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

Nicolas Alcala[1] and Noah A. Rosenberg[2]

[1]Genomic Epidemiology Branch, International Agency for Research on Cancer/World Health Organization, Lyon, 69372, France
[2]Department of Biology, Stanford University, Stanford, CA 94305-5020, USA

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}.

©2021 The authors

https://doi.org/XXX

Manuscript compiled: Monday 25th October, 2021

Subject Areas:
statistical genetics

Keywords:
allele frequency, chimpanzee, genetic differentiation, migration, population structure

Author for correspondence:
Nicolas Alcala
e-mail: alcalan@iarc.fr

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]:

$$F_{ST} = \frac{H_T - H_S}{H_T}. \tag{1}$$

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

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, 4, 5, 6, 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 $\frac{1}{2}$ 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 $0 < M < \frac{1}{2}$, and decreases from 1 to 0 for $\frac{1}{2} \leq M < 1$ in the manner reported by MARUKI *et al.* [10] for biallelic loci. EDGE and 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 $K = I = 2$ in a different direction, ALCALA and 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 behavior as a function of M: the upper bound has a maximum of 1 if and only if $M = \frac{k}{K}$, for integers k with $\lceil \frac{K}{2} \rceil \leq k \leq K - 1$. 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 $F_{ST} < 2\sqrt{2} - 2 \approx 0.8284$.

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 and 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 and 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 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, 13, 14, 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 $p_{k,i}$, with sum $\sigma_i = \sum_{k=1}^{K} p_{k,i}$ across subpopulations. Each allele frequency $p_{k,i}$ lies in $[0, 1]$. Within subpopulations, allele frequencies sum to 1: for each k, $\sum_{i=1}^{\infty} p_{k,i} = 1$. Hence, σ_i lies in $[0, K]$, and $\sum_{i=1}^{\infty} \sigma_i = K$. 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 nonzero 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 $p_{k,i}$ and associated quantities M and σ_i as parametric values, and not as estimates computed from data.

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

$$F_{ST} = \frac{\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{\infty} p_{k,i}^2 - \sum_{i=1}^{\infty}\left(\sum_{k=1}^{K}\frac{p_{k,i}}{K}\right)^2}{1 - \sum_{i=1}^{\infty}\left(\sum_{k=1}^{K}\frac{p_{k,i}}{K}\right)^2}. \tag{2}$$

Hence, our goal is, for fixed $\sigma_1 = KM$, $0 < \sigma_1 < K$, to identify the matrices $(p_{k,i})_{K\times\infty}$, with $p_{k,i}$ in $[0, 1]$, $\sum_{i=1}^{\infty} p_{k,i} = 1$ and $\frac{1}{K}\sum_{k=1}^{K} p_{k,1} = \sigma_1/K = M$, that minimize and maximize F_{ST} in eq. 2.

Note that we adopt the interpretation of F_{ST} 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 and ROSENBERG [14] for a discussion of interpretations of F_{ST} when studying its mathematical properties.

Table 1 Studies describing the mathematical constraints on F_{ST}.

Reference	Number of alleles	Number of subpopulations	Variable in terms of which constraints are reported*
LONG and KITTLES [8]	unspecified value ≥ 2	fixed finite value ≥ 2	H_S
ROSENBERG *et al.* [11]	2	2	δ
HEDRICK [9]	unspecified value ≥ 2	fixed finite value ≥ 2	H_S
MARUKI *et al.* [10]	2	2	H_S, M
JAKOBSSON *et al.* [12]	unspecified value ≥ 2	2	H_T, M
EDGE and ROSENBERG [13]	fixed finite value ≥ 2	2	H_T, M
ALCALA and ROSENBERG [14]	2	fixed finite value ≥ 2	M
This paper	unspecified value ≥ 2	fixed finite value ≥ 2	M

H_S and H_T denote the within-subpopulation and total heterozygosities, respectively. δ denotes the absolute difference in the frequency of a specific allele between two subpopulations, and M denotes the frequency of the most frequent allele in the total population. Instead of heterozygosities H_S or H_T, some studies consider homozygosities $1 - H_S$ or $1 - H_T$.

3. Mathematical constraints

(a) Lower bound of F_{ST}

Bounds on F_{ST} in terms of the frequency of the most frequent allele can be written with respect to M or σ_1, noting that M ranges in $(0,1)$ and σ_1 ranges in $(0,K)$. For the lower bound, from eq. 2, for any choice of σ_1, $F_{ST} = 0$ can be achieved. Consider $(\sigma_1, \sigma_2, \ldots)$ with σ_i in $[0,K)$ for each k, $\sigma_i \geq \sigma_j$ for $i \leq j$, $\sum_{i=1}^{\infty} \sigma_i = K$, and $\sigma_1 > 0$ and $\sigma_2 > 0$. We set $p_{k,i} = \sigma_i / K$ for all subpopulations k and alleles i; this choice yields $F_{ST} = 0$.

$F_{ST} = 0$ implies that the numerator of eq. 2, $H_T - H_S$, is zero. This numerator can be written $(\frac{1}{K^2}) \sum_{i=1}^{\infty} (K \sum_{k=1}^{K} p_{k,i}^2 - \sigma_i^2)$. The Cauchy-Schwarz inequality guarantees that $K \sum_{k=1}^{K} p_{k,i}^2 \geq \sigma_i^2$, with equality if and only if $p_{1,i} = p_{2,i} = \ldots = p_{K,i} = \sigma_i / K$. Applying the Cauchy-Schwarz inequality to all alleles i, the numerator of eq. 2 is zero only if for all i, $(p_{1,i}, p_{2,i}, \ldots, p_{K,i}) = (\sigma_i/K, \sigma_i/K, \ldots, \sigma_i/K)$.

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 $F_{ST} = 0$. The lower bound on F_{ST} is equal to 0 irrespective of M or σ_1, for any value of the number of subpopulations K.

(b) Upper bound of F_{ST}

To derive the upper bound on F_{ST} in terms of $M = \sigma_1 / K$, we must maximize F_{ST} in eq. 2, assuming that σ_1 and K are constant. The computations are performed in the Appendix; we write the main result as a function of σ_1, noting that it can be converted into a function of M by replacing σ_1 with KM.

