Mathematical constraints on $F_{ST}$: multiallelic markers in arbitrarily many populations

Nicolas Alcala$^1$ and Noah A. Rosenberg$^2$

$^1$Rare Cancers Genomics Team (RCG), Genetic Epidemiology Branch (GEM), International Agency for Research on Cancer/World Health Organization, Lyon 69008, France
$^2$Department of Biology, Stanford University, Stanford, CA 94305-5020, USA

Received: 22 July 2021
Accepted: 23 October 2021

Interpretations of values of the $F_{ST}$ measure of genetic differentiation rely on an understanding of its mathematical constraints. Previously, it has been shown that $F_{ST}$ values computed from a biallelic locus in a set of multiple populations and $F_{ST}$ values computed from a multiallelic locus in a pair of populations are mathematically constrained as a function of the frequency of the allele that is most frequent across populations. We generalize from these cases to report here the mathematical constraint on $F_{ST}$ given the frequency $M$ of the most frequent allele at a multiallelic locus in a set of multiple populations. Using coalescent simulations of an island model of migration with an infinitely-many-alleles mutation model, we argue that the joint distribution of $F_{ST}$ and $M$ helps in disentangling the separate influences of mutation and migration on $F_{ST}$. Finally, we show that our results explain a puzzling pattern of microsatellite differentiation: the lower $F_{ST}$ in an interspecific comparison between humans and chimpanzees than in the comparison of chimpanzee populations. We discuss the implications of our results for the use of $F_{ST}$.

This article is part of the theme issue ‘Celebrating 50 years since Lewontin’s apportionment of human diversity’.

1. Introduction

Multiallelic loci such as microsatellites and haplotype assignments are used to study genetic differentiation in a variety of fields, ranging from ecology and conservation genetics to anthropology and human genomics. Genetic differentiation is often measured for multiallelic loci using the multiallelic extension of Wright’s fixation index $F_{ST}$ [1]

$$F_{ST} = \frac{H_T - H_S}{H_T}. \tag{1.1}$$

For a polymorphic multiallelic locus with $l$ distinct alleles in a set of $K$ subpopulations, denoting by $p_i$ the frequency of allele $i$ in subpopulation $k$, $H_S = 1 - \left(1/K\right) \sum_{k=1}^{K} \sum_{i=1}^{l} p_{ik}^2$, and $H_T = 1 - \left(1/\sum_{k=1}^{K} p_{k}\right)^2$.

$F_{ST}$ values are known to be smaller for multiallelic than for biallelic loci [2]. One reason invoked to explain this difference is that within-subpopulation heterozygosity $H_S$ mathematically constrains the maximal value of $F_{ST}$ to be below 1, and the constraint is stronger when $H_S$ is high. This phenomenon was noticed concurrently in simulation-based, empirical and theoretical studies [3–7], and the mathematical constraints describing the dependence were subsequently clarified [8,9].

Studies have found that the maximal value of $F_{ST}$ can be viewed as constrained not only by functions of the within-subpopulation allele frequency distribution such as $H_S$, but alternatively by aspects of the global allele...
frequency distribution across subpopulations. For a biallelic locus in $K=2$ subpopulations, Maruki et al. [10] showed that the maximal $F_{ST}$ as a function of the frequency $M$ of the most frequent allele decreases as $M$ increases from $1/2$ to 1 (see also [11]). Generalizing the biallelic case to arbitrarily many alleles, Jakobsson et al. [12] showed that for multiallelic loci with an unspecified number of distinct alleles, the maximal $F_{ST}$ increases from 0 to 1 as a function of $M$ if $0 < M < 1/2$, and decreases from 1 to 0 for $1/2 < M < 1$ in the manner reported by Maruki et al. [10] for biallelic loci. Edge & Rosenberg [13] generalized these results to the case of a fixed finite number of alleles, showing that the maximal $F_{ST}$ differs slightly from the unspecified case when the fixed number of distinct alleles is an odd number.

Generalizing the simplest case of $K = l = 2$ in a different direction, Alcala & Rosenberg [14] considered biallelic loci in the case of a fixed number of subpopulations $K \geq 2$. We showed that the maximal value of $F_{ST}$ displays a peculiar behaviour as a function of $M$: the upper bound has a maximum of 1 if and only if $M = k/K$, for integers $k$ with $[K/2] \leq k \leq K$. The constraints on the maximal value of $F_{ST}$ dissipate as $K$ tends to infinity, even though for any fixed $K$, there always exists a value of $M$ for which $F_{ST} < 2\sqrt{2} \approx 2.8284$.

Relating $F_{ST}$ to its maximum as a function of $M$ helps explain surprising phenomena that arise during population-genetic data analysis. For example, Jakobsson et al. [12] showed that stronger constraints on $F_{ST}$ could explain the low $F_{ST}$ values seen in pairs of African human populations. They also found that such constraints could explain the lower $F_{ST}$ values seen in high-diversity multiallelic loci compared to lower-diversity loci—microsatellites compared to single-nucleotide polymorphisms. Alcala & Rosenberg [14] showed that constraints on the maximal $F_{ST}$ could explain the lower $F_{ST}$ values between human populations seen when computing $F_{ST}$ pairwise rather than from all populations simultaneously.

In this study, we characterize the relationship between $F_{ST}$ and the frequency $M$ of the most frequent allele, for a multiallelic locus and an arbitrary specified value of the number of subpopulations $K$. We derive the mathematical upper bound on $F_{ST}$ in terms of $M$, extending the biallelic result of Alcala & Rosenberg [14] to the multiallelic case, and providing the most comprehensive description of the mathematical constraints on $F_{ST}$ in terms of $M$ to date (table 1).

To assist in interpreting the new bound, we simulate the joint distribution of $F_{ST}$ and $M$ in the island migration model, describing its properties as a function of the number of subpopulations, the migration rate and a mutation rate. The $K$-subpopulation upper bound on $F_{ST}$ in terms of $M$ facilitates an explanation of counterintuitive aspects of inter-species genetic differentiation. We discuss the importance of the results for applications of $F_{ST}$ more generally.

## 2. Model

Our goal is to derive the range of values that $F_{ST}$ can take—the lower and upper bounds on $F_{ST}$—as a function of the frequency $M$ of the most frequent allele for a multiallelic locus, when the number of subpopulations $K$ is a fixed finite value greater than or equal to 2. We follow previous studies [12–15] in describing notation and constructing the scenario.

We consider a polymorphic locus with an unspecified number of distinct alleles, in a setting with $K$ subpopulations contributing equally to the total population. We denote the frequency of allele $i$ in subpopulation $k$ by $p_{ki}$, with sum $\sigma_i = \sum_{k=1}^{K} p_{ki}$ across subpopulations. Each allele frequency $p_{ki}$ lies in $[0, 1]$. Within subpopulations, allele frequencies sum to 1: for each $k$, $\sum_{i=1}^{\sigma_i} p_{ki} = 1$. Hence, $\sigma_i$ lies in $[0, K]$, and $\sum_{i=1}^{\sigma_i} \sigma_i = K$. We number alleles from most to least frequent, so $\sigma_i \geq \sigma_j$ for $i \leq j$.

Because by assumption the locus is polymorphic, $\sigma_i < K$ for each $i$. Alleles 1 and 2 have non-zero frequency in at least one subpopulation, not necessarily the same one; we have $\sigma_1 > 0$ and $\sigma_2 > 0$. We denote the mean frequency of the most frequent allele across subpopulations by $M = \sigma_1/K$. We then have $0 < M < 1$. We treat the allele frequencies $p_{ki}$ and associated quantities $M$ and $\sigma_i$ as parametric values, and not as estimates computed from data.

