_{ki}$ for $i = 1, 2, 3$. In this case, $J_k = 0.345$. (B) New homozygosity $J_k^+$ (Equation 12) upon addition of an independent SNP with $q_k = 0.3$. In this case, $J_k^+ = 0.1035$. (C) Dot product $D_{12}$ between two populations at a haplotype locus with $p_{11} = 0.4$, $p_{12} = 0.35$, and $p_{13} = 0.25$ as in (A) and (B), and $p_{21} = 0.2$, $p_{22} = 0.3$, and $p_{23} = 0.5$. The dot product $D_{12}$ is represented geometrically by the total area of rectangles with side lengths $p_{1i}$ and $p_{2i}$ for $i = 1, 2, 3$. In this case, $D_{12} = 0.31$. (D) New dot product $D_{12}^+$ (Equation 13) upon addition of an unlinked SNP with $q_1 = 0.3$ and $q_2 = 0.4$. In this case, $D_{12}^+ = 0.1674$. For all plots, the total shaded area equals the value of homozygosity (A and B) or the dot product (C and D). The dashed lines in (B) and (D) represent the boundaries of the solid areas in (A) and (C), respectively. Pop., population.

Considering all $q$, the maximal $F_{ST}$ occurs at $F_{ST}^+(0) = (J_1 + J_2 - 2D_{12})/(4 - J_1 - J_2 - 2D_{12})$ and the minimum occurs at $F_{ST}^+(\frac{1}{2}) = (J_1 + J_2 - 2D_{12})/(8 - J_1 - J_2 - 2D_{12})$.

***Subcase: the SNP minor allele occurs only in one population:*** We now consider the subcase in which the SNP minor allele is private to one population, assuming $q_1 = 0$ without loss of generality. The SNP splits some haplotypes into distinct new haplotypes in population 2 only, reducing allele sharing between populations. Therefore, unlike in the previous case in which adding a SNP always decreases $F_{ST}$, this case might be expected to increase $F_{ST}$.

Inserting $q_1 = 0$ and $q_2 = 2q$ into Equations 16 and 17, and applying Equation 6, yields

$$F_{ST}^+(q) = \frac{J_1 + [1 - 4q(1-2q)]J_2 - 2(1-2q)D_{12}}{4 - J_1 - [1 - 4q(1-2q)]J_2 - 2(1-2q)D_{12}}. \tag{22}$$

Equation 22 can also be derived from Equation 20, inserting $j_1 = 1, j_2 = 1 - 4q(1-2q)$, and $d_{12} = 1 - 2q$.

The influence on $F_{ST}^+$ (Equation 22) of the SNP minor allele frequency $q$ depends on the value of $D_{12}$. If $D_{12} = 0$, then the two populations share no haplotypes; they are maximally diverged at the haplotype locus. In this case, $F_{ST}^+$ becomes:

$$F_{ST}^+(q) = \frac{J_1 + [1 - 4q(1-2q)]J_2}{4 - J_1 - [1 - 4q(1-2q)]J_2}. \tag{23}$$

The function in Equation 23 is symmetric in $q$ across $q = \frac{1}{4}$, as for each $a$, $0 \leq a \leq \frac{1}{4}$, $F_{ST}^+(\frac{1}{4} + a) = F_{ST}^+(\frac{1}{4} - a) = [J_1 + (\frac{1}{2} + 8a^2)J_2]/[4 - J_1 - (\frac{1}{2} + 8a^2)J_2]$. It is minimized at $q = \frac{1}{4}$, and maximized at $q = 0$ and $q = \frac{1}{2}$ (Figure 5A). The maximum value is the value of haplotype $F_{ST}$ prior to the addition of a SNP and the minimum is $(J_1 + \frac{1}{2}J_2)/(4 - J_1 - \frac{1}{2}J_2)$. Thus, if the populations are maximally diverged at the haplotype locus in the sense that they share no haplotypes, then adding a SNP whose minor allele appears in only one population always decreases $F_{ST}$, with two exceptions. If the SNP is monomorphic in each population, with either $(q_1, q_2) = (0, 0)$ or $(q_1, q_2) = (0, 1)$, then the $F_{ST}$ value remains the same.

If $D_{12} > 0$ and we disregard the case of a monomorphic haplotype locus with $J_1 = J_2 = D_{12} = 1$, then the two populations share at least one haplotype and therefore admit the possibility of increased divergence through decreased allele sharing. To understand the effect of the minor allele frequency ($q$) on whether $F_{ST}$ increases or decreases, we examine the derivative of Equation 22 and assess the monotonicity of $F_{ST}^+$ with increasing $q$.

From Appendix C, for fixed $J_1$, $J_2$, and $D_{12}$, $F_{ST}^+(q)$ has a critical point in the permissible region for $q$ if and only if the root $q^*$ of the derivative $\frac{d}{dq}F_{ST}^+(q)$ satisfies $0 \leq q^* \leq \frac{1}{2}$, where

$$q^* = \frac{1}{2}\left(1 - \frac{1}{D_{12}} + \sqrt{\frac{1}{D_{12}^2} - \frac{1}{D_{12}} - \frac{2 - J_1 - J_2}{2J_2}}\right). \tag{24}$$

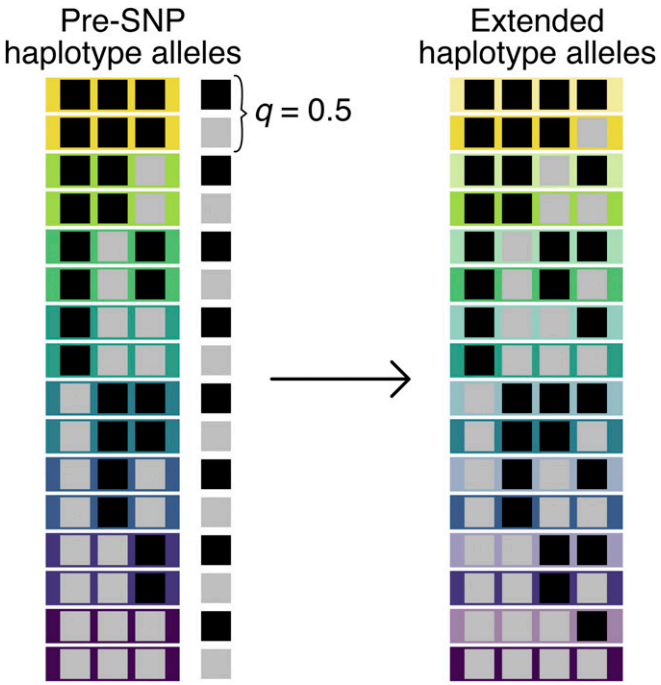


**Figure 3** Schematic of the process of creating an extended haplotype locus by adding a SNP to a set of existing haplotypes in a population, in the special case in which the SNP and haplotype alleles are in linkage equilibrium. Colors represent different haplotypes, gray and black represent the two SNP alleles, and color intensity in the right panel differentiates between the two extended haplotype alleles corresponding to a single haplotype allele in the left panel. The case shown here is specifically the situation described by Equation 28, in which haplotypes are constructed from SNPs that all have the same allele frequencies. In this case, the SNP minor allele has frequency $q = 0.5$.

We find that $q^* \geqslant 0$ if

$$D_{12} \leqslant \frac{2J_2}{2 - J_1 + J_2}, \tag{25}$$

and that $q^* \leqslant \frac{1}{2}$ if

$$\frac{1}{D_{12}} \geqslant \frac{J_1 + J_2 - 2}{2J_2}. \tag{26}$$

Equation 26 always holds, as its left-hand side is positive and its right-hand side is negative.

If Equation 25 holds, then we can see that the critical point $q^*$ is a local minimum: owing to Equation 25, at $q = 0$, the numerator of $\frac{d}{dq}F_{ST}^{+}(q)$ (Equation 39), and hence the derivative itself, is less than or equal to 0. Hence, if Equation 25 holds, then $F_{ST}$ decreases as $q$ increases from 0 to $q^*$ and increases as $q$ increases from $q^*$ to $\frac{1}{2}$. If Equation 25 fails, then the derivative has positive numerator at $q = 0$, and no critical points occur in $[0, \frac{1}{2}]$. $F_{ST}$ then increases with $q$ on $[0, \frac{1}{2}]$.

