Genetics: Advance Online Publication, published on November 19, 2012 as 10.1534/genetics.112.144758

The relationship between F_{ST} and the frequency of the most frequent allele

Mattias Jakobsson*
Department of Evolutionary Biology and Science for Life Laboratory, Uppsala University

Michael D. Edge
Department of Biology, Stanford University

Noah A. Rosenberg
Department of Biology, Stanford University

October 24, 2012

F_{ST} is frequently used as a summary of genetic differentiation among groups. It has been suggested that F_{ST} depends on the allele frequencies at a locus, as it exhibits a variety of peculiar properties related to genetic diversity: higher values for biallelic single-nucleotide polymorphisms (SNPs) than for multiallelic microsatellites, low values among high-diversity populations viewed as substantially distinct, and low values for populations that differ primarily in their profiles of rare alleles. A full mathematical understanding of the dependence of F_{ST} on allele frequencies, however, has been elusive. Here, we examine the relationship between F_{ST} and the frequency of the most frequent allele, demonstrating that the range of values that F_{ST} can take is restricted considerably by the allele frequency distribution. For a two-population model, we derive strict bounds on F_{ST} as a function of the frequency M of the allele with highest mean frequency between the pair of populations. Using these bounds, we show that for a value of M chosen uniformly between 0 and 1 at a multiallelic locus whose number of alleles is left unspecified, the mean maximum F_{ST} is $\sim$0.3585. Further, F_{ST} is restricted to values much less than 1 when M is low or high, and the contribution to the maximum F_{ST} made by the most frequent allele is on average $\sim$0.4485. Using bounds on homozygosity that we have previously derived as functions of M, we describe strict bounds on F_{ST} in terms of the homozygosity of the total population, finding that the mean maximum F_{ST} given this homozygosity is $1 - \ln 2 \approx 0.3069$. Our results provide a conceptual basis for understanding the dependence of F_{ST} on allele frequencies and genetic diversity, and for interpreting the roles of these quantities in computations of F_{ST} from population-genetic data. Further, our analysis suggests that many unusual observations of F_{ST}, including the relatively low F_{ST} values in high-diversity human populations from Africa and the relatively low estimates of F_{ST} for microsatellites compared to SNPs, can be understood not as biological phenomena associated with different groups of populations or classes of markers but rather as consequences of the intrinsic mathematical dependence of F_{ST} on the properties of allele frequency distributions.

Differentiation among groups is one of the fundamental subjects of the field of population genetics. Comparisons of the level of variation among subpopulations with the level of variation in the total population have been employed frequently in population-genetic theory, in statistical methods for data analysis, and in empirical studies of distributions of genetic variation. Wright's (WRIGHT 1951) fixation indices, and F_{ST} in particular, have been central to this effort.

Wright's F_{ST} was originally defined as the correlation between two randomly sampled gametes from the same subpopulation when the correlation of two randomly sampled gametes from the total population is set to zero. Several definitions of F_{ST} or F_{ST}-like quantities are now available, relying on a variety of different conceptual formulations but all measuring some aspect of population differentiation (e.g. CHARLESWORTH 1998; HOLSINGER and WEIR 2009). Many authors have claimed that one or another formulation of F_{ST} is affected by levels of genetic diversity or by allele frequencies, either because the range of F_{ST} is restricted by these quantities or because these quantities affect the degree to which F_{ST} reflects population differentiation (e.g. CHARLESWORTH 1998; NAGYLAKI 1998; HEDRICK 1999, 2005; LONG and KITTLES 2003; RYMAN and LEIMAR 2008; JOST 2008; LONG 2009; MEIRMANS and HEDRICK 2011). For example, NAGYLAKI (1998) and HEDRICK (1999) argued that measures of F_{ST} may be poor measures of genetic differentiation when the level of diversity is high. CHARLESWORTH (1998) suggested that F_{ST} can be inflated when diversity is low, arguing that F_{ST} might not be appropriate for comparing loci with substantially different levels of variation. In a provocative recent article, JOST (2008) has used the diversity dependence of forms of F_{ST} to question their utility as differentiation measures at all.

One definition that is convenient for mathematical assessment of the relationship of an F_{ST}-like quantity and allele frequencies is the quantity labeled G_{ST} by NEI (1973), which for a given locus measures the difference between the heterozygosity of the total (pooled) population, h_T, and the mean heterozygosity across subpopulations, h_S, divided by the heterozygosity of

*Corresponding author: Uppsala University, Norbyvägen 18D, SE-752 36, Uppsala, Sweden. Email: mattias.jakobsson@ebc.uu.se

the total population:

$$G_{ST} = \frac{h_T - h_S}{h_T}. \quad (1)$$

In terms of the homozygosity of the total population, $H_T = 1 - h_T$, and the mean homozygosity across subpopulations, $H_S = 1 - h_S$, we can write

$$G_{ST} = \frac{H_S - H_T}{1 - H_T}. \quad (2)$$

The WAHLUND (1928) principle guarantees that $H_S \geq H_T$, and therefore, because $H_S \leq 1$ and for a polymorphic locus with finitely many alleles, $0 < H_T < 1$, G_{ST} lies in the interval [0,1].

Using G_{ST} for their definition of F_{ST}, HEDRICK (1999, 2005) and LONG and KITTLES (2003) pointed out that because $h_T < 1$, F_{ST} cannot exceed the mean homozygosity across subpopulations, H_S:

$$F_{ST} = 1 - h_S/h_T < 1 - h_S = H_S. \quad (3)$$

HEDRICK (2005) obtained this result by considering a set of K equal-sized subpopulations, in which each allele is private to a single subpopulation. In the limit as $K \to \infty$, a stronger upper bound on F_{ST} as a function of H_S and K reduces to eq. 3 (see also JIN and CHAKRABORTY (1995) and LONG and KITTLES (2003)).

While HEDRICK (1999, 2005) and LONG and KITTLES (2003) have clarified the relationship between F_{ST} and the mean homozygosity H_S across subpopulations, their approaches do not easily illuminate the connection between F_{ST} and allele frequencies themselves. A formal understanding of the relationship between F_{ST} and allele frequencies would make it possible to more fully understand the behavior of F_{ST} in situations where markers of interest differ substantially in allele frequencies or levels of genetic diversity. Our recent work on the relationship between homozygosity and the frequency of the most frequent allele (ROSENBERG and JAKOBSSON 2008; REDDY and ROSENBERG 2012) provides a mathematical approach for formal investigation of bounds on population-genetic statistics in terms of allele frequencies. In this paper, we therefore seek to thoroughly examine the dependence of F_{ST} on allele frequencies by investigating the upper bound on F_{ST} in terms of the frequency M of the most frequent allele across a pair of populations. We derive bounds on F_{ST} given the frequency of the the most frequent allele and bounds on the frequency of the most frequent allele given F_{ST}. We consider loci with arbitrarily many alleles in a pair of subpopulations. Using theory for the bounds on homozygosity given the frequency of the most frequent allele, we obtain strict bounds on F_{ST} given the homozygosity of the total population. Our analysis clarifies the relationships among F_{ST}, allele frequencies, and homozygosity, providing explanations for peculiar observations of F_{ST} that can be attributed to allele-frequency dependence.

MODEL

We examine a polymorphic locus with at least two alleles in a setting with K subpopulations that contribute equally to a total population. Denote the number of distinct alleles by I, the frequency of allele i in population k by p_{ki}, and the mean frequency of allele i across populations by $\bar{p}_i = \frac{1}{K}\sum_{k=1}^{K} p_{ki}$. We primarily report our results in terms of homozygosities, which can be easily transformed into heterozygosities.

We consider F_{ST} formulated as a property of nonnegative numbers between 0 and 1 such that within populations, the allele frequencies sum to 1 ($\sum_{i=1}^{I} p_{ki} = 1$ for each k). This formulation is the same as the formulation of Nei's G_{ST}, which we hereafter denote by F. We have (NEI 1973):

$$F = \frac{h_T - h_S}{h_T} = \frac{H_S - H_T}{1 - H_T}$$

where

$$H_T = \sum_{i=1}^{I} \bar{p}_i^2 = \sum_{i=1}^{I} \left(\frac{1}{K} \sum_{k=1}^{K} p_{ki} \right)^2$$

and

$$H_S = \frac{1}{K} \sum_{k=1}^{K} \left(\sum_{i=1}^{I} p_{ki}^2 \right) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{I} p_{ki}^2.$$

The assumption that the locus is polymorphic guarantees that $H_T < 1$. The assumption that I, the number of distinct alleles at the locus, is finite guarantees that $H_T > 0$ (and hence $H_S > 0$ because $H_S \geq H_T$). Thus, $0 < H_T < 1$ and $0 < H_S \leq 1$.

We assume that all allele frequencies are the parametric allele frequencies of the population under consideration. Thus, the frequency of an allele is the probability of drawing the allele from the parametric frequency distribution; homozygosity is then the probability that two independent random draws carry the same allelic type, and heterozygosity is the probability that two independent random draws carry different allelic types. We emphasize that in our formulation, F, H_T, and H_S are functions of the parametric allele frequencies, and our interest is in the properties of these functions and their relationships with the allele frequencies; we do not investigate their estimation from data, nor do we consider how evolutionary models affect the underlying allele frequencies involved in their computation.

Two populations: We focus on the case of $K = 2$ subpopulations. In this case, the allele frequencies are denoted p_{1i} for population 1 and p_{2i} for population 2. For each i from 1 to I, let $\sigma_i = p_{1i} + p_{2i}$ be the sum across populations of the frequency of allele i. Each σ_i lies in $(0, 2)$, and the number of alleles I counts only those alleles with $\sigma_i > 0$. We denote $\bar{p}_i = \sigma_i/2$. Without loss of generality, we place the alleles in decreasing order, such that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_I$. We denote the frequency of the most frequent allele in the total pooled population by $M = \sigma_1/2$, and we find it convenient to express some results in terms of σ_1 and others in terms of M. Because $\sum_{i=1}^{I} \sigma_i = 2$ and each σ_i is positive, we have $1/I \leq M < 1$.