Equation (1.1) expresses $F_{ST}$ as a ratio involving within-subpopulation heterozygosity, $H_2$, and total heterozygosity, $H_T$, with $0 \leq H_2 < 1$ and $0 \leq H_T < 1$. Because we assume the locus is polymorphic, $H_T > 0$. We write equation (1.1) in terms of allele frequencies, permitting the number of distinct alleles to be arbitrarily large

$$F_{ST} = \frac{(1/K) \sum_{k=1}^{K} \sum_{i=1}^{\sigma_i} p_{ki}^2 - \sum_{i=1}^{\sigma_i} (\sum_{k=1}^{K} p_{ki}/K)^2}{1 - \sum_{i=1}^{\sigma_i} (\sum_{k=1}^{K} p_{ki}/K)^2}.$$  (2.1)
Hence, our goal is, for fixed $\sigma_1 = KM, 0 < \sigma_1 < K$, to identify the matrices ($\theta_{,j}$) with $\theta_{,j} \in [0, 1], \sum_{j=1}^{K} \theta_{,j} = 1$ and $(1/K) \sum_{j=1}^{K} \theta_{,j} = \sigma_1/\sigma_1 = M$, that minimize and maximize $\text{FST}_i$ in equation (2.1).

Note that we adopt the interpretation of $\text{FST}$ as a ‘statistic’ that describes a mathematical function of allele frequencies rather than as a ‘parameter’ that describes coancestry of individuals in a population [e.g. 16]. See Alcala & Rosenberg [14] for a discussion of interpretations of $\text{FST}$ when studying its mathematical properties.

### 3. Mathematical constraints

#### (a) Lower bound of $\text{FST}$

Bounds on $\text{FST}$ in terms of the frequency of the most frequent allele can be written with respect to $M$ or $\sigma_1$, noting that $M$ ranges in $(0, 1)$ and $\sigma_1$ ranges in $(0, K)$. For the lower bound, from equation (2.1), for any choice of $\sigma_1, \text{FST} = 0$ can be achieved. Consider ($\sigma_1, \sigma_2, \ldots$) with $\sigma_1 \in [0, K)$ for each $k, \sigma_k \geq \sigma_1$ for $i \leq j$, $\sum_{i=1}^{K} \sigma_i = K$, and $\sigma_1 > 0$ and $\sigma_2 > 0$. We set $p_{i,j} = \sigma_i/K$ for all subpopulations $k$ and alleles $i$; this choice yields $\text{FST} = 0$.

$\text{FST} = 0$ implies that the numerator of equation (2.1), $H_T - H_S$, is zero. This numerator can be written as $(1/K^2) \sum_{i=1}^{K} (K \sum_{j=1}^{J} \theta_{,j} - \sigma_1^2)$. The Cauchy–Schwarz inequality guarantees that $K \sum_{j=1}^{J} \theta_{,j} \geq \sigma_1^2$, with equality if and only if $\theta_{1,j} = \theta_{2,j} = \ldots = \theta_{K,j} = \sigma_1 / K$. Applying the Cauchy–Schwarz inequality to all alleles $i$, the numerator of equation (2.1) is zero only if for all $i$, $\theta_{1,j} \theta_{2,j} \ldots \theta_{K,j} = (\sigma_i/K, \sigma_i/K, \ldots, \sigma_i/K)$.

Thus, we can conclude that the allele frequency matrices in which all $K$ subpopulations have identical allele frequency vectors are the only matrices for which $\text{FST} = 0$. The lower bound on $\text{FST}$ is equal to 0 irrespective of $M$ or $\sigma_1$, for any value of the number of subpopulations $K$.

#### (b) Upper bound of $\text{FST}$

To derive the upper bound on $\text{FST}$ in terms of $M = \sigma_1/K$, we must maximize $\text{FST}$ in equation (2.1), assuming that $\sigma_1$ and $K$ are constant. The computations are performed in appendix A; we write the main result as a function of $\sigma_1$, noting that it can be converted into a function of $M$ by replacing $\sigma_1$ with $KM$.

In theorem A.1, we treat the case in which $\sigma_1$ has an integer value. For non-integer $\sigma_1$, theorem A.2 shows that the maximal $\text{FST}$ requires that (i) the sum of squared allele frequencies across alleles and subpopulations, $S = \sum_{j=1}^{K} \sum_{i=1}^{J} \theta_{,j}^2$, is maximal, and (ii) alleles $i = 2, 3, \ldots$ are each present in at most one subpopulation, but allele 1 might be present in more than one subpopulation. We then separately maximize $\text{FST}$ as a function of $\sigma_1$ for $\sigma_1$ in $(0, 1)$ and non-integer $\sigma_1$ in $(1, K)$.

These two cases differ in that allele 1 appears in a single subpopulation in the former case, and it must appear in at least two subpopulations in the latter.

The maximal $\text{FST}$ as a function of $\sigma_1$ for $\sigma_1$ in $(0, 1)$ is

$$\text{FST} \leq \begin{cases} \frac{1}{K} & \text{for } K = 1, \\
\frac{(K - 1)[1 - \sigma_1(f - 1) - f\sigma_1]}{K - 1 - f\sigma_1} & \text{for } K > 1,
\end{cases}$$

where $f = \lceil \sigma_1^{-1} \rceil$. Here, $\lfloor x \rfloor$ denotes the smallest integer greater than or equal to $x$, $\lceil x \rceil$ denotes the greatest integer less than or equal to $x$, and $\{ x \} = x - \lfloor x \rfloor$ denotes the fractional part of $x$. Note that for an integer choice of $\sigma_1$, the maximum from equation (3.1) and the limits as $\sigma_1$ tends to the integer from above and below all equal 1, so that the maximum as a function of $\sigma_1$ is continuous.

From appendix A, $\text{FST}$ reaches its upper bound for integer $\sigma_1$ when allele 1 has frequency 1 in each of $\sigma_1$ subpopulations, and when in each of the remaining $K - \sigma_1$ subpopulations, an allele other than allele 1 has frequency 1. These alleles of frequency 1 need not be private, although they can be any identity relationships among them are permissible, provided that when summing frequencies across subpopulations, none of these alleles has a sum that exceeds $\sigma_1$. The locus can have as few as $\lceil K\sigma_1^{-1} \rceil$ alleles of non-zero frequency and as many as $K - \sigma_1 + 1$.

For $\sigma_1$ in interval $(0, 1)$, $\text{FST}$ is maximal when each allele is present in only a single subpopulation, and when each subpopulation has exactly 1 allele with a non-zero frequency: $J - 1$ alleles at frequency $\sigma_1$ and one allele at frequency $1 - (J - 1)\sigma_1 \leq \sigma_1$. Because each subpopulation has $J$ distinct alleles and no alleles are shared across subpopulations, this upper bound requires that the locus has $KJ$ alleles of non-zero frequency.

For non-integer $\sigma_1$ in $(1, K)$, $\text{FST}$ reaches its maximum when there are $\lceil \sigma_1 \rceil$ subpopulations in which the most frequent allele has frequency 1, a single subpopulation in which it has frequency $\sigma_1$, and a private allele has frequency $1 - \sigma_1$, and $K - \lceil \sigma_1 \rceil - 1$ subpopulations each with a different private allele at frequency 1. Only the most frequent allele is shared across subpopulations, and a single subpopulation displays polymorphism. At the maximum, $K - \lceil \sigma_1 \rceil + 1$ alleles have non-zero frequency.

#### (c) Properties of the upper bound

Figure 1 shows the maximal value of $\text{FST}$ in terms of $M = \sigma_1/K$ for various values of the number of subpopulations, $K$. We describe a number of properties of this upper bound.

