Published in final edited form as:
J Math Biol. 2020 December ; 81(6-7): 1217–1250. doi:10.1007/s00285-020-01531-9.

On the heterozygosity of an admixed population

Simina M. Boca[*], **Lucy Huang**[†], **Noah A. Rosenberg**[‡]

[*]Innovation Center for Biomedical Informatics, Department of Oncology, Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University Medical Center, Washington, DC 20007, United States.

[†]Bioinformatics Graduate Program, University of Michigan, Ann Arbor, MI 48109, United States.

[‡]Department of Biology, Stanford University, Stanford, CA 94305, United States.

Abstract

In this study, we consider admixed populations through their *expected heterozygosity*, a measure of genetic diversity. A population is termed *admixed* if its members possess recent ancestry from two or more separate sources. As a result of the fusion of source populations with different genetic variants, admixed populations can exhibit high levels of genetic diversity, reflecting contributions of their multiple ancestral groups. For a model of an admixed population derived from K source populations, we obtain a relationship between its heterozygosity and its proportions of admixture from the various source populations. We show that the heterozygosity of the admixed population is at least as great as that of the least heterozygous source population, and that it potentially exceeds the heterozygosities of *all* of the source populations. The admixture proportions that maximize the heterozygosity possible for an admixed population formed from a specified set of source populations are also obtained under specific conditions. We examine the special case of $K = 2$ source populations in detail, characterizing the maximal admixture in terms of the heterozygosities of the two source populations and the value of F_{ST} between them. In this case, the heterozygosity of the admixed population exceeds the maximal heterozygosity of the source groups if the divergence between them, measured by F_{ST}, is large enough, namely above a certain bound that is a function of the heterozygosities of the source groups. We present applications to simulated data as well as to data from human admixture scenarios, providing results useful for interpreting the properties of genetic variability in admixed populations.

Keywords

Admixture; allele frequencies; heterozygosity; population genetics

Corresponding author: smb310@georgetown.edu.

Publisher's Disclaimer: This Author Accepted Manuscript is a PDF file of an unedited peer-reviewed manuscript that has been accepted for publication but has not been copyedited or corrected. The official version of record that is published in the journal is kept up to date and so may therefore differ from this version.

1 Introduction

Admixed populations are populations that possess ancestry from multiple source groups. They result from the fusion of populations that have long been separated, in processes such as long-distance migration and hybrid-zone formation at population boundaries.

Several features of ancestry and allele frequencies are characteristic of admixed populations (Chakraborty, 1986; Long, 1991; Verdu & Rosenberg, 2011; Gravel, 2012). In an admixed population, the values of allele frequencies are typically intermediate between those of the various sources. Unlike in a mixture that pools individuals taken from separate populations, in an admixed population, alleles from different sources cooccur within individuals. The contributions from the source populations are each large enough that most members of an admixed population have ancestry in more than one source group.

In admixed populations, the history of mating among populations is recent enough that time has not yet eroded differences among admixed individuals in their relative proportions of ancestry. This feature of high levels of variability in admixture proportions has been central to studies of admixed populations. Investigations of such phenomena as the timing and contributions of the source populations (Verdu & Rosenberg, 2011; Gravel, 2012), the effect of admixture levels on assortative mating patterns (Risch *et al.*, 2009; Zou *et al.*, 2015), and the genetic basis of traits in admixed populations (Buerkle & Lexer, 2008; Zhu *et al.*, 2008) all make use of variation in levels of admixture levels across admixed individuals.

A second aspect of variability in admixed populations is potentially of interest: the variability of alleles as captured by genetic diversity measures. The effect of admixture in contributing to increased genetic diversity, however, is not simple. For example, in a study of the genetics of populations founded by relatively small groups, Mooney *et al.* (2018) examined genetic diversity in admixed and non-admixed populations, some of which were regarded as founder populations. Mooney *et al.* (2018) observed that genetic diversity was relatively high in multiple admixed populations of Latin America. This pattern was observed even for populations that, on the basis of small population size and past history of isolation, might have been expected to have relatively low levels of genetic diversity.

Here, to deepen understanding of the relationship between admixture and genetic variability, we focus in admixed populations on levels of genetic diversity computed from allele frequencies, rather than on variability among individuals in admixture proportions. For a model of an admixed population with K source groups, we derive a relationship between genetic diversity, as measured by heterozygosity, and proportions of admixture drawn from the various source populations. The model is the same model we have previously used to examine the genetic differentiation between admixed populations and their source groups, as measured by F_{ST} (Boca & Rosenberg, 2011). We show that for all values of the admixture contributions from the source populations, the heterozygosity of the admixed population is greater than or equal to the smallest of the source population heterozygosities. We further examine the maximal values of the heterozygosity of the admixed population over the space of possible admixture proportions. We consider in more detail special cases with $K = 2$ and $K = 3$ source populations, providing explicit results for $K = 2$ in terms of relatively few parameters. Finally, we use simulations and example analyses from human population data to illustrate the mathematical results.

2 Notation and model

We consider a model with $K \geqslant 2$ source populations and an admixed population arising from these sources. A single polymorphic locus is considered, with $J \geqslant 2$ alleles, such that each of the J alleles appears in at least one of the K source populations.

In Sections 2.1, 2.2, and 2.3, respectively, we define the expected heterozygosity and the fixation index, and we provide a result about relationships between fixation indices and heterozygosities. In Section 2.4, we introduce the admixture model. Notation is summarized in Table 1.

2.1 Expected heterozygosity

The expected heterozygosity is a measure of genetic diversity, giving the probability that two alleles randomly drawn from a population differ in type.

Definition 1. The *expected heterozygosity* in a population for a given locus with J distinct alleles is defined as $H = 1 - \sum_{j=1}^{J} p_j^2$, where p_j is the frequency of allelic type j.

We denote by p_{kj} the frequency of allelic type j, $1 \leqslant j \leqslant J$, in source population k, $1 \leqslant k \leqslant K$, with $0 \leqslant p_{kj} \leqslant 1$. We denote by H_k the expected heterozygosity of source population k at a locus. We have $0 \leqslant H_k < 1$, with $H_k = 0$ if and only if source population k has only a single allelic type of nonzero frequency. For fixed J, the maximal value of H_k is $1 - \frac{1}{J}$, attained when all J alleles have the same frequency, namely $\frac{1}{J}$ (Reddy & Rosenberg, 2012, Lemma 4). We refer to expected heterozygosity simply as heterozygosity.

2.2 Fixation index

The fixation index F_{ST} is a measure of genetic divergence among a set of subpopulations. In its general form, it is computed from H_S, the mean of the heterozygosities of the subpopulations, and H_T, the heterozygosity of a population formed by pooling the subpopulations into a single "total" population.

Definition 2. The *fixation index*, F_{ST} is defined as $F_{ST} = (H_T - H_S)/H_T$, where H_T is the heterozygosity of the total population and H_S is the mean heterozygosity across subpopulations.

The fixation index can be regarded as a measure of genetic divergence between two populations, with $F_{k\ell}$ denoting the value of F_{ST} between source populations k and ℓ For its calculation, the two subpopulations have the same contribution to the overall population, so that they are weighted equally in producing the total population. We assume that when pooled together, the two subpopulations produce a polymorphic population. In other words, for each (k, ℓ), we disallow the case in which there is some allelic type $1 \leqslant j \leqslant J$ for which

$p_{kj} = p_{\ell j} = 1$. Our assumption that pooling any two populations produces a polymorphic population avoids a denominator of 0 in the formula for $F_{k\ell}$.

For this pairwise scenario, $H_S = (H_k + H_\ell)/2$, $H_T = 1 - \sum_{j=1}^{J} \left[\left(p_{kj} + p_{\ell j}\right)/2\right]^2$, and

$$F_{k\ell} = \frac{\left[1 - \sum_{j=1}^{J} \left(\frac{p_{kj} + p_{\ell j}}{2}\right)^2\right] - \frac{H_k + H_\ell}{2}}{1 - \sum_{j=1}^{J} \left(\frac{p_{kj} + p_{\ell j}}{2}\right)^2}. \quad (1)$$

We can observe by the Cauchy-Schwarz inequality that $0 \leqslant F_{k\ell} \leqslant 1$, with $F_{k\ell} = 0$ requiring $p_{kj} = p_{\ell j}$ for all j. $F_{k\ell} = 1$ requires $H_S = H_k = H_\ell = 0$.

2.3 The fixation index in relation to the heterozygosities

We will need a result on the relationship between the fixation index for source populations k and ℓ, $F_{k\ell}$, and the heterozygosities of those source populations, H_k and H_ℓ. We first introduce a quantity, $C_{k\ell}$, the probability that, when randomly drawing one allele from population k and one allele from population ℓ, the two alleles differ in type. For population k, let $\underline{p_k}$ denote a $J \times 1$ column vector of its allele frequencies. $C_{k\ell}$ can then be written as 1 minus the dot product of the allele frequency vectors of populations k and ℓ:

$$C_{k\ell} = 1 - \underline{p_k}' \cdot \underline{p_\ell} = 1 - \sum_{j=1}^{J} p_{kj} p_{\ell j}. \quad (2)$$

This quantity is a generalization of heterozygosity to two populations, as $H_k = C_{kk}$. Because we exclude the case in which populations k and ℓ are fixed for the same allelic type, $C_{k\ell}$ strictly exceeds 0, so that $0 < C_{k\ell} \leqslant 1$. The upper bound of 1 is achieved if populations k and ℓ share no allelic types in common.