The behavior of Equation 22 as a function of $q$ appears in Figure 5. In Figure 5A, $J_1 = J_2 = 0.5$, and $D_{12}$ ranges over its permissible space from 0 to 0.5 (Equation 7). Equation 25 is always satisfied. As $D_{12}$ increases, allele sharing between populations increases, and the range of $q$ at which the

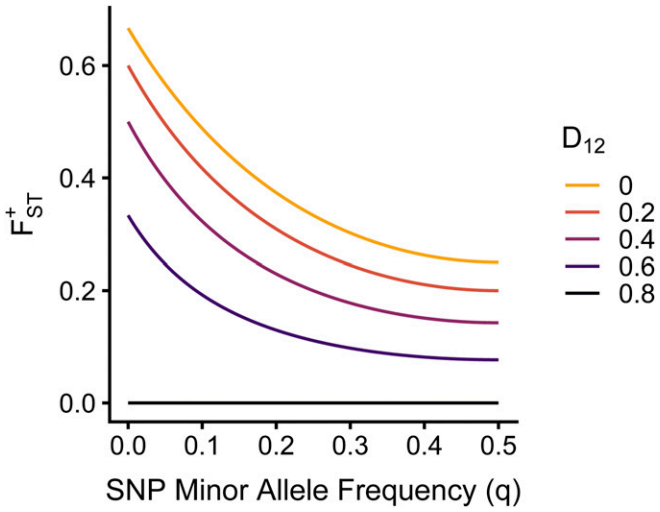


**Figure 4** $F_{ST}^{+}$ as a function of SNP minor allele frequency ($q$) for the case in which the SNP minor allele has the same frequency in both populations (Equation 21). The haplotypes have $J_1 = J_2 = 0.8$, with $D_{12}$ ranging from 0 to 0.8, leading to haplotype $F_{ST}$ values (represented in the plot by $q = 0$) ranging from 0.67 for $D_{12} = 0$ to 0 for $D_{12} = 0.8$. All values of $D_{12}$ in this range are permitted by Equation 7, as $J_1 = J_2$. $F_{ST}^{+}$ (Equation 21) decreases monotonically from the haplotype $F_{ST}$ at $q = 0$ to a minimum value at $q = 0.5$, except if haplotype $F_{ST}$ equals zero, in which case the SNP has no effect on $F_{ST}$.

population-specific SNP increases $F_{ST}$ by decreasing allele sharing expands in turn.

In Figure 5B, $J_1 = 0.5, D_{12} = 0.25$, and $J_2$ ranges from 0.2 to 1. Equation 7 is always satisfied for these values of $J_2$. Equation 25 is satisfied for all $J_2$ values considered, except 0.2. For the $J_1, J_2$, and $D_{12}$ shown, except at $J_2 = 0.2$, $F_{ST}^{+}$ (Equation 22) has a local minimum at $q^*$ (Equation 24). For $J_2 = 0.2$, Equation 25 is not satisfied, and $F_{ST}^{+}$ increases monotonically with increasing $q$. As $J_2$ increases from 0.2 to 1 for fixed $J_1 = 0.5$ and $D_{12} = 0.25$, the range of minor allele frequencies $q$ for which an added population-specific allele increases $F_{ST}$ gets smaller.

In summary, the effect of adding a private SNP depends on $q$. For large $q$, $F_{ST}$ increases. For small $q$, $F_{ST}$ only increases if the haplotype locus has large $D_{12}$ (Figure 5A) or if the population with the minor allele has low homozygosity at the haplotype locus (Figure 5B).

***Subcase: multiple SNPs with the same allele frequencies:*** The third subcase we consider is the construction of haplotypes from independent SNPs, with equivalent frequencies for all SNPs. Therefore, each SNP has the same values for $j_1$, $j_2$, and $d_{12}$. For one of these SNPs, the "haplotype" $F_{ST}$ is $(j_1 + j_2 - 2d_{12})/(4 - j_1 - j_2 - 2d_{12})$ (Equation 6). If we now add another independent SNP with the same properties, then using Equation 20, we obtain

$$F_{ST}^{+} = \frac{j_1^2 + j_2^2 - 2d_{12}^2}{4 - j_1^2 - j_2^2 - 2d_{12}^2}. \tag{27}$$

Figure 3 provides a schematic of this case for one of the populations $k$, considering a SNP with minor allele frequency

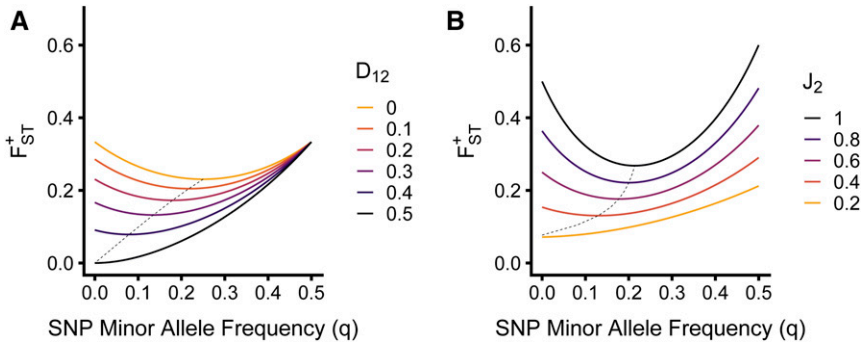


**Figure 5** $F_{ST}^{+}$ as a function of SNP minor allele frequency ($q$) for the case in which the SNP minor allele appears only in population 2 (Equation 22). (A) $J_1$ and $J_2$ are fixed and both equal 0.5. $D_{12}$ is varied from 0 to 0.5, leading to haplotype $F_{ST}$ values (occurring at $q = 0$) ranging from 0.33 to 0. All values of $D_{12}$ in this range are permitted by Equation 7, as $J_1 = J_2$. For all values of $D_{12}, F_{ST}^{+}$ (Equation 22) starts at the haplotype $F_{ST}$ at $q = 0$, then decreases to a minimum value at $q = q^*$ (Equation 24), then increases to a minimum value of $\frac{1}{3}$ at $q = \frac{1}{2}$. (B) $J_1$ and $D_{12}$ are fixed, with $J_1 = 0.5$ and $D_{12} = 0.25$. $J_2$ is varied from 0.2 to 1, leading to haplotype $F_{ST}$ values (occurring at $q = 0$) ranging from 0.07 to 0.5. If $J_1$ is fixed at 0.5, then $D_{12}$ must be less than $\sqrt{0.5J_2}$ unless $J_2$ also equals 0.5 (Equation 7). Setting $D_{12} = 0.25$ ensures $D_{12} < \sqrt{0.5J_2}$ holds for all $J_2 > 0.125$, which covers the range used here for $J_2$. The value of $J_2$ affects the shape of $F_{ST}^{+}$ (Equation 22); smaller values of $J_2$ result in monotonically increasing $F_{ST}^{+}$ with $q$, and larger values result in a decrease followed by an increase, as seen in (A). In both (A) and (B), the dashed line tracks the local minimum given by $q^*$ (Equation 24).

$q_k = 0.5$. By induction, $F_{ST}$ for the extended haplotype locus constructed by concatenation of $n$ independent SNPs with the same allele frequencies is

$$F_{ST}^{+n} = \frac{j_1^n + j_2^n - 2d_{12}^n}{4 - j_1^n - j_2^n - 2d_{12}^n}. \tag{28}$$

We plot Equation 28 as a function of $n$ with $j_1, j_2$, and $d_{12}$ fixed. In Figure 6A, $F_{ST}^{+n}$ appears as a function of $n$ for fixed $j_1$ and $j_2$ at each of several values of $d_{12}$. For each $d_{12}$, a decline occurs in $F_{ST}^{+n}$ with increasing $n$. Figure 6B plots $F_{ST}^{+n}$ as a function of $n$ for fixed $j_1$ and $d_{12}$ at each of several $j_2$ values. As in Figure 6A, for each $j_2$, $F_{ST}^{+n}$ decreases with increasing $n$.