Let $\delta_i = |p_{1i} - p_{2i}|$ be the absolute difference between p_{1i} and p_{2i}. We can write the homozygosity of the total population as

$$H_T = \sum_{i=1}^{I} \bar{p}_i^2 = \frac{1}{4} \sum_{i=1}^{I} \sigma_i^2,$$

and the mean homozygosity across subpopulations as

$$H_S = \frac{1}{2} \sum_{k=1}^{2} \sum_{i=1}^{I} p_{ki}^2 = \frac{1}{2} \sum_{i=1}^{I} (p_{1i}^2 + p_{2i}^2).$$

We then have (BOCA and ROSENBERG 2011)

$$F = \frac{\sum_{i=1}^{I} \delta_i^2}{4 - \sum_{i=1}^{I} \sigma_i^2}. \quad (4)$$

In other words, F can be computed solely using the allele frequency sums and differences between the two populations.

BOUNDS ON F

Our goal is to study the relationship between F and M in the general case of I alleles in two populations. For convenience, we write F as a function of σ_1, keeping in mind that $\sigma_1/2 = M$, and we begin by considering the special case in which $I = 2$.

Bounds on F for two alleles: This case has two alleles, with frequencies p_{11} and p_{12} in population 1, and p_{21} and p_{22} in population 2 (Table 1). The frequency of the second allele is $p_{12} = 1 - p_{11}$ in population 1 and $p_{22} = 1 - p_{21}$ in population 2. Using eq. 4, we have a simple expression for F (WEIR 1996; ROSENBERG *et al.* 2003):

$$\begin{aligned} F &= \frac{\delta_1^2 + [(1 - p_{11}) - (1 - p_{21})]^2}{4 - \sigma_1^2 - [(1 - p_{11}) + (1 - p_{21})]^2} \\ &= \frac{\delta_1^2}{\sigma_1(2 - \sigma_1)}. \end{aligned} \quad (5)$$

We determine the upper and lower bounds of F in terms of the frequency of the most frequent allele $M = \sigma_1/2$. Because the alleles are arranged to satisfy $\sigma_1 \geq \sigma_2$ and because $\sigma_1 + \sigma_2 = 2$, σ_1 must lie in $[1, 2)$. For the lower bound on F as a function of σ_1, we note that if allele 1 has the same frequency in both populations, then $p_{11} = p_{21} = \sigma_1/2$. The frequency of allele 2 will also be the same in the two populations, $p_{12} = p_{22} = 1 - \sigma_1/2$, and δ_1 and δ_2 will both equal zero. For these allele frequencies, we see that $H_S = H_T$, and it is clear from eq. 5 that $F(\sigma_1) \geq 0$ for all values of σ_1 in $[1, 2)$, with equality if and only if $p_{11} = p_{21} = \sigma_1/2$.

Table 1: Notation for two alleles in two populations.

	Allele		
Population	1	2	Sum
1	p_{11}	p_{12}	1
2	p_{21}	p_{22}	1
Sum	σ_1	σ_2	2
Absolute difference	δ_1	δ_2	-

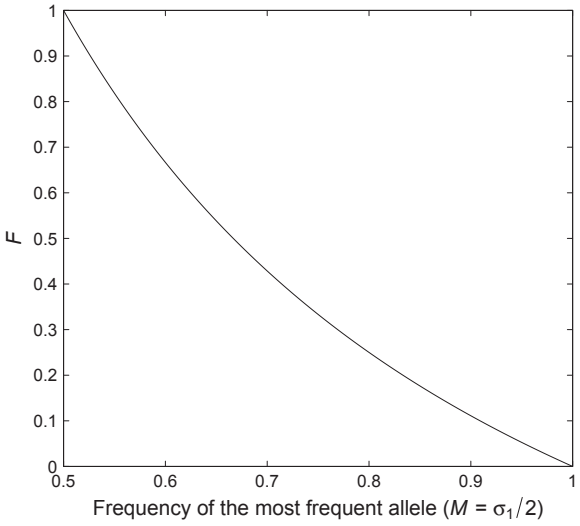

Figure 1: The upper bound on F as a function of the frequency M of the most frequent allele, for the two-allele case. The upper bound is computed from eq. 7. The lower bound on F is 0 for all values of M.

For the upper bound, we first note that because $\delta_1 = 2p_{11} - \sigma_1$ when $p_{11} \geq p_{21}$ and $\delta_1 = 2p_{21} - \sigma_1$ when $p_{21} \geq p_{11}$,

$$\delta_1^2 \leq (2 - \sigma_1)^2, \quad (6)$$

with equality if and only if $p_{11} = 1$ or $p_{21} = 1$. Using eqs. 5 and 6, we have

$$F(\sigma_1) \leq \frac{(2 - \sigma_1)^2}{\sigma_1(2 - \sigma_1)} = \frac{2 - \sigma_1}{\sigma_1}.$$

Thus, the upper bound on F as a function of σ_1 is achieved when the allele frequencies of the two populations differ as much as possible, that is, when $(p_{11}, p_{21}) = (1, \sigma_1 - 1)$ or $(p_{11}, p_{21}) = (\sigma_1 - 1, 1)$. The bounds on F are

$$F \in \left[0, \frac{2 - \sigma_1}{\sigma_1}\right]. \quad (7)$$

Figure 1 shows the upper bound as a function of the most frequent allele, illustrating a monotonic decline from $q(1/2) = 1$ to $q(1) = 0$.

Lower bound on F for an unspecified number of alleles: For any number of alleles I and any set of σ_i, by noting that the denominator of F in eq. 4 is positive and that the numerator is $\sum_{i=1}^{I} \delta_i^2 \geq 0$, we see that eq. 4 takes the value of zero if and only if for each i, $p_{1i} = p_{2i} = \sigma_i/2$. Thus, the lower bound on F as a function of σ_1 is achieved when the allele frequencies are the same in both populations for all I alleles. Thus, $F = 0$ is attainable for any value of σ_1 in $(0, 2)$.

Upper bound on F for an unspecified number of alleles: The upper bound on F as a function of σ_1 has different properties for $\sigma_1 \in (0,1)$ and for $\sigma_1 \in [1,2)$. We begin with $\sigma_1 \in (0,1)$.

Using eq. 4, we can rearrange $F(\sigma_1)$ to obtain

$$F(\sigma_1) = -1+2\frac{2-2\sum_{i=1}^{I} p_{1i}p_{2i}}{(4-\sum_{i=1}^{I} p_{1i}^2-\sum_{i=1}^{I} p_{2i}^2)-2\sum_{i=1}^{I} p_{1i}p_{2i}}. \tag{8}$$

As we assume that the locus of interest is polymorphic, both the numerator and denominator in the fraction in eq. 8 are positive. Fix $\sum_{i=1}^{I} p_{1i}^2$ and $\sum_{i=1}^{I} p_{2i}^2$. Because the same quantity $2\sum_{i=1}^{I} p_{1i}p_{2i}$ is subtracted in the numerator and denominator from quantities that must exceed it (2 in the numerator, $4-\sum_{i=1}^{I} p_{1i}^2-\sum_{i=1}^{I} p_{2i}^2$ in the denominator), the fraction is maximized when $\sum_{i=1}^{I} p_{1i}p_{2i}$ is minimized, that is, when $\sum_{i=1}^{I} p_{1i}p_{2i}=0$. In other words, given σ_1, for fixed $\sum_{i=1}^{I} p_{1i}^2$ and $\sum_{i=1}^{I} p_{2i}^2$, $F(\sigma_1)$ is maximal when each allele is found only in one of the two subpopulations.

To complete the maximization of $F(\sigma_1)$ as a function of σ_1, it remains to maximize $\sum_{i=1}^{I} p_{1i}^2$ and $\sum_{i=1}^{I} p_{2i}^2$. These two maximizations can be performed separately, as no allele appears in both subpopulations. Further, by symmetry, $\sum_{i=1}^{I} p_{1i}^2$ and $\sum_{i=1}^{I} p_{2i}^2$ must have the same maximum.

Define $J = \lceil \sigma_1^{-1} \rceil$. The number of alleles I is unspecified; we search for an upper bound over all possible values $I \geq 2$ and discover that the maximum occurs when each subpopulation has $I = J$ distinct alleles. Because $p_{1i}+p_{2i} \leq \sigma_1$ and because for each i, at the maximum of $F(\sigma_1)$, each allele has either $p_{1i}=0$ or $p_{2i}=0$, it suffices to maximize $\sum_{i=1}^{I} p_{1i}^2$ subject to $\sum_{i=1}^{I} p_{1i}=1$ and $p_{1i} \leq \sigma_1$ for all i. This maximization is the same problem considered in Lemma 3 of ROSENBERG and JAKOBSSON (2008), which demonstrates that the maximum occurs if and only if the locus has $J-1$ alleles of frequency σ_1, and one remaining allele of frequency $1-(J-1)\sigma_1$.