We can rewrite eq. 1 as

$$F_{k\ell} = \frac{2C_{k\ell} - H_k - H_\ell}{2C_{k\ell} + H_k + H_\ell}. \quad (3)$$

If $F_{k\ell} < 1$, then we can solve for $C_{k\ell}$:

$$C_{k\ell} = \left(\frac{H_k + H_\ell}{2}\right)\left(\frac{1 + F_{k\ell}}{1 - F_{k\ell}}\right). \quad (4)$$

Recall that $F_{k\ell} = 1$ implies $H_k = H_\ell = 0$, so that populations k and ℓ each have only a single allelic type with nonzero frequency. We have excluded the case in which the two populations are fixed for the same allelic type; hence, they must be fixed for different allelic types, and $C_{k\ell} = 1$ in eq. 2.

We have previously shown by the Cauchy-Schwarz inequality that $1 - \sqrt{(1 - H_k)(1 - H_\ell)} \leqslant C_{k\ell} \leqslant 1$ (Mehta *et al.*, 2019, eq. 7). Equality in the lower bound requires $p_{kj} = p_{\ell j}$ for all j, and hence $H_k = H_\ell$. Rewriting this inequality with eq. 4, we obtain the allowable space of $F_{k\ell}$ given H_k, $H_\ell \in [0, 1)$:

$$F_{k\ell} \in \left[\frac{2 - H_k - H_\ell - 2\sqrt{(1 - H_k)(1 - H_\ell)}}{2 + H_k + H_\ell - 2\sqrt{(1 - H_k)(1 - H_\ell)}}, \frac{2 - H_k - H_\ell}{2 + H_k + H_\ell}\right]. \quad (5)$$

The lower limit is achieved if and only if the two populations k and ℓ are identical, with $H_k = H_\ell$ and $p_{kj} = p_{\ell j}$ for all j. The upper limit is achieved if and only if populations k and ℓ share no allelic types in common. This result adds to the understanding of constraints on F_{ST} placed by genetic diversity (Nagylaki, 1998; Hedrick, 1999; Long & Kittles, 2003; Rosenberg *et al.*, 2003; Hedrick, 2005; Boca & Rosenberg, 2011; Maruki *et al.*, 2012; Jakobsson *et al.*, 2013; Edge & Rosenberg, 2014; Alcala & Rosenberg, 2017, 2019; Mehta *et al.*, 2019). We use the allowable region to constrain our examples to permissible values of (H_k, H_ℓ, F_{ST}).

Appendix A of Mehta *et al.* (2019) shows that given H_k and H_ℓ in [0, 1), if the number of distinct alleles J is not fixed, then we can choose allele frequency vectors $\underline{p_k}$ and $\underline{p_\ell}$ such that each $C_{k\ell}$ value in $[1 - \sqrt{(1 - H_k)(1 - H_\ell)}, 1]$ is achievable. The lower bound is achievable only if $H_k = H_\ell$. Hence, each value in the interval in eq. 5 for F_{ST} is also achievable by some pair $\underline{p_k}$ and $\underline{p_\ell}$, the lower bound only if $H_k = H_\ell$.

2.4 Admixture model

We use an admixture model that describes current patterns of variation in an admixed population, rather than mechanistic dynamics. This model follows a commonly used approach, treating allele frequencies in the admixed population as linear combinations of those of the source populations (e.g. Pritchard *et al.*, 2000; Boca & Rosenberg, 2011).

In our K-source-population model, $K \geqslant 2$, we follow Section 2.2 in assuming that no two populations are fixed for the same allelic type. We now make a stronger assumption that no two populations are identical, so that for each (k, ℓ), some j exists for which $p_{kj} \neq p_{\ell j}$. Further, it is convenient to assume that no source population can have its vector of allele frequencies written as the linear combination of vectors of allele frequencies of other source populations; otherwise, an admixed population would not have a unique representation as a linear combination of sources. We thus assume that not only are no two source populations identical, no source can be described as an admixture of two or more of the other sources.

Note that the assumption that no population is a linear combination of the others also excludes linear combinations with one or more negative coefficients. Because the maximal number of vectors of length J that can be linearly independent is J, the linear independence assumption implies $J \geqslant K$. A succinct way of describing the assumption is that if we define the $J \times K$ matrix of allele frequencies in the source populations,

$$P = \begin{pmatrix} p_{11} & p_{21} & \cdots & p_{K1} \\ p_{12} & p_{22} & \cdots & p_{K2} \\ \cdots & \cdots & \cdots & \cdots \\ p_{1J} & p_{2J} & \cdots & p_{KJ} \end{pmatrix} = \left(\underline{p_1}, \underline{p_2}, \ldots, \underline{p_K}\right), \tag{6}$$

then we assume that P has rank K.

For the admixed population generated from the K source populations, we denote by γ_k the admixture fraction for source population k; for each k with $1 \leqslant k \leqslant K$, fraction γ_k of the ancestry of the admixed population, $0 \leqslant \gamma_k \leqslant 1$, derives from source k. We denote by $\underline{\gamma}$ the K×1 column vector of admixture fractions. This vector lies in the simplex Δ^{K-1}, the set of all vectors of K nonnegative entries with $\sum_{k=1}^{K} \gamma_k = 1$.

The frequency of allele j in the admixed population is denoted $\bar{p}_j$. By the linear combination assumption,

$$\bar{p}_j = \sum_{k=1}^{K} \gamma_k p_{kj}\,. \tag{7}$$

In the special case that $\gamma_k = \frac{1}{K}$ for each K, the admixed population is equivalent to the "pooled population" used in defining the fixation index F_{ST} among the K populations.

3 General case: *K* source populations

Our goal is to study the heterozygosity of the admixed population. Using Definition 1 with eq. 7, we compute the heterozygosity for the admixed population, which we denote by H_{adm}:

$$H_{\text{adm}} = 1 - \sum_{j=1}^{J} \bar{p}_j^2 = 1 - \sum_{j=1}^{J} \left(\sum_{k=1}^{K} \gamma_k p_{kj} \right)^2. \tag{8}$$

The heterozygosity of the admixed population can be written in terms of the heterozygosities of the source populations and the dot products of the allele frequencies. Using eq. 4 in eq. 8, we have:

$$H_{\text{adm}} = \sum_{k=1}^{K} \gamma_k^2 H_k + 2 \sum_{k=1}^{K-1} \sum_{\ell=k+1}^{K} \gamma_k \gamma_\ell C_{k\ell} \tag{9}$$

$$= \sum_{k=1}^{K} \gamma_k^2 H_k + \sum_{k=1}^{K-1} \sum_{\ell=k+1}^{K} \gamma_k \gamma_\ell (H_k + H_\ell) \left(\frac{1 + F_{k\ell}}{1 - F_{k\ell}} \right). \tag{10}$$

The last simplification can be made only for $F_{k\ell} \neq 1$; if $F_{k\ell} = 1$, then eq. 9 is used, or, as noted after eq. 4, $(H_k + H_\ell)(1 + F_{k\ell})/(1 - F_{k\ell})$ is understood to equal 2.

With the formula for H_{adm} established, we now explore how H_{adm} varies in relation to the admixture fractions γ. Given the allele frequencies P, we determine the range of H_{adm} over the space of possible values of γ. We write H_m for the smallest heterozygosity among the source populations, $H_m = \min_{k \in \{1, 2, \dots, K\}} H_k$, and H_M for the largest heterozygosity among the source populations, $H_M = \max_{k \in \{1, 2, \dots, K\}} H_k$.

3.1 Minimum of H_{adm} in terms of the ancestry proportions

For the minimum of H_{adm} over vectors $(\gamma_1, \gamma_2, \dots, \gamma_K)$, we can immediately observe from the form of eq. 10 that for a fixed set of source population allele frequencies P, H_{adm} is minimized as a function of the admixture fractions when the admixed population consists of only one of the source populations.

Proposition 3. The minimum of H_{adm} as a function of the ancestry proportions γ is $H_m = \min_{k \in \{1, 2, \dots, K\}} H_k$, the smallest heterozygosity among the source populations, and it is obtained when the admixed population consists solely of that source population.

Proof. To obtain this result, we use eq. 10 and the fact that $H_k \geqslant H_m$ for all k:

$$\begin{aligned} H_{\text{adm}} &= \sum_{k=1}^{K} \gamma_k^2 H_k + \sum_{k=1}^{K-1} \sum_{\ell=k+1}^{K} \gamma_k \gamma_\ell (H_k + H_\ell) \left(\frac{1 + F_{k\ell}}{1 - F_{k\ell}} \right) \\ &\geqslant \sum_{k=1}^{K} \gamma_k^2 H_m + \sum_{k=1}^{K-1} \sum_{\ell=k+1}^{K} 2 \gamma_k \gamma_\ell H_m \\ &= \left(\sum_{k=1}^{K} \gamma_k \right)^2 H_m = H_m . \end{aligned}$$

Because equality is achieved when $\gamma_m = 1$ and $\gamma_k = 0$ for all $k \neq m$, we have shown that the minimal value of H_{adm} as a function of the ancestry proportions is H_m. □

The result finds that nonzero admixture inflates heterozygosity at least above the level seen in the least heterozygous source. It applies whether or not $H_1, H_2, \dots, H_K$ are mutually distinct. If two or more of $H_1, H_2, \dots, H_K$ are tied for the minimal heterozygosity H_m, then the minimum of H_{adm} is achieved at each vector associated with complete ancestry from one of the minimally heterozygous populations.