One special case has $q_1 = 0$ and $j_1 = 1$, so that population 1 is monomorphic for all SNPs. The SNPs are polymorphic in population 2, with $q_2 > 0$. Then $j_1^n = 1, d_{12}^n = (1-q_2)^n$, and

$$F_{ST}^{+n} = \frac{1 + [1-2q_2(1-q_2)]^n - 2(1-q_2)^n}{4 - 1 - [1-2q_2(1-q_2)]^n - 2(1-q_2)^n} \to \frac{1}{3}, \tag{29}$$

with the limit taken as $n \to \infty$. The same limit occurs for $q_2 = 0$ and $q_1 > 0$ (Figure 6B, $j_2 = 1$). Otherwise, if both $q_1 > 0$ and $q_2 > 0$, then every term raised to the $n$th power in Equation 28 is less than 1, and $F_{ST}^{+n} \to 0$ as $n \to \infty$ (Figure 6).

We can conclude that if haplotypes are constructed by concatenating SNPs that all have the same allele frequencies, then $F_{ST}$ generally decreases with haplotype length. It has limit 0 in most cases and limit $\frac{1}{3}$ if one population is monomorphic for all SNPs.

## Application to data

To evaluate the empirical applicability of our theoretical results, we examined $F_{ST}$ calculated on human SNP haplotypes. We used phased SNP data from Pemberton *et al.* (2012); the data contain 938 individuals from 53 populations from the Human Genome Diversity Panel (HGDP), with a total of 640,034 genome-wide autosomal SNPs.

Our theoretical results are applicable to $F_{ST}$ calculated in pairs of populations. For this empirical application, we treated the seven geographical regions in the HGDP data set—Africa, Europe, Middle East, Central and South Asia, East Asia, Oceania, and America—as "populations." To obtain a set of haplotypes for a region, we pooled all sampled haplotypes from every individual in every population in that region.

### *Haplotype construction*

We constructed haplotypes from collections of $n$ SNPs obtained in two different ways, choosing windows of size $n_{\max} = 30$ SNPs. First, we drew 10,000 sets of $n_{\max}$ random SNPs without replacement from the entire set of SNPs, requiring all pairs of SNPs in a set to be separated by at least 5 Mb or to be located on different chromosomes. Each "haplotype" started with the first SNP in the set, and subsequent "haplotypes" were constructed by sequentially appending the remaining SNPs in the set.

The purpose of this first "random SNPs" procedure was to create "haplotypes" from SNPs that were not likely to be physically linked, a situation that accords with the assumptions of our theoretical computations. The value of $n_{\max} = 30$ SNPs was chosen to be large enough that most haplotypes in a data set were likely to be distinct: for instance, at $n = 30$, the first random SNP set for the Europe/East Asia pair had 607 unique haplotypes in a sample of size 774 (387 individuals). In this circumstance, $F_{ST}$ is effectively zero (Figure 7A). The distance threshold of 5 Mb was chosen to exceed the scale of tens to hundreds of kilobases for LD decay in humans (Patil *et al.* 2001; Gabriel *et al.* 2002; Wall and Pritchard 2003).

In our second "SNP window" approach for constructing haplotypes, we randomly chose 10,000 starting SNPs without replacement, each with at least $n_{\max} - 1$ SNPs between it and the chromosome end, as measured in order of increasing SNP position. Each haplotype started with the first SNP in the set, and subsequent haplotypes were constructed by sequentially appending remaining SNPs in the set. The purpose of this procedure was to test the theory on a situation in which the assumption of SNP independence is violated due to likely LD of neighboring SNPs.

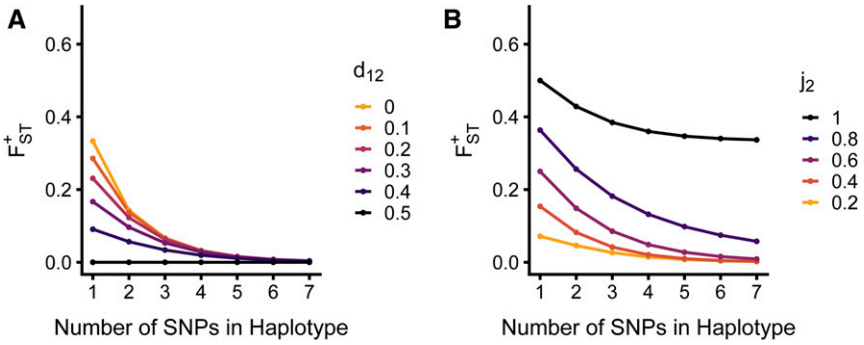


**Figure 6** $F_{ST}^{+n}$ as a function of $n$, the number of SNPs for the case in which all SNPs have the same allele frequencies (Equation 28). (A) All SNPs have $j_1 = j_2 = 0.5$, with $d_{12}$ ranging from 0 to 0.5, leading to SNP $F_{ST}$ values ranging from 0.33 to 0. All values of $d_{12}$ in this range are permitted by Equation 7. (B) All SNPs have $j_1 = 0.5$ and $d_{12} = 0.25$, with $j_2$ ranging from 0.2 to 1, leading to SNP $F_{ST}$ values ranging from 0.07 to 0.5. If $j_1$ is fixed at 0.5, then $d_{12}$ must be less than $\sqrt{0.5j_2}$ unless $j_2$ also equals 0.5 (Equation 7). Setting $d_{12} = 0.25$ ensures $d_{12} < \sqrt{0.5j_2}$ holds for all $j_2 > 0.125$, which covers the range used here. For both plots, $F_{ST}^{+}$ (Equation 28) decreases monotonically as the number of SNPs increases. For $j_2 < 1$, it decreases to 0.

### *General observations*

Figure 7A plots the observed $F_{ST}$ between Europe and East Asia, regions with relatively large samples in the data set—157 and 230 individuals, respectively—as a function of haplotype length. The $F_{ST}$ decay with haplotype length is faster for sets of random SNPs than for neighboring windows of SNPs. This result accords with the fact that LD in SNP windows maintains haplotype homozygosity over larger numbers of SNPs than in the case of the largely independent random SNP sets. We observe that the mean $F_{ST}$ across SNP windows is greatest for $n = 2$, after which it decays. This pattern accords with the claim that as haplotypes increase in length, haplotype homozygosity decreases and the maximal $F_{ST}$ in terms of homozygosity decreases, so that empirical $F_{ST}$ values decrease.

To evaluate the agreement of our theoretical results with observed $F_{ST}$ values, for each haplotype of length $n \geqslant 2$ SNPs, we used Equation 20 to compute a predicted $F_{ST}^{+}$ from the haplotype frequencies of the nested set of $n - 1$ SNPs and the allele frequencies of the $n$th SNP. The theoretical $F_{ST}^{+}$ produces the same qualitative decay with haplotype length and the same peak at a small number of SNPs $(n = 2)$ as was seen for the empirical values (Figure 7B).

For each SNP set and haplotype length, we computed the ratio of the difference between observed and theoretically predicted values of $F_{ST}$ and the theoretical value, a quantity we term "rescaled error." For a particular SNP set and haplotype length, rescaled error is:

$$R = \frac{F_{ST} - F_{ST}^{+}}{F_{ST}^{+}}. \tag{30}$$

Values of rescaled error (Equation 30) as a function of haplotype length for the SNP sets in Figure 7, A and B, appear in Figure 7C. The rescaled error is small for small $n$, increasing with $n$. Our theoretical predictions are therefore more accurate for short haplotypes. Owing to the generally low $F_{ST}$ values recorded for longer haplotypes (Figure 7A), the absolute magnitude of the poorer predictions for longer haplotypes is relatively small. For $2 \leqslant n \leqslant 14$, the prediction is more accurate for random SNP sets than for SNP windows. Interestingly, for $n \geqslant 15$, the prediction is instead more accurate for the neighboring SNP windows, despite the fact that the prediction is designed for SNP sets with no LD. This change in accuracy might be explained by the fact that SNP windows of a particular length produce $F_{ST}$ values similar to those of random SNP sets of smaller length (Figure 7A), so that our predictions remain reasonably accurate for longer SNP windows than in the case of random SNP sets.