#### (i) Piecewise structure of the upper bound

First, we observe that the upper bound has a piecewise structure.

For $M < 1/K$, the upper bound depends on $J = \lceil \sigma_1^{-1} \rceil = \lceil 1/(KM) \rceil$. As KM increases in $(0, 1)$, each decrement in the integer value of $\lceil 1/(KM) \rceil$ produces a distinct ‘piece’ with domain $[1/(Kj), 1/(K(j - 1))]$, for integers $j \geq 2$. Within each interval $[1/(Kj), 1/(K(j - 1))]$, $J$ has the constant value $j$.

At $M = 1/K$, the upper bound has its first transition between cases. For $M > 1/K$, the upper bound depends on $\lceil \sigma_1 \rceil = \lceil KM \rceil$. As KM increases in $[1, K)$, each increment in
[KM] also produces a distinct piece of the domain. For each k from 1 to K−1, [KM] = k for M in (k/K, (k+1)/K).

Counting the intervals of the domain, we see that an infinite number of distinct intervals occur for M in (0, 1/K), and K−1 intervals occur for M in (1/K, 1). Within intervals, the function describing the upper bound is smooth.

(ii) Behaviour of the upper bound for M = 1/K, 2/K, ..., (K−1)/K
The upper bound is equal to 1 at M = 1/K, 2/K, ..., (K−1)/K. For M in (0, 1/K), setting the numerator and denominator equal in equation (3.1), we find that the upper bound is never equal to 1. For M in (1/K, 1), the upper bound is equal to 1 if and only if σ₁ = 0, that is, if and only if σ₁ is an integer and M = k/K for k = 2, 3, ..., K − 1.

Hence, noting that the upper bound is equal to 1 at M = 1/K, we conclude that the upper bound can equal 1 if and only if M = k/K for integers k = 1, 2, ..., K − 1. For fixed K, the upper bound on FST has exactly K − 1 maxima at which FST can equal 1, at M = 1/K, 2/K, ..., (K−1)/K. We can conclude that FST is unconstrained within the unit interval only for a finite set of values of the frequency M of the most frequent allele. The size of this set increases with the number of subpopulations K.

(iii) Behaviour of the upper bound for M in (0, 1/K)
For M in (0, 1/K), we can compute the value of the upper bound at the transition points between distinct pieces of the domain, namely values of 1/(jK) for integers j ≥ 2. Applying equation (3.1), we observe that at M = 1/(jK), the upper bound has value (K−1)/(jK−1). In other words, the upper bound touches the curve

\[ q^*(M) = \frac{(K-1)M}{1-M}. \]  

This curve is represented in figure 1 as a dashed line.

Note that for K = 2, the special case considered by Jakobsson et al. [12], equation (3.2) reduces to \( q^*(M) = M/(1-M) = \sigma_1/(2-\sigma_1) \), which matches equation 21 from Jakobsson et al. [12]. In fact, setting K = 2, equation (3.1) for M in (0, 1/K) reduces to the K = 2 upper bound on FST in eqn 9 of [12].

(iv) Behaviour of the upper bound for M in (1/K, 1)
Because the upper bound is a smooth function on each interval of its domain, and because it possesses maxima at interval boundaries M = 1/K, 2/K, ..., (K−1)/K, it must possess local minima in intervals [k/K, (k+1)/K) for k = 1, 2, ..., K − 2. Indeed, such minima are visible in figure 1 in cases with K = 3, K = 6, K = 40 and K = 100; for K = 2, only one maximum occurs, so that there is no interval between a pair of maxima in which a minimum can occur. Note that because we restrict attention to M in (0, 1), we do not count the point at M = 1 and FST = 0 as a local minimum.

4. Joint distribution of M and FST under an evolutionary model
So far, we have described the mathematical constraint imposed on FST by M without respect to the frequency with which particular values of M arise in evolutionary scenarios. As an assessment of the bounds in evolutionary models can illuminate the settings in which they are most salient in population-genetic data analysis [9,14,17–20], we simulated the joint distribution of FST and M under an island migration model, relating the distribution to the mathematical bounds on FST. This analysis considers allele frequency distributions, and hence values of M and FST generated by evolutionary models. The simulation approach is modified from [14,15].

(a) Simulations
We simulated alleles under a coalescent model, using the software MS [21]. We considered a total population of KN diploid individuals subdivided into K subpopulations of size N. At each generation, a proportion m of the individuals in a subpopulation originated outside the subpopulation. Thus, the scaled migration rate is 4Nm, and it corresponds to twice the number of individuals in a subpopulation that originate elsewhere. We considered the island model [22–24], in which migrants have the same probability m/(K−1) of coming from any other specific subpopulation. We used an infinitely-many-alleles model; mutations occur at rate μ, and the scaled mutation rate is 4Nμ.

We examined three values of K (2, 6, 40), three values of 4Nμ (0.1, 1, 10) and three values of 4Nm (0.1, 1, 10). Note that in MS, time is scaled in units of 4N generations, and there is no need to specify subpopulation sizes N. MS simulates an infinitely-many-sites model, where each mutation occurs at a new site; each haplotype is a new allele, so that each mutation creates a new allele. For our analysis, we are concerned only with the allelic categories and not with the simulated sequences; thus, although the simulation follows the infinitely-many-sites model, the analysis treats simulated datasets as having been generated under an infinitely-many-alleles model.

For each parameter triplet (K, 4Nμ, 4Nm), we performed 1000 replicate simulations, sampling 100 sequences per
subpopulation in each replicate. We computed $F_{ST}$ values from the parametric allele (haplotype) frequencies. MS commands appear in electronic supplementary material. File S1; note that the simulation approach here uses the standard method of simulating MS with a specified mutation rate $\theta = 4N\mu$, whereas in our previous analyses of biallelic cases [14,15], we had employed the alternative approach of requiring the upper bound on $F_{ST}$ in terms of $M$ (equation 3.1); the black point plots the mean values of $M$ and $F_{ST}$. Colours represent the density of loci, estimated using a Gaussian kernel density estimate with a bandwidth of 0.02, with density set to 0 outside of the bounds. Loci are simulated using coalescent software MS, assuming an island model of migration and an infinitely-many-alleles mutation model. Each panel considers 1000 replicate simulations, with 100 lineages sampled per subpopulation. Electronic supplementary material, figures S1 and S2 present similar results for migration rates $M$.

(b) Impact of the mutation rate
For fixed migration rate $4Nm$ and number of subpopulations $K$, the main impact of the mutation rate is on the frequency $M$ of the most frequent allele. For $K = 2$, under weak mutation ($4N\mu = 0.1$), the joint distribution of $M$ and $F_{ST}$ is highest in the high-$M$ region, for all values of $4Nm$ (figure 2a,d,g). Although most simulation replicates produce $M > 1/2$ with an upper bound on $F_{ST}$ less than one, this set of parameter values does give rise to replicates near the peak at $(M, F_{ST}) = (1/2, 1)$.

Under intermediate mutation ($4N\mu = 1$), the increased mutation rate tends to decrease $M$, shifting the joint distribution to lower values of $M$ for all values of $4Nm$ (figure 2b,e,h). Finally, under strong mutation ($4N\mu = 10$), the joint distribution of $M$ and $F_{ST}$ is highest in the low-$M$ region, for all values of $4Nm$ (figure 2c,f,i). In this region, the upper bound on $F_{ST}$ is most strongly constrained, leading to low $F_{ST}$ values.