A consequence of Proposition 3 is that if all K populations have the same heterozygosity H_m—for example, in cases where the different alleles have distinct frequencies and each population has an allele frequency vector that is a permutation of the vectors for the other populations—then $H_{\text{adm}} > H_m$ for all ancestry vectors γ with two or more nonzero entries. In particular, note that $F_{k\ell} > 0$ for each (k, ℓ), $k \neq \ell$, by the assumption that each pair of source populations has distinct allele frequencies. Hence, $(H_k + H_\ell)(1 + F_{k\ell})/(1 - F_{k\ell}) > 2H_m$ for each (k, ℓ), $k \neq \ell$. Because at least one product $\gamma_k \gamma_\ell$ is positive, the inequality $\gamma_k \gamma_\ell (H_k + H_\ell)(1 + F_{k\ell})/(1 - F_{k\ell}) \geqslant 2\gamma_k \gamma_\ell H_m$ is strict for at least one (k, ℓ), so that $H_{\text{adm}} > (\sum_{k=1}^{K} \gamma_k)^2 H_m = H_m$. This same reasoning shows that if two or more populations are tied with heterozygosity H_m, then $H_{\text{adm}} > H_m$ for each $\underline{\gamma}$ with two or more nonzero entries.

We note that the result $H_{\text{adm}} \geqslant \min_{k \in \{1,2,\ldots,K\}} H_k$ for all $\underline{\gamma} \in \Delta^{K-1}$ in Proposition 3 can be quickly obtained from the classic Wahlund principle, by which the heterozygosity of a population formed by mixing populations 1, 2, …, K, with proportion γ_k of the mixed population taken from population k, $0 \leqslant \gamma_k \leqslant 1$, is greater than or equal to the mean of the K population heterozygosities (e.g. Rosenberg & Calabrese, 2004, Theorem 2). The heterozygosity of the population mixture in the setting of the Wahlund principle is the same as the heterozygosity of the admixed population in our scenario. Thus, in our notation, setting $\gamma_k = \frac{1}{K}$ for all k, the Wahlund principle gives $H_{\text{adm}} \geqslant \frac{1}{K}\sum_{k=1}^{K} H_k$. Because the mean $\frac{1}{K}\sum_{k=1}^{K} H_k$ is greater than or equal to the minimum $\min_{k \in \{1,2,\ldots,K\}} H_k$, it immediately follows that $H_{\text{adm}} \geqslant \min_{k \in \{1,2,\ldots,K\}} H_k$.

3.2 Maximum of H_{adm} in terms of the ancestry proportions

To obtain the maximum of H_{adm} over the space of values of $\underline{\gamma}$, we write eq. 9 as a quadratic form:

$$H_{\text{adm}}(\underline{\gamma}) = \underline{\gamma}' A \underline{\gamma}\,.$$

Here, $\underline{\gamma}'$ represents the transpose of the column vector $\underline{\gamma}$ and A is the $K \times K$ symmetric matrix with the H_k on the diagonal and the $C_{k\ell}$ off the diagonal:

$$A = \begin{pmatrix} H_1 & C_{12} & \ldots & C_{1K} \\ C_{12} & H_2 & \ldots & C_{2K} \\ \ldots & \ldots & \ldots & \ldots \\ C_{1K} & C_{2K} & \ldots & H_K \end{pmatrix} = \underline{1}\underline{1}' - \begin{pmatrix} \sum_{j=1}^{J} p_{1j}^2 & \sum_{j=1}^{J} p_{1j}p_{2j} & \ldots & \sum_{j=1}^{J} p_{1j}p_{Kj} \\ \sum_{j=1}^{J} p_{1j}p_{2j} & \sum_{j=1}^{J} p_{2j}^2 & \ldots & \sum_{j=1}^{J} p_{2j}p_{Kj} \\ \ldots & \ldots & \ldots & \ldots \\ \sum_{j=1}^{J} p_{1j}p_{Kj} & \sum_{j=1}^{J} p_{2j}p_{Kj} & \ldots & \sum_{j=1}^{J} p_{Kj}^2 \end{pmatrix} \tag{11}$$
$$= \underline{1}\underline{1}' - P'P,$$

where P is the $J \times K$ allele frequency matrix (eq. 6) and $\underline{1}$ is a $K \times 1$ vector of ones.

Maximizing H_{adm} in terms of $\underline{\gamma}$ is equivalent to finding $\max_{\underline{\gamma} \in \Delta^{K-1}} \underline{\gamma}' A \underline{\gamma}$ subject to $\underline{1}'\underline{\gamma} = 1$. We denote by $\underline{\gamma}_{\text{arg max}}$ the location of the maximal value of H_{adm}. We first observe that $\underline{\gamma}_{\text{arg max}}$ is sometimes interior to the simplex, and that it sometimes lies at a vertex. In other words, for a fixed set of sources, a population nontrivially admixed among the sources can sometimes have a higher heterozygosity than all of the sources, but sometimes, *no* population admixed among the sources has higher heterozygosity than all the sources.

Proposition 4. Consider the case of K source populations, $K \geqslant 2$.

(i) There exists some collection of source population allele frequencies P and some collection of admixture proportions $\underline{\gamma}$ for which the heterozygosity of the admixed population exceeds the heterozygosity H_M of the most heterozygous source population.

(ii) There exists some collection of source population allele frequencies P for which *no* collection of admixture proportions $\underline{\gamma}$ produces an admixed population with heterozygosity greater than the heterozygosity H_M of the most heterozygous source population.

Proof. (i) Consider K populations, each with different allele frequencies, but identical heterozygosity: $\underline{p_k} \neq \underline{p_\ell}$ for $k \neq \ell$ but $H_k = H$ for $k = 1, 2, \ldots, K$. Suppose that a locus has $K + 1$ distinct alleles, and that the allele frequencies are $\underline{p_1} = \left(\frac{1}{2}, \frac{1}{2}, 0, 0, \ldots, 0\right)$, $\underline{p_2} = \left(\frac{1}{2}, 0, \frac{1}{2}, 0, \ldots, 0\right), \ldots, \underline{p_K} = \left(\frac{1}{2}, 0, 0, \ldots, 0, \frac{1}{2}\right)$. By eq. 9, $H_{\text{adm}} = \frac{3}{4} - \frac{1}{4}\sum_{k=1}^{K}\gamma_k^2$, which is minimized if and only if $\sum_{k=1}^{K}\gamma_k^2 = 1$ or $\underline{\gamma} = \underline{e_k}$ for some k. The minimal value of H_{adm} is thus $\frac{1}{2}$, all other values of the admixture proportions resulting in $H_{\text{adm}} > H = \frac{1}{2}$.

(ii) Consider K populations and a locus with K distinct alleles. Suppose that the number of distinct alleles at the locus is k for population k, with $\underline{p_k} = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$. Hence, $H_k = 1 - \frac{1}{k}$ and, in particular, $H_1 < \ldots < H_K$. We show that $H_{\text{adm}} \leqslant H_K$ irrespective of $\underline{\gamma}$.

By eq. 9,

$$H_{\text{adm}} = 1 - \left(\gamma_1 + \frac{\gamma_2}{2} + \ldots + \frac{\gamma_K}{K}\right)^2 - \ldots - \left(\frac{\gamma_K}{K}\right)^2.$$

By the Cauchy-Schwarz inequality:

$$\left[\left(\gamma_1 + \frac{\gamma_2}{2} + \ldots + \frac{\gamma_K}{K}\right)^2 + \ldots + \left(\frac{\gamma_K}{K}\right)^2\right]K \geqslant \left(\gamma_1 + \frac{\gamma_2}{2}2 + \ldots + \frac{\gamma_K}{K}K\right)^2 = \left(\sum_{k=1}^{K}\gamma_k\right)^2 = 1.$$

Thus, $H_{\text{adm}} \leqslant 1 - \frac{1}{K} = H_K$. □

The proof is constructive, exhibiting example source groups for which specific features are obtained. In part (i), each source has an allele that is not present in the other sources, and a nontrivially admixed population—which possesses all of these private alleles—is necessarily more heterozygous than each source. For part (ii), we have a sequence of increasingly heterozygous source populations, each with one additional allele, and no population admixed among them is more heterozygous than the most heterozygous source. Other constructive examples are possible, with, for example, low heterozygosities but distinct alleles across populations generating additional examples along the lines of Proposition 4i.

Note that it is trivial to see that in general, $\max_{\underline{\gamma} \in \Delta^{K-1}} H_{\text{adm}}(\underline{\gamma}) \geqslant \max\{H_1, \ldots, H_K\}$: the K source populations simply correspond to the K vertices of the simplex. This result that the maximal H_{adm} is at least is great as the heterozygosity of the most heterozygous source population immediately implies $\max_{\underline{\gamma} \in \Delta^{K-1}} H_{\text{adm}}(\underline{\gamma}) \geqslant \frac{1}{K}\sum_{k=1}^{K} H_k$.

Having established that the maximum can be at a vertex or an interior point of the simplex—a trivial admixed population consisting only of a single source population, or a population admixed among all the sources—we now provide a general theorem. The theorem gives the location of the maximum when it lies in the interior of Δ^{K-1}, rather than on the boundary, assuming a condition applies on the allele frequencies. The proof is in Appendix 1, making use of a general constrained quadratic optimization procedure.