### *Correlation between observations and theory*

To study the change in $F_{ST}$ as SNPs are added to a haplotype locus, we considered the value of $F_{ST}$ with increasing haplotype length for each collection of $n_{\max} = 30$ SNPs. For each collection of SNPs, random SNPs or SNP windows, we obtained a "trajectory" of $F_{ST}$: the values of $F_{ST}$ as a function of the number of SNPs used to construct haplotypes for each $n$ from 1 to $n_{\max}$. We then compared the observed $F_{ST}$ for haplotypes of length $n$ to the theoretical $F_{ST}^{+}$ obtained by using Equation 20 on the set of haplotypes with length $n - 1$ together with the $n$th SNP.

In each trajectory, we also compared the observed $F_{ST}$ for haplotypes of length $n$ to a value of $F_{ST}$ drawn with replacement from the set of all observed values of $F_{ST}$ for haplotypes of length $n$. These random draws were designed to serve as a null model of $F_{ST}$ as a function of haplotype length, where the value of $F_{ST}$ depends only on haplotype length without regard to values of $F_{ST}$ for previous entrants in the trajectory from $n = 2$ to $n = n_{\max}$.

Table 2 displays correlation coefficients between observed $F_{ST}$ values, and both theoretical values obtained from Equation 20 and null model values drawn from the empirical distribution of $F_{ST}$. The correlations are computed between sets of 290,000 sets of paired values, 10,000 SNP sets and 29 values per SNP set $(n = 2, 3, \ldots, 30)$. The value of $n = 1$ was not used because $F_{ST}^{+}$ in Equation 20 only applies for $n \geqslant 2$. The correlations between observed and theoretical values range from 0.96 to 1.00 for random SNP sets, and from 0.94 to 0.98 for SNP windows, compared to 0.24–0.47 and 0.07–0.23 for the correlation between observed and null values for random SNP sets and SNP windows, respectively.

Supplemental Material, Figure S1 plots representative results from Table 2 for the Europe/East Asia pair of regions.

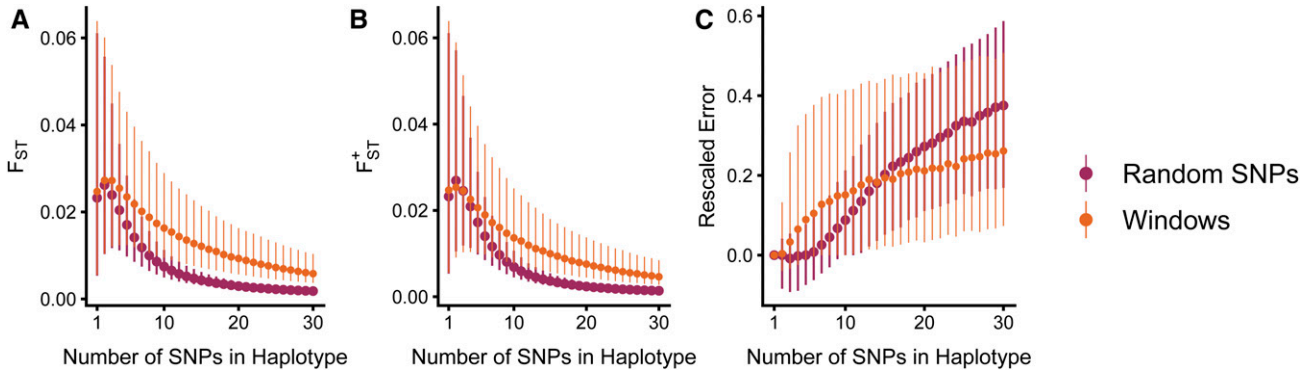


**Figure 7** $F_{ST}$ for collections of random SNPs and windows of neighboring SNPs, as a function of the number of SNPs considered. (A) Median observed $F_{ST}$. (B) Median theoretical $F^+_{ST}$. (C) Median rescaled error (Equation 30). The median is taken across 10,000 SNP sets. For $n \geqslant 2$ SNPs, the rescaled error is computed as the absolute difference between the observed $F_{ST}$ and the $F^+_{ST}$ predicted from Equation 20 with the allele frequencies of the $n$th SNP, and the values of $J_1, J_2$, and $D_{12}$ of the haplotype locus for the $n-1$ initial SNPs, normalized by the predicted $F^+_{ST}$. The plot considers as the two populations the data from Europe and East Asia. Error bars denote first to third quartiles, considering 10,000 SNP sets.

As expected, theoretical values of $F^+_{ST}$ match observed values more closely for random SNP sets than for SNP windows. However, the SNP windows produce results that are comparable to the random SNP results, indicating that our theoretical results are reasonable in situations in which the assumption of linkage equilibrium does not hold. For both methods of haplotype construction, the theoretical results dramatically outperform the null model results, indicating that the theory predicts substantial additional information about haplotype-based $F_{ST}$ compared with null predictions.

### *Trajectories as observations*

For each collection of $n_{\max} = 30$ SNPs, considering the 29 values from $n = 2$ to 30, we fit a linear regression of observed $F_{ST}$ on the theoretical prediction from Equation 20 and computed the corresponding $r^2$ statistic for goodness-of-fit. The purpose of this analysis was to treat each trajectory as a separate observation with its own $r^2$, in contrast to grouping them as in Table 2 and Figure S1.

For the Europe/East Asia pair, Figure S2 plots $r^2$ distributions across 10,000 trajectories for theoretical and null models, for both random SNPs and SNP windows. The fit of the theoretical values is substantially closer compared to that of the null values. The fit is also closer for random SNP trajectories compared to window trajectories (Figure S2).

Figure 8 displays the median $r^2$ trajectories for each category of result in Figure S2 for the Europe/East Asia pair. Figure 8 reveals a distinction between the null and theoretical results; the theoretical model (Figure 8, A and C) closely matches observations for shorter haplotypes but consistently underestimates the value of $F_{ST}$ for longer haplotypes. In contrast, the null model (Figure 8, B and D) produces a poor fit for shorter haplotypes but is less consistently biased for longer haplotypes. This observation provides more detail about the observation in Figure 7 that rescaled error (Equation 30) is higher for longer haplotypes than for shorter haplotypes; in particular, the longer-haplotype $F_{ST}$ is underestimated.

Figure 9 plots example trajectories as a function of the frequency $M$ of the most frequent haplotype instead of haplotype length, together with the upper bound on $F_{ST}$ given $M$ (Jakobsson *et al.* 2013). The haplotype locus starts with one SNP, with major allele frequency at least $\frac{1}{2}$. As more SNPs are added, $M$ either stays the same (if one SNP allele does not cooccur with the previous most frequent haplotype) or decreases (if both SNP alleles cooccur with the previous most frequent haplotype). Increasing haplotype length first increases the upper bound on $F_{ST}$, increasing the potential for an increase in $F_{ST}$ to occur upon addition of a SNP. Once $M$ decreases below $\frac{1}{2}$ increasing the haplotype length decreases the $F_{ST}$ upper bound, generally forcing $F_{ST}$ to decrease. In aggregate, these properties of the upper bound of $F_{ST}$ as a function of $M$ can explain the tendency of $F_{ST}$ to increase upon addition of the first few SNPs before decreasing with more SNPs, as seen in Figure 7A.

### *Error and LD*

We expected that the primary cause of deviation of observed values from theoretical values was greater LD in SNP windows than in random SNP sets. LD has been detected in these SNP data for nearby SNPs, decaying quickly so that it is unexpected for random SNP pairs [see Jakobsson *et al.* (2008), Figure 2 and Li *et al.* (2008), Figure 3].

To assess the effect of LD on rescaled error, Figure 10 plots rescaled error (Equation 30) against a multiallelic $D'$ measure of LD (Hedrick 1987) for European SNP–haplotype pairs. This quantity, which we term $D'_1$, measures the deviation of extended haplotype allele frequencies from linkage equilibrium, and is plotted for each SNP–haplotype pair. For each SNP set, for each $n$ from 2 to $n_{\max}$, we computed $D'$ between the haplotype locus of length $n-1$ and the SNP. For East Asia, we denote the quantity analogous to $D'_1$ in Europeans by $D'_2$.