(c) Impact of the migration rate
For fixed mutation rate $4N\mu$ and number of subpopulations $K$, the impact of the migration rate is seen primarily in the $F_{ST}$ values rather than the values of $M$. Under weak migration ($4Nm = 0.1$), subpopulations are differentiated, and the joint distribution of $M$ and $F_{ST}$ is highest near the upper bound on $F_{ST}$ in terms of $M$ (figure 2a–c).

Under intermediate migration ($4Nm = 1$), differentiation between subpopulations decreases, and the joint density of $M$ and $F_{ST}$ is highest at lower values of $F_{ST}$ (figure 2d–f). Under strong migration ($4Nm = 10$), the joint density of $M$ and $F_{ST}$ nears the lower bound (figure 2g–i).

(d) Impact of the number of subpopulations
In figure 1, the number of subpopulations changes the shape of the region in which $F_{ST}$ is permitted to range as a function of $M$. Thus, in simulations, the impact of the number of subpopulations $K$ is observed in cases in which a change in $K$
3.1). The coloured points represent the mean \( z \) locus for the set, or \( 4Nm \) and different numbers of subpopulations \( K \). (a) \( K = 2 \), (b) \( K = 6 \) and (c) \( K = 40 \). The black solid lines represent the upper bound on \( F_{ST} \) in terms of \( M \) (equation 3.1). The coloured points represent the mean \( M \) and mean \( F_{ST} \) where colours correspond to values of \( 4Nm \). These points are taken from figures 2 and electronic supplementary material, S1 and S2.

permits \( F_{ST} \) to expand its range within the unit square for \((M, F_{ST})\). For each of the nine choices of \((4N\mu, 4Nm)\), figure 3 summarizes the means observed for \((M, F_{ST})\) in figures 2 and electronic supplementary material, S1 and S2, corresponding to \( K = 2, K = 6 \) and \( K = 40 \), respectively.

The number of subpopulations generally increases \( F_{ST} \) for fixed \( 4N\mu \) and \( 4Nm \). For example, the mean \( F_{ST} \) can be substantially larger for \( K = 6 \) than for \( K = 2 \). Consider \((4N\mu, 4Nm) = (0.1, 0.1)\). For \( K = 2 \), the mean \( F_{ST} \) is near its upper bound (figure 3a); for \( K = 6, F_{ST} \) is not as close to the bound (figure 3b). However, because the upper bound for \( K = 6 \) exceeds that for \( K = 2 \), the mean \( F_{ST} \) is nevertheless larger in the case of \( K = 6 \).

5. Example: humans and chimpanzees

We now use our theoretical results to examine genetic differentiation in humans and chimpanzees. Because humans and chimpanzees are distinct species, we might expect a genetic differentiation measure such as \( F_{ST} \) to produce a greater value for a computation between them than for a computation among populations within one or the other. Indeed, studies of multiallelic loci do find that adding chimpanzees to data on multiple human populations increases the value of \( F_{ST} \) [8,25]. However, we will see that \( F_{ST} \) has a more subtle pattern when considering data on multiple chimpanzee populations, and that our theoretical computations explain a surprising result.

We examine data on 246 multiallelic microsatellite loci assembled by Pemberton et al. [26] from several studies of worldwide human populations and a study of chimpanzees [27]. We consider \( F_{ST} \) comparisons both between humans and chimpanzees and among populations of chimpanzees. For the human data, we consider all 5795 individuals in the dataset, and for the chimpanzee data, we consider 84 chimpanzee individuals from six populations: one bonobo population, and five common chimpanzee populations (Central, Eastern, Western, hybrid and captive).

In the data analysis, we perform a computation to summarize the relationship of \( F_{ST} \) to the upper bound. For a set of \( Z \) loci, denote by \( F_z \) and \( M_z \) the values of \( F_{ST} \) and \( M \) at locus \( z \). The mean \( F_{ST} \) for the set, or \( F_{ST} \), is

\[
\bar{F}_{ST} = \frac{1}{Z} \sum_{z=1}^{Z} F_z.
\]  

Using equation (3.1), we can compute the corresponding maximum \( F_{ST} \) given the observed \( \sigma_z = KM_{Lz}, z = 1, 2, \ldots, Z \).

Denoting this quantity by \( F_{max, z} \), we have

\[
\frac{F_{ST}}{F_{max}} = \frac{1}{Z} \sum_{z=1}^{Z} \frac{F_z}{F_{max,z}}.
\]  

\( F_{ST}/F_{max} \) measures the proximity of the \( F_{ST} \) values to their upper bounds: it ranges from 0, if \( F_{ST} \) values at all loci equal 0, to 1, if \( F_{ST} \) values at all loci equal their upper bounds.

We computed the parametric allele frequencies for each subpopulation—the human and chimpanzee groups for the human–chimpanzee comparison, and chimpanzee subpopulations for the comparison of chimpanzees—averaging across subpopulations to obtain the frequency \( M \) of the most frequent allele. We then computed \( F_{ST} \) and the associated upper bound for each locus, averaging across loci to obtain the overall \( F_{ST} \) and \( F_{ST}/F_{max} \) for the full microsatellite set (equations (5.1) and (5.2)).

Surprisingly, given the longer evolutionary time between humans and chimpanzees than among chimpanzee populations, the \( F_{ST} \) value is significantly greater when comparing chimpanzee populations (\( F_{ST} = 0.16 \)) than when comparing humans and chimpanzees (\( F_{ST} = 0.10; p = 4.2 \times 10^{-14} \), Wilcoxon rank sum test). The explanation for this result can be found in the properties of the upper bound on \( F_{ST} \) given \( M \).

Values of \( M \) are similar in the two comparisons (figure 4a,b). However, \( K \) differs, equaling 2 for the human–chimpanzee comparison and 6 for the comparison of chimpanzee subpopulations. Because the theoretical range of \( F_{ST} \) is seen to be smaller for \( F_{ST} \) values computed among smaller sets of subpopulations than among larger sets (figure 1), the \( F_{ST} \) values among chimpanzees possess a larger range. For example, the maximal \( F_{ST} \) at the mean \( M \) of 0.27 observed in pairwise comparisons is 0.34 for \( K = 2 \) (red segment in figure 4a), whereas the maximal \( F_{ST} \) at the mean \( M \) of 0.36 observed for six chimpanzee populations is 0.93 for \( K = 6 \) (figure 4b). Given the stronger constraint in pairwise calculations than in calculations with more subpopulations, it is not unexpected that pairwise \( F_{ST} \) values would be smaller than those in a 6-region computation. A high \( F_{ST} \) among chimpanzees compared to between humans and chimpanzees is a by-product of mathematical constraints on \( F_{ST} \).

Interestingly, the effect of \( K \) on \( F_{ST} \) is largely eliminated when each \( F_{ST} \) value is normalized by the associated maximum given \( K \) and \( M \) (figure 4c). The normalization leads to higher values for human–chimpanzee comparisons than among chimpanzee subpopulations (\( F_{ST}/F_{max} = 0.32 \) and 0.20, respectively; \( p = 1.1 \times 10^{-9} \), Wilcoxon rank sum test),
Figure 4. $F_{ST}$ values for comparisons involving humans and chimpanzees based on multiallelic microsatellite loci. (a) $F_{ST}$ between humans and chimpanzees, considering $K=2$ subpopulations (humans, chimpanzees). (b) $F_{ST}$ among $K=6$ chimpanzee subpopulations. In (a,b), colours represent the number of points in a neighbourhood of radius 0.03; red points indicate the mean results for the case of $K$ restriction to $M$ and vertical red segments indicate the permissible range of $F_{ST}$ at the mean $M$. (c) $F_{ST}$ computed using equation (2.1), and $F_{ST}/F_{max}$ computed using equations (2.1) and (3.1). Each point plotted represents one locus.

as expected from the greater evolutionary distance between humans and chimpanzees compared to that among chimpanzees.