Theorem 5. Suppose that $\underline{1}'(P'P)^{-1}\underline{1} \neq 1$. Suppose also that $\frac{A^{-1}\underline{1}}{\underline{1}'A^{-1}\underline{1}} \in \Delta^{K-1}$. Then the maximum of H_{adm} as a function of the ancestry proportions $\underline{\gamma} \in \Delta^{K-1}$ is attained at $\underline{\gamma_{\text{arg max}}} = \underline{\gamma}^*$, where:

$$\underline{\gamma}^* = \frac{A^{-1}\underline{1}}{\underline{1}'A^{-1}\underline{1}} = \frac{(P'P)^{-1}\underline{1}}{\underline{1}'(P'P)^{-1}\underline{1}}.$$

The maximum is equal to:

$$H_{\text{adm}}(\underline{\gamma}^*) = \frac{1}{\underline{1}'A^{-1}\underline{1}} = 1 - \frac{1}{\underline{1}'(P'P)^{-1}\underline{1}}.$$

If $\frac{A^{-1}\underline{1}}{\underline{1}'A^{-1}\underline{1}} \notin \Delta^{K-1}$, then $\underline{\gamma_{\text{arg max}}}$ lies on the boundary of the set $\{\underline{\gamma} : \underline{1}'\underline{\gamma} = 1 \text{ and } \underline{\gamma} \in \Delta^{K-1}\}$.

The "boundary" of a set R is the set of points in R for which a neighborhood around them always contains both points in R and points in the complement of R. For simplex Δ^{K-1}, the boundary includes all points for which at least one of the K coordinates is 0, with the vertices occurring at locations where all of the coordinates except one are 0.

The following corollary, also proven in Appendix 1, further describes the possible locations of the maximal H_{adm}. Note that if the maximum is not at $\underline{\gamma}^*$, then it lies at a point that has some elements equal to 0, the nonzero subvector having a similar form to $\underline{\gamma}^*$, but in a lower number of dimensions. Thus, the maximum can occur in a scenario in which the admixture involves only a strict subset of the source populations.

Consider a nonempty subset $\mathcal{S} \subset \{1, 2, \ldots, K\}$. Define by $A_{\mathcal{S}}$ the $|\mathcal{S}| \times |\mathcal{S}|$ matrix that has diagonal terms H_k for each $k \in \mathcal{S}$ and off-diagonal terms $C_{k\ell}$for each distinct $k, \ell \in \mathcal{S}$. Additionally, denote by $P_{\mathcal{S}}$ the matrix consisting of the columns of P corresponding to the subset $\mathcal{S}$. $P_{\mathcal{S}}$ contains the allele frequencies for the source populations in $\mathcal{S}$.

Corollary 6. Suppose that $\underline{1}'(P'_{\mathcal{S}}P_{\mathcal{S}})^{-1}\underline{1} \neq 1$ for all nonempty $\mathcal{S} \subset \{1, 2, \ldots, K\}$. Then the maximum of H_{adm} as a function of the ancestry proportions $\underline{\gamma} \in \Delta^{K-1}$ is attained at a point that has nonzero elements for some nonempty subset of the source populations

$\mathcal{S}^* \subset \{1, 2, \ldots, K\}$. The nonzero subvector of ancestry proportions at the location of the maximum is equal to $\underline{\gamma}_{\mathcal{S}^*} = \frac{A_{\mathcal{S}^*}^{-1}\underline{1}}{\underline{1}' A_{\mathcal{S}^*}^{-1}\underline{1}}$.

In particular, note that $\underline{\gamma}_{\text{arg max}} = \underline{\gamma}^*$ corresponds to $\mathcal{S}^* = \{1, 2, \ldots, K\}$: all source populations contribute nonzero admixture fractions. The K vertices of the simplex Δ^{K-1} correspond to the cases of $\mathcal{S}^* = \{k\}$, at which only one source population contributes. $\mathcal{S}$ has $2^K - 1$ nonempty subsets, each representing a distinct collection of source populations.

4 *K* = 2 source populations

With general results established for the case of arbitrary K, we now focus on the simplest case, with $K = 2$ source populations contributing to the admixed population.

We continue to exclude the scenario in which the allele frequencies for the two source populations are identical, so that we assume $\underline{p_1} \neq \underline{p_2}$. Noting that $\gamma_2 = 1 - \gamma_1$, we can consider H_{adm} in terms of a single admixture coefficient γ_1, the admixture fraction of the first population, with $\gamma_1 \in [0, 1]$. Using eqs. 9 and 10 with this substitution, we obtain:

$$\begin{aligned} H_{\text{adm}} &= \gamma_1^2 H_1 + (1-\gamma_1)^2 H_2 + 2\gamma_1(1-\gamma_1)C_{12} && (12) \\ &= \gamma_1^2 H_1 + (1-\gamma_1)^2 H_2 + \gamma_1(1-\gamma_1)(H_1+H_2)\frac{1+F_{12}}{1-F_{12}} && (13) \\ &= \gamma_1^2(H_1 + H_2 - 2C_{12}) - 2\gamma_1(H_2 - C_{12}) + H_2\,. && (14) \end{aligned}$$

In particular, we note from eq. 13 that H_{adm} is increasing as a function of F_{12}.

From eq. 14, we can see that H_{adm} is concave down in γ_1. We have $d^2H_{\text{adm}}/d\gamma_1^2 = 2(H_1 + H_2 - 2C_{12})$. By Definition 1 and eq. 2, $2(H_1 + H_2 - 2C_{12}) = -2\sum_{j=1}^{J}(p_{1j} - p_{2j})^2$. Because $\underline{p_1} \neq \underline{p_2}$, $p_{1j} \neq p_{2j}$ for at least one choice of j, and hence $d^2H_{\text{adm}}/d\gamma_1^2 < 0$. By symmetry, H_{adm} is also concave down in γ_2.

To illustrate eq. 13, for H_1 and H_2 fixed, Figure 1 plots the concave-down H_{adm} as a function of γ_1 for a variety of values of F_{12}. We observe that for each value of F_{12} considered, the minimum of H_{adm} occurs at $(\gamma_1, \gamma_2) = (0, 1)$, reflecting the result of Proposition 3 that the minimum occurs when the admixed population consists solely of the less heterozygous source population. In accord with the fact that in eq. 13, H_{adm} increases for fixed H_1, H_2, and γ_1 with increasing F_{12}, the value at the maximum increases with increasing F_{12}. The location of the maximum lies at a value of $\gamma_1 \geqslant \frac{1}{2}$, decreasing with increasing F_{12}. This location has a pattern where for larger values of F_{12}, it lies interior to the unit interval, and for smaller values of F_{12}, it occurs when the admixed population consists solely of the more heterozygous source population. We now consider this pattern in more detail.

4.1 Minimum and maximum of H_{adm} in terms of the ancestry proportions

Applying the results from Section 3.1 on the minimum and maximum of H_{adm} as a function of $\underline{\gamma}$, by Proposition 3, H_{adm} has minimum $\min\{H_1, H_2\}$. The maximum can occur in one of three locations.

Proposition 7. Consider two source populations with distinct allele frequencies, $\underline{p_1} \neq \underline{p_2}$. As a function of γ_1, H_{adm} is maximized at $\gamma_1 = \gamma_1^*$, where γ_1^* takes one of three forms.

(i) If $H_1 < C_{12}$ and $H_2 < C_{12}$, then $\gamma_1^* \in (0, 1)$ satisfies

$$\gamma_1^* = \frac{C_{12} - H_2}{2(C_{12} - H_S)} = \frac{1}{2} + \frac{H_1 - H_2}{8(H_T - H_S)}, \tag{15}$$

and H_{adm} has maximum equal to

$$H_{\text{adm}}(\gamma_1^*) = \frac{C_{12}^2 - H_1 H_2}{2(C_{12} - H_S)} = H_T + \frac{(H_1 - H_2)^2}{16(H_T - H_S)}. \tag{16}$$

(ii) If $H_1 < C_{12}$ and $H_2 \geqslant C_{12}$, then $\gamma_1^* = 0$ and H_{adm} has maximum H_2.

(iii) If $H_1 \geqslant C_{12}$ and $H_2 < C_{12}$, then $\gamma_1^* = 1$ and H_{adm} has maximum H_1.

An elementary proof appears in Appendix 2. The locations specified in Proposition 7 accord with Theorem 5 and Corollary 6. For $K = 2$, the result of Theorem 5 gives $\underline{\gamma}^* = \frac{A^{-1}\underline{1}}{\underline{1}'A^{-1}\underline{1}} = \left(\frac{C_{12} - H_2}{2(C_{12} - H_S)}, \frac{C_{12} - H_1}{2(C_{12} - H_S)}\right)$, where

$$A = \begin{pmatrix} H_1 & C_{12} \\ C_{12} & H_2 \end{pmatrix}.$$

The locations in Corollary 6 are $\gamma_1^* = \frac{A_1^{-1}}{A_1^{-1}} = 1$ and $\gamma_2^* = 0$, and $\gamma_1^* = 0$ and $\gamma_2^* = \frac{A_2^{-1}}{A_2^{-1}} = 1$.

We now give two corollaries of Proposition 7, providing more features of the maximal H_{adm} for specific cases. Proofs appear in Appendix 2. In accord with the observation in Figure 1 that the maximal H_{adm} lies at a value of $\gamma_1 \geqslant \frac{1}{2}$ in an example with $H_1 \geqslant H_2$, Corollary 8 demonstrates $\gamma_1^* \geqslant \frac{1}{2}$ if and only if $H_1 \geqslant H_2$.