Figure 10, A and B, which consider random SNP sets and SNP windows, respectively, are split by quartile of values of $D'_2$. Increasing LD in one or both populations increases the rescaled error. This pattern is clear for SNP windows (Figure

**Table 2 Correlations between theoretical and observed values of $F_{ST}$ upon the addition of a SNP to a set of haplotypes, compared to correlations between observed values with those produced by a null model**

| Region 1 | Region 2 | Random SNPs | | SNP windows | |
|---|---|---|---|---|---|
| | | Theoretical | Null | Theoretical | Null |
| Africa | Europe | 0.9930 | 0.4375 | 0.9685 | 0.2318 |
| Africa | Middle East | 0.9923 | 0.4251 | 0.9684 | 0.2321 |
| Africa | Central/South Asia | 0.9926 | 0.4289 | 0.9669 | 0.2340 |
| Africa | East Asia | 0.9948 | 0.4428 | 0.9727 | 0.2173 |
| Africa | Oceania | 0.9945 | 0.4399 | 0.9761 | 0.1642 |
| Africa | America | 0.9957 | 0.4699 | 0.9739 | 0.1898 |
| Europe | Middle East | 0.9691 | 0.2353 | 0.9429 | 0.0892 |
| Europe | Central/South Asia | 0.9823 | 0.2754 | 0.9578 | 0.1177 |
| Europe | East Asia | 0.9936 | 0.3786 | 0.9709 | 0.1596 |
| Europe | Oceania | 0.9921 | 0.3756 | 0.9741 | 0.0974 |
| Europe | America | 0.9930 | 0.3959 | 0.9713 | 0.1028 |
| Middle East | Central/South Asia | 0.9809 | 0.3059 | 0.9544 | 0.1315 |
| Middle East | East Asia | 0.9937 | 0.3900 | 0.9709 | 0.1639 |
| Middle East | Oceania | 0.9919 | 0.3881 | 0.9735 | 0.1017 |
| Middle East | America | 0.9934 | 0.4067 | 0.9708 | 0.1070 |
| Central/South Asia | East Asia | 0.9925 | 0.3636 | 0.9677 | 0.1400 |
| Central/South Asia | Oceania | 0.9911 | 0.3665 | 0.9731 | 0.0857 |
| Central/South Asia | America | 0.9921 | 0.3804 | 0.9700 | 0.0854 |
| East Asia | Oceania | 0.9926 | 0.3414 | 0.9756 | 0.0868 |
| East Asia | America | 0.9933 | 0.3384 | 0.9732 | 0.0749 |
| Oceania | America | 0.9952 | 0.3896 | 0.9765 | 0.0900 |

For this computation, 290,000 paired values are compared, as every haplotype length from 2 to 30 is considered for each of 10,000 random or neighboring window SNP sets.

10B), for which increasing $D_1'$ (within a plot) and $D_2'$ (moving left to right across plots) produce greater rescaled error. As LD increases, the model becomes less accurate, so that rescaled error increases.

The magnitude of the influence of LD on rescaled error is relatively small. When we separate SNP windows into quartiles by the physical distance between SNPs $n-1$ and $n$, representing four quartiles expected to have different LD levels, we see little difference among quartiles in the rescaled error (Figure S3).

### *Data availability*

See Pemberton *et al.* (2012) for the data used in this study. Supplemental material available at FigShare: https://doi.org/10.25386/genetics.8792594.

## Discussion

We have derived the value of $F_{ST}$ that is obtained when a haplotype locus is augmented by a SNP (Figure 1B), focusing on the situation in which the SNP is in linkage equilibrium with the haplotype locus. Three special cases we studied theoretically—a SNP with the same allele frequencies in both populations (Figure 4), a SNP whose minor allele appears only in one of the populations (Figure 5), and haplotype loci that are constructed from SNPs that all have the same allele frequencies (Figure 6)—suggest a general pattern: $F_{ST}$ is likely to decrease when a SNP is added to a haplotype locus, even if the SNP itself has a high value of $F_{ST}$. Our empirical results using human SNP data corroborate this conclusion (Figure 7A).

The relationship between $F_{ST}$ and the within-population homozygosities and dot product of allele frequencies between populations assists in understanding the effect on $F_{ST}$ of adding a SNP to a haplotype locus. $F_{ST}$ decreases both by a reduction in the within-population homozygosities and by an increase in the between-population allele sharing. Adding a SNP to a haplotype locus necessarily decreases homozygosities within populations by subdividing each allele of the haplotype locus. The addition of the SNP might or might not increase between-population allele sharing; if it does decrease allele sharing, then it might not do so sufficiently to overcome decreases in homozygosity, and $F_{ST}$ might still decrease. We have found that a decrease in allele sharing through differing SNP allele frequencies in the two populations only increases $F_{ST}$ compared to the haplotype locus if the SNP allele frequencies differ greatly between the two populations, the two populations are very similar in their frequencies at the haplotype locus, or they have high diversity at the haplotype locus.

In our $F_{ST}$ trajectories, as more SNPs are added to SNP windows, $F_{ST}$ approaches 0. Typically, the first few SNPs enable an increase in $F_{ST}$ as the frequency of the most frequent haplotype across the population pair decreases toward $\frac{1}{2}$, the value that permits the greatest $F_{ST}$ (Figure 9). With enough SNPs, the extended haplotype locus becomes too heterozygous within populations for any population divergence information to be gleaned from $F_{ST}$.

Because $F_{ST}$ has a systematic length dependence, a useful data analysis strategy is to not restrict attention to a single length and to report entire "profiles" of $F_{ST}$ in terms of haplotype length. For example, Figure S4 examines the

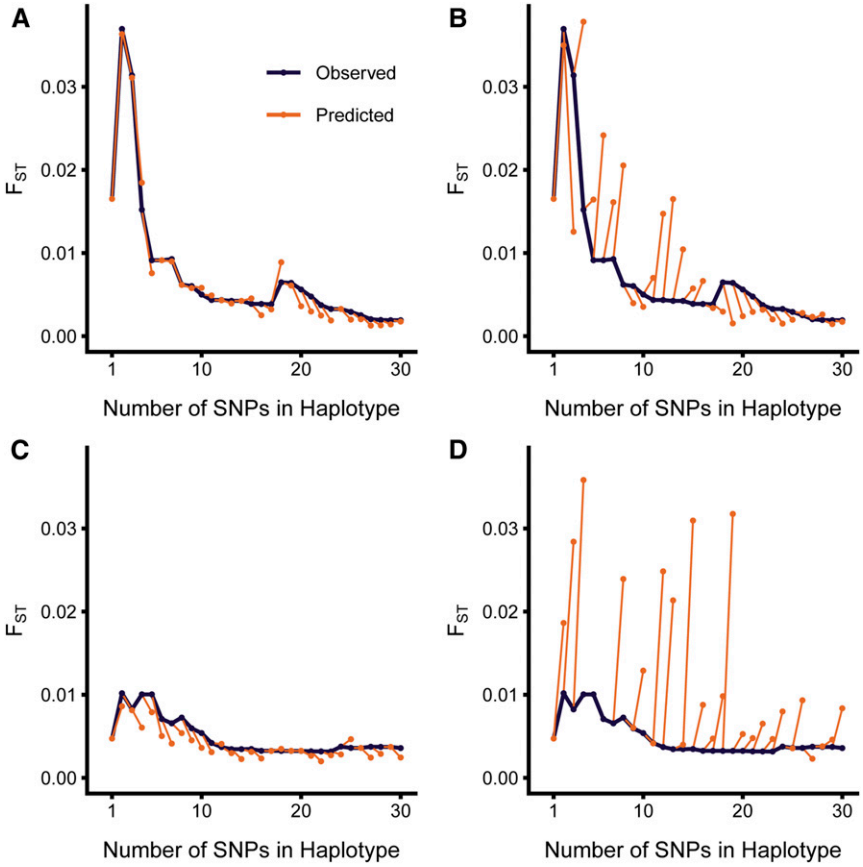


**Figure 8** Example trajectories of observed, theoretical, and null values of $F_{ST}$ for random SNP sets and SNP windows. (A) Random SNP sets, theory. (B) Random SNP sets, null model. (C) SNP windows, theory. (D) SNP windows, null model. For each number of SNPs $n$, $1 \leqslant n_{\max} - 1$, a prediction is made for $F_{ST}^{+}$ on the basis of a theoretical or null model. The prediction is indicated by an orange line from $(n, F_{ST})$ to $(n+1, F_{ST}^{+})$. The trajectories shown are those with median (5000th lowest) $r^2$ values in the observed *vs.* theoretical $F_{ST}$ comparison distributions that appear in Figure S2.

dependence of $F_{ST}$ on haplotype length for various population pairs. Some of the lines representing different comparisons cross, indicating that the length affects which of a pair of comparisons has a larger value. In other cases, lines have the same relative position irrespective of the length considered. If $F_{ST}$ profiles are computed for multiple population pairs, and the same pairs have larger values across multiple lengths, then relative values can potentially be regarded as robust.