6. Discussion

We have analysed the range of values that $F_{ST}$ can take as a function of the frequency $M$ of the most frequent allele at a multiallelic locus, for an arbitrary value of the number of subpopulations $K$. We showed that $F_{ST}$ can span the full unit interval only for a finite set of values of $M$, at $M = k/K$ for integers $k$ in $[1, K - 1]$. For all other $M$, $F_{ST}$ necessarily lies below 1. The number of subpopulations $K$ enlarges the range of values that $F_{ST}$ can take as it increases.

This study provides the most complete relationship between $F_{ST}$ and $M$ obtained to date, generalizing previous results for the case of $K=2$ subpopulations [12] and for a restriction to $l=2$ alleles [14]. Interestingly, the maximal $F_{ST}$ we have obtained merges patterns observed in these previous studies. Fixing $K=2$, we obtain the upper bound on $F_{ST}$ in terms of $M$ that was reported by Jakobsson et al. [12]. As $K$ increases, the piecewise pattern seen by Jakobsson et al. [12] for the maximal $F_{ST}$ in the $K=2$ case for $M$ in $[0, 1/2)$ is observed in the multiallelic case for $M$ in $[0, 1/K)$. The decay from $(M, F_{ST}) = (1/2, 1)$ to $(M, F_{ST}) = (1, 0)$ seen by Jakobsson et al. [12] for $K=2$ is observed for $M$ in the decay from $(K-1)/K, 1$ to $(1, 0)$ for arbitrary $K$.

The allele frequency values for which the upper bound is reached for $M$ in $[0, 1/K)$ generalize those seen for the case of $K=2$ and $M$ in $[0, 1/2)$ [12]. The upper bound is reached when all alleles are private, each subpopulation has as many alleles as possible at frequency $KM$, and at most one additional allele. The allele frequency values for which the upper bound is reached for $M$ in $((K-1)/K, 1)$ also generalize those seen for $K=2$ and $M$ in $[1/2, 1)$: the maximum is reached when the most frequent allele is fixed in all subpopulations except one, and a single private allele is present in this last subpopulation.

The results from Alcala & Rosenberg [14] for $l=2$ produce a more constrained upper bound on $F_{ST}$ than for arbitrary $I$, with the domain of $M$ restricted to $[1/2, 1)$. Nevertheless, many properties of the maximal $F_{ST}$ we observe for unspecified $I$ and $M$ in $(1/K, 1)$ are similar to those seen for $l=2$ and $M$ in $[1/2, 1)$: finitely many peaks at points $M = k/K$, local minima between the peaks, and an increase in coverage of the unit square for $(M, F_{ST})$ as $K$ increases. The maximal $F_{ST}$ functions for $M$ in $((K-1)/K, 1)$ for unspecified $I$ and for $l=2$ agree, as the number of alleles required to maximize $F_{ST}$ in this interval in the case of unspecified $I$ is simply equal to 2.

In assuming that the number of alleles is unspecified, we found that the number of distinct alleles needed for achieving the maximal $F_{ST}$ is $K\lfloor \sigma_1^{-1} \rfloor$ for $M$ in $[0, 1/K)$ and $K - \lfloor \sigma_1 \rfloor - 1$ for non-integer $M$ in $[1/K, 1)$; the maximum can be achieved with each number of distinct alleles in $[(\lfloor K\sigma_1^{-1} \rfloor - K - \sigma_1 + 1) / M = 1/2, 2/K, ..., (K-1)/K)$. With a fixed maximal number of distinct alleles, such as in the $l=2$ case of Alcala & Rosenberg [14] with $K$ specified and in the $K=2$ case with $I$ specified [13], the upper bound on $F_{ST}$ is less than or equal to that seen in the corresponding unspecified-$I$ case. For $K=2$, specifying $I$ has a relatively small effect in reducing the maximal value of $F_{ST}$ [13]. As in Edge & Rosenberg [13], specifying $I$ in the case of larger values of $K$ is expected to have the greatest impact on the $F_{ST}$ upper bound at the lowest end of the domain for $M$.

In coalescent simulations, we found that the joint distribution of $M$ and $F_{ST}$ within their permissible space can help separate the impact of mutation and migration. Although the dependence of $F_{ST}$ on mutation and migration rates has been long documented, the symmetric effects of mutation and migration under the island model [22] illustrate the difficulty in separating their effects. Under the island model, allele frequency $M$ is informative about the scaled mutation rate $4Nu$, and comparing the value of $F_{ST}$ to its maximum given $M$ is informative about the scaled migration rate $4Nm$. Adding a dimension that is more sensitive to mutation than to migration—$M$ in our case—enables the separation of their effects. Other statistics, such as total heterozygosity $H_T$ or within-subpopulation heterozygosity $H_C$, have the potential to play a similar role [20].

Our results can inform data analyses. In particular, we caution users to examine upper bounds on $F_{ST}$ to assess how mathematical constraints influence observations. As the constraints are strongest for $K=2$, this step is valuable in pairwise comparisons; it is also useful when the frequency
M of the most frequent allele can be small in relation to the number of populations K, such as for high-diversity forensic [28] and immunological [29] loci in human populations. Visual inspection of the values of M and FST within their bounds can suggest that constraints have an effect. \( \text{FST}_{\text{M}}/\text{F}_{\text{max}} \) can provide a helpful summary by evaluating the proximity of FST values to their maxima.

Further, joint use of M along with FST could be useful in various applications of FST, such as in inference of model parameters by approximate Bayesian computation [30] and machine learning [31]. FST outlier tests to detect local adaptation from multiallelic loci [32] could search for FST values that represent outliers not in the distribution of FST values, but rather, outliers in relation to associated upper bounds. Computing null distributions for FST conditional on M could enhance the approach.

In an example data analysis, we have shown that taking into account mathematical constraints on FST can help understand puzzling FST behaviour. In our example, FST at a set of loci was higher when comparing K = 6 chimpanzee populations than when comparing humans and chimpanzees (K = 2), even though the same loci were used and the mean value for M was similar in the two comparisons. A comparison of FST values to their respective maxima explained these counterintuitive results.

We note that analyses of FST in relation to M differ from analyses of FST in relation to within-subpopulation statistics \( H_2 \) and \( J_2 = 1 - H_2 \), such as those performed in deriving the influential Hedrick’s \( c_\text{ST} \) [9] and Jost's D [33] statistics. We have previously shown that for biallelic loci in K subpopulations, for fixed M, the statistics \( c_\text{ST} \), \( c_{\text{ST}'} \), and D are all maximized at the same set of allele frequency values [15]. Although the normalizations of FST used to produce \( c_\text{ST} \), \( c_{\text{ST}'} \), and D lead to statistics that are unconstrained in the unit interval as functions of \( H_2, c_{\text{ST}} \), and D continue to be constrained as functions of M. A statistic that instead normalizes FST by its maximum as a function of M, a statistic of the total population, captures aspects of the allele frequency dependence of FST that differ from those captured by normalizations by functions of within-subpopulation statistics.