Lemma 3 of ROSENBERG and JAKOBSSON (2008) yields $1-\sigma_1(J-1)(2-J\sigma_1)$ for each of the two maxima, on $\sum_{i=1}^{I} p_{1i}^2$ and on $\sum_{i=1}^{I} p_{2i}^2$. We then conclude

$$F(\sigma_1) \leq \frac{1-\sigma_1(J-1)(2-J\sigma_1)}{1+\sigma_1(J-1)(2-J\sigma_1)}, \tag{9}$$

with equality if and only if the locus has $2J$ alleles, J of which occur only in the first subpopulation and the other J of which occur only in the second population, and each subpopulation has $J-1$ alleles of frequency σ_1 and one allele of frequency $1-(J-1)\sigma_1$. Because $J = \lceil \sigma_1^{-1} \rceil$, we have

$$F(\sigma_1) \leq \frac{1-\sigma_1(\lceil \sigma_1^{-1} \rceil-1)(2-\lceil \sigma_1^{-1} \rceil\sigma_1)}{1+\sigma_1(\lceil \sigma_1^{-1} \rceil-1)(2-\lceil \sigma_1^{-1} \rceil\sigma_1)}. \tag{10}$$

For the case of $\sigma_1 \in [1,2)$, we separate terms in eq. 4 for the first and subsequent alleles:

$$F(\sigma_1) = \frac{\delta_1^2+\sum_{i=2}^{I}\delta_i^2}{4-\sigma_1^2-\sum_{i=2}^{I}\sigma_i^2}. \tag{11}$$

The upper bound on F, given σ_1, occurs when δ_1^2, $\sum_{i=2}^{I}\delta_i^2$, and $\sum_{i=2}^{I}\sigma_i^2$ are maximized. To maximize δ_1^2, note that as in the two-allele case (eq. 6), for $\sigma_1 \in [1,2)$, $\delta_1^2 \leq (2-\sigma_1)^2$, with equality if and only if $p_{11}=1$ or $p_{21}=1$.

Next, for any i, $\delta_i \leq \sigma_i$, with equality if and only if $p_{1i}=0$ or $p_{2i}=0$. Then

$$\begin{aligned}\sum_{i=2}^{I}\delta_i^2 &\leq \sum_{i=2}^{I}\sigma_i^2 \\ &\leq \left(\sum_{i=2}^{I}\sigma_i\right)^2 \\ &= (2-\sigma_1)^2,\end{aligned} \tag{12}$$

where the last step follows from the fact that $\sum_{i=1}^{I}\sigma_i = 2$. Equality in the second step requires that among the σ_i with $i \geq 2$, only one can be positive, namely σ_2, by the assumption that the alleles are labeled in decreasing order of frequency. Thus, equality occurs in both inequalities if and only if $\sigma_2 = 2-\sigma_1$ and either p_{12} or p_{22} is 0.

We have therefore found that given $\sigma_1 \in [1,2)$, δ_1^2, $\sum_{i=2}^{I}\sigma_i^2$, and $\sum_{i=2}^{I}\delta_i^2$ are all maximized under exactly the same conditions—when $(p_{11},p_{12},p_{21},p_{22}) = (1,0,\sigma_1-1,2-\sigma_1)$ or $(\sigma_1-1,2-\sigma_1,1,0)$. Replacing the terms δ_1^2, $\sum_{i=2}^{I}\delta_i^2$ and $\sum_{i=2}^{I}\sigma_i^2$ in eq. 11 using inequalities 6 and 12, we have

$$\begin{aligned}F(\sigma_1) &\leq \frac{(2-\sigma_1)^2+(2-\sigma_1)^2}{4-\sigma_1^2-(2-\sigma_1)^2} \\ &= \frac{2-\sigma_1}{\sigma_1},\end{aligned} \tag{13}$$

with equality if and only if $p_{11}=1$ or $p_{21}=1$ and $\sigma_2 = 2-\sigma_1$. This result matches the two-allele case: when $\sigma_1 \in [1,2)$, the case of an unspecified number of alleles reduces to the case of two alleles.

Summarizing our results, the bounds of F are:

$$F \in \begin{cases} [0, Q(\sigma_1)] & 0 < \sigma_1 < 1 \\ \\ [0, q(\sigma_1)] & 1 \leq \sigma_1 < 2, \end{cases} \tag{14}$$

where

$$Q(\sigma_1) = \frac{1-\sigma_1(\lceil \sigma_1^{-1} \rceil-1)(2-\lceil \sigma_1^{-1} \rceil\sigma_1)}{1+\sigma_1(\lceil \sigma_1^{-1} \rceil-1)(2-\lceil \sigma_1^{-1} \rceil\sigma_1)} \tag{15}$$

$$q(\sigma_1) = \frac{2-\sigma_1}{\sigma_1}. \tag{16}$$

Note that the upper bound on F is continuous at $\sigma_1 = 1$, as $\lim_{\sigma_1 \to 1} Q(\sigma_1) = q(1) = 1$.

The upper bound on F is shown as the solid line in Figure 2. The plot illustrates that the upper bound on $F(\sigma_1)$ has a piecewise structure on $(0,1)$, with changes in shape occurring

when σ_1 is equal to the reciprocal of an integer. Similarly to the bounds examined by ROSENBERG and JAKOBSSON (2008), for each $J \geq 2$, $Q(\sigma_1)$ is monotonically increasing on the interval $[1/J, 1/(J-1))$, where $\lceil \sigma_1^{-1} \rceil$ has the constant value J. Further, $Q(\sigma_1)$ is continuous at the boundaries $1/J$ between intervals, with $Q(1/J) = 1/(2J-1)$. On $[1,2)$, the upper bound has a simple monotonic decline according to $q(\sigma_1)$.

PROPERTIES OF THE UPPER BOUND ON F

The region between 0 and the upper bound on F exactly circumscribes the set of possible values of F as a function of σ_1, as the upper bound is strict. We now explore a series of features of the upper bound on F as a function of σ_1.

The space between the upper and lower bounds on F: The mean maximum F across the range of possible frequencies for the most frequent allele gives a sense of the maximal F attainable on average, when M is uniformly distributed. This mean can be obtained by evaluating the area of the region between the lower and upper bounds on F.

Because the lower bound on F is zero over the entire interval $\sigma_1 \in (0,2)$, we only need to determine the area A under the upper bound on F. We integrate $Q(\sigma_1)$ for $\sigma_1 \in (0,1)$ and $q(\sigma_1)$ for $\sigma_1 \in [1,2)$,

$$A = \int_0^1 Q(\sigma_1)\, d\sigma_1 + \int_1^2 q(\sigma_1)\, d\sigma_1. \tag{17}$$

The first integral can be computed as a sum over intervals $[1/J, 1/(J-1))$ for $J \geq 2$. On each such interval, $\lceil \sigma_1^{-1} \rceil$ has a fixed value of J. We then have

$$\int_0^1 Q(\sigma_1)\, d\sigma_1 = \sum_{J=2}^{\infty} \int_{\frac{1}{J}}^{\frac{1}{J-1}} \frac{1-\sigma_1(J-1)(2-J\sigma_1)}{1+\sigma_1(J-1)(2-J\sigma_1)}\, d\sigma_1.$$

In the APPENDIX, we show that

$$\int_0^1 Q(\sigma_1)\, d\sigma_1 = -1 + \sum_{J=2}^{\infty} \frac{\ln\left(\frac{\sqrt{(J-1)(2J-1)}+1}{\sqrt{(J-1)(2J-1)}-1}\right)}{\sqrt{(J-1)(2J-1)}}. \tag{18}$$

By numerically evaluating the sum in eq. 18, we obtain an approximation $\int_0^1 Q(\sigma_1)\, d\sigma_1 \approx 0.3307808$.

The second term in eq. 17 is

$$\begin{aligned} \int_1^2 q(\sigma_1)\, d\sigma_1 &= \int_1^2 \frac{2-\sigma_1}{\sigma_1}\, d\sigma_1 \\ &= 2\ln 2 - 1, \end{aligned} \tag{19}$$

and the area under $q(\sigma_1)$ for $\sigma_1 \in [1,2)$ is $\sim$0.3862944.

Summing the values for the two integrals, the area A under the upper bound on F is $\sim$0.7170751. Considering F as a function of $M = \sigma_1/2$ rather than σ_1, F is confined to a region with area $\sim$0.3585376. This area under the curve is the mean maximal value of F across the space of values of M, and it is substantially less than 1. Thus, on average, F is constrained within a narrow range, and across most of the space of possible values for the frequency of the most frequent allele, F cannot achieve large values. For example, only over half the range—for M between $1/4$ and $3/4$—is it possible for F to exceed $1/3$.

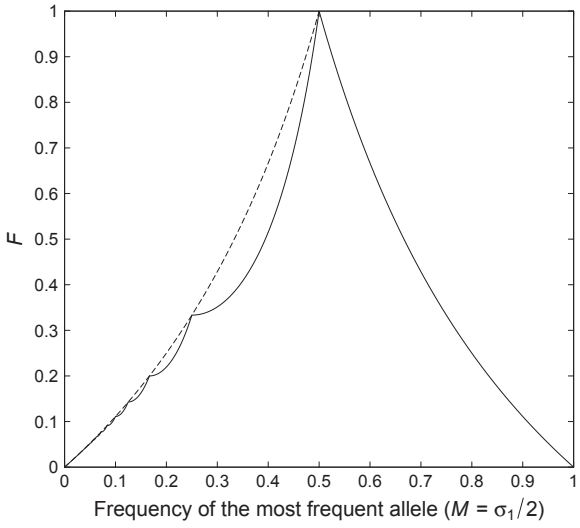

Figure 2: The upper bound on F (solid line) as a function of the frequency M of the most frequent allele, for the general case of any number of alleles. The upper bound is computed from eqs. 15 and 16. The dashed line shows eq. 21, which the upper bound touches when $M = 1/(2J)$ for integers $J \geq 2$. The lower bound on F is 0 for all values of M.