This study augments recent attempts to analyze how population-genetic statistics change as the unit of analysis extends from a single SNP to a haplotype locus (*e.g.*, Morin *et al.* 2009; Gattepaille and Jakobsson 2012; Duforet-Frebourg *et al.* 2015; García-Fernández *et al.* 2018). In particular, our approach follows Gattepaille and Jakobsson (2012), who compared a statistic for ancestry information for two loci combined and treated as a single "haplotype locus" to the information content of the loci individually. We show how a two-locus framework can be used iteratively to examine haplotype loci on larger numbers of SNPs.

We have considered a particular form of $F_{ST}$, following recent work on the dependence of $F_{ST}$ on allele frequencies (Jakobsson *et al.* 2013; Edge and Rosenberg 2014; Alcala and Rosenberg 2017), by treating $F_{ST}$ as a function computed from allele frequencies rather than as a parameter of an evolutionary model. In our perspective, $F_{ST}$ values at different haplotype lengths are not expected to be equal, either numerically or conceptually. In an alternative and widely used perspective in which $F_{ST}$ is treated as an evolutionary parameter (*e.g.*, Holsinger and Weir 2009), haplotype loci of different lengths represent different scales

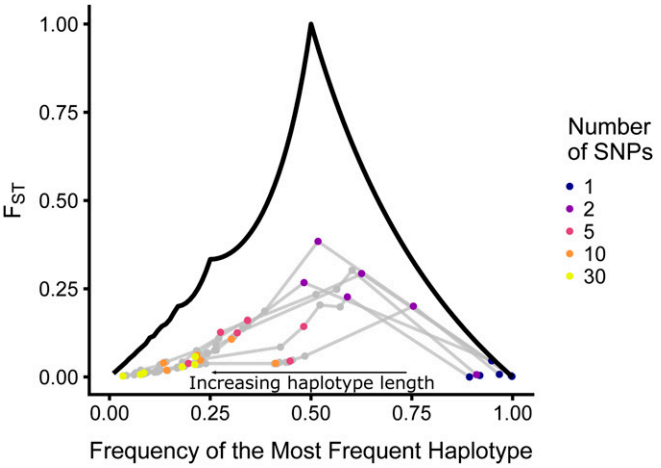


**Figure 9** Example trajectories of observed $F_{ST}$ as haplotype length increases, viewed as a function of the frequency of the most frequent haplotype. As the haplotype length increases, the frequency of the most frequent allele decreases, moving the trajectory from right to left. The solid black curve indicates the upper bound on $F_{ST}$ given the frequency of the most frequent allele for an infinite number of alleles [from Jakobsson *et al.* (2013)]. $F_{ST}$ values associated with numbers of SNPs other than 1, 2, 5, 10, and the maximum of 30 appear in gray.

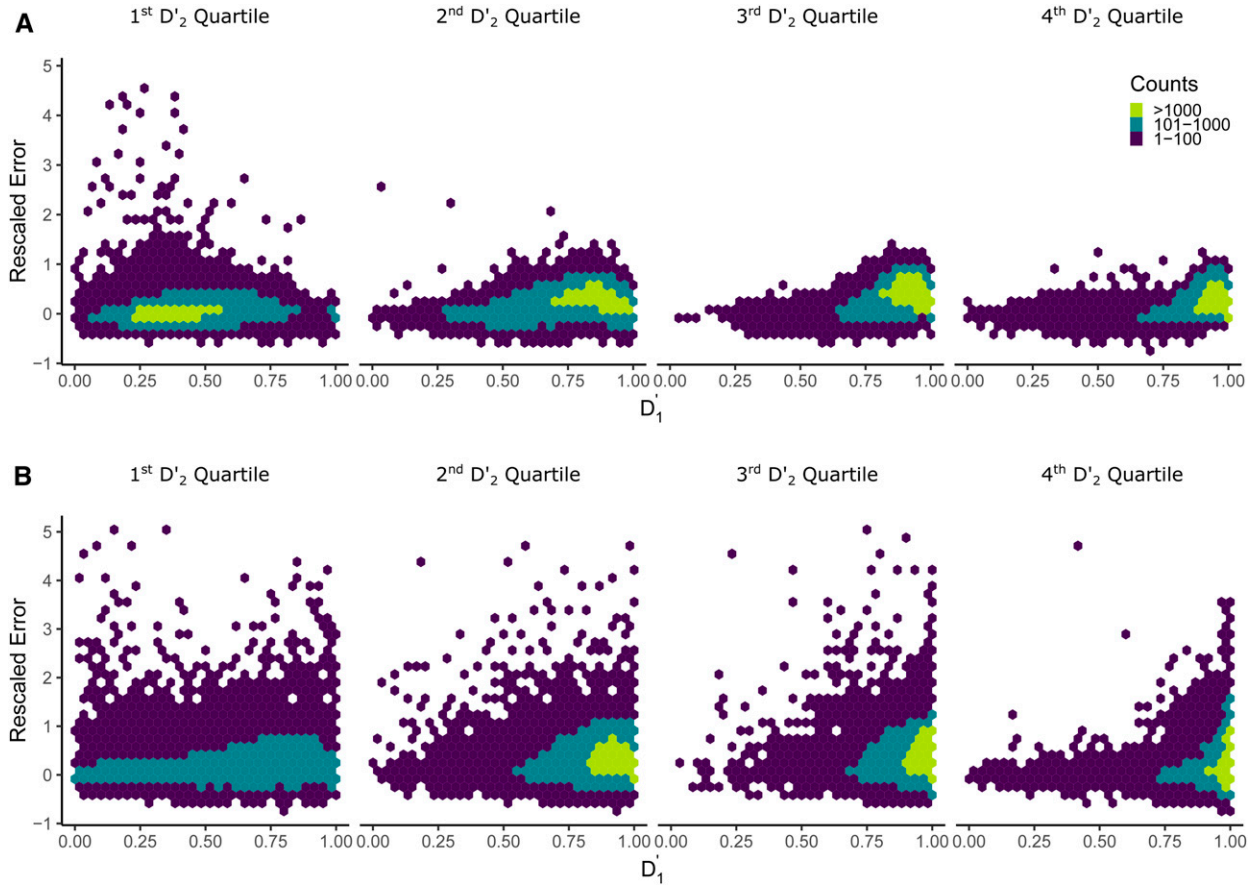


**Figure 10** Rescaled error (Equation 30) *vs*. linkage disequilibrium ($D_1'$ and $D_2'$). (A) Random SNP sets. (B) SNP windows. For both panels, four plots represent four increasing quartiles of $D_2'$ from left to right. The four plots in a row together represent 290,000 data points, 10,000 SNP sets and 29 values for the number of SNPs $(2, 3, \ldots, 30)$, with the exception that those data points yielding a rescaled error greater than 5 are omitted. Data presented here use Europe and East Asia as regions 1 and 2, respectively, so that $D_1'$ and $D_2'$ represent linkage disequilibrium in Europe and East Asia, respectively.

for investigating the same underlying parameter. Thus, haplotype-based $F_{ST}$ methods that consider each locus in the haplotype as part of a sum or average (Excoffier *et al.* 1992; Hudson *et al.* 1992) are expected to be less sensitive to haplotype length than in our case, in which haplotype loci of increasing lengths can be viewed as loci with an increasing mutation rate due to the larger number of SNP sites at which mutations can occur.