In Theorem 1, we treat the case in which σ_1 has an integer value. For non-integer σ_1, Theorem 2 shows that the maximal F_{ST} requires that (i) the sum of squared allele frequencies across alleles and subpopulations, $S = \sum_{i=1}^{\infty} \sum_{k=1}^{K} p_{k,i}^2$, 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 F_{ST} as a function of σ_1 for σ_1 in $(0,1)$ and non-integer σ_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 F_{ST} as a function of σ_1 for σ_1 in $(0,K)$ is

$$F_{ST} \leq \begin{cases} 1, & \sigma_1 = 1, 2, \ldots, K-1, \\ \dfrac{(K-1)[1 - \sigma_1(J-1)(2 - J\sigma_1)]}{K - [1 - \sigma_1(J-1)(2 - J\sigma_1)]}, & 0 < \sigma_1 < 1, \\ \dfrac{K(K-1) - \sigma_1^2 + \lfloor \sigma_1 \rfloor - 2(K-1)\{\sigma_1\} + (2K-1)\{\sigma_1\}^2}{K(K-1) - \sigma_1^2 - \lfloor \sigma_1 \rfloor + 2\sigma_1 - \{\sigma_1\}^2}, & \\ & \text{non-integer } \sigma_1,\ 1 < \sigma_1 < K, \end{cases} \tag{3}$$

where $J = \lceil \sigma_1^{-1} \rceil$. Here, $\lceil x \rceil$ denotes the smallest integer greater than or equal to x, $\lfloor x \rfloor$ 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 σ_1, the maximum from eq. 3 and the limits as σ_1 tends to the integer from above and below all equal 1, so that the maximum as a function of σ_1 is continuous.

From the Appendix, F_{ST} reaches its upper bound for integer σ_1 when allele 1 has frequency 1 in each of σ_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 σ_1. The locus can have as few as $\lceil K\sigma_1^{-1} \rceil$ alleles of nonzero frequency and as many as $K - \sigma_1 + 1$.

For σ_1 in interval $(0,1)$, F_{ST} is maximal when each allele is present in only a single subpopulation, and when each subpopulation has exactly J alleles with a nonzero frequency: $J - 1$ alleles at frequency σ_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 nonzero frequency.

For non-integer σ_1 in $(1,K)$, F_{ST} reaches its maximum when there are $\lfloor \sigma_1 \rfloor$ 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 $K - \lfloor \sigma_1 \rfloor - 1$ 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, $K - \lfloor \sigma_1 \rfloor + 1$ alleles have nonzero frequency.

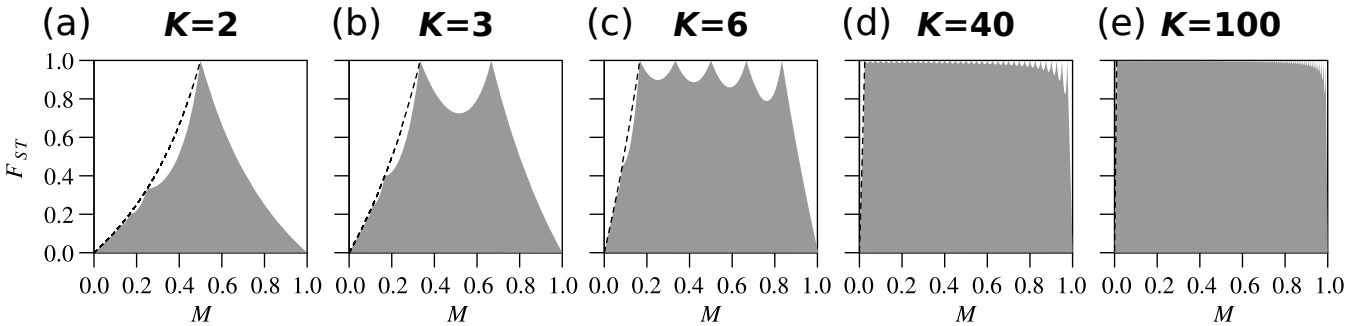

Figure 1 Bounds on F_{ST} as a function of the frequency of the most frequent allele, M, for a multiallelic locus, for each of several different numbers of subpopulations K. (a) $K = 2$. (b) $K = 3$. (c) $K = 6$. (d) $K = 40$. (e) $K = 100$. The gray region represents the space between the upper and lower bounds on F_{ST}. The dashed line represents the curve that the jagged maximal F_{ST} touches when $M < \frac{1}{K}$, computed from eq. 4. The upper bound is computed from eq. 3; for each K, the lower bound is 0 for all values of M.

(c) Properties of the upper bound

Figure 1 shows the maximal value of F_{ST} 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.

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

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

At $M = \frac{1}{K}$, the upper bound has its first transition between cases. For $M > \frac{1}{K}$, the upper bound depends on $\lfloor \sigma_1 \rfloor = \lfloor KM \rfloor$. As KM increases in $[1, K)$, each increment in $\lfloor KM \rfloor$ also produces a distinct piece of the domain. For each k from 1 to $K-1$, $\lfloor KM \rfloor = k$ for M in $[\frac{k}{K}, \frac{k+1}{K})$.

Counting the intervals of the domain, we see that an infinite number of distinct intervals occur for M in $(0, \frac{1}{K})$, and $K-1$ intervals occur for M in $(\frac{1}{K}, 1)$. Within intervals, the function describing the upper bound is smooth.

Behavior of the upper bound for $M = \frac{1}{K}, \frac{2}{K}, \ldots \frac{K-1}{K}$. The upper bound is equal to 1 at $M = \frac{1}{K}, \frac{2}{K}, \ldots \frac{K-1}{K}$. For M in $(0, \frac{1}{K})$, setting the numerator and denominator equal in eq. 3, we find that the upper bound is never equal to 1. For M in $(\frac{1}{K}, 1)$, the upper bound is equal to 1 if and only if $\{\sigma_1\} = 0$, that is, if and only if σ_1 is an integer and $M = \frac{k}{K}$ for $k = 2, 3, \ldots, K-1$.

Hence, noting that the upper bound is equal to 1 at $M = \frac{1}{K}$, we conclude that the upper bound can equal 1 if and only if $M = \frac{k}{K}$ for integers $k = 1, 2, \ldots, K-1$. For fixed K, the upper bound on F_{ST} has exactly $K-1$ maxima at which F_{ST} can equal 1, at $M = \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}$. We can conclude that F_{ST} 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.

Behavior of the upper bound for M in $(0, \frac{1}{K})$. For M in $(0, \frac{1}{K})$, we can compute the value of the upper bound at the transition points between distinct pieces of the domain, namely values of $\frac{1}{Kj}$ for integers $j \geq 2$. Applying eq. 3, we observe that at $M = \frac{1}{Kj}$, the upper bound has value $\frac{K-1}{Kj-1}$. In other words, the upper bound touches the curve

$$q^*(M) = \frac{(K-1)M}{1-M}. \tag{4}$$

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

Note that for $K = 2$, the special case considered by JAKOBSSON *et al.* [12], eq. 4 reduces to $q^*(M) = M/(1-M) = \sigma_1/(2-\sigma_1)$, which matches eq. 21 from JAKOBSSON *et al.* [12]. In fact, setting $K = 2$, eq. 3 for M in $(0, \frac{1}{K})$ reduces to the $K = 2$ upper bound on F_{ST} in eq. 9 of [12].