Jagged points touch a simple curve: For $\sigma_1 \in [1,2)$, the upper bound on F is a smooth function $q(\sigma_1)$. For $\sigma_1 \in (0,1)$, however, the upper bound is a jagged curve. At $\sigma_1 = 1/J$ for any integer $J \geq 2$, that is, at the "jagged points" where the upper bound is not differentiable, $Q(\sigma_1)$ coincides with the reflection of $q(\sigma_1)$ across the line $\sigma_1 = 1$. We have

$$Q(\sigma_1 = 1/J) = \frac{1}{2J-1}, \tag{20}$$

because $\lceil \sigma_1^{-1} \rceil = J$ when $\sigma_1 = 1/J$. Thus, for $\sigma_1 = 1/J$, $Q(\sigma_1)$ touches the curve

$$q^*(\sigma_1) = \frac{\sigma_1}{2-\sigma_1}. \tag{21}$$

The dashed line in Figure 2 plots $q^*(\sigma_1)$ on $(0,1)$.

Because $q^*(\sigma_1)$ on $(0,1)$ is the reflection of $q(\sigma_1)$ on $[1,2)$ across the line $\sigma_1 = 1$, the area under $q^*(\sigma_1)$ on $(0,1)$ is the same as the area of $q(\sigma_1)$ on $[1,2)$, or $2\ln 2 - 1$. Thus, on the interval $(0,1)$, the space between $q^*(\sigma_1)$ and $Q(\sigma_1)$ is

$$(2\ln 2 - 1) - \int_0^1 Q(\sigma_1)\, d\sigma_1 \approx 0.0555136. \tag{22}$$

The contribution made by M to the upper bound on F: We denote by $F_1(\sigma_1)$ the contribution of the most frequent allele to $F(\sigma_1)$. By this quantity, we mean the term in $F(\sigma_1)$ contributed by the difference between populations in the frequency of the most frequent allele. From eq. 4, $F(\sigma_1)$ can be written

$$F(\sigma_1) = \sum_{i=1}^{I} \frac{\delta_i^2}{4 - \sum_{j=1}^{I} \sigma_j^2}. \tag{23}$$

If the ith term in the summation is denoted $F_i(\sigma_1)$, our interest is in the value of $F_1(\sigma_1)$ obtained at the set of allele frequencies that maximizes $F(\sigma_1)$.

For σ_1 in the interval $(0, 1)$, defining $\lceil \sigma_1^{-1} \rceil = J$, the maximum has $2J - 2$ alleles with frequency σ_1 and two alleles with frequency $1-(J-1)\sigma_1$: $J-1$ alleles with frequency σ_1 and one allele with frequency $1 - (J-1)\sigma_1$ in each subpopulation. The value of δ_1^2 at the maximum is σ_1^2. Denoting the contribution $F_1(\sigma_1)$ to $F(\sigma_1)$ at the maximum by $Q_1(\sigma_1)$, we have

$$Q_1(\sigma_1) = \frac{\sigma_1^2}{2 + 2\sigma_1(\lceil \sigma_1^{-1} \rceil - 1)(2 - \lceil \sigma_1^{-1} \rceil \sigma_1)}. \tag{24}$$

In the APPENDIX, we evaluate $\int_0^1 Q_1(\sigma_1)\, d\sigma_1$. The expression is unwieldy, but it provides a numerical approximation $\int_0^1 Q_1(\sigma_1)\, d\sigma_1 \approx 0.1284522$.

For $\sigma_1 \in [1, 2)$, at the maximum of $F(\sigma_1)$, $\delta_1^2 = (2 - \sigma_1)^2$, and we have

$$\begin{aligned} q_1(\sigma_1) &= \frac{(2-\sigma_1)^2}{4 - \sigma_1^2 - (2-\sigma_1)^2} \\ &= \frac{2 - \sigma_1}{2\sigma_1}. \end{aligned} \tag{25}$$

The area under $q_1(\sigma_1)$ is

$$\int_1^2 q_1(\sigma_1)\, d\sigma_1 = \ln 2 - \frac{1}{2} \approx 0.1931472.$$

Summing the areas under $Q_1(\sigma_1)$ and $q_1(\sigma_1)$, the total area B under F_1 as σ_1 ranges from 0 to 2 is

$$B = \int_0^1 Q_1(\sigma_1)\, d\sigma_1 + \int_1^2 q_1(\sigma_1)\, d\sigma_1 \approx 0.3215994.$$

If we instead consider $M = \sigma_1/2$, we find that F_1 is confined to $\sim$0.1607997 of the space of possible pairs of values (M, F). The fraction of the area A under the upper bound on F contributed by the most frequent allele over the entire interval $\sigma_1 \in (0, 2)$ is $B/A \approx 0.4484877$. This quantity can be interpreted as the mean contribution of the most frequent allele to the maximum value of F, and it indicates a substantial role for the most frequent allele. Indeed, for $\sigma_1 \in [1, 2)$, $q_1(\sigma_1)/q(\sigma_1) = 1/2$. The contribution made by the most frequent allele to the upper bound on F appears in Figure 3.

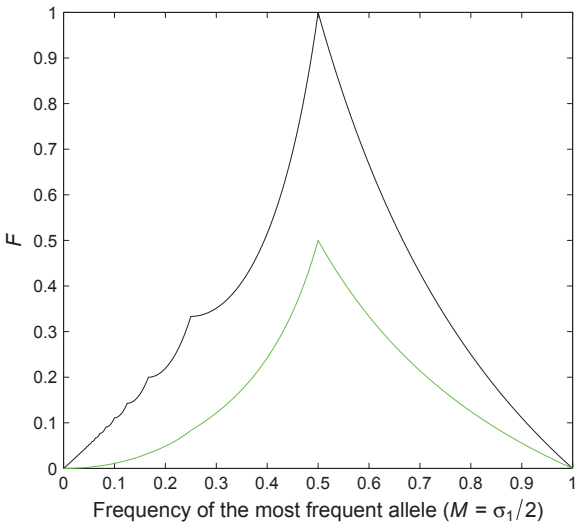

Figure 3: The contribution to F, at the upper bound, that is made by the most frequent allele (green line). The contribution by the most frequent allele is computed from eqs. 24 and 25. For comparison, the upper bound on F is shown as a black line.

BOUNDS ON M

Our derivation of the bounds on F as functions of the frequency M of the most frequent allele enables us to provide bounds on M as functions of F by taking the inverse of the functions $q(\sigma_1)$ and $Q(\sigma_1)$. For $0 < F < 1$, we show that the bounds on the frequency of the most frequent allele in terms of F are

$$\sigma_1 \in \left[\frac{1}{\lceil \frac{1+F}{2F} \rceil} \left(1 + \sqrt{\frac{\left(2\lceil \frac{1+F}{2F} \rceil - 1\right)F - 1}{(\lceil \frac{1+F}{2F} \rceil - 1)(F+1)}}\right), \frac{2}{1+F} \right]. \tag{26}$$

At the trivial case of $F = 1$, σ_1 must equal 1, and for $F = 0$, σ_1 lies in the open interval $(0, 2)$.

Bounds on σ_1 for two alleles: We first consider the two-allele case. By definition of σ_1, regardless of the value of F, σ_1 can be no smaller than 1, and when $\sigma_1 = 1$, $F(\sigma_1) = \delta_1^2$. For any $F \in [0, 1]$, it is possible to choose allele frequencies p_{11} and p_{21} so that $\delta_1 = |p_{11} - p_{21}| = \sqrt{F}$ and $\sigma_1 = p_{11} + p_{21} = 1$. We simply set $p_{11} = (1 + \sqrt{F})/2$ and $p_{21} = (1 - \sqrt{F})/2$. Thus, the lower bound of $\sigma_1(F) = 1$ can be achieved across the full domain $F \in [0, 1]$.

For the upper bound on σ_1, recall that the upper bound on F in terms of σ_1 (eq. 7) is a continuous monotonically decreasing function on the interval $\sigma_1 \in [1, 2)$. We can therefore obtain the upper bound on σ_1 as the inverse of this function. Thus, for $F \in [0, 1]$, the bounds on σ_1 are:

$$\sigma_1(F) \in \left[1, \frac{2}{1+F}\right]. \tag{27}$$

The corresponding bounds on $M = \sigma_1/2$ appear in Figure 4.

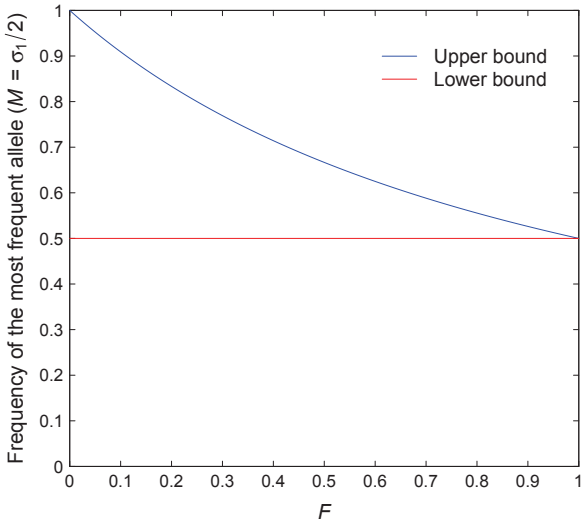

Figure 4: The upper and lower bounds on the frequency M of the most frequent allele as functions of F, for the two-allele case. The bounds are computed from eq. 27.

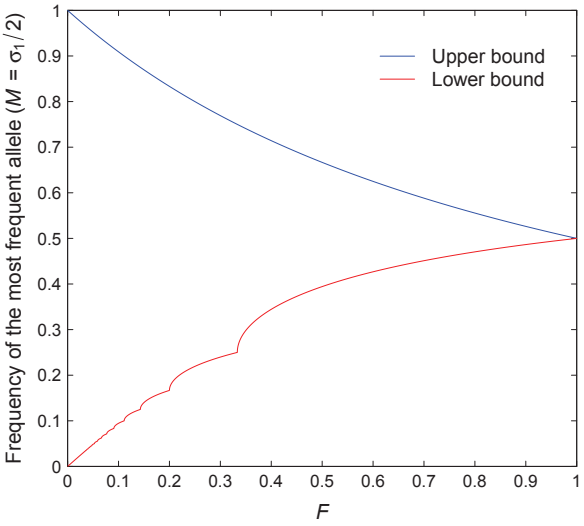

Figure 5: The upper and lower bounds on the frequency M of the most frequent allele as functions of F, for the general case of any number of alleles. The bounds are computed from eqs. 29 and 28.