We note that although the scenario of interest assumes that the appended locus is biallelic, much of our theoretical analysis applies if the locus is multiallelic (Appendix B). Our main theoretical analysis focuses on the situation in which an added SNP is in linkage equilibrium with the haplotype locus (Equation 20). Indeed, we have found that the theory is least accurate when substantial LD is present (Figure 10). However, our more general theoretical result (Equation 14) does not assume linkage equilibrium and could be used for explicit linkage models that permit LD. Theoretical predictions of the values of the SNP allele frequencies for specific haplotypes $q_{ki}$ under these alternative models could be used in the same way that we used the assumption of $q_{ki} = q_k$ in the case of linkage equilibrium.

The assumption of linkage equilibrium between the SNP and haplotype locus nevertheless produces reasonably accurate predictions about $F_{ST}$ even under circumstances in which linkage equilibrium is not expected (Figure 7, Figure 8, Figure 10, Table 2, and Figures S1–S3). Although the LD level might be smaller in the data we examined than in dense DNA sequence data, the general robustness to the presence of some LD suggests that our results can apply in approximate form to the general situations we have studied in data from human populations.

## Acknowledgments

Support was provided by National Institutes of Health grant R01 HG005855, National Science Foundation grant DBI-1458059, and a Graduate Fellowship from the Stanford Center for Computational, Evolutionary, and Human Genomics.

## Literature Cited


Alcala, N., and N. A. Rosenberg, 2017 Mathematical constraints on $F_{ST}$: biallelic markers in arbitrarily many populations. Genetics 206: 1581–1600. https://doi.org/10.1534/genetics.116.199141

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan *et al.*, 1998 Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. 63: 595–612. https://doi.org/10.1086/301977

Duforet-Frebourg, N., L. M. Gattepaille, M. G. B. Blum, and M. Jakobsson, 2015 HaploPOP: a software that improves population assignment by combining markers into haplotypes. BMC Bioinformatics 16: 242. https://doi.org/10.1186/s12859-015-0661-6

Edge, M. D., and N. A. Rosenberg, 2014 Upper bounds on $F_{ST}$ in terms of the frequency of the most frequent allele and total homozygosity: the case of a specified number of alleles. Theor. Popul. Biol. 97: 20–34. https://doi.org/10.1016/j.tpb.2014.08.001

Excoffier, L., P. E. Smouse, and J. M. Quattro, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131: 479–491.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. Science 296: 2225–2229. https://doi.org/10.1126/science.1069424

García-Fernández, C., J. A. Sánchez, and G. Blanco, 2018 SNP-haplotypes: an accurate approach for parentage and relatedness inference in gilthead sea bream (*Sparus aurata*). Aquaculture 495: 582–591. https://doi.org/10.1016/j.aquaculture.2018.06.019

Gattepaille, L. M., and M. Jakobsson, 2012 Combining markers into haplotypes can improve population structure inference. Genetics 190: 159–174. https://doi.org/10.1534/genetics.111.131136

Hanson, M. A., B. S. Gaut, A. O. Stec, S. I. Fuerstenberg, M. M. Goodman *et al.*, 1996 Evolution of anthocyanin biosynthesis in maize kernels: the role of regulatory and enzymatic loci. Genetics 143: 1395–1407.

Hedrick, P. W., 1987 Gametic disequilibrium measures: proceed with caution. Genetics 117: 331–341.

Holsinger, K. E., and B. S. Weir, 2009 Genetics in geographically structured populations: defining, estimating and interpreting $F_{ST}$. Nat. Rev. Genet. 10: 639–650. https://doi.org/10.1038/nrg2611

Hudson, R., D. Boos, and N. Kaplan, 1992 A statistical test for detecting geographic subdivision. Mol. Biol. Evol. 9: 138–151. https://doi.org/10.1093/oxfordjournals.molbev.a040703

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451: 998–1003. https://doi.org/10.1038/nature06742

Jakobsson, M., M. D. Edge, and N. A. Rosenberg, 2013 The relationship between $F_{ST}$ and the frequency of the most frequent allele. Genetics 193: 515–528. https://doi.org/10.1534/genetics.112.144758

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104. https://doi.org/10.1126/science.1153717

Morin, P. A., K. K. Martien, and B. L. Taylor, 2009 Assessing statistical power of SNPs for population structure and conservation studies. Mol. Ecol. Resour. 9: 66–73. https://doi.org/10.1111/j.1755-0998.2008.02392.x

Nei, M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA 70: 3321–3323. https://doi.org/10.1073/pnas.70.12.3321

Oleksyk, T. K., G. W. Nelson, P. An, J. B. Kopp, and C. A. Winkler, 2010 Worldwide distribution of the *MYH9* kidney disease susceptibility alleles and haplotypes: evidence of historical selection in Africa. PLoS One 5: e11474. https://doi.org/10.1371/journal.pone.0011474

Patil, N., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi *et al.*, 2001 Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294: 1719–1723. https://doi.org/10.1126/science.1065573

Pemberton, T. J., D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg *et al.*, 2012 Genomic patterns of homozygosity in worldwide human populations. Am. J. Hum. Genet. 91: 275–292. https://doi.org/10.1016/j.ajhg.2012.06.014

Rocha, L. A., D. R. Robertson, J. Roman, and B. W. Bowen, 2005 Ecological speciation in tropical reef fishes. P Roy Soc Lond B Bio 272: 573–579. https://doi.org/10.1098/2004.3005

San Lucas, F. A., N. A. Rosenberg, and P. Scheet, 2012 Haploscope: a tool for the graphical display of haplotype structure in populations. Genet. Epidemiol. 36: 17–21. https://doi.org/10.1002/gepi.20640

Sjöstrand, A. E., P. Sjödin, and M. Jakobsson, 2014 Private haplotypes can reveal local adaptation. BMC Genet. 15: 61. https://doi.org/10.1186/1471-2156-15-61

Slatkin, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. 58: 167–175. https://doi.org/10.1017/S0016672300029827

Wall, J. D., and J. K. Pritchard, 2003 Haplotype blocks and linkage disequilibrium in the human genome. Nat. Rev. Genet. 4: 587–597. https://doi.org/10.1038/nrg1123

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. 15: 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x


*Communicating editor: G. Coop*

## Appendix

### Appendix A: Bounds on $D_{12}$

Here we derive the upper bound on $D_{12}$ for a locus with frequencies $p_{1i}$ and $p_{2i}$ in populations 1 and 2 (Equation 5), when $J_1$ and $J_2$ (Equation 4) are treated as fixed quantities in $(0, 1]$, permitting the number of distinct alleles at the locus to be arbitrarily large. Because we are concerned with nonnegative allele frequencies, $D_{12} \geqslant 0$.

By the Cauchy–Schwarz inequality, $D_{12} \leqslant \sqrt{J_1 J_2}$, with equality if and only if one allele frequency distribution is a scalar multiple of the other. Because allele frequency distributions must sum to 1, the equality $D_{12} = \sqrt{J_1 J_2}$ occurs if and only if the two allele frequency distributions are identical, with $p_{1i} = p_{2i}$ for all $i$. This condition implies $J_1 = J_2 = D_{12}$.

If $J_1 \neq J_2$, then no pair of allele frequency distributions satisfies $D_{12} = \sqrt{J_1 J_2}$. However, we can construct a pair of allele frequency distributions, each with a finite number of alleles, such that $D_{12}$ is arbitrarily close to $\sqrt{J_1 J_2}$.

Choose $\epsilon > 0$, $\epsilon \ll J_1$ and $\epsilon \ll J_2$. Suppose $J_1 \neq 1$ and $J_2 \neq 1$. Let $K$ be an integer with

$$K \geqslant \max\left(\left\lceil J_1^{-1}\right\rceil - 1, \left\lceil J_2^{-1}\right\rceil - 1\right). \tag{31}$$

Then $K \geqslant 1, J_1(K+1) - 1 \geqslant 0$, and $J_2(K+1) - 1 \geqslant 0$.