Behavior of the upper bound for M in $(\frac{1}{K}, 1)$. Because the upper bound is a smooth function on each interval of its domain, and because it possesses maxima at interval boundaries $M = \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}$, it must possess local minima in intervals $[\frac{k}{K}, \frac{k+1}{K})$ for $k = 1, 2, \ldots, K-2$. 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 $F_{ST} = 0$ as a local minimum.

4. Joint distribution of M and F_{ST} under an evolutionary model

So far, we have described the mathematical constraint imposed on F_{ST} 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, 18, 19, 20], we simulated the joint distribution of F_{ST} and M under an island migration model, relating the distribution to the mathematical bounds on F_{ST}. This analysis considers allele frequency distributions, and hence values of M and F_{ST}, 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

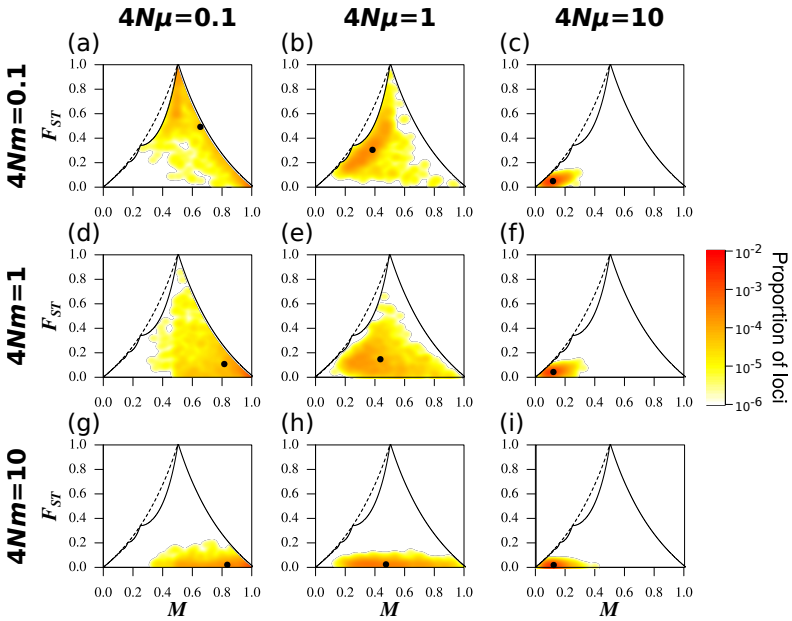

Figure 2 Joint density of the frequency M of the most frequent allele and F_{ST} in the island migration model with $K = 2$ subpopulations, for different scaled migration rates $4Nm$ and mutation rates $4N\mu$. (a) $4N\mu = 0.1$, $4Nm = 0.1$. (b) $4N\mu = 1$, $4Nm = 0.1$. (c) $4N\mu = 10$, $4Nm = 0.1$. (d) $4N\mu = 0.1$, $4Nm = 1$. (e) $4N\mu = 1$, $4Nm = 1$. (f) $4N\mu = 10$, $4Nm = 1$. (g) $4N\mu = 0.1$, $4Nm = 10$. (h) $4N\mu = 1$, $4Nm = 10$. (i) $4N\mu = 10$, $4Nm = 10$. The black solid line represents the upper bound on F_{ST} in terms of M (eq. 3); the black point plots the mean values of M and F_{ST}. Colors 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 1,000 replicate simulations, with 100 lineages sampled per subpopulation. Figures S1 and S2 present similar results for $K = 6$ and $K = 40$ subpopulations, respectively.

individuals in a subpopulation that originate elsewhere. We considered the island model [22, 23, 24], in which migrants have the same probability $\frac{m}{K-1}$ to come from any specific other subpopulation. We used an infinitely-many-alleles model; mutations occur at rate μ, and the scaled mutation rate is $4N\mu$.

We examined three values of K (2, 6, 40), three values of $4N\mu$ (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 data sets as having been generated under an infinitely-many-alleles model.

For each parameter triplet $(K, 4N\mu, 4Nm)$, we performed 1,000 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 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 simulated datasets to possess exactly one segregating site.

Figure 2 shows the joint distribution of M and F_{ST} for the nine values of $(4N\mu, 4Nm)$ in the case of $K = 2$. Figures S1 and S2 provide similar figures for $K = 6$ and $K = 40$, respectively.

(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 $4Nm$ (Fig. 2A, D, G). Although most simulation replicates produce $M > \frac{1}{2}$ with an upper bound on F_{ST} less than one, this set of parameter values does give rise to replicates near the peak at $(M, F_{ST}) = (\frac{1}{2}, 1)$.

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 $4Nm$ (Fig. 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 $4Nm$ (Fig. 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 (Fig. 2A, B, 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} (Fig. 2D, E, F). Under strong migration ($4Nm = 10$), the joint density of M and F_{ST} nears the lower bound (Fig. 2G, H, 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 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, 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 (Fig. 3A); for $K = 6$, F_{ST} is not as close to the bound (Fig. 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

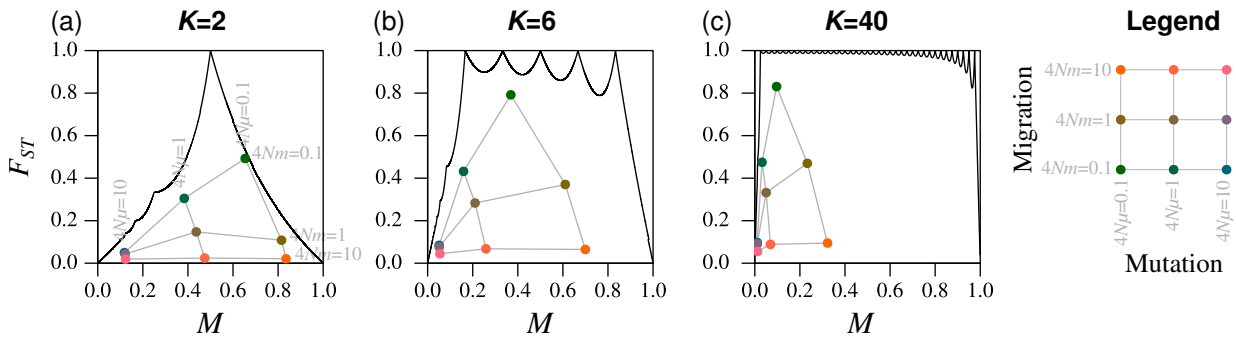

Figure 3 Mean frequency M of the most frequent allele and mean F_{ST} in the island migration model, for different scaled migration rates $4Nm$ and mutation rates $4N\mu$ and different numbers of subpopulations K. (a) $K = 2$. (b) $K = 6$. (c) $K = 40$. The black solid line represents the upper bound on F_{ST} in terms of M (eq. 3). The colored points represent the mean M and mean F_{ST}, where colors correspond to values of $4Nm$. These points are taken from Figures 2, S1, and S2.