Lower bound on σ_1 for an unspecified number of alleles: For the general case, we obtain lower and upper bounds on F, considering all possible choices for the number of distinct alleles. It is useful to first recall that the function $Q(\sigma_1)$ for the upper bound on F for $\sigma_1 \in (0,1)$ is monotonically increasing, while the function $q(\sigma_1)$ for the upper bound on F for $\sigma_1 \in [1,2)$ is monotonically decreasing. We can therefore invert $Q(\sigma_1)$ and $q(\sigma_1)$, so that the lower bound on σ_1 as a function of F is obtained by solving $Q(\sigma_1) = F$ for σ_1, and the upper bound by solving $q(\sigma_1) = F$ for σ_1. For the lower bound, we perform the inversion piecewise. For integers $J \geq 2$, if $\sigma_1 \in [1/J, 1/(J-1))$, then $Q(\sigma_1) \in [1/(2J-1), 1/(2J-3))$. Therefore, for $J \geq 2$, if $Q \in [1/(2J-1), 1/(2J-3))$, then the lower bound on σ_1 lies in $[1/J, 1/(J-1))$. For this interval on Q, $\lceil (1+Q)/(2Q) \rceil = J$, and in this region, the lower bound on σ_1, which we term $r(F)$, also satisfies $\lceil r(F) \rceil = J$. We solve eq. 10 for σ_1 for $Q \in [1/(2J-1), 1/(2J-3))$, where both $\lceil \sigma_1^{-1} \rceil$ and $\lceil (1+Q)/(2Q) \rceil$ are equal to J:

$$r(F) = \frac{1}{J}\left(1 + \sqrt{\frac{(2J-1)F - 1}{(J-1)(1+F)}}\right), \tag{28}$$

A negative root is discarded because it yields values that are incompatible with the definition that $\sigma_1 \geq \sigma_i$ for all $i > 1$. The upper and lower bounds appear in Figure 5.

Upper bound on σ_1 for an unspecified number of alleles: From eq. 13 and Figure 2, we see that for any $F \in [0,1]$, the upper bound on σ_1 is greater than or equal to 1. Because eq. 13 is continuous and monotonically decreasing, we can take the inverse of this function to compute the upper bound on σ_1 as a function of F. The upper bound $R(F)$ on σ_1 is

$$R(F) = \frac{2}{1+F}, \tag{29}$$

the same upper bound as for the two-allele case (eq. 27).

F AND HOMOZYGOSITY OF THE TOTAL POPULATION

The relationship between F and the frequency of the most frequent allele can be used together with the relationship between homozygosity and the frequency of the most frequent allele (ROSENBERG and JAKOBSSON 2008; REDDY and ROSENBERG 2012), to find a relationship between F and homozygosity, again in the setting of two populations. The homozygosity that we consider, H in ROSENBERG and JAKOBSSON (2008), corresponds to the homozygosity of the total pooled population H_T. We first note that given any $H_T \in (0,1)$, the lower bound on F is zero. For example, for any H_T, $F = 0$ is obtained by using the equality condition in Theorem 1ii of ROSENBERG and JAKOBSSON (2008) to specify a list of allele frequencies with sum of squares H_T and then assigning that same list of frequencies to both of the component subpopulations.

Upper bound on F given H_T for an unspecified number of alleles: ROSENBERG and JAKOBSSON (2008) showed that

the value of H_T constrains the frequency M of the most frequent allele to a narrow range. We have already determined the upper bound on F as a function of M. Thus, we can obtain an upper bound on F as a function of H_T by taking the maximum value of the upper bound over the range of possible values of M allowed under the results of ROSENBERG and JAKOBSSON (2008) for a given value of H_T. This approach does not guarantee that the upper bound on F that we obtain in terms of H_T is strict; nevertheless, the approach happens to produce a strict bound for $H_T \in [1/2, 1)$. For $H_T \in (0, 1/2)$, it is possible to produce a strict bound by writing F in terms of H_T.

To obtain the bound for $H_T \in (0, 1/2)$, we substitute $\sigma_i^2 - 4p_{1i}p_{2i}$ for δ_i^2 in eq. 4 to write

$$F = \frac{H_T - \sum_{i=1}^{I} p_{1i}p_{2i}}{1 - H_T}. \tag{30}$$

Because $\sum_{i=1}^{1} p_{1i}p_{2i} \geq 0$, we obtain the bound

$$F \leq \frac{H_T}{1 - H_T}. \tag{31}$$

Given H_T, equality is obtained in eq. 31 when $\sum_{i=1}^{I} p_{1i}p_{2i} = 0$. In other words, for $H_T \in (0, 1/2)$, F is maximized when each allele occurs in only one of the two populations. To see that the upper bound is strict, note that when $\sum_{i=1}^{I} p_{1i}p_{2i} = 0$, labeling the homozygosities of the two populations by H_1 and H_2, $H_T = (H_1 + H_2)/4$. As $H_T < 1/2$, $2H_T < 1$, and we can choose $H_1 = H_2 = 2H_T$. Using the equality condition in Theorem 1ii of ROSENBERG and JAKOBSSON (2008), we can specify a set L of exactly $\lceil (2H_T)^{-1} \rceil$ allele frequencies whose sum of squares is H_T. We then construct a set of $2\lceil (2H_T)^{-1} \rceil$ alleles. In population 1, the first $\lceil (2H_T)^{-1} \rceil$ alleles in the set have exactly the allele frequencies in L and the next $\lceil (2H_T)^{-1} \rceil$ alleles have frequency 0. In population 2, the first $\lceil (2H_T)^{-1} \rceil$ alleles have frequency 0, and the next $\lceil (2H_T)^{-1} \rceil$ alleles have the frequencies in L.

For $H_T \in [1/2, 1)$, $H_T/(1 - H_T) \geq 1$, so eq. 31 provides only the trivial bound of $F \leq 1$, and another approach is needed. For any $H_T \in [1/2, 1)$, using Theorem 1ii of ROSENBERG and JAKOBSSON (2008), $M \geq 1/2$. For $M \geq 1/2$, the upper bound on F as a function of σ_1 is monotonically decreasing in σ_1, and consequently, the upper bound on F as a function of H_T is obtained by evaluating $q(\sigma_1)$ at the smallest value of σ_1 permitted by H_T. Theorem 1ii of ROSENBERG and JAKOBSSON (2008) indicates that this smallest allowed σ_1 satisfies

$$\sigma_1/2 = M = \frac{1}{\lceil H_T^{-1} \rceil} \left(1 + \frac{\sqrt{\lceil H_T^{-1} \rceil H_T - 1}}{\sqrt{\lceil H_T^{-1} \rceil - 1}} \right).$$

By replacing $\sigma_1/2$ in eq. 16 with this expression, we have

$$\begin{aligned} F &\leq \frac{\lceil H_T^{-1} \rceil}{1 + \frac{\sqrt{\lceil H_T^{-1} \rceil H_T - 1}}{\sqrt{\lceil H_T^{-1} \rceil - 1}}} - 1 \\ &= \frac{1 - \sqrt{2H_T - 1}}{1 + \sqrt{2H_T - 1}}, \end{aligned} \tag{32}$$

where the last step follows from the fact that $\lceil H_T^{-1} \rceil = 2$ when $H_T \in [1/2, 1)$.

For $H_T \in [1/2, 1)$, the set of allele frequencies that achieves the minimum M as a function of H_T and the set that achieves the maximum F as a function of M coincide. Given H_T, M is minimized by setting $\bar{p}_1 = (1 + \sqrt{2H_T - 1})/2$, $\bar{p}_2 = (1 - \sqrt{2H_T - 1})/2$, and $\bar{p}_i = 0$ for all $i \geq 2$. If these mean frequencies are distributed between the two populations such that $(p_{11}, p_{12}, p_{21}, p_{22}) = (1, 0, \sqrt{2H_T - 1}, 1 - \sqrt{2H_T - 1})$ or $(\sqrt{2H_T - 1}, 1 - \sqrt{2H_T - 1}, 1, 0)$, then the upper bound on F is achieved.

Figure 6 shows our upper bound on F as a function of the total homozygosity H_T. If H_T is low, and particularly if H_T is high, then F is restricted to small values. High values of F are only possible when H_T is near $1/2$. In fact, using eqs. 31 and 32, F can only exceed $1/2$ if H_T lies in $(1/3, 5/9)$.

The space between the upper and lower bounds on F given H_T: In the same manner as in our investigation of the bounds on F as a function of M, we evaluate the area of the region between the upper and lower bounds on F to find the mean maximum F across the range of possible values of H_T.

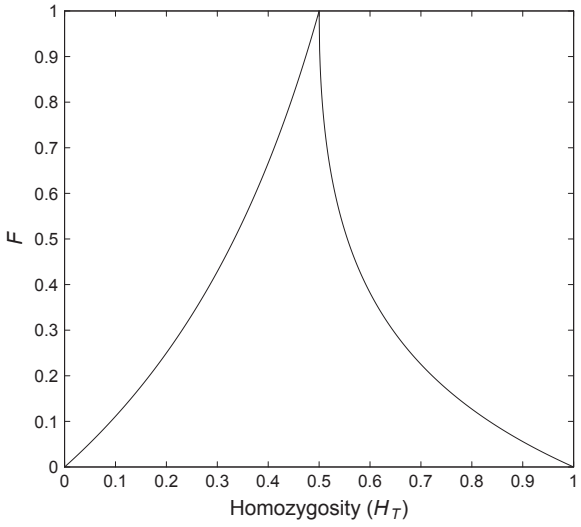

Figure 6: The upper bound on F as a function of H_T. The upper bound is computed from eqs. 31 and 32. The lower bound on F is 0 for all values of H_T.