In human populations, efforts to understand FST patterns trace in large part to Lewontin’s foundational FST-like variance-partitioning computation [34], in which it was seen that among-population differences (analogous to FST) were small relative to within-population differences (analogous to 1 − FST). Studies using loci with different numbers of alleles, loci with different frequencies for the most frequent allele, and samples with different numbers of subpopulations have varied to some extent in their numerical estimates of FST [14,35–38]. Mathematical results on FST bounds provide part of the explanation for these differences: they establish that each dataset differing in the character of its loci and subpopulation set has its own distinctive interval in which its associated FST calculation could potentially land. Hence, each dataset can give rise to a numerically distinct value not due to features of the underlying human biology, but rather, due to different constraints on the FST measure itself. FST bounds contribute to explaining quantitative variation in variance-partitioning computations—in which, although numerical values differ, the within-population component of genetic variation consistently predominates. The mathematics serves to support the qualitative claim that worldwide human genetic differentiation measurements represented by FST-like statistics have low values—as was argued by Lewontin 50 years ago.

Data accessibility. Data are publicly available as described in the references cited. MS commands are provided in electronic supplementary material [39].

Authors’ contributions. N.A. and N.A.R. designed the study. N.A. analysed the data. N.A. and N.A.R. wrote the manuscript.

Both authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Competing interests. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

Funding. Support was provided by NIH grant no. R01 HG005855, NSF grant no. BCS-2116322 and a France-Stanford Center for Interdisciplinary Studies grant.

Acknowledgements. We thank Kent Holsinger for comments on the manuscript and Maike Morrison for helpful conversations.

Appendix A. Proof of equation (3.1)

This appendix derives the upper bound on FST as a function of \( \sigma_1 \) (equation 3.1). First, we separate the case of integer values of \( \sigma_1 \). Next, for non-integer values of \( \sigma_1 \), we reduce the problem of maximizing FST to the problem of maximizing the sum of squared allele frequencies across alleles and subpopulations, \( S = \sum_{k=1}^{\infty} \sum_{j=1}^{\sigma_k} p_{i,j}^2 \).

Next, we maximize S as a function of \( \sigma_1 \), separately for \( \sigma_1 \) in (0, 1) and for non-integer \( \sigma_1 \) in (1, K).

(a) A useful expression for \( F_{\text{ST}} \)

Suppose \( K \geq 2 \) is a specified integer. Suppose \( \sigma_1 \) is a fixed value, with \( 0 < \sigma_1 < K \). We leave the number of alleles l unspecified. For each \( i \leq 1 \), we write \( \sigma_i = \sum_{k=1}^{\sigma_1} p_{i,j} \), with \( \sigma_i \geq \sigma_j \) for each \( i \) and \( j \) with \( i \leq j \). For convenience, \( \sigma_1 \) is taken to mean both the function that computes the sum \( \sum_{k=1}^{\sigma_1} p_{i,j} \) for a specified set of values of the \( p_{i,j} \) and a fixed value for that sum.

For each \((k, i)\) with \( 1 \leq k \leq K \) and \( i \geq 1 \), \( p_{k,i} \) lies in \([0, 1]\), and \( \sum_{i=1}^{\sigma_1} p_{k,i} = 1 \) for all \( k \), \( 1 \leq k \leq K \). Define FST as in equation (2.1). We seek to maximize FST over all possible sets of values of the \( p_{k,i} \) with a fixed value \( \sigma_1 \) for the sum \( \sum_{k=1}^{K} p_{1,k} \) . Note that because \( \sigma_1 < K \) and \( \sum_{k=1}^{K} \sum_{i=1}^{\sigma_1} p_{i,k} = \sum_{i=1}^{\sigma_1} \sigma_i = K \), it follows that \( \sigma_2 > 0 \).

Denote the sum of squared frequencies of allele 1 across subpopulations, \( \sum_{k=1}^{K} p_{1,k}^2 \), by \( S_1 \). Denote \( S = \sum_{k, j=1}^{K} p_{k,j}^2 = \sum_{k=1}^{K} \sum_{i=1}^{\sigma_1} p_{k,i}^2 \) for the corresponding sum of squared frequencies of all alleles. We express equation (2.1) in terms of \( \sigma_1, S_1 \) and S:

\[
F_{\text{ST}} = \frac{(K - 1)S + S_1 - \sigma_1^2 - 2 \sum_{k=1}^{K} \sum_{i=1}^{\sigma_1} k - 1 p_{k,i} p_{k,i}}{K^2 - S + S_1 - \sigma_1^2 - 2 \sum_{k=1}^{K} \sum_{i=1}^{\sigma_1} k - 1 p_{k,i} p_{k,i}}.
\]

(A 1)

By construction of equation (2.1), the denominator of equation (A 1) lies in \([0, K^2]\), as \( 0 < H_1 < 1 \) from the fact that \( \sigma_2 > 0 \). The numerator lies in \([0, K^2]\), as \( 0 \leq H_1 \leq H_2 < 1 \), so that \( 0 \leq H_T - H_2 < 1 \). FST lies in \([0, 1]\), as \( 0 \leq H_2 \) and \( 0 < H_1 \) imply \( 0 \leq (H_T - H_2)/H_T \leq 1 \).
(b) The case of integer values of $\sigma_i$

In equation (A 1), the numerator is less than or equal to the denominator, with equality if and only if $K = S = 1$. Equality in turn requires that for each $k$, there exists some $i$ for which $p_{k,i} = 1$, a condition that can be achieved only if $\sigma_i$ is an integer.

**Theorem A.1.** Suppose $\sigma_i$ is an integer value, 1, 2, ..., $K - 1$. $F_{ST} = 1$ if and only if (i) $p_{k,1} = 1$ in each of $\sigma_i$ subpopulations, and (ii) for each of the $K - \sigma_i$ remaining subpopulations, there exists a value of $i \geq 2$ with $p_{k,i} = 1$.

**Proof.** $F_{ST} = 1$ if and only if $S = K$, and $S = K$ if and only if for each $k$, there exists an associated $i$ with $p_{k,i} = 1$. For a fixed integer value of $\sigma_i$, $p_{k,1} = 1$ in exactly $\sigma_i$ subpopulations. $\blacksquare$

Note that any set of equivalence relationships can exist among the values of $i$ associated with the $K - \sigma_i$ subpopulations in which $p_{k,1} = 0$, provided that none of these values of $i$ is associated with more than $\sigma_i$ subpopulations. For example, these values of $i$ can be mutually distinct, or groups of them with size as large as $\sigma_i$ can be mutually equal.

(c) Non-integer values of $\sigma_i$

For non-integer $\sigma_i$, the numerator of equation (A 1) is strictly less than the denominator. Hence, if the other quantities in equation (A 1) are fixed, then $F_{ST}$ decreases with increasing $2 \sum_{i=2}^{\infty} \sum_{k=1}^{k-1} \sum_{l=1}^{K} p_{k,l} p_{l,j}$. We have the following theorem.

**Theorem A.2.** Suppose $\sigma_i$ is not an integer. $F_{ST}$ satisfies

$$F_{ST} \leq \frac{(K-1)S + S_i - \sigma_i^2}{K^2 - S + S_i - \sigma_i^2}$$

equality requiring that for each $i \geq 2$, there exists at most one value of $k$ for which $p_{k,i} > 0$.

**Proof.** Because $\sum_{i=2}^{\infty} \sum_{k=1}^{k-1} \sum_{l=1}^{K} p_{k,l} p_{l,j}$ is subtracted in both the numerator and the denominator of equation (A 1), and because the numerator is strictly less than the denominator for non-integer $\sigma_i$, $F_{ST}$ can be bounded above by minimizing this term. Because $p_{k,i} \geq 0$ for all $(k, i)$, each sum $\sum_{k=1}^{K} \sum_{l=1}^{K} p_{k,l} p_{l,j}$ is bounded below by zero. Setting the sum to 0 for all $i \geq 2$ gives the upper bound in equation (A 2).