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 6 populations: one bonobo population, and 5 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 $\bar{F}_{ST}$, is

$$\bar{F}_{ST} = \frac{1}{Z}\sum_{z=1}^{Z} F_z. \tag{5}$$

Using eq. 3, we can compute the corresponding maximum F_{ST} given the observed $\sigma_z = KM_z$, $z = 1, 2, \ldots, Z$. Denoting this quantity by $F_{\max,z}$, we have

$$\overline{F_{ST}/F_{\max}} = \frac{1}{Z}\sum_{z=1}^{Z} \frac{F_z}{F_{\max,z}}. \tag{6}$$

$\overline{F_{ST}/F_{\max}}$ 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 $\bar{F}_{ST}$ and $\overline{F_{ST}/F_{\max}}$ for the full microsatellite set (eqs. 5 and 6).

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 ($\bar{F}_{ST} = 0.16$) than when comparing humans and chimpanzees ($\bar{F}_{ST} = 0.10$; $p = 4.2 \times 10^{-14}$, 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 (Fig. 4A, 4B). 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 (Fig. 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 $K = 6$ (Fig. 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 byproduct 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 (Fig. 4C). The normalization leads to higher values for human–chimpanzee comparisons than among chimpanzee subpopulations ($\overline{F_{ST}/F_{\max}} = 0.32$ and 0.20, respectively; $p = 1.1 \times 10^{-9}$, Wilcoxon rank sum test), as expected from the greater evolutionary distance between humans and chimpanzees compared to that among chimpanzees.

6. Discussion

We have analyzed 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 $M = \frac{k}{K}$ 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 $I = 2$ alleles [14]. Interestingly, the maximal F_{ST} we have obtained merges patterns observed in these previous studies. Fixing $K =$

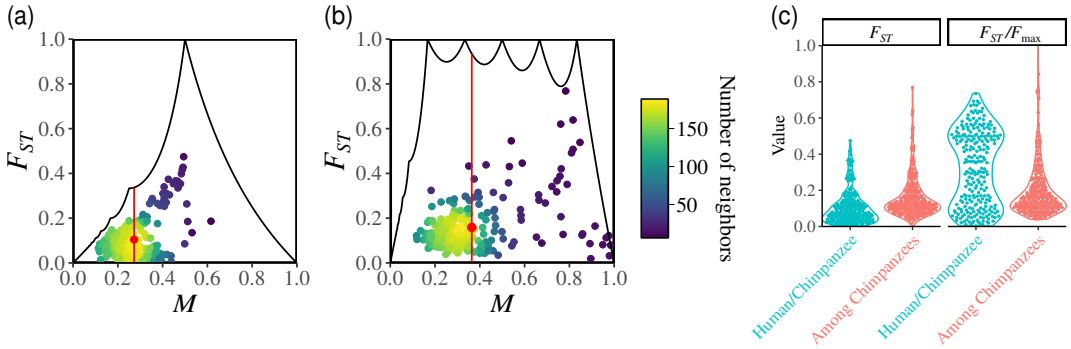

Figure 4 F_{ST} values for comparisons involving humans and chimpanzees based on multiallelic microsatellite loci. (a) F_{ST} between humans and chimpanzees, considering $K = 2$ subpopulations (humans, chimpanzees). (b) F_{ST} among $K = 6$ chimpanzee subpopulations. In (a) and (b), colors represent the number of points in a neighborhood of radius 0.03; red points indicate the mean M and F_{ST}, and vertical red segments indicate the permissible range of F_{ST} at the mean M. (c) F_{ST}, computed using eq. 2, and $F_{ST}/F_{\max}$, computed using eqs. 2 and 3. Each point plotted represents one locus.

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 $(0, \frac{1}{2})$ is observed in the multiallelic case for M in $(0, \frac{1}{K})$. The decay from $(M, F_{ST}) = (\frac{1}{2}, 1)$ to $(M, F_{ST}) = (1, 0)$ seen by JAKOBSSON *et al.* [12] for $K = 2$ is observed for M in the decay from $(\frac{K-1}{K}, 1)$ to $(1, 0)$ for arbitrary K.

The allele frequency values for which the upper bound is reached for M in $(0, \frac{1}{K})$ generalize those seen for the case of $K = 2$ and M in $(0, \frac{1}{2})$ [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 $(\frac{K-1}{K}, 1)$ also generalize those seen for $K = 2$ and M in $(\frac{1}{2}, 1)$: 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 and ROSENBERG [14] for $I = 2$ produce a more constrained upper bound on F_{ST} than for arbitrary I, with the domain of M restricted to $(\frac{1}{2}, 1)$. Nevertheless, many properties of the maximal F_{ST} we observe for unspecified I and M in $(\frac{1}{K}, 1)$ are similar to those seen for $I = 2$ and M in $(\frac{1}{2}, 1)$: finitely many peaks at points $M = \frac{k}{K}$, 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 $(\frac{K-1}{K}, 1)$ for unspecified I and for $I = 2$ 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 $K\lceil\sigma_1^{-1}\rceil$ for M in $(0, \frac{1}{K})$ and $K - \lfloor\sigma_1\rfloor + 1$ for non-integer M in $(\frac{1}{K}, 1)$; the maximum can be achieved with each number of distinct alleles in $[\lceil K\sigma_1^{-1}\rceil, K - \sigma_1 + 1]$ for M equal to $\frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}$. With a fixed maximal number of distinct alleles, such as in the $I = 2$ case of ALCALA and 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 and 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 $4N\mu$, 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_S, 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 F_{ST} within their bounds can suggest that constraints have an effect. $F_{ST}/F_{\max}$ can provide a helpful summary by evaluating the proximity of F_{ST} values to their maxima.

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

In an example data analysis, we have shown that taking into account mathematical constraints on F_{ST} can help understand puzzling F_{ST} behavior. In our example, F_{ST} 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 F_{ST} values to their respective maxima explained these counterintuitive results.