Because the lower bound on F is zero over the entire interval $H_T \in (0, 1)$, it suffices to evaluate the area A under the upper bound on F. This area is

$$A = \int_0^{1/2} \frac{H_T}{1 - H_T} dH_T + \int_{1/2}^1 \frac{1 - \sqrt{2H_T - 1}}{1 + \sqrt{2H_T - 1}} dH_T. \quad (33)$$

The first term has indefinite integral $-H_T - \ln(1 - H_T)$ and evaluates to $\ln 2 - 1/2$. The second term has indefinite integral $-H_T + 2\sqrt{2H - 1} - 2\ln(1 + \sqrt{2H - 1})$ and evaluates to $3/2 - 2\ln 2$, so that $A = 1 - \ln 2 \approx 0.3068528$.

Note that F is substantially more constrained when $H_T \in [1/2, 1)$ than when $H_T \in (0, 1/2)$. The difference between the areas under the upper bound for $H_T \in (0, 1/2)$ and for $H_T \in [1/2, 1)$ is $3\ln 2 - 2 \approx 0.0794415$, a sizeable fraction of the sum of the two areas. Twice the difference in areas, or $6\ln 2 - 4 \approx 0.1588831$, is the expectation of the difference between the maximum value of F for a value of H_T chosen uniformly at random from $(0, 1/2)$ and the maximum value of F for a value of H_T chosen uniformly at random from $[1/2, 1)$.

APPLICATION TO DATA

We illustrate the bounds on F, M, and H_T for a series of examples using human polymorphism data from ROSENBERG *et al.* (2005) and LI *et al.* (2008). For each example, for each locus, we assume that the allele frequencies in the data sets are parametric allele frequencies. The parametric allele frequencies are obtained in each of a pair of populations, and they are then averaged to obtain parametric allele frequencies for the total population. F, M, and H_T are then computed. The data set of ROSENBERG *et al.* (2005) considers 1048 individuals genotyped for 783 microsatellites, and the data set of LI *et al.* (2008) considers 938 unrelated individuals genotyped for single-nucleotide polymorphisms (SNPs); for all analyses, we restrict our attention to the 935 individuals found in both data sets. For the LI *et al.* (2008) data, we only examine 640,034 SNPs studied by PEMBERTON *et al.* (2012).

Example 1: Africans and Native Americans: Our first example considers microsatellites in 101 Africans and 63 Native Americans, and it is chosen to illustrate a relatively wide range of values of F, M, and H_T. Figure 7 shows F and M, demonstrating that for the comparison of Africans and Native Americans, $F < 0.1$ for most of the 783 loci. The mean value of F is 0.05 with standard deviation 0.06, and the mean value of M is 0.37 with standard deviation 0.11.

Similarly, Figure 8 plots F and H_T for the 783 loci. The mean H_T is 0.25 with standard deviation 0.08. In both Figures 7 and 8, relatively few loci approach the upper bound on F.

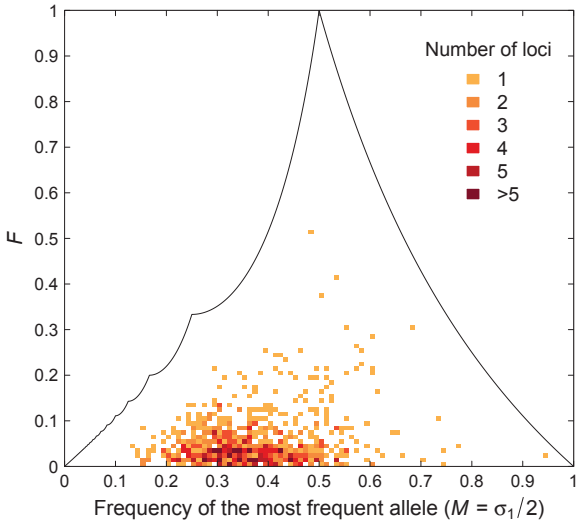

Figure 7: F and the frequency of the most frequent allele (M) for 101 Africans and 63 Native Americans. At each of 783 microsatellite loci, allele frequencies are computed separately for the two population groups, and the total allele frequency is the average of the two group frequencies. Each bin has size 0.01×0.01, and the upper bound on F as a function of M is shown for comparison.

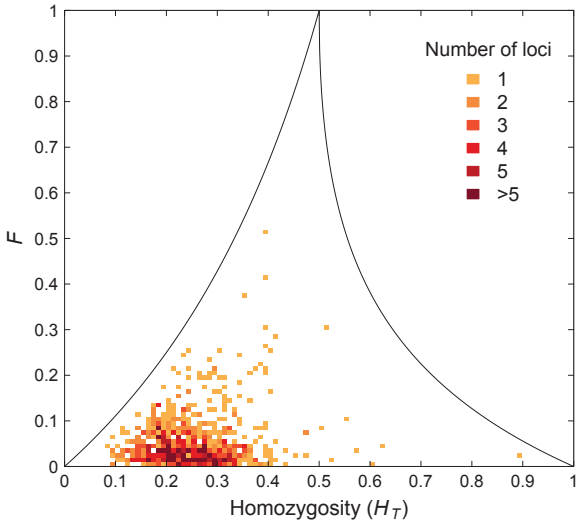

Figure 8: F and homozygosity (H_T) for 101 Africans and 63 Native Americans. At each of 783 microsatellite loci, allele frequencies are computed separately for the two population groups, and the total allele frequency is the average of the two group frequencies. Each bin has size 0.01×0.01, and the upper bound on F as a function of H_T is shown for comparison.

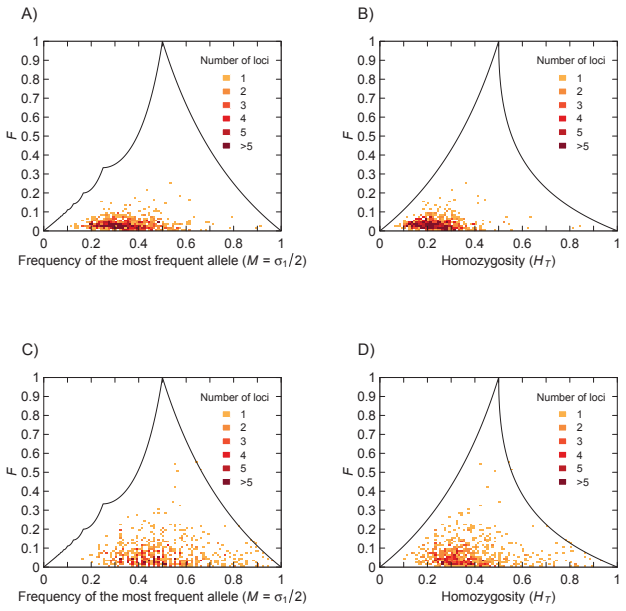

Figure 9: Relationships among F, M and H_T, for pairs of African and Native American populations. (A) F and M for 21 Yoruba and 15 Mbuti Pygmy individuals. (B) F and H_T for 21 Yoruba and 15 Mbuti Pygmy individuals. (C) F and M for 7 Colombian and 14 Pima individuals. (D) F and H_T for 7 Colombian and 14 Pima individuals. In each plot, at each of 783 microsatellite loci, allele frequencies are computed separately for the two populations, and the total allele frequency is the average of the two population frequencies. Each bin has size 0.01×0.01, and the upper bound on F is shown for comparison.

Example 2: High-diversity and low-diversity populations: The bounds on F as a function of M and H_T indicate that genetic diversity in a pair of populations has a strong effect on the value of F between them. To illustrate this point, we compare the values of F obtained from two populations each with high within-population diversity to those obtained from two populations with lower within-population diversity.

The Yoruba and Mbuti Pygmy populations are two African populations with high genetic diversity; the Colombian and Pima populations are Native American populations with lower diversity. Figure 9A shows F and M computed from the Yoruba and Mbuti Pygmy populations, and Figure 9B shows F and H_T. The mean value of F is 0.04 with standard deviation 0.03, the mean value of M is 0.35 with standard deviation 0.11, and the mean value of H_T is 0.24 with standard deviation 0.08.

By contrast, in corresponding plots for the less diverse Colombian and Pima populations, higher values of F, M, and H_T are apparent (Figure 9, panels C and D). In particular, because M and H_T tend to be nearer to $1/2$, larger values of F are possible. The mean values of M and H_T are much closer to $1/2$ than in the African groups; the mean M is 0.50 with standard deviation 0.15, and the mean H_T is 0.38 with standard deviation 0.15. As is suggested by the fact that F can attain its largest values when M and H_T lie near $1/2$, the mean value of F for the Native American groups is nearly twice as high as in the African groups (mean 0.07, standard deviation 0.07).

Example 3: Single-nucleotide polymorphisms: Our third example considers SNPs in the same set of Africans and Native Americans for which microsatellites were examined in Figures 7 and 8. Figure 10 shows the joint distribution of F and M as well as the mean and median of F for intervals of M ranging from $1/2$ to 1 with width 0.01. Mean values of F decrease with M for $M \in (1/2, 1)$, and this decrease is correlated with the decreasing value of the bound on F as a function of M ($r = 0.94$). Compared with the mean, the median value of F is less correlated with the value of the bound, though it also declines with increasing M ($r = 0.77$).

For biallelic markers, for $M > 1/2$, at least one of the two alleles must appear in both populations, and the upper bound on F occurs when one of the populations has only one allele. In Figure 10, for high values of M, more SNPs approach the upper bound on F than for low values of M. This result indicates that SNPs with high values of M are more likely to have an allele found in one but not the other of the two populations.