Corollary 8. Consider two source populations with distinct allele frequencies, $\underline{p_1} \neq \underline{p_2}$. As a function of γ_1, H_{adm} is maximized at $\gamma_1^* \geqslant \frac{1}{2}$ if and only if $H_1 \geqslant H_2$.

A second corollary is that the maximal H_{adm} is always at least as great as H_T.

Corollary 9. Consider two source populations with distinct allele frequencies, $\underline{p_1} \neq \underline{p_2}$. Then $H_{\text{adm}}(\gamma_1^*) \geqslant H_T$, with equality occurring if $H_1 = H_2$.

We can also succinctly describe the region where γ_1^* lies interior to (0, 1).

Corollary 10. Consider two source populations with distinct allele frequencies, $\underline{p_1} \neq \underline{p_2}$. γ_1^* lies in (0, 1) if and only if the following inequality holds:

$$F_{12} > \frac{|H_1 - H_2|}{2(H_1 + H_2) + |H_1 - H_2|}. \tag{17}$$

This corollary is proven in Appendix 2. Note that if $H_1 + H_2$ is fixed, then the right-hand side of eq. 17 increases with $|H_1 - H_2|$, from a minimum of 0 when $H_1 = H_2$ to a maximum of $\frac{1}{3}$ as $|H_1 - H_2|$ approaches $H_1 + H_2$. Thus, in accord with the observation in Section 3.1 that $H_{\text{adm}} > H$ for all nontrivial admixtures of equal-heterozygosity source populations, the maximal H_{adm} exceeds $\max\{H_1, H_2\}$ over a broader range of F_{12} values if $|H_1 - H_2|$ is small rather than large. Moreover, if $F_{12} > \frac{1}{3}$, then eq. 17 necessarily holds. Hence, irrespective of H_1 and H_2, if the source populations are distant enough that $F_{12} > \frac{1}{3}$, then the maximal heterozygosity exceeds the heterozygosities of the source populations.

4.2 Special case of *J* = 2 alleles

For $K = 2$ sources, when the locus has only $J = 2$ allelic types, further simplifications are possible, as results can be stated in terms of frequencies of one specific allele. We substitute $p_{12} = 1 - p_{11}$ and $p_{22} = 1 - p_{21}$.

Proposition 11. Consider two source populations with distinct allele frequencies, $\underline{p_1} \neq \underline{p_2}$. For a biallelic locus, H_{adm} is maximized at $\gamma_1 = \gamma_1^*$, where γ_1^* takes one of three forms.

(i) If $p_{11} > \frac{1}{2} > p_{21}$ or $p_{21} > \frac{1}{2} > p_{11}$, then $\gamma_1^* \in (0, 1)$ satisfies

$$\gamma_1^* = \frac{1 - 2p_{21}}{2(p_{11} - p_{21})}, \tag{18}$$

and H_{adm} has maximum equal to

$$H_{\text{adm}}(\gamma_1^*) = \frac{1}{2}. \tag{19}$$

(ii) If $\frac{1}{2} \geqslant p_{21} > p_{11}$ or $p_{11} > p_{21} \geqslant \frac{1}{2}$, then $\gamma_1^* = 0$ and H_{adm} has maximum H_2.

(iii) If $\frac{1}{2} \geqslant p_{11} > p_{21}$ or $p_{21} > p_{11} \geqslant \frac{1}{2}$, then $\gamma_1^* = 1$ and H_{adm} has maximum H_1.

The result is proven in Appendix 2. The unit square representing possible values of the location of the maximum appears in Figure 2. It has six nonoverlapping regions: in Proposition 11, each of the three cases generates two disjoint subsets of $[0,1]^2$. A smooth gradient exists for regions in case (i). However, an abrupt transition occurs at the line $p_{21} = p_{11}$ between case-(ii) regions where $\gamma_1^* = 0$ and case-(iii) regions where $\gamma_1^* = 1$. Note that the $p_{21} = p_{11}$ line, where the two populations have equal allele frequencies, is disallowed.

5 Simulations

We illustrate properties of H_{adm} by simulating population sets for different values of K and J. Given a value of K, we generated allele frequency vectors for the K source populations from independent and identically distributed symmetric multivariate J-dimensional Dirichlet distributions with a common concentration parameter $\alpha = 1$. This distribution corresponds to a uniform distribution on the simplex Δ^{J-1}. A number of mathematical results can be obtained in this Dirichlet setting; these appear in Appendix 3.

First, for $K = 2$ and $K = 3$, we assessed the probability that the maximal H_{adm} over possible admixture vectors $\underline{\gamma}$ occurs interior to the simplex Δ^{K-1}, rather than on its boundary. This computation gives the probability that the heterozygosity-maximizing admixture vector contains nonzero contributions from all K source populations. We considered $2 \leqslant J \leqslant 30$ for $K = 2$ and $3 \leqslant J \leqslant 30$ for $K = 3$, recalling the condition $J \geqslant K$ for the K allele frequency vectors to be linearly independent.

For each (K, J), we ran 10,000 simulation replicates. In each replicate, to determine the location of the maximum, we applied Theorem 5 and Corollary 6 to identify the locations specified for each choice $\mathcal{S}$ of the nonempty subset of the K populations with nonzero allele frequencies. Among these 2^{K-1} locations, excluding those outside the simplex Δ^{K-1}, we identified the point with the largest H_{adm}. Note that in each replicate, we observed that the $\underline{1}'(P_{\mathcal{S}}' P_{\mathcal{S}})^{-1}\underline{1} \neq 1$ condition of Corollary 6 was satisfied for each $\mathcal{S}$.

Figure 3 finds that, for both $K = 2$ and $K = 3$, the maximum of H_{adm} is increasingly likely to be in the interior of the simplex as the number of distinct alleles, J, increases. For $K = 3$, we also observe that the probability that H_{adm} is maximized on an edge, corresponding to nonzero contributions from two of three sources, exceeds the probability that it is maximized at a vertex, with only one contributing source.

Next, we assessed the probability $\mathbb{P}[H_{\text{adm}} > \max\{H_1, \ldots, H_K\}]$ in a scenario in which both the allele frequency vectors $\underline{p_k}$ and the admixture fractions $\underline{\gamma}$ were chosen from independent Dirichlet distributions. We simulated the $\underline{p_k}$ as before, additionally simulating $\underline{\gamma}$ from a K-dimensional symmetric Dirichlet-(1, 1, …, 1) distribution. For each (K, J) with $K = 2, 3, 4, 5$ and $J = 2, 3, \ldots, 30$, we simulated 50,000 replicate populations. Note that here, unlike in Section 2.4, we impose no restrictions on linear combinations of allele frequency vectors from the source populations, so that it is not necessarily true that $J \geqslant K$.

The fraction of replicates with $\mathbb{P}[H_{\text{adm}} > \max\{H_1, \ldots, H_K\}]$ appears in Figure 4. We see that this fraction increases with K: for an admixture involving more populations, the probability is larger that the admixed population exceeds all source populations in heterozygosity. This probability also increases with J.

For $(K, J) = (2, 2)$, Proposition 17 in Appendix 3 obtains the probability analytically, $\mathbb{P}[H_{\text{adm}} > \max\{H_1, H_2\}] = 1 - \log 2 \approx 0.307$. Following this result, the $K = 2$ curve in Figure 4 begins near (2, 0.307).

Figure 5 provides further detail on H_{adm} in the $K = 2$ case by graphing H_{adm} versus γ_1 for 10 simulation replicates chosen at random for each of three values of J. The figure illustrates that H_{adm} is a concave-down quadratic polynomial in γ_1, as in eq. 14. Averaging across replicates, by examining the figure panels from left to right, we can also observe that $\mathbb{E}[H_{\text{adm}}]$ increases as a function of J, as in Corollary 16 of Appendix 3. For $J = 2$, as in Proposition 11, the possible values of H_{adm} at the maximum are H_1, H_2, and $\frac{1}{2}$.

6 Application to data

Next, we illustrate the mathematical results using data from human populations. As multiallelic loci satisfy $J \geqslant K$ with both $K = 2$ and $K = 3$, we focus on a multiallelic data example. First, we begin with a simpler biallelic data set whose set of individuals overlaps with the multiallelic data set, illustrating our maximal heterozygosity results in the case of $K = 2$ source populations. For both data sets, we treat allele frequencies, heterozygosities, and F_{ST} values computed from the data as parametric values rather than estimates.

6.1 Biallelic loci: K = 2 source populations

We consider the single-nucleotide polymorphism (SNP) data of Li *et al.* (2008), as employed by Pemberton *et al.* (2012) in phased form with no missing data. In this data set, which contains 640,034 autosomal SNPs, we consider Europeans and Native Americans as putative source populations for an admixed population, considering the 156 Europeans and 63 Native Americans in the data. We drop from consideration the 32,989 SNPs with identical allele frequencies in the two populations; 32,888 of these are monomorphic.

We select 20 loci at random from the data set for illustration. Treating γ_1 as the fraction of European ancestry in an admixed population and $1 - \gamma_1$ as the fraction of Native American ancestry, for each locus, the plot for H_{adm} versus γ_1 appears in Figure 6. Following Proposition 3, the minimum of H_{adm} lies either at $\gamma_1 = 0$ or at $\gamma_1 = 1$ for all loci. For 3 of the 20 loci, the maximum lies in the interior of the unit interval (case (i) of Proposition 11); 8 loci have the maximum at $\gamma_1 = 0$, representing membership in the less heterozygous Native American population (case (ii)); and 9 loci have the maximum at $\gamma_1 = 1$, representing membership in the more heterozygous European population (case (iii)). Following Proposition 11i, at each locus for which the maximum lies in the interior, the maximum is equal to $\frac{1}{2}$.