We note that analyses of F_{ST} in relation to M differ from anal-

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

In human populations, efforts to understand F_{ST} patterns trace in large part to Lewontin's foundational F_{ST}-like variance-partitioning computation [34], in which it was seen that among-population differences (analogous to F_{ST}) were small relative to within-population differences (analogous to $1 - F_{ST}$). 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 F_{ST} [14, 35, 36, 37, 38]. Mathematical results on F_{ST} bounds provide part of the explanation for these differences: they establish that each data set differing in the character of its loci and subpopulation set has its own distinctive interval in which its associated F_{ST} calculation could potentially land. Hence, each data set 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 F_{ST} measure itself. F_{ST} 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 F_{ST}-like statistics have low values—as was argued by Lewontin fifty years ago.

Data accessibility All data are publicly available (see cited references).

Authors' contributions NA and NAR designed the study. NA analysed the data. NA and NAR wrote the manuscript.

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 R01 HG005855, NSF grant 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.

References

[1] NEI, M., 1973 Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences USA **70**: 3321–3323.

[2] HOLSINGER, K. E., and B. S. WEIR, 2009 Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nature Reviews Genetics **10**: 639–650.

[3] JIN, L., and R. CHAKRABORTY, 1995 Population structure, stepwise mutations, heterozygote deficiency and their implications in dna forensics. Heredity **74**: 274–285.

[4] CHARLESWORTH, B., 1998 Measures of divergence between populations and the effect of forces that reduce variability. Molecular Biology and Evolution **15**: 538–543.

[5] NAGYLAKI, T., 1998 Fixation indices in subdivided populations. Genetics **148**: 1325–1332.

[6] HEDRICK, P. W., 1999 Highly variable loci and their interpretation in evolution and conservation. Evolution **53**: 313–318.

[7] BALLOUX, F., H. BRÜNNER, N. LUGON-MOULIN, J. HAUSSER, and J. GOUDET, 2000 Microsatellites can be misleading: an empirical and simulation study. Evolution **54**: 1414–1422.

[8] LONG, J. C., and R. A. KITTLES, 2003 Human genetic diversity and the nonexistence of biological races. Human Biology **75**: 449–471.

[9] HEDRICK, P. W., 2005 A standardized genetic differentiation measure. Evolution **59**: 1633–1638.

[10] MARUKI, T., S. KUMAR, and Y. KIM, 2012 Purifying selection modulates the estimates of population differentiation and confounds genome-wide comparisons across single-nucleotide polymorphisms. Molecular Biology and Evolution **29**: 3617–3623.

[11] ROSENBERG, N. A., L. M. LI, R. WARD, and J. K. PRITCHARD, 2003 Informativeness of genetic markers for inference of ancestry. American Journal of Human Genetics **73**: 1402–1422.

[12] 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.

[13] 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. Theoretical Population Biology **97**: 20–34.

[14] ALCALA, N., and N. A. ROSENBERG, 2017 Mathematical constraints on F_{ST}: biallelic markers in arbitrarily many populations. Genetics **206**: 1581–1600.

[15] ALCALA, N., and N. A. ROSENBERG, 2019 G'_{ST}, Jost's D, and F_{ST} are similarly constrained by allele frequencies: A mathematical, simulation, and empirical study. Molecular Ecology **28**: 1624–1636.

[16] WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. Evolution **38**: 1358–1370.

[17] WHITLOCK, M. C., 2011 G'_{ST} and D do not replace F_{ST}. Molecular Ecology **20**: 1083–1091.

[18] ROUSSET, F., 2013 Exegeses on maximum genetic differentiation. Genetics **194**: 557–559.

[19] ALCALA, N., J. GOUDET, and S. VUILLEUMIER, 2014 On the transition of genetic differentiation from isolation to panmixia: what we can learn from G_{ST} and D. Theoretical Population Biology **93**: 75–84.

[20] WANG, J., 2015 Does G_{ST} underestimate genetic differentiation from marker data? Molecular Ecology **24**: 3546–3558.

[21] HUDSON, R. R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics **18**: 337–338.

[22] MARUYAMA, T., 1970 Effective number of alleles in a subdivided population. Theoretical Population Biology **1**: 273–306.

[23] WAKELEY, J., 1998 Segregating sites in Wright's island model. Theoretical Population Biology **53**: 166–174.

[24] FU, R., A. E. GELFAND, and K. E. HOLSINGER, 2003 Exact moment calculations for genetic models with migration, mutation, and drift. Theoretical Population Biology **63**: 231–243.

[25] LONG, J. C., 2009 Update to Long and Kittles's "Human genetic diversity and the nonexistence of biological races" (2003): fixation on an index **81**: 799–803.

[26] PEMBERTON, T. J., M. DEGIORGIO, and N. A. ROSENBERG, 2013 Population structure in a comprehensive genomic data set on human microsatellite variation. G3: Genes, Genomes, Genetics **3**: 891–907.

[27] BECQUET, C., N. PATTERSON, A. C. STONE, M. PRZEWORSKI, and D. REICH, 2007 Genetic structure of chimpanzee populations **3**: e66.

[28] ALGEE-HEWITT, B. F., M. D. EDGE, J. KIM, J. Z. LI, and N. A. ROSENBERG, 2016 Individual identifiability predicts population identifiability in forensic microsatellite markers. Current Biology **26**: 935–942.

[29] MARÓSTICA, A., K. NUNES, E. CASTELLI, N. SILVA, B. S. WEIR, *et al.*, 2022 Population structure in the MHC region. Philosophical Transactions of the Royal Society of London B **377**: xx.

[30] BEAUMONT, M. A., 2010 Approximate Bayesian computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics **41**: 379–406.

[31] SCHRIDER, D. R., and A. D. KERN, 2018 Supervised machine learning for population genetics: a new paradigm. Trends in Genetics **34**: 301–312.

[32] HOBAN, S., J. L. KELLEY, K. E. LOTTERHOS, M. F. ANTOLIN, G. BRADBURD, *et al.*, 2016 Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. American Naturalist **188**: 379–397.

[33] JOST, L., 2008 G_{ST} and its relatives do not measure differentiation. Molecular Ecology **17**: 4015–4026.

[34] LEWONTIN, R. C., 1972 The apportionment of human diversity. Evolutionary Biology **6**: 381–398.

[35] BROWN, R. A., and G. J. ARMELAGOS, 2001 Apportionment of racial diversity: a review. Evolutionary Anthropology **10**: 34–40.