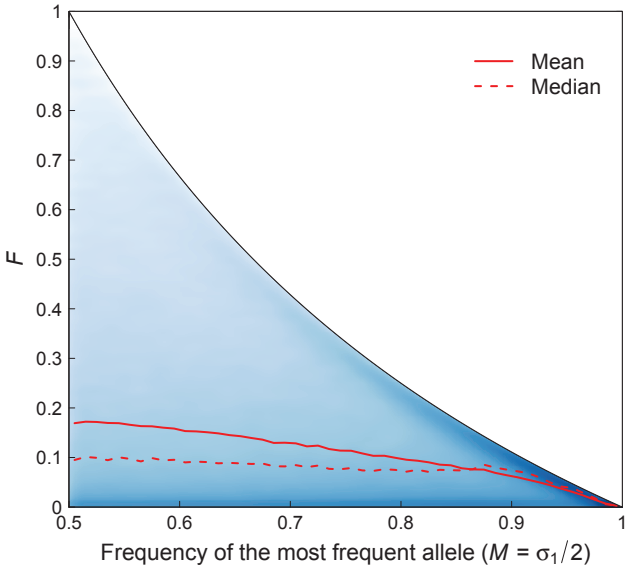

Figure 10: Smoothed scatterplot of F as a function of M for 101 Africans and 63 Native Americans, using single-nucleotide polymorphism (SNP) data. The shading reflects a two-dimensional kernel density estimate using a Gaussian kernel with bandwidth set to 0.007; the density was set to 0 outside the bounds on F as a function of M. For each of 640,034 SNP loci, allele frequencies are computed separately for the two population groups, and the total allele frequency is the average of the two group frequencies. The mean and median of F are computed for 50 bins of width 0.01 ranging from $M = 1/2$ to $M = 1$. The upper bound on F as a function of H_T is shown for comparison.

DISCUSSION

The range of F depends on the level of diversity in the markers considered. In this paper, we have further shown that not only does diversity constrain the range of F, the frequency of the most frequent allele has a strong influence on the values that F can take. When the frequency of the most frequent allele is small *or* large, F is restricted to small values far from one (Figure 2). In fact, considering all possible values of M, F is restricted on average to only $\sim$35.85% of the space of possibilities. This extreme reduction in range for F can be viewed as a consequence of our result that about half of the contribution to the maximal F arises from the most frequent allele (exactly half for $\sigma_1 \in [1, 2)$). Using results from ROSENBERG and JAKOBSSON (2008) on the relationship between homozygosity and the frequency of the most frequent allele, we have described a link between F and homozygosity of the total population (H_T) via separate relationships of F and homozygosity to the frequency of the most frequent allele. F is restricted by H_T even further than by M, to only $\sim$30.69% of the space of possibilities.

Our work extends knowledge of the connection between F and genetic diversity, providing a framework interpreting a variety of features of values of F measured in population-genetic data. We have presented empirical computations that illuminate recently observed phenomena in human population genetics. In particular, even without a formal understanding of the ways in which evolutionary processes and the population-genetic models that encode them give rise to values of M, H_T, and F, the mathematical constraints linking these quantities can aid in interpreting the patterns found in the data.

Low F_{ST} values in human populations from Africa: Estimates of F_{ST} in human populations have been low in Africa compared with other geographic regions, such as among Native Americans (e.g. ROSENBERG *et al.* 2002; TISHKOFF *et al.* 2009). This pattern appears to belie the extensive genetic differentiation known to exist among African populations. For example, using microsatellite loci, TISHKOFF *et al.* (2009) identified a number of genetically distinctive subgroups of African populations despite confirming that F_{ST} in Africa has an unexpectedly small value. The apparent discrepancy between the extensive genetic differentiation among populations in Africa and counterintuitively low values of F_{ST} can be explained using our results. Because Africa has high within-population genetic diversity—including microsatellite homozygosities well below $1/2$ in many populations (Figure S2B of TISHKOFF *et al.* (2009))—the maximum F_{ST} for comparisons of African populations at microsatellite loci is relatively constrained compared with the maximum F_{ST} for comparisons of groups that have less within-population diversity and mean homozygosities nearer $1/2$. Figure 9 shows that F_{ST} values comparing African populations are more constrained by M and H_T than are those comparing Native American populations. Thus, the observation for microsatellites of low F_{ST} in African populations can be attributed to high within-population genetic diversities.

That F_{ST} is more tightly constrained for high-diversity populations than for populations with $H_T \approx 1/2$ has an additional consequence. When considering two pairs of populations with the same F_{ST} value and $H_T < 1/2$, it is likely that a pair of populations with higher within-group diversity is more differentiated than is a pair of populations with relatively low within-group diversity. In other words, the higher the level of genetic diversity within a population, the greater the extent to which raw values of F_{ST} underpredict the intuitive level of differentiation among subpopulations; the result of TISHKOFF *et al.* (2009) exactly follows this pattern.

Lower F_{ST} values for microsatellites than for SNPs: Computations of F_{ST} in human populations have generally found that F_{ST} estimates based on multiallelic loci such as microsatellites are lower than those obtained from biallelic loci such as SNPs (e.g ROSENBERG *et al.* 2002; LI *et al.* 2008). This observation is apparent in the difference between F_{ST}-like computations from nearly the same sets of individuals for microsatellites and for SNPs. When separating human populations into seven geographic regions and computing the within-population component of genetic variation, a quantity analogous to $1-F_{ST}$, ROSENBERG *et al.* (2002) obtained an estimate of 0.941 with microsatellites, whereas LI *et al.* (2008) obtained 0.889 with SNPs. Our results provide a simple explanation for this difference. The SNPs of LI *et al.* (2008) each have only two alleles, so for each locus, the frequency of the most frequent allele is at least $1/2$; further, the minor alleles tend to be common, such that many of the loci have M near $1/2$. By contrast, the microsatellites in the study of ROSENBERG *et al.* (2002) have $\sim$12 alleles on average, so M is typically smaller than $1/2$ and often much smaller (ROSENBERG and JAKOBSSON 2008). Thus, for microsatellites, because of lower frequencies of the most frequent allele and higher levels of genetic diversity, the maximum value of F is substantially more constrained than the corresponding maximum of F for SNPs (Figure 2). We can explain the difference in the magnitudes of the ROSENBERG *et al.* (2002) and LI *et al.* (2008) F_{ST} values via this phenomenon.

Recently, attention has increasingly focused on biallelic sites for which the rarer allele has low frequency (KEINAN and CLARK 2012; NELSON *et al.* 2012; TENNESSEN *et al.* 2012). In our terms, these are sites for which the frequency of the most frequent allele, M, is high. Because F is tightly constrained for high values of M, we might expect that when F_{ST} is calculated using sites with rare minor alleles, small F_{ST} values will be produced. Indeed, Figure 10 shows that when F is used to compare Africans with Native Americans at SNP loci, mean values of F decrease as M increases from $1/2$ to 1.

Conclusions: Measures of F_{ST} have often been used for making inferences about such phenomena as population structure, migration patterns, and range expansions. However, we have found that without a proper understanding of the dependence of F_{ST} on diversity and allele frequencies, F_{ST} can potentially produce puzzling or misleading results. We have described mathematical relationships between F_{ST}, the frequency

of the most frequent allele, and homozygosity that are useful for interpreting the properties of differentiation measures when features of allele frequencies and diversity statistics vary across loci or populations—as they inevitably do in typical scenarios.

Beginning with CHARLESWORTH (1998), NAGYLAKI (1998), and HEDRICK (1999), recent studies have noted that F_{ST} is constrained by diversity, and the issue was described as early as in the work of Sewall Wright (WRIGHT 1978, p. 82). JOST (2008) generated new interest in the dependence of F_{ST} on diversity, illustrating that the dependence can produce substantial discord between intuitions about and measurements of differentiation levels. JOST (2008) also used a multiplicative definition of diversity to propose a pair of new differentiation indices that have the feature of reaching their maximum value if and only if each allele is private to a single subpopulation. In our view, the key to choosing and applying measures of differentiation lies not in "fixation on an index" (LONG 2009), be it F_{ST}, the measures of JOST (2008), or other indices that have recently been proposed (MEIRMANS and HEDRICK 2011), but in developing an understanding of the ways in which possible statistics relate both to intuitive aspects of differentiation and to mathematical features of allele frequencies and genetic diversity. In this context, F_{ST} remains of particular interest on the basis of its long history of use in population genetics and its connection to features of biological models (WHITLOCK 2011). Our examples provide only a few among many ways in which the mathematical properties we have obtained for F_{ST} can be used to interpret its behavior in the analysis of empirical data.

We thank S. Boca and J. VanLiere for numerous discussions of this work. Financial support was provided by the Swedish Research Council Formas, the Erik Philip Sörensen Foundation, the Burroughs Wellcome Fund, a Stanford Graduate Fellowship, and US National Institutes of Health grant GM081441.

REFERENCES

BOCA, S. M., and N. A. ROSENBERG, 2011 Mathematical properties of F_{st} between admixed populations and their parental source populations. Theor. Pop. Biol. **80**: 208–216.

CHARLESWORTH, B., 1998 Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. **15**: 538–543.

HEDRICK, P. W., 1999 Perspective: highly variable loci and their interpretation in evolution and conservation. Evolution **53**: 313–318.

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

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

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

JOST, L., 2008 G_{ST} and its relatives do not measure differentiation. Mol. Ecol. **17**: 4015–4026.

KEINAN, A., and A. G. CLARK, 2012 Recent explosive human population growth has resulted in an excess of rare genetic variants. Science **336**: 740–743.

LI, J. Z., D. M. ABSHER, H. TANG, A. M. SOUTHWICK, A. M. CASTO, *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science **319**: 1100–1104.

LONG, J. C., 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.

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

MEIRMANS, P. G., and P. W. HEDRICK, 2011 Assessing population structure: F_{ST} and related measures. Mol. Ecol. Resources **11**: 5–18.