Examining all 607,045 loci, 19% have the maximum in the interior, 27% at γ_1=0, and 54% at $\gamma_1 = 1$. That more loci have the maximum at $\gamma_1 = 1$ than $\gamma_1 = 0$ is expected from the fact that European populations generally have greater heterozygosity than Native American populations (e.g. Pemberton *et al.*, 2013).

6.2 Multiallelic loci: $K = 2$ source populations

For our multiallelic data set, we follow Boca & Rosenberg (2011) in considering data from Wang *et al.* (2008) on 678 microsatellite loci typed in 160 Europeans, 463 Native Americans, 123 Africans, and 249 individuals from admixed Mestizo populations. To represent Mestizo populations under our model, we use Europeans and Native Americans as source populations in the $K = 2$ case, also including Africans for $K = 3$.

As we did in the biallelic data set, we select 20 loci at random from Wang *et al.* (2008), choosing the same loci as in Boca & Rosenberg (2011). Again treating γ_1 as the fraction of European ancestry and $1 - \gamma_1$ as the fraction of Native American ancestry in an admixed population, for each locus, the plot for H_{adm} versus γ_1 appears in Figure 7. Comparing Figures 7 and 6, we see that the maximum of H_{adm} lies in the interior of the unit interval for γ_1 more often for the multiallelic than for the biallelic loci. Indeed, examining all 678 loci, 53% have the maximum in the interior—a greater number than for the SNPs. The fraction with the maximum at $\gamma_1 = 1$ is 39%, and 8% have the maximum at $\gamma_1 = 0$.

The Dirichlet model in Corollary 16 in Appendix 3 and Figures 3 and 5 predicts a dependence of the location of the maximum on the number of distinct alleles of a locus, with the probability that the maximum lies in the interior increasing with the number of distinct alleles. The multiallelic data produce a trend in the same direction as this prediction. The mean numbers of distinct alleles are 9.36, 10.40, and 10.75, for the loci with γ_1^* at 0, 1, and in (0, 1), respectively (one-way ANOVA, $P = 0.008$, F test, 2 df). The mean number of distinct alleles for the loci with the maximum on either boundary is 10.24, smaller than the mean of 10.74 for those with the mean in the interior ($P = 0.03$, two-tailed t test).

6.3 Comparison of predicted H_{adm} to observed H_{adm}

We next compare predicted and observed H_{adm} values for the 678 loci for the admixed Mestizo population. In this approach, we used estimated locus-wise values of γ_1 in the Mestizo population together with locus-wise heterozygosities in the European and Native American populations to “predict” locus-wise Mestizo heterozygosities. The prediction is compared to the observed heterozygosity value to examine if our formulas for the heterozygosity of an admixed population are reflected in actual heterozygosities in an admixed group.

This computation follows a similar computation of Boca & Rosenberg (2011). The estimated admixture fractions, computed for the same data, are taken from Schroeder *et al.* (2009), who obtained them by a maximum likelihood approach (Millar, 1987) that does not take into account source population heterozygosities. Using these estimates, locus-wise heterozygosity estimates in the source populations, and locus-wise F_{ST} values calculated from allele frequencies in the source populations, we predicted H_{adm} with eq. 13.

The predicted and observed H_{adm} values for individual loci are compared in Figure 8. In general, the observation closely matches the prediction (Figure 8A), with the correlation between the observed and predicted H_{adm} values equaling 0.978 (Figure 8B). For 56% of the 678 loci, the prediction provides an underestimate of the observed value.

6.4 K = 3 source populations

We now consider the European, Native American, and African populations as the source populations, using γ_1 for the proportion of European ancestry, γ_2 for Native American ancestry, and γ_3 for African ancestry. We select 3 loci for illustration, choosing the same ones as in a similar analysis of Boca & Rosenberg (2011).

Plots for H_{adm} over the unit simplex for $(\gamma_1, \gamma_2, \gamma_3)$ appear in Figure 9. Each plot depicts H_{adm} as a function of $(\gamma_1, \gamma_2, \gamma_3)$ for a specific locus. The three panels show the possible locations of the maximal value of H_{adm}: in the first panel, the maximum lies in the interior of the simplex; in the second panel, at a vertex, and in the third panel, on an edge.

Considering all 678 loci, 15% have the maximum in the interior of the region, with $\gamma_1 > 0$, $\gamma_2 > 0$, and $\gamma_3 > 0$. The fractions with the maximum on an edge are 20% for a maximum on the edge with $\gamma_1 = 0$, 26% on the $\gamma_2 = 0$ edge, and 5% on the $\gamma_3 = 0$ edge. The fractions with the maximum at a vertex are 27% for the vertex (0, 0, 1), 2% for (0, 1, 0), and 5% for (1, 0, 0). The observations that (0, 0, 1) is the vertex with the largest number of maxima and (1, 0, 1) is the edge with the most maxima accord with the fact that African populations have generally higher heterozygosity than European populations, which in turn have higher heterozygosity than Native American populations (e.g. Pemberton *et al.*, 2013).

7 Discussion

We have considered the heterozygosity H_{adm} of an admixed population in terms of the admixture fractions of the source populations, and their heterozygosities and F_{ST} values at a locus. We have derived formulas describing H_{adm} in relation to these quantities (eqs. 8–10). In particular, we showed that H_{adm} is minimized over the set of possible admixture coefficient vectors when the admixed population consists of only one of the source populations (Proposition 3): an admixed population is at least as heterozygous as the least heterozygous source population. The maximal H_{adm} is more complicated, as its heterozygosity can either exceed or equal that of the most heterozygous source population (Proposition 4).

In studying the possible locations of the maximal H_{adm} for a fixed set of source populations, we found that the maximum can lie either in the interior of the region describing the allowable values of the admixture fractions—in which case all source populations contribute to the admixed population—or on the boundary, where one or more source populations does not contribute to the admixed population (Propositions 4–6, Figures 1–3). Simulations under a Dirichlet model for allele frequencies suggest that the maximal value of H_{adm} lies with increasing frequency in the interior of the allowable region as K and J increase (Figure 4).

For K = 2 source populations, we obtained further results, in particular showing that H_{adm} is a concave-down quadratic polynomial in the admixture coefficient γ_1 (eqs. 12–14). We obtained an analytical expression for the maximal heterozygosity of an admixture of a specific pair of source populations in terms of H_1, H_2, and the F_{ST} value between the two populations (Proposition 7). For fixed values of H_1, H_2, and the admixture fraction γ_1, H_{adm} is increasing as a function of F_{ST} (eq. 13, Figure 1). If $H_1 > H_2$, then the admixture fraction in source population 1 that maximizes H_{adm} is greater than $\frac{1}{2}$ (Proposition 7), meaning that at the maximal heterozygosity of the admixed population, the contribution of the more heterozygous source population exceeds that of the less heterozygous one. Interestingly, for the $K = 2$ case with $J = 2$ allelic types, if the location of the maximal value lies in (0, 1), then heterozygosity at the maximum is always $\frac{1}{2}$ (Proposition 11 and Figure 5): irrespective of the allele frequencies of the source populations, a linear combination (γ_1, γ_2) always exists so that the admixed population has frequencies of $\frac{1}{2}$ for both alleles.

For $K = 2$ source populations, a key result is that the maximal value of H_{adm} exceeds the larger of the two source population heterozygosities if and only if F_{ST} exceeds a bound defined by those heterozygosities (Corollary 10). Thus, with all other quantities equal, combining source populations that are more rather than less divergent is more likely to lead to an admixed population with heterozygosity exceeding those of the source populations. To obtain this result, it was important to utilize bounds on F_{ST} that constrain its values within a possibly narrow region of the unit interval, particularly for high-heterozygosity loci.

In multiallelic human data, we observed that for heterozygosities and F_{ST} values for putative sources of Mestizo populations, the maximal H_{adm} was more likely to be in the interior of the unit simplex or on an edge rather than at a vertex (Figures 7 and 9). This result indicates that the heterozygosities and F_{ST} values of these populations lie in a parameter range for which admixed populations are frequently more heterozygous than all their source populations. Examining heterozygosities of 267 worldwide populations in Table S20 of Pemberton *et al.* (2013), the 13 Mestizo populations all have heterozygosities exceeding all 29 Native American populations, and 4 have heterozygosities exceeding all 8 European populations. Interestingly, the 10 most heterozygous populations among the 267 include all five admixed populations involving a source population from the high-heterozygosity region of Africa: a Cape Mixed Ancestry group from South Africa, and four African-American populations. Thus, our mathematical results predicting that admixed populations often exceed all their source populations in heterozygosity are reflected in admixed human groups.

For $K = 2$, our model successfully predicted the heterozygosities in an admixed population from the source population heterozygosities, F_{ST} between the source populations, and the estimated admixture coefficient $\hat{\gamma}_1$ (Figure 8). Because H_{adm} is not necessarily monotonic in γ_1, however, the reverse problem of using H_{adm} to estimate γ_1 is problematic—unlike for the monotonically varying F_{ST} between an admixed population and one of the source populations (Boca & Rosenberg, 2011, Theorem 3). Given H_{adm}, source population heterozygosities H_1 and H_2, and F_{ST} between the source populations, two solutions to eq. 13 might exist for γ_1—so that although H_{adm} can be predicted from γ_1, it is inadvisable to proceed in the reverse direction to estimate γ_1 from the heterozygosity of an admixed population.