[36] RUVOLO, M., and M. T. SEIELSTAD, 2001 "The apportionment of human diversity" 25 years later. In R. S. Singh, C. S. Krimbas, D. B. Paul and J. Beatty, editors, *Thinking about Evolution: Historical, Philosophical, and Political Perspectives*. Cambridge University Press, Cambridge, 141–151.

[37] NOVEMBRE, J., 2022 The background and legacy of Lewontin's apportionment of diversity. Philosophical Transactions of the Royal Society of London B **377**: xx.

[38] SHEN, H., and M. W. FELDMAN, 2022 Diversity and its causes: Lewontin against racism and biological determinism. Philosophical Transactions of the Royal Society of London B **377**: xx.

[39] ROSENBERG, N. A., and M. JAKOBSSON, 2008 The relationship between homozygosity and the frequency of the most frequent allele. Genetics **179**: 2027–2036.

Appendix. Proof of eq. 3

This appendix derives the upper bound on F_{ST} as a function of σ_1 (eq. 3). First, we separate the case of integer values of σ_1. Next, for non-integer values of σ_1, we reduce the problem of maximizing F_{ST} to the problem of maximizing the sum of squared allele frequencies across alleles and subpopulations, $S = \sum_{i=1}^{\infty}\sum_{k=1}^{K} p_{k,i}^2$. Next, we maximize S as a function of σ_1, separately for σ_1 in $(0,1)$ and for non-integer σ_1 in $(1,K)$.

A useful expression for F_{ST}

Suppose $K \geq 2$ is a specified integer. Suppose σ_1 is a fixed value, with $0 < \sigma_1 < K$. We leave the number of alleles I unspecified. For each $i \geq 1$, we write $\sigma_i = \sum_{k=1}^{K} p_{k,i}$, with $\sigma_i \geq \sigma_j$ for each i and j with $i \leq j$. For convenience, σ_1 is taken to mean both the function that computes the sum $\sum_{k=1}^{K} p_{k,1}$ for a specified set of values of the $p_{k,i}$ 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 $\sum_{i=1}^{\infty} p_{k,i} = 1$ for all k, $1 \leq k \leq K$. Define F_{ST} as in eq. 2. We seek to maximize F_{ST} over all possible sets of values of the $p_{k,i}$ with a fixed value σ_1 for the sum $\sum_{k=1}^{K} p_{k,1}$. Note that because $\sigma_1 < K$ and $\sum_{k=1}^{K}\sum_{i=1}^{\infty} p_{k,i} = \sum_{i=1}^{\infty}\sigma_i = K$, it follows that $\sigma_2 > 0$.

Denote the sum of squared frequencies of allele 1 across subpopulations, $\sum_{k=1}^{K} p_{k,1}^2$, by S_1. Denote $S = \sum_{i=1}^{\infty}\sum_{k=1}^{K} p_{k,i}^2 = \sum_{k=1}^{K}\sum_{i=1}^{\infty} p_{k,i}^2$ for the corresponding sum of squared frequencies of all alleles. We express eq. 2 in terms of σ_1, S_1, and S:

$$F_{ST} = \frac{(K-1)S + S_1 - \sigma_1^2 - 2\sum_{i=2}^{\infty}\sum_{k=1}^{K-1}\sum_{\ell=k+1}^{K} p_{k,i}p_{\ell,i}}{K^2 - S + S_1 - \sigma_1^2 - 2\sum_{i=2}^{\infty}\sum_{k=1}^{K-1}\sum_{\ell=k+1}^{K} p_{k,i}p_{\ell,i}}. \tag{A.1}$$

By construction of eq. 2, the denominator of eq. A.1 lies in $(0, K^2)$, as $0 < H_T < 1$ from the fact that $\sigma_2 > 0$. The numerator lies in $[0, K^2)$, as $0 \leq H_S \leq H_T < 1$, so that $0 \leq H_T - H_S < 1$. F_{ST} lies in $[0,1]$, as $0 \leq H_S$ and $0 < H_T$ imply $0 \leq (H_T - H_S)/H_T \leq 1$.

The case of integer values of σ_1

In eq. A.1, the numerator is less than or equal to the denominator, with equality if and only if $K = S = \sum_{k=1}^{K}\sum_{i=1}^{\infty} p_{k,i}^2$. This 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 σ_1 is an integer.

Theorem 1. *Suppose σ_1 is an integer value, $1, 2, \ldots, K-1$. $F_{ST} = 1$ if and only if (i) $p_{k,1} = 1$ in each of σ_1 subpopulations, and (ii) for each of the $K - \sigma_1$ 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 σ_1, $p_{k,1} = 1$ in exactly σ_1 subpopulations. □

Note that any set of equivalence relationships can exist among the values of i associated with the $K - \sigma_1$ subpopulations in which $p_{k,1} = 0$, provided that none of these values of i is associated with more than σ_1 subpopulations. For example, these values of i can be mutually distinct, or groups of them with size as large as σ_1 can be mutually equal.

Non-integer values of σ_1

For non-integer σ_1, the numerator of eq. A.1 is strictly less than the denominator. Hence, if the other quantities in eq. A.1 are fixed, then F_{ST} decreases with increasing $2\sum_{i=2}^{\infty}\sum_{k=1}^{K-1}\sum_{\ell=k+1}^{K} p_{k,i}p_{\ell,i}$. We have the following theorem.

Theorem 2. *Suppose σ_1 is not an integer. F_{ST} satisfies*

$$F_{ST} \leq \frac{(K-1)S + S_1 - \sigma_1^2}{K^2 - S + S_1 - \sigma_1^2}, \tag{A.2}$$

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

Proof. Because $2\sum_{i=2}^{\infty}\sum_{k=1}^{K-1}\sum_{\ell=k+1}^{K} p_{k,i}p_{\ell,i}$ is subtracted in both the numerator and the denominator of eq. A.1, and because the numerator is strictly less than the denominator for non-integer σ_1, F_{ST} can be bounded above by minimizing this term. Because $p_{k,i} \geq 0$ for all (k,i), each sum $\sum_{k=1}^{K-1}\sum_{\ell=k+1}^{K} p_{k,i}p_{\ell,i}$ is bounded below by zero. Setting the sum to 0 for all $i \geq 2$ gives the upper bound in eq. A.2.

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

By Theorem 2, to maximize F_{ST} for fixed non-integer σ_1, we must maximize the quantity in eq. A.2. It suffices to consider sets of values of $p_{k,i}$ in which for each $i \geq 2$, at most one value of k has $p_{k,i} > 0$.