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

NEI, M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA **70**: 3321–3323.

NELSON, M. R., D. WEGMANN, M. G. EHM, D. KESSNER, P. S. JEAN, *et al.*, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science **337**: 100–104.

PEMBERTON, T. J., D. ABSHER, M. W. FELDMAN, R. M. MYERS, N. A. ROSENBERG, *et al.*, 2012 Genomic patterns of homozygosity in worldwide human populations. Am. J. Hum. Genet. : in press.

REDDY, S. B., and N. A. ROSENBERG, 2012 Refining the relationship between homozygosity and the frequency of the most frequent allele. J. Math. Biol. **64**: 87–108.

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

ROSENBERG, N. A., L. M. LI, R. WARD, and J. K. PRITCHARD, 2003 Informativeness of genetic markers for inference of ancestry. Am. J. Hum. Genet. **73**: 1402–1422.

ROSENBERG, N. A., S. MAHAJAN, S. RAMACHANDRAN, C. ZHAO, J. K. PRITCHARD, *et al.*, 2005 Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. **1**: 660–671.

ROSENBERG, N. A., J. K. PRITCHARD, J. L. WEBER, H. M. CANN, K. K. KIDD, *et al.*, 2002 Genetic structure of human populations. Science **298**: 2381–2385.

RYMAN, N., and O. LEIMAR, 2008 Effect of mutation on genetic differentiation among nonequilibrium populations. Evolution **62**: 2250–2259.

TENNESSEN, J. A., A. W. BIGHAM, T. D. O'CONNOR, W. FU, E. E. KENNY, *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science **337**: 64–69.

TISHKOFF, S. A., F. A. REED, F. R. FRIEDLAENDER, C. EHRET, A. RANCIARO, *et al.*, 2009 The genetic structure and history of Africans and African Americans. Science **324**: 1035–1044.

WAHLUND, S., 1928 Zusammensetzung von Populationen und Korrelationerscheinungen vom Standpunkt der Vererbungslehre aus Betrachtet. Hereditas **11**: 65–106.

WEIR, B. S., 1996 *Genetic Data Analysis II*. Sinauer, Sunderland, MA.

WHITLOCK, M. C., 2011 G'_{ST} and D do not replace F_{ST}. Mol. Ecol. **20**: 1083–1091.

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugen. **15**: 323–354.

WRIGHT, S., 1978 *Evolution and the Genetics of Populations Volume 4: Variability within and among Natural Populations*. University of Chicago Press, Chicago.

APPENDIX

The appendix provides the derivations of two integrals described in the main text.

Integral $\int_0^1 Q(\sigma_1)\,d\sigma_1$ (eq. 18): To obtain $\int_0^1 Q(\sigma_1)\,d\sigma_1$, we first note that for any integer $k \geq 1$, $\lceil \sigma_1^{-1} \rceil = k+1$ if $1/(k+1) \leq \sigma_1 < 1/k$. We have

$$\begin{aligned}\int_0^1 Q(\sigma_1)\,d\sigma_1 &= \int_0^1 \frac{1+\sigma_1(\lceil\sigma_1^{-1}\rceil-1)(\lceil\sigma_1^{-1}\rceil\sigma_1-2)}{1-\sigma_1(\lceil\sigma_1^{-1}\rceil-1)(\lceil\sigma_1^{-1}\rceil\sigma_1-2)}\,d\sigma_1 \\ &= \sum_{k=1}^{\infty}\int_{\frac{1}{k+1}}^{\frac{1}{k}} \frac{1+k\sigma_1((k+1)\sigma_1-2)}{1-k\sigma_1((k+1)\sigma_1-2)}\,d\sigma_1 \\ &= \sum_{k=1}^{\infty}\int_{\frac{1}{k+1}}^{\frac{1}{k}} \Big[-1-\frac{2}{-1+k(k+1)\sigma_1^2-2k\sigma_1}\Big]\,d\sigma_1.\end{aligned}$$

Defining $D=\sqrt{k+2k^2}$, we then have

$$\begin{aligned}\int_0^1 Q(\sigma_1)\,d\sigma_1 &= \sum_{k=1}^{\infty}\int_{\frac{1}{k+1}}^{\frac{1}{k}} \Big[-1+\frac{D^{-1}}{\sigma_1+(k+D)^{-1}}-\frac{D^{-1}}{\sigma_1-(-k+D)^{-1}}\Big]\,d\sigma_1 \\ &= \sum_{k=1}^{\infty}\Big[-\sigma_1+D^{-1}\big[\ln(1+k\sigma_1+D\sigma_1)-\ln(-1-k\sigma_1+D\sigma_1)\big]\Big]_{\frac{1}{k+1}}^{\frac{1}{k}} \\ &= \sum_{k=1}^{\infty}\Big(-\frac{1}{k}+\frac{1}{k+1}\Big)+\sum_{k=1}^{\infty}D^{-1}\ln\Big[\frac{(1+k\frac{1}{k}+\frac{D}{k})(-1-k\frac{1}{k+1}+\frac{D}{k+1})}{(-1-k\frac{1}{k}+\frac{D}{k})(1+k\frac{1}{k+1}+\frac{D}{k+1})}\Big] \\ &= -1+\sum_{k=1}^{\infty}D^{-1}\ln\Big(\frac{D+1}{D-1}\Big)\end{aligned} \tag{34}$$

Integral $\int_0^1 Q_1(\sigma_1)\,d\sigma_1$ (with Q_1 as in eq. 24): To obtain $\int_0^1 Q_1(\sigma_1)\,d\sigma_1$, we first note that for any integer $k \geq 1$, $\lceil \sigma_1^{-1} \rceil = k+1$ when $\sigma_1 \in [\frac{1}{k+1}, \frac{1}{k})$. We have

$$\begin{aligned}\int_0^1 Q_1(\sigma_1)\,d\sigma_1 &= \sum_{k=1}^{\infty}\int_{\frac{1}{k+1}}^{\frac{1}{k}} \frac{\sigma_1^2}{-2k(1+k)\sigma_1^2+4k\sigma_1+2}\,d\sigma_1 \\ &= \sum_{k=1}^{\infty}\int_{\frac{1}{k+1}}^{\frac{1}{k}} \Big[-\frac{1}{2k(1+k)}-\frac{2k\sigma_1+1}{2k(1+k)(k(1+k)\sigma_1^2-2k\sigma_1-1)}\Big]\,d\sigma_1.\end{aligned}$$

The second term can be decomposed, defining

$$A=\frac{1+\frac{3k+1}{2k\sqrt{2+1/k}}}{2k(1+k)^2} \quad \text{and} \quad B=\frac{1-\frac{3k+1}{2k\sqrt{2+1/k}}}{2k(1+k)^2}.$$

We have

$$\begin{aligned}
\int_0^1 Q_1(\sigma_1)\,d\sigma_1 &= \sum_{k=1}^{\infty}\int_{\frac{1}{k+1}}^{\frac{1}{k}}\Big[-\frac{1}{2k(1+k)}-\frac{A}{\sigma_1-\frac{1+\sqrt{2+1/k}}{1+k}}-\frac{B}{\sigma_1-\frac{1-\sqrt{2+1/k}}{1+k}}\Big]\,d\sigma_1\\
&= \sum_{k=1}^{\infty}\Big[-\frac{\sigma_1}{2k(1+k)}+\frac{1}{2k(1+k)^2}\Big[-\ln(-1-\sqrt{2+1/k}+\sigma_1+n\sigma_1)-\ln(-1+\sqrt{2+1/k}+\sigma_1+n\sigma_1)\Big]\\
&\quad+\frac{1}{2k(1+k)^2}\frac{3k+1}{2k\sqrt{2+1/k}}\Big[-\ln(-1-\sqrt{2+1/k}+\sigma_1+n\sigma_1)+\ln(-1+\sqrt{2+1/k}+\sigma_1+n\sigma_1)\Big]\Big]_{\frac{1}{k+1}}^{\frac{1}{k}}\\
&= \sum_{k=1}^{\infty}\Big(-\frac{1}{2k^2(1+k)}+\frac{1}{2k(1+k)^2}\Big)\\
&\quad+\sum_{k=1}^{\infty}\frac{1}{2k(1+k)^2}\ln\Big[\frac{\big(-1-\sqrt{2+1/k}+\frac{1}{1+k}+\frac{k}{1+k}\big)\big(-1+\sqrt{2+1/k}+\frac{1}{1+k}+\frac{k}{1+k}\big)}{\big(-1-\sqrt{2+1/k}+\frac{1}{k}+\frac{k}{k}\big)\big(-1+\sqrt{2+1/k}+\frac{1}{k}+\frac{k}{k}\big)}\Big]\\
&\quad+\sum_{k=1}^{\infty}\frac{3k+1}{4k^2(1+k)^2\sqrt{2+1/k}}\ln\Big[\frac{\big(-1-\sqrt{2+1/k}+\frac{1}{1+k}+\frac{k}{1+k}\big)\big(-1+\sqrt{2+1/k}+\frac{1}{k}+\frac{k}{k}\big)}{\big(-1-\sqrt{2+1/k}+\frac{1}{k}+\frac{k}{k}\big)\big(-1+\sqrt{2+1/k}+\frac{1}{1+k}+\frac{k}{1+k}\big)}\Big]\\
&= \frac{9-\pi^2}{6}+\sum_{k=1}^{\infty}\frac{1}{2k(1+k)^2}\ln\Big[1+\frac{1}{2k^2+k-1}\Big]+\sum_{k=1}^{\infty}\frac{3k+1}{4k^2(1+k)^2\sqrt{2+1/k}}\ln\Big[1+\frac{2}{k\sqrt{2+1/k}-1}\Big] \qquad (35)
\end{aligned}$$