We note that we have assumed $J \geqslant K$: the number of alleles is greater than or equal to the number of populations. While the results are suited to biallelic markers for $K = 2$, they apply primarily to multiallelic markers. Thus, in addition to the microsatellite loci we have used, we can use them with haplotype loci, for which each distinct haplotype over a length of genome is regarded as a separate allele (Mehta *et al.*, 2019), and haplotype clusters, for which haplotypes are grouped into a fixed number of clusters and each individual is assigned a haplotype cluster membership at each site in the genome (San Lucas *et al.*, 2012).

Our approach has followed the study of F_{ST} and admixture from Boca & Rosenberg (2011), and it shares similar limitations. The model assumes source population allele frequencies are known rather than estimated, and it considers population-level rather than individual-level admixture. It relies on patterns of variation from a single time point and does not incorporate mechanistic admixture processes or a bottleneck at the founding of the admixed population; strong genetic drift since the onset of admixture might interfere with the linear combination assumption for allele frequencies in the admixed population. Despite these limitations, the observed H_{adm} values and those predicted under our model are correlated in the Mestizo example (Figure 8B), indicating that the model captures key features relevant to the relationship between admixture and heterozygosity. Thus, the empirical results suggest that assessing this relationship in the mathematical formulations we have presented can be useful for understanding the genetics of admixed populations.

Acknowledgments.

Rohan Mehta provided assistance with the SNP data. We thank two reviewers for comments on the manuscript. Support was provided by NIH grant HG005855 and NSF grant BCS-1515127.

Appendix 1. Proofs for arbitrary K: Theorem 5 and Corollary 6

For the proof of Theorem 5, we first show (i) that $P'P$ and A are both invertible under the conditions stated in the theorem, and that:

$$\frac{1}{\underline{1}'A^{-1}\underline{1}} = 1 - \frac{1}{\underline{1}'(P'P)^{-1}\underline{1}}.$$

We then (ii) use constrained optimization via Lagrange multipliers to obtain the maximum of $\underline{\gamma}'A\underline{\gamma}$ subject to $\underline{1}'\underline{\gamma} = 1$. This step consists of the first-derivative test to find a stationary point, coupled with the second-derivative test, in Lemma 12, to show that the stationary point defines a local maximum. Finally, we (iii) show that this means that the overall maximum is either at the local maximum $\underline{\gamma}^*$ as described in the statement of the theorem or on the boundary of the set $\{\underline{\gamma} : \underline{1}'\underline{\gamma} = 1 \text{ and } \underline{\gamma} \in \Delta^{K-1}\}$.

Proof of Theorem 5 (i) Because P is a $J \times K$ matrix with column rank K, $K \times K$ matrix $P'P$ is positive definite. As a positive definite matrix, $P'P$ is invertible and $(P'P)^{-1}$ is also positive definite (Graybill, 1976, pp. 21–22).

To show that $A = \underline{11}' - P'P$ is invertible, we use the Sherman-Morrison formula for the inverse of a rank-one update of an invertible matrix (Horn & Johnson, 2012, pp. 18–19).

This formula states that for an invertible square $n \times n$ matrix X and $n \times 1$ column vectors $\underline{y}$ and $\underline{z}$, $X + \underline{y}\underline{z}'$ is invertible if and only if $1 + \underline{z}' X^{-1}\underline{y} \neq 0$, with:

$$\left(X + \underline{yz'}\right)^{-1} = X^{-1} - \frac{X^{-1}\underline{yz'}X^{-1}}{1 + \underline{z}'X^{-1}\underline{y}}.$$

Because we assumed $\underline{1}'(P'P)^{-1}\underline{1} \neq 1$, the Sherman-Morrison formula applies with $-(P'P)$ in the role of X, and $K \times 1$ column vectors $\underline{1}$ in the role of $\underline{y}$ and $\underline{z}$. A has inverse:

$$A^{-1} = \frac{(P'P)^{-1}\underline{1}\underline{1}'(P'P)^{-1}}{\underline{1}'(P'P)^{-1}\underline{1} - 1} - (P'P)^{-1}. \tag{20}$$

Left-multiplying by $\underline{1}'$ and right-multiplying by $\underline{1}$, we obtain

$$\frac{1}{\underline{1}'A^{-1}\underline{1}} = 1 - \frac{1}{\underline{1}'(P'P)^{-1}\underline{1}}.$$

Because $(P'P)^{-1}$ is positive definite, $\underline{1}'(P'P)^{-1}\underline{1} > 0$ by definition, and because $\underline{1}'(P'P)^{-1}\underline{1} \neq 1$ by assumption, we conclude that $\frac{1}{\underline{1}'A^{-1}\underline{1}}$ is always defined.

(ii) To maximize $\underline{\gamma}'A\underline{\gamma}$ subject to $\underline{1}'\underline{\gamma} = 1$, we use Lagrange multipliers. Let $f(\underline{\gamma}) = \underline{\gamma}'A\underline{\gamma}$, and let $g(\underline{\gamma}) = \underline{1}'\underline{\gamma}$. The Lagrange function is defined as:

$$\Lambda(\underline{\gamma}, \lambda) = f(\underline{\gamma}) + \lambda[g(\underline{\gamma}) - 1].$$

Denoting by $\underline{0}$ is a column vector of length K, we solve a system of equations for $\underline{\gamma}$ and λ,

$$\left(\frac{\delta\Lambda(\underline{\gamma}, \lambda)}{\delta\underline{\gamma}}, \frac{\delta\Lambda(\underline{\gamma}, \lambda)}{\delta\lambda}\right) = (\underline{0}, 0). \tag{21}$$

Eq. 21 includes K equations $\delta\Lambda(\underline{\gamma}, \lambda)/\delta\gamma_k = 0$ for $1 \leqslant k \leqslant K$.

A is symmetric, so we have

$$\begin{aligned} \frac{\delta f(\underline{\gamma})}{\delta\underline{\gamma}} &= \frac{\delta\left(\underline{\gamma}'A\underline{\gamma}\right)}{\delta\underline{\gamma}} = (A + A')\underline{\gamma} = 2A\underline{\gamma} \\ \frac{\delta g(\underline{\gamma})}{\delta\underline{\gamma}} &= \underline{1}. \end{aligned}$$

For the derivatives of the Lagrange function, we have:

$$\left(\frac{\delta\Lambda(\underline{\gamma}, \lambda)}{\delta\underline{\gamma}}, \frac{\delta\Lambda(\underline{\gamma}, \lambda)}{\delta\lambda}\right) = \left(2A\underline{\gamma} + \lambda\underline{1}, \underline{1}'\underline{\gamma} - 1\right).$$

Setting the derivatives with respect to $\underline{\gamma}$ to $\underline{0}$ leads to:

$$(\underline{\gamma}, \lambda) = \left(-\frac{\lambda}{2}A^{-1}\underline{1}, -\frac{2}{\underline{1}'A^{-1}\underline{1}}\right).$$

Hence, the solution for $\underline{\gamma}$ is:

$$\underline{\gamma}^* = \frac{A^{-1}\underline{1}}{\underline{1}'A^{-1}\underline{1}}.$$

Because $\underline{\gamma}'A\underline{\gamma}$ is a differentiable function of $\underline{\gamma}$, its maximum on Δ^{K-1} can occur either on the boundary or at a critical point. The following lemma shows that the critical point $\underline{\gamma}^* = \frac{A^{-1}\underline{1}}{\underline{1}'A^{-1}\underline{1}}$ is a local maximum.

Lemma 12. The critical point $\underline{\gamma}^* = \frac{A^{-1}\underline{1}}{\underline{1}'A^{-1}\underline{1}}$ is a local maximum of H_{adm} seen as a function of $\underline{\gamma}$ on Δ^{K-1}, under the conditions stated in Theorem 5.

Proof. To show that $\underline{\gamma}^*$ is a local maximum, we use the second-derivative test for constrained optimization (e.g. Magnus & Neudecker, 2007, p. 155). This test considers the bordered Hessian matrix, representing the matrix of second derivatives of the Lagrange function Λ with respect to λ and the components of $\underline{\gamma}$:

$$F = \begin{pmatrix} \frac{\delta^2\Lambda}{\delta\lambda^2} & \left(\frac{\delta^2\Lambda}{\delta\underline{\gamma}\delta\lambda}\right)' \\ \frac{\delta^2\Lambda}{\delta\underline{\gamma}\delta\lambda} & \frac{\delta^2\Lambda}{\delta\underline{\gamma}^2} \end{pmatrix} = \begin{pmatrix} 0 & \left(\frac{\delta g}{\delta\underline{\gamma}}\right)' \\ \frac{\delta g}{\delta\underline{\gamma}} & \frac{\delta^2\Lambda}{\delta\underline{\gamma}^2} \end{pmatrix} = \begin{pmatrix} 0 & \underline{1}' \\ \underline{1} & 2A \end{pmatrix}.$$

We must consider the principal minors—determinants of matrices in the upper-left corner—of F. We denote the upper-left corner matrix of size $r \times r$ of F by F_r, for $r = 2, 3, \ldots, K$. The principal minors are the $\det(F_r)$. Using the definition of A from eq. 11, we obtain

$$F_r = \begin{pmatrix} 0 & 1 & 1 & \ldots & 1 \\ 1 & 2H_1 & 2C_{12} & \ldots & 2C_{1r} \\ 1 & 2C_{12} & 2H_2 & \ldots & 2C_{2r} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2C_{1r} & 2C_{2r} & \ldots & 2H_r \end{pmatrix}$$

A sufficient condition for the critical point to be a local maximum is for $(-1)^r \det(F_r) > 0$ for each r (Magnus & Neudecker, 2007, p. 155). We now show that this condition is satisfied.