For the equality condition, $\sum_{i=2}^{\infty} \sum_{k=1}^{k-1} \sum_{l=1}^{K} p_{k,l} p_{l,j} = 0$ if and only if all products $p_{k,l} p_{l,j}$ are zero—that is, if and only if for each $i \geq 2$, at most one value of $k$ has $p_{k,i} > 0$. $\blacksquare$

By theorem A.2, to maximize $F_{ST}$ for fixed non-integer $\sigma_i$, we must maximize the quantity in equation (A 2). It suffices to consider sets of values of $p_{k,j}$ in which for each $i \geq 2$, at most one value of $k$ has $p_{k,i} > 0$.

(d) The case of (non-integer) $\sigma_i$ in $(0, 1)$

In this section, we find the set of values of the $p_{k,j}$ that maximize $F_{ST}$ for $\sigma_i$ in $(0, 1)$. We proceed in two steps. (i) We show that for $\sigma_i$ in $(0, 1)$, the maximal $F_{ST}$ occurs at a set of $p_{k,j}$ values for which all alleles are private: that is, for each $i \geq 1$, $p_{k,j} > 0$ for at most one value of $k$. (ii) We determine the set of $p_{k,j}$ values that, with all alleles private, maximizes $F_{ST}$.

(i) In equation (A 2), note that $\sigma_i^2 - S_i = 2 \sum_{k=1}^{K} \sum_{l=1}^{k-1} p_{k,l} p_{l,j}$. Because $\sigma_i^2 - S_i$ is subtracted from both numerator and denominator in equation (A 2), the quantity in equation (A 2) is maximal when $\sigma_i^2 - S_i$ is minimal. In other words, the upper bound on $F_{ST}$ is maximal if and only if $2 \sum_{k=1}^{K} \sum_{l=1}^{k-1} p_{k,l} p_{l,j}$ is minimal.

Because $\sigma_i < 1$, a minimum of 0 for $2 \sum_{k=1}^{K} \sum_{l=1}^{k-1} p_{k,l} p_{l,j}$ is achieved if and only if there is a single value $k = k'$ at which $p_{k,i} = \sigma_i$ so that $p_{k,i} = 0$ for all $k \neq k'$. We then have $\sigma_i^2 = S_i$, and from equation (A 2),

$$F_{ST} \leq \frac{(K-1)S}{K^2 - S}.$$  \hspace{1cm} (A 3)

Each allele is private, and because allele 1 is the most frequent, $p_{k,1}$ lies in $[0, \sigma_i]$ for all $(k, i)$.

(ii) The problem of finding the set of $p_{k,j}$ values that maximizes $F_{ST}$ has now been reduced to the problem of maximizing the right-hand side of equation (A 3), with the constraint that all alleles are private. Because the numerator in equation (A 3) increases with $S$ and the denominator decreases with $S$, the maximum is achieved if and only if $S$ achieves its maximal value. In other words, we seek to maximize $S = 2 \sum_{k=1}^{K} \sum_{l=1}^{k-1} p_{k,l} p_{l,j}$ with the constraints $\sum_{k=1}^{K} p_{k,1} = 1$ and $p_{k,1} \leq \sigma_i$ for each $(k, i)$ with $1 \leq k \leq K$ and $i \geq 2$. Because each allele is private, the maximum is achieved by separately maximizing each $\sum_{l=1}^{K} p_{k,l} p_{l,j}$ with constraints $\sum_{l=1}^{K} p_{k,l} = 1$ and $p_{k,1} \leq \sigma_i$.

This maximization is precisely that of lemma 3 of Rosenberg & Jakobsson [40]. Applying the lemma, the maximum is achieved with $p_{k,1} = p_{k,2} = \ldots = p_{k,l-1} = \sigma_i$, $p_{k,l} = 1 - (1 - \sigma_i)$, and $p_{k,1} = 0$ for $i > j$, where $j = \lceil \sigma_i \rceil$. It satisfies $\sum_{k=1}^{K} \sum_{l=1}^{K} p_{k,l} p_{l,j} \leq 1 - (1 - \sigma_i)$ in (2) (2 - $\sigma_i$)]. In other words, each subpopulation $k$ possesses $j$ private alleles with frequency $\sigma_i$ and one private allele with frequency $1 - (1 - \sigma_i)$. Hence, $S \leq K(1 - \sigma_i)(1 - (2 - \sigma_i))$, so that equation (A 3) leads to equation (3.1) for $\sigma_i$ in (0, 1).

(e) The case of non-integer $\sigma_i$ in $(1, K)$

This section finds the set of values of the $p_{k,j}$ that maximizes $F_{ST}$ for non-integer $\sigma_i$ in $(1, K)$. For non-integer $\sigma_i = \sum_{k=1}^{K} p_{k,1}$ in $(1, K)$, because $0 < p_{k,1} \leq 1$ for all $k$, $p_{k,1} > 0$ for at least two values of $k$. Writing $S^* = S - S_i$, equation (A 2) can be rewritten

$$F_{ST} \leq \frac{KS_1 + (K-1)S^* - \sigma_i^2}{K^2 - S^* - \sigma_i^2}.$$  \hspace{1cm} (A 4)

Because the numerator increases with $S_1$ and because the numerator increases with $S^*$ and the denominator decreases with $S^*$, the upper bound on $F_{ST}$ is greatest when both $S_1$ and $S^*$ are maximized subject to $\sum_{k=1}^{K} p_{k,1} = 1$ for each $k$ and $\sum_{k=1}^{K} p_{k,1} \leq \sigma_i$ for each $i$. If $S_1$ and $S^*$ can be simultaneously maximized at the same set of values of the $p_{k,j}$, then this set of values of the $p_{k,j}$ achieves the maximal $F_{ST}$.

We proceed in three steps. (i) First, we find the set of values of the $p_{k,j}$ that maximizes $S_1$. (ii) Next, we find the set of values that maximizes $S^*$. (iii) We then conclude that because the same set maximizes both $S_1$ and $S^*$ separately, this set achieves the upper bound in equation (A 4), and hence in equation (A 2).

(i) We first maximize $S_1$ for fixed non-integer $\sigma_i$ in $(1, K)$. More precisely, we seek to maximize $S_1 = \sum_{k=1}^{K} p_{k,1}^2$ with constraints $\sum_{k=1}^{K} p_{k,1} = \sigma_i$ and $p_{k,1} \leq 1$ for each $k$ from 1 to $K$. This
maximization is precisely that performed in theorem 1 from Alcala & Rosenberg [14], a corollary of lemma 3 of Rosenberg & Jakobsson [40]. Applying the theorem, the maximum is achieved by setting $p_{1,1} = p_{2,1} = \ldots = p_{i,1} = 1$, $p_{1,i+1} = \ldots = p_{i,i+1} = 1$, and $p_{k,1} = 0$ for all $k > i + 1$. The maximal value of $S_1$ is $\left\lceil \frac{1}{1} \right\rceil + \left\lceil \frac{1}{1} \right\rceil + \ldots + \left\lceil \frac{1}{1} \right\rceil + \left\lceil \frac{1}{1} \right\rceil$. 

(ii) Next, we maximize $S^* = \sum_{i=2}^{\infty} \sum_{k=1}^{s_i} p_{k,i}$. Because, by theorem A.2, all alleles with $i \geq 2$ are private at the set of values of the $p_{k,i}$ that maximizes $F_{ST}$ for fixed non-integer $s_i$, each non-zero $p_{k,i}$ for $i \geq 2$ is equal to the associated $s_i$. The sum of the frequencies of all alleles across all subpopulations is $\sum_{i=2}^{\infty} s_i = K$, so that $\sum_{i=2}^{\infty} s_i K = K - s_1$. The problem of maximizing $S^*$ is the problem of maximizing $S^* = \sum_{i=2}^{\infty} \sum_{k=1}^{s_i} s_i$ with the constraints $\sum_{i=2}^{\infty} s_i K = K - s_1$ and $s_i \leq 1$ for each $i$ from 2 to $\infty$. This maximization is achieved by performed in equation (3.1).