Consider the allele frequency distributions defined by

$$\begin{aligned} p_{11} &= \sqrt{J_1} - \epsilon_1 \\ p_{1i} &= \frac{1 - \sqrt{J_1}}{K} + \frac{\epsilon_1}{K} \\ p_{21} &= \sqrt{J_2} - \epsilon_2 \\ p_{2i} &= \frac{1 - \sqrt{J_2}}{K} + \frac{\epsilon_2}{K}, \end{aligned} \tag{32}$$

where $i$ ranges from 2 to $K+1$, and

$$\begin{aligned} \epsilon_1 &= \tfrac{1}{K+1}\left[\sqrt{J_1}(K+1) - 1 - \sqrt{K[J_1(K+1) - 1]}\right] \\ \epsilon_2 &= \tfrac{1}{K+1}\left[\sqrt{J_2}(K+1) - 1 - \sqrt{K[J_2(K+1) - 1]}\right]. \end{aligned} \tag{33}$$

Note that $\epsilon_1, \epsilon_2 > 0 : \sqrt{J_1}(K+1) - 1 > J_1(K+1) - 1 \geqslant 0$, so that when we add $KJ^2 + KJ$ to the inequality $(K+1)\left(\sqrt{J} - 1\right)^2 > 0$, rearrange terms, and take the square root, we obtain that $\epsilon_1 > 0$. Because $\epsilon_1 \leqslant \sqrt{J_1} - \frac{1}{K+1}$, we have $p_{11} \geqslant p_{1i}$ for all $i > 1$. Analogously, $p_{21} \geqslant p_{2i}$ for all $i > 1$. Thus, alleles are placed in descending order of frequency in both populations.

It is straightforward to calculate $\sum_{i=1}^{K+1} p_{1i} = \sum_{i=1}^{K+1} p_{2i} = 1, \sum_{i=1}^{K+1} p_{1i}^2 = J_1$, and $\sum_{i=1}^{K+1} p_{2i}^2 = J_2$. The dot product $D_{12} = \sum_{i=1}^{K+1} p_{1i} p_{2i}$ between the two allele frequency distributions exceeds the product $p_{11} p_{21}$, so that:

$$\begin{aligned} D_{12} &> \left(\sqrt{J_1} - \epsilon_1\right)\left(\sqrt{J_2} - \epsilon_2\right) \\ &> \sqrt{J_1 J_2} - \epsilon_1 - \epsilon_2. \end{aligned} \tag{34}$$

Choose $K$ large enough that

$$K > \max\left[\frac{\left(2 + \epsilon - 2\sqrt{J_1}\right)^2}{\epsilon\left(4\sqrt{J_1} - \epsilon\right)}, \frac{\left(2 + \epsilon - 2\sqrt{J_2}\right)^2}{\epsilon\left(4\sqrt{J_2} - \epsilon\right)}\right]. \tag{35}$$

From Equation 33, solving $\sqrt{J_1}(K+1) - 1 - \sqrt{K[J_1(K+1) - 1]} = (K+1)\frac{\epsilon}{2}$ for $K$, we find that for $K$ exceeding the root $\left(2 + \epsilon - 2\sqrt{J_1}\right)^2 / [\epsilon(4\sqrt{J_1} - \epsilon)]$, $\epsilon_1 < \frac{\epsilon}{2}$. Similarly, $\epsilon_2 < \frac{\epsilon}{2}$, so that $D_{12} > \sqrt{J_1 J_2} - \epsilon$. Thus, given $J_1, J_2$ in $(0, 1)$, allele frequency distributions exist for which $D_{12}$ is equal to or arbitrarily close to $\sqrt{J_1 J_2}$, with equality possible if and only if $J_1 = J_2$.

The case in which one but not the other homozygosity equals 1 remains. For $J_1 = 1$ and $J_2 \neq 1$, we set $p_{11} = 1$. We set $p_{21}$ and $p_{2i}$ as in Equation 32 for $2 \leqslant i \leqslant K+1$, with $\epsilon_2$ as in Equation 33, and with $K > (2+\epsilon-2\sqrt{J_2})^2/[\epsilon(4\sqrt{J_2}-\epsilon)]$. Then $D_{12} > p_{11}p_{21} = \sqrt{J_2}-\epsilon_2 > \sqrt{J_2}-\epsilon$. A similar argument holds for $J_2 = 1$ and $J_1 \neq 1$.

## Appendix B: Multiallelic Loci with Linkage Equilibrium

Here, we relax the requirement that the appended "SNP" locus must be biallelic. We show that under linkage equilibrium between the appended locus and the haplotype locus, Equations 18–20 continue to hold for multiallelic loci. Suppose, as before, that there are $I$ distinct haplotype alleles, and $M \geqslant 2$ distinct alleles of the additional multiallelic locus. In population $k$, we can write the frequency of the extended haplotype allele that contains haplotype $i$ and additional multiallelic locus allele $m$ analogously to Equations 8 and 9 as

$$p_{k,i,m} = p_{k,i}p_{k,m|i}, \tag{36}$$

where $p_{k,i}$ is the frequency of haplotype allele $i$ in population $k$ and $p_{k,m|i}$ is the frequency of multiallelic locus allele $m$ on haplotype allele $i$ in population $k$.

Under linkage equilibrium, $p_{k,m|i} = p_{k,m}$. We can then proceed, as with Equations 12 and 13, to obtain $J_k^+$ and $D_{12}^+$, as in Equations 18 and 19:

$$J_k^+ = \sum_{i=1}^{I}\sum_{m=1}^{M} p_{k,i,m}^2 = \sum_{i=1}^{I}\sum_{m=1}^{M}\left(p_{k,i}p_{k,m|i}\right)^2 = \sum_{i=1}^{I} p_{k,i}^2 \sum_{m=1}^{M} p_{k,m}^2 = j_k J_k \tag{37}$$

$$D_{12}^+ = \sum_{i=1}^{I}\sum_{m=1}^{M} p_{1,i,m}p_{2,i,m} = \sum_{i=1}^{I}\sum_{m=1}^{M} p_{1,i}p_{1,m|i}p_{2,i}p_{2,m|i} = \sum_{i=1}^{I} p_{1,i}p_{2,i}\sum_{m=1}^{M} p_{1,m}p_{2,m} = d_{12}D_{12}, \tag{38}$$

where $j_k$ and $d_{12}$ are the homozygosity in population $k$ and the allele frequency dot product, respectively, of the additional multiallelic locus.

Using $J_k^+$ and $D_{12}^+$ from Equations 37 and 38 in Equation 6 produces Equation 20.

## Appendix C: Roots of the Derivative $\frac{d}{dq}F_{ST}^+(q)$ in the Case that the Minor Allele of the SNP Occurs Only in One Population and $D_{12}>0$

We use the derivative $\frac{d}{dq}F_{ST}^+(q)$ to determine conditions under which $F_{ST}^+(q)$ has a critical point in the permissible region for $q$, $0 \leqslant q \leqslant \frac{1}{2}$. Using Equation 22,

$$\frac{d}{dq}F_{ST}^+(q) = \frac{64J_2D_{12}q^2 - 64J_2(D_{12}-1)q - 8[(J_1-J_2-2)D_{12}+2J_2]}{[8J_2q^2-4(J_2+D_{12})q+J_1+J_2+2D_{12}-4]^2}. \tag{39}$$

To find the roots of Equation 39, we first show that there are no discontinuities over the range of $q$ with which we are concerned. The quantity $8J_2q^2 - 4(J_2+D_{12})q + J_1 + J_2 + 2D_{12} - 4$ in the denominator is negative for $0 \leqslant q \leqslant \frac{1}{2}$: at $q = 0$, its value is $J_1 + J_2 + 2D_{12} - 4$, which is negative for a polymorphic locus because $J_1$, $J_2$, and $D_{12}$ cannot simultaneously equal one; at $q = \frac{1}{2}$, its value is $J_1 + J_2 - 4 < 0$. As a quadratic with positive leading term, it then has no roots in $[0, \frac{1}{2}]$. The denominator is therefore never zero and Equation 39 has no discontinuities.

Consequently, the roots of Equation 39 are roots of the numerator. As a quadratic in $q$, the numerator of Equation 39 has two roots. One root, termed $q^*$, appears in Equation 24; the other root subtracts rather than adds the term with the square root, and because $0 < D_{12} < 1$ it cannot be positive. Hence, if and only if $0 \leqslant q^* \leqslant \frac{1}{2}$, for fixed $J_1$, $J_2$, and $D_{12}$, $F_{ST}^+(q)$ has a critical point in the permissible region for $q$.