The case of (non-integer) σ_1 in $(0,1)$

In this section, we find the set of values of the $p_{k,i}$ that maximize F_{ST} for σ_1 in $(0,1)$. We proceed in two steps. (i) We show that for σ_1 in $(0,1)$, the maximal F_{ST} occurs at a set of $p_{k,i}$ values for which *all* alleles are private: that is, for each $i \geq 1$, $p_{k,i} > 0$ for at most one value of k. (ii) We determine the set of $p_{k,i}$ values that, with all alleles private, maximizes F_{ST}.

(i) In eq. A.2, note that $\sigma_1^2 - S_1 = 2\sum_{k=1}^{K-1}\sum_{\ell=k+1}^{K} p_{k,1}p_{\ell,1}$. Because $\sigma_1^2 - S_1$ is subtracted from both numerator and denominator in eq. A.2, the quantity in eq. A.2 is maximal when $\sigma_1^2 - S_1$ is minimal. In other words, the upper bound on F_{ST} is maximal if and only if $2\sum_{k=1}^{K-1}\sum_{\ell=k+1}^{K} p_{k,1}p_{\ell,1}$ is minimal.

Because $\sigma_1 < 1$, a minimum of 0 for $2\sum_{k=1}^{K-1}\sum_{\ell=k+1}^{K} p_{k,1}p_{\ell,1}$ is achieved if and only if there is a single value $k = k'$ at which $p_{k',1} = \sigma_1$, so that $p_{k,1} = 0$ for all $k \neq k'$. We then have $\sigma_1^2 = S_1$, and from eq. A.2,

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

Each allele is private, and because allele 1 is the most frequent, $p_{k,i}$ lies in $[0, \sigma_1]$ for all (k,i).

(ii) The problem of finding the set of $p_{k,i}$ values that maximizes F_{ST} has now been reduced to the problem of maximizing the right-hand side of eq. A.3, with the constraint that all alleles are private. Because the numerator in eq. 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 $S = \sum_{k=1}^{K}\sum_{i=1}^{\infty} p_{k,i}^2$, with the constraints $\sum_{i=1}^{\infty} p_{k,i} = 1$ and $p_{k,i} \leq \sigma_1$ for each (k,i) with $1 \leq k \leq K$ and $i \geq 1$. Because each allele is private, the maximum is achieved by separately maximizing each $\sum_{i=1}^{\infty} p_{k,i}^2$ with constraints $\sum_{i=1}^{\infty} p_{k,i} = 1$ and $p_{k,i} \leq \sigma_1$.

This maximization is precisely that of Lemma 3 of ROSENBERG and JAKOBSSON [39]. Applying the lemma, the maximum is achieved with $p_{k,1} = p_{k,2} = \ldots = p_{k,J-1} = \sigma_1$, $p_{k,J} = 1 - (J-1)\sigma_1$, and $p_{k,i} = 0$ for $i > J$, where $J = \lceil \sigma_1^{-1} \rceil$. It satisfies $\sum_{i=1}^{\infty} p_{k,i}^2 \leq 1 - \sigma_1(J-1)(2 - J\sigma_1)$. In other words, each subpopulation k possesses $J-1$ private alleles with frequency σ_1 and one private allele with frequency $1 - (J-1)\sigma_1$. Hence, $S \leq K[1 - \sigma_1(J-1)(2 - J\sigma_1)]$, so that eq. A.3 leads to eq. 3 for σ_1 in $(0,1)$.

The case of non-integer σ_1 in $(1,K)$

This section finds the set of values of the $p_{k,i}$ that maximizes F_{ST} for non-integer σ_1 in $(1,K)$. For non-integer $\sigma_1 = \sum_{k=1}^{K} p_{k,1}$ in $(1,K)$, because $0 \leq p_{k,1} \leq 1$ for all k, $p_{k,1} > 0$ for at least two values of k. Writing $S^* = S - S_1$, Eq. A.2 can be rewritten

$$F_{ST} \leq \frac{KS_1 + (K-1)S^* - \sigma_1^2}{K^2 - S^* - \sigma_1^2}. \tag{A.4}$$

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 $\sum_{i=1}^{\infty} p_{k,i} = 1$ for each k and $\sum_{k=1}^{K} p_{k,i} \leq \sigma_1$ for each i. If S_1 and S^* can be simultaneously maximized at the same set of values of the $p_{k,i}$, then this set of values of the $p_{k,i}$ achieves the maximal F_{ST}.

We proceed in three steps. (i) First, we find the set of values of the $p_{k,i}$ 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 eq. A.4, and hence in eq. A.2.

(i) We first maximize S_1 for fixed non-integer σ_1 in $(1,K)$. More precisely, we seek to maximize $S_1 = \sum_{k=1}^{K} p_{k,1}^2$ with constraints $\sum_{k=1}^{K} p_{k,1} = \sigma_1$ and $p_{k,1} \leq 1$ for each k from 1 to K. This maximization is precisely that performed in Theorem 1 from ALCALA and ROSENBERG [14], a corollary of Lemma 3 of ROSENBERG and JAKOBSSON [39]. Applying the theorem, the maximum is achieved by setting $p_{1,1} = p_{2,1} = \ldots = p_{\lfloor\sigma_1\rfloor,1} = 1$, $p_{\lfloor\sigma_1\rfloor+1,1} = \{\sigma_1\}$, and $p_{k,1} = 0$ for all $k > \lfloor\sigma_1\rfloor + 1$. The maximal value of S_1 is $\{\sigma_1\}^2 + \lfloor\sigma_1\rfloor$.

(ii) Next, we maximize $S^* = \sum_{i=2}^{\infty}\sum_{k=1}^{K} p_{k,i}^2$. Because, by Theorem 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 σ_1, each nonzero $p_{k,i}$ for $i \geq 2$ is equal to the associated σ_i. The sum of the frequencies of all alleles across all subpopulations is $\sum_{i=1}^{\infty}\sigma_i = K$, so that $\sum_{i=2}^{\infty}\sigma_i = K - \sigma_1$. The problem of maximizing S^* is the problem of maximizing $S^* = \sum_{i=2}^{\infty}\sigma_i^2$ with the constraints $\sum_{i=2}^{\infty}\sigma_i = K - \sigma_1$ and $\sigma_i \leq 1$ for each i from 2 to ∞. This maximization is again that performed in Lemma 3 of ROSENBERG and JAKOBSSON [39]. Applying the lemma, the maximum is achieved by setting $\sigma_2 = \sigma_3 = \ldots = \sigma_{K-\lfloor\sigma_1\rfloor} = 1$, $\sigma_{K-\lfloor\sigma_1\rfloor+1} = 1 - \{\sigma_1\}$, and $\sigma_i = 0$ for $i > K - \lfloor\sigma_1\rfloor + 1$. The maximum is $(1 - \{\sigma_1\})^2 + (K - \lfloor\sigma_1\rfloor - 1)$.