Substituting the maximal values of $S_1$ and $S^*$ into equation (3.1), for non-integer $s_1$ in $(1, K)$, we obtain the maximal $F_{ST}$ in terms of $s_1$ shown in equation (3.1).

References

1. Nei M. 1973 Analysis of gene diversity in subdivided populations. Proc. Natl Acad. Sci. USA 70, 3231–3233. (doi:10.1073/pnas.70.12.3231)
2. Holsinger KE, Weir BS. 2009 Genetics in geographically structured populations: defining, estimating and interpreting $F_{ST}$. Nat. Rev. Genet. 10, 639–650. (doi:10.1038/nrg26111)
3. Jin L, Chakraborty R. 1995 Population structure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics. Heredity 74, 274–285. (doi:10.1038/hdy.1995.41)
4. Charlesworth B. 1998 Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. 15, 538–543. (doi:10.1093/molbev/msv015)
5. Nagylaki T. 1998 Fixation indices in subdivided populations. Genetics 148, 1325–1332. (doi:10.1093/genetics/148.3.1325)
6. Hedrick PW. 1999 Highly variable loci and their interpretation in evolution and conservation. Evolution 53, 313–318. (doi:10.1111/j.1558-5646.1999.tb03767.x)
7. Balloux F, Brünnner H, Lugon-Moulin N, Haussler J, Goudet J. 2000 Microsatellites can be misleading: an empirical and simulation study. Evolution 54, 1414–1422. (doi:110.1111/j.0016-6560.2000.tb05737.x)
8. Long JC, Kittles RA. 2003 Human genetic diversity and the nonexistence of biological races. Hum. Biol. 75, 449–471. (doi:10.1353/hub.2003.0058)
9. Hedrick PW. 2005 A standardized genetic differentiation measure. Evolution 59, 1633–1638. (doi:110.1111/j.0016-6560.2005.tb01614.x)
10. Maruki T, Kumar S, Kim Y. 2012 Purifying selection modulates the estimates of population differentiation and confounds genome-wide comparisons across single-nucleotide polymorphisms. Mol. Biol. Evol. 29, 3617–3623. (doi:10.1093/molbev/msq337)
11. Rosenberg NA, Li LM, Ward R, Pritchard JK. 2003 Informativeness of genetic markers for inference of ancestry. Am. J. Hum. Genet. 73, 1402–1422. (doi:10.1093/aajg/304061)
12. Jakobsson M, Edge MD, Rosenberg NA. 2013 The relationship between $F_{ST}$ and the frequency of the most frequent allele. Genetics 193, 515–528. (doi:10.1534/genetics.112.144758)
13. Edge MD, Rosenberg NA. 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. (doi:10.1101/j.pdb.2014.08.001)
14. Alcala N, Rosenberg NA. 2017 Mathematical constraints on $F_{ST}$: biallelic markers in arbitrarily many populations. Genetics 206, 1581–1600. (doi:10.1093/genetics/116.1999141)
15. Alcala N, Rosenberg NA. 2019 $F_{ST}$, Jost’s $D$, and $F_{ST}$ are similarly constrained by allele frequencies: a mathematical, simulation, and empirical study. Mol. Ecol. 28, 1624–1636. (doi:10.1111/mec15000)
16. Weir BS, Cockerham CC. 1984 Estimating $F$-statistics for the analysis of population structure. Evolution 38, 1358–1370. (doi:10.1111/j.1558-5646.1984.tb00567.x)
17. Whitlock MC. 2011 $G_{ST}$ and $D$ do not replace $F_{ST}$. Mol. Biol. Evol. 20, 1083–1091. (doi:110.1111/j.1558-7009.2010.004996.x)
18. Rousset F. 2013 Exeeges on maximum genetic differentiation. Genetics 194, 557–559. (doi:10.1101/153443.1152132)
19. Alcala N, Goudet J, Vaillamont S. 2014 On the transition of genetic differentiation from isolation to panmixia: what we can learn from $G_{ST}$ and $D$. Theor. Popul. Biol. 93, 75–84. (doi:10.1016/j.tpb.2014.02.003)
20. Wang J. 2015 Does $G_{ST}$ underestimate genetic differentiation from marker data? Mol. Ecol. 24, 3546–3558. (doi:10.1111/mec.13204)
21. Hudson RR. 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18, 337–338. (doi:10.1093/bioinformatics/18.2.337)
22. Marumya T. 1970 Effective number of alleles in a subdivided population. Theor. Popul. Biol. 1, 273–306. (doi:10.1006/tpbi.1997.1535)
23. Wakeley J. 1998 Segregating sites in Wright’s island model. Theor. Popul. Biol. 53, 166–174. (doi:10.1006/tpbi.1998.1535)
24. Fu R, Gelfand AE, Holsinger KE. 2003 Exact moment calculations for genetic models with migration, mutation, and drift. Theor. Popul. Biol. 63, 231–243. (doi:10.1016/j.theo.2003.09.003)
25. Long JC. 2009 Update to Long and Kittles’s ‘Human genetic diversity and the nonexistence of biological races’ (2003): fixation on an index. Hum. Biol. 81, 799–803. (doi:10.1353/hub.2009.0162)
26. Pemberton TJ, DeGiorgio M, Rosenberg NA. 2013 Population structure in a comprehensive genomic data set on human microsatellite variation. Genet. Mol. Biol. 36, 891–907. (doi:10.1534/g3.113.0065728)
27. Becquet C, Patterson N, Stone AC, Przeworski M, Reich D. 2007 Genetic structure of chimpanzee populations. *PloS Genet.* 3, e66. (doi:10.1371/journal.pgen.0030066)

28. Algee-Hewitt BF, Edge MD, Kim J, Li JZ, Rosenberg NA. 2016 Individual identifiability predicts population identifiability in forensic microsatellite markers. *Curr. Biol.* 26, 935–942. (doi:10.1016/j.cub.2016.01.065)

29. Maróstica AS, Nunes K, Castelli EC, Silva NSB, Weir BS, Goudet J, Meyer D. 2022 How HLA diversity is apportioned: influence of selection and relevance to transplantation. *Phil. Trans. R. Soc. B* 377, 20200420. (doi:10.1098/rstb.2020.0420)

30. Beaumont MA. 2010 Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* 41, 379–406. (doi:10.1146/annurev-ecolsys-102209-144621)

31. Schrider DR, Kern AD. 2018 Supervised machine learning for population genetics: a new paradigm. *Trends Genet.* 34, 301–312. (doi:10.1016/j.tig.2017.12.005)

32. Hoban S et al. 2016 Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am. Nat.* 188, 379–397. (doi:10.1086/688018)

33. Jost L. 2008 *GST* and its relatives do not measure differentiation. *Mol. Ecol.* 17, 4015–4026. (doi:10.1111/j.1365-294X.2008.03887.x)

34. Lewontin RC. 1972 The apportionment of human diversity. *Am. Nat.* 8, 381–398. (doi:10.1007/978-1-4684-9063-3_14)

35. Brown RA, Armelagos GJ. 2001 Apportionment of racial diversity: a review. *Am. J. Phys. Anthropol.* 10, 34–40. (doi:10.1002/ajpa.10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<34::AID-AJPA10230-66055(2001)10:1<