(iii) S_1 is maximized at a set of $p_{k,i}$ for which $\lfloor\sigma_1\rfloor$ subpopulations are fixed for allele 1, allele 1 has frequency $\{\sigma_1\}$ in one subpopulation, and allele 1 has frequency 0 in all other subpopulations. S^* is maximized at a set of $p_{k,i}$ for which $K - \lfloor\sigma_1\rfloor - 1$ subpopulations are fixed, each for a distinct allele i with $i \geq 2$, one subpopulation possesses a distinct allele $i \geq 2$ with frequency $1 - \{\sigma_1\}$, and all $\lfloor\sigma_1\rfloor$ other subpopulations possess no alleles $i \geq 2$ of nonzero frequency.

The upper bound in eq. A.4 depends on both S_1 and S^*, each of which depends on the $p_{k,i}$. Were the set of values of the $p_{k,i}$ that maximizes S_1 and the set of values of the $p_{k,i}$ that maximizes S^* to differ, additional work would be required to find the set of values of the $p_{k,i}$ that maximizes F_{ST}. However, we now observe that S_1 and S^* can be simultaneously maximized at the

same set of values of $p_{k,i}$, so that the same set of values of the $p_{k,i}$ maximizes S_1 and S^* and hence F_{ST}. In particular, $\lfloor \sigma_1 \rfloor$ subpopulations are fixed for allele 1, each of $K - \lfloor \sigma_1 \rfloor - 1$ subpopulations is fixed for its own private allele, and a single subpopulation possesses allele 1 with frequency $\{\sigma_1\}$ and a private allele with frequency $1 - \{\sigma_1\}$. The number of alleles of nonzero frequency is $K - \lfloor \sigma_1 \rfloor + 1$. Only the most frequent allele is shared by more than one subpopulation, and a single subpopulation possesses more than one allele of nonzero frequency.

Substituting the maximal values of S_1 and S^* into eq. A.4, for non-integer σ_1 in $(1, K)$, we obtain the maximal F_{ST} in terms of σ_1 shown in eq. 3.

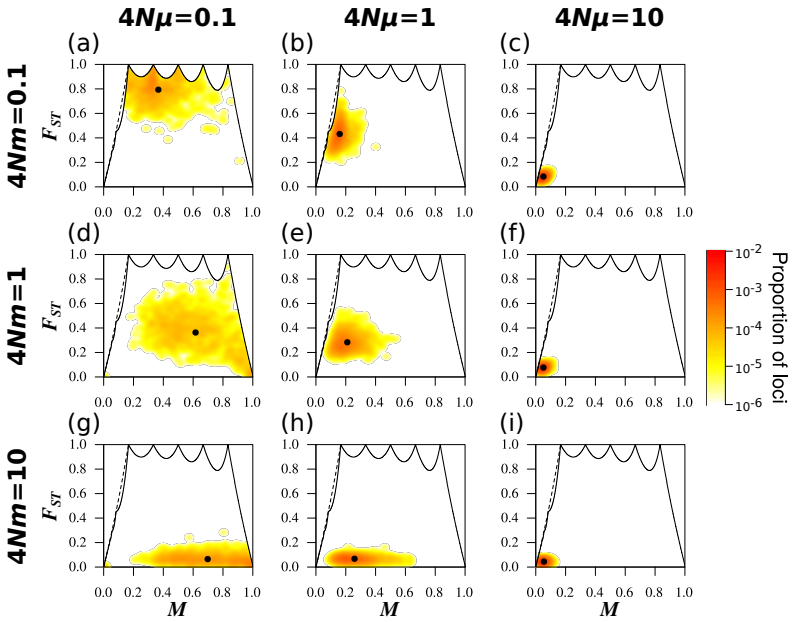

Figure S1 Joint density of the frequency M of the most frequent allele and F_{ST} in the island migration model with $K = 6$ subpopulations, for different scaled migration rates $4Nm$ and mutation rates $4N\mu$. The simulation procedure and figure design follow Figure 2.

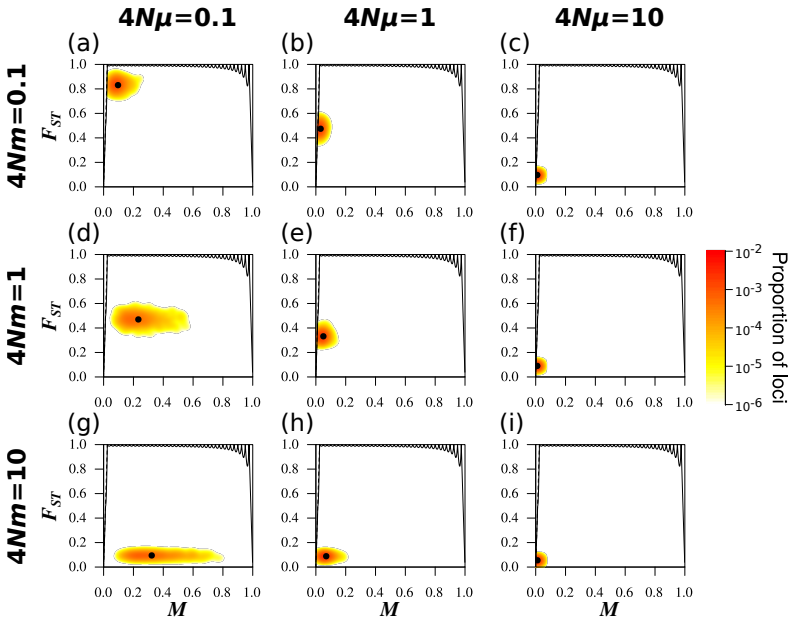

Figure S2 Joint density of the frequency M of the most frequent allele and F_{ST} in the island migration model with $K = 40$ subpopulations, for different scaled migration rates $4Nm$ and mutation rates $4N\mu$. The simulation procedure and figure design follow Figure 2.

Supplementary File S1: MS commands

We applied MS, specifying the scaled mutation and migration parameters. We performed the simulations for $K = 2$, $K = 6$, and $K = 40$ subpopulations. For each command, we replace x by the desired $4N\mu$ value and y by the desired $4Nm$ value.

$K = 2$

```
./ms 200 1000 -t x -I 2 100 100 y
```

$K = 6$

```
./ms 600 1000 -t x -I 6 100 100 100 100 100 100 y
```

$K = 40$

```
./ms 4000 1000 -t x -I 40 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
    100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 y
```