Using the fact that multiplying a row or column of a matrix by a scalar multiplies the determinant by that scalar, we multiply rows 2 through $r + 1$ by -1 and get

$$\det(F_r) = \det\begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 2H_1 & 2C_{12} & \dots & 2C_{1r} \\ 1 & 2C_{12} & 2H_2 & \dots & 2C_{2r} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2C_{1r} & 2C_{2r} & \dots & 2H_r \end{pmatrix} = (-1)^r \det\begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ -1 & -2H_1 & -2C_{12} & \dots & -2C_{1r} \\ -1 & -2C_{12} & -2H_2 & \dots & -2C_{2r} \\ \vdots & \vdots & \vdots & \vdots & \dots \\ -1 & -2C_{1r} & -2C_{2r} & \dots & -2H_r \end{pmatrix}.$$

Using the fact that adding a multiple of a row or column to another row does not change the determinant, we add −2 times the first column to each of the remaining columns. We also multiply the first column by −1. We then have

$$(-1)^r \det(F_r) = (-1)^{2r+1} \det\begin{pmatrix} 0 & \underline{1_r}' \\ \underline{1}_r & 2M_r \end{pmatrix} = -\det\begin{pmatrix} 0 & \underline{1_r}' \\ \underline{1}_r & 2M_r \end{pmatrix}, \qquad (22)$$

where M_r is the $r \times r$ matrix consisting of the upper-left corner of matrix $P'P$, and $\underline{1}_r$ is the column vector of length r consisting of 1s.

We now apply a result for the determinant of partitioned matrices (Graybill, 1976, pp. 19–20). If W is invertible, then

$$\det\begin{pmatrix} X & Y \\ Z & W \end{pmatrix} = \det(W)\det\left(X - YW^{-1}Z\right).$$

Applying this result to eq. 22, we obtain

$$\begin{aligned} (-1)^r \det(F_r) &= -\det(2M_r)\det\left(-\underline{1_r}'(2M_r)^{-1}\underline{1}_r\right) \\ &= -\left[2^r \det(M_r)\right]\left[\left(-\frac{1}{2}\right)\underline{1_r}'M_r^{-1}\underline{1}_r\right] \\ &= 2^{r-1}\det(M_r)\left(\underline{1_r}'M_r^{-1}\underline{1}_r\right). \end{aligned}$$

Because $P'P$ is positive definite, M_r is also positive definite. To demonstrate this result, note that because $\underline{x}'P'P\underline{x} > 0$ for each nonzero column vector $\underline{x}$, $\underline{x}'P'P\underline{x} > 0$ for each nonzero $\underline{x}$ with $x_k = 0$ for $k > r$. Because M_r is positive definite, $\det(M_r) > 0$ and M_r^{-1} is also positive definite, leading to $\underline{1_r}'M_r^{-1}\underline{1}_r > 0$. We conclude

$$(-1)^r \det(F_r) > 0,$$

so that the critical point is the location of a local maximum. □

Concluding the proof of Theorem 5. Returning to part (iii) of the proof, following Lemma 12, if $\underline{\gamma}^* = \frac{A^{-1}\underline{1}}{\underline{1}'A^{-1}\underline{1}}$ is interior to the simplex Δ^{K-1}, then H_{adm} is maximal at $\underline{\gamma} = \underline{\gamma}*$, with maximum $H(\underline{\gamma}) = \frac{1}{\underline{1}'A^{-1}\underline{1}}$. This value is the reciprocal of the sum of the elements of A^{-1}. If $\underline{\gamma}*$ is not interior to Δ^{K-1}, then the maximum lies on the boundary of Δ^{K-1}.

Finally, we note that $\gamma^* = \frac{(P'P)^{-1}\underline{1}}{\underline{1}'(P'P)^{-1}\underline{1}}$ by using eq. 20. □

Proof of Corollary 6. In Theorem 5, the maximum of H_{adm} occurs either in the interior of the simplex Δ^{K-1} or on its boundary, $\{\gamma : \underline{1}'\gamma = 1 \text{ and } \gamma \in \Delta^{K-1}\}$.

The boundary of the simplex is the union of K faces, which are themselves $(K-2)$-simplices. If the maximum lies on the boundary of Δ^{K-1}, then without loss of generality, we can permute the labels of the source populations so that $\gamma_K = 0$.

We drop column K from matrix P and apply Theorem 5 with this new $J \times (K-1)$ matrix, $P_{\{1,\ldots,K-1\}}$, which has rank $K-1$. By assumption, $\underline{1}'\left(P'_{\{1,\ldots,K-1\}}P_{\{1,\ldots,K-1\}}\right)^{-1}\underline{1} \neq 1$.

We then apply Theorem 5 to $P_{\{1,\ldots,K-1\}}$. The maximum of H_{adm} occurs either at the point $\gamma_{\mathcal{S}}$, where $\mathcal{S} = \{1, 2, \ldots, K-1\}$, or on the boundary of the set $\{\gamma : \underline{1}'\gamma = 1 \text{ and } \gamma \in \Delta^{K-2}\}$.

We repeat this method of descent, decrementing the dimension (and permuting population labels without loss of generality) until we reach the case of only two source populations. A final application of Theorem 5 then finds that H_{adm} is maximized either interior to the 1-simplex—the line connecting vertices (1, 0) and (0, 1)—or at one of these vertices. □

Appendix 2. Proofs for K = 2: Propositions 7–11

Proof of Proposition 7. We maximize the quadratic polynomial in eqs. 12–14 over $\gamma \in [0, 1]$. The maximum occurs at the unique critical point or on the boundary of the interval.

Setting the derivative of eq. 14 with respect to γ_1 to 0, we find that the critical point is

$$(\gamma_1^*, H_{\mathrm{adm}}) = \left(\frac{C_{12} - H_2}{2(C_{12} - H_S)}, \frac{C_{12}^2 - H_1H_2}{2(C_{12} - H_S)}\right). \tag{23}$$

Because the leading coefficient of eq. 14 is negative for $\underline{p_1} \neq \underline{p_2}$, the critical point is a maximum. Hence, if $(C_{12} - H_2)/[2(C_{12} - H_S)] \in (0, 1)$, then the maximum of H_{adm} on the interval [0, 1] lies at $\gamma_1 = (C_{12} - H_2)/[2(C_{12} - H_S)]$. Otherwise, the maximum lies either at $\gamma_1 = 0$, in which case it equals H_2, or at $\gamma_1 = 1$, in which case it equals H_1.

The conditions describing the location of the maximum can be written in terms of H_1, H_2, and C_{12}. Because the denominator of γ_1^* in eq. 23 is always positive for $\underline{p_1} \neq \underline{p_2}$ (Section 4), $\gamma_1^* \in (0, 1)$ becomes equivalent to $C_{12} > H_1$ and $C_{12} > H_2$, the former inequality arising from the condition $\gamma_1^* < 1$ and the latter from the condition $\gamma_1^* > 0$.

If the requirement $C_{12} > H_1$ and $C_{12} > H_2$ for $\gamma_1^* \in (0, 1)$ fails, then the maximum occurs on the boundary of the unit interval. We have $H_{\mathrm{adm}}(0) = H_2$ and $H_{\mathrm{adm}}(1) = H_1$. Thus, the maximum lies at $\gamma_1 = 0$ if $H_2 > H_1$ and at $\gamma_1 = 1$ if $H_1 > H_2$.

If $C_{12} > H_1$ and $C_{12} > H_2$ do not both hold, then one of them must hold, as we showed in Section 4 that $2C_{12} > H_1 + H_2$. Combining the fact that either $C_{12} > H_1$ or $C_{12} > H_2$ holds with the observation that $H_2 > H_1$ leads to a maximum at $\gamma_1 = 0$ and $H_1 > H_2$ leads to a maximum at $\gamma_1 = 1$, we complete the characterization of the three cases.

Note that the three cases in the statement of the proposition capture all possible values of (H_1, H_2, C_{12}). By the Cauchy-Schwarz inequality, $(1 - C_{12})^2 \leqslant (1 - H_1)(1 - H_2)$, with equality requiring $\underline{p_1} = \underline{p_2}$. Hence, with $\underline{p_1} \neq \underline{p_2}$ assumed, either $1 - C_{12} < 1 - H_1$ and $1 - C_{12} \geqslant 1 - H_2$ (case (ii)), $1 - C_{12} < 1 - H_2$ and $1 - C_{12} \geqslant 1 - H_1$ (case (iii)), or both $1 - C_{12} < 1 - H_1$ and $1 - C_{12} < 1 - H_2$ (case (i)).

Alternative expressions in terms of H_1, H_2, and F_{12} can be derived by noting that $H_S = \frac{1}{2}(H_1 + H_2)$, $H_1 H_2 = H_S^2 - [(H_1 - H_2)/2]^2$ and $C_{12} = H_S(1 + F_{12})/(1 - F_{12})$, the latter simply restating eq. 4 (recalling $C_{12} = 1$ for $F_{12} = 1$). Thus, we have

$$\gamma_1^* = \frac{C_{12} - H_2}{2(C_{12} - H_S)} = \frac{1}{2} + \frac{H_1 - H_2}{4\frac{F_{12}}{1 - F_{12}}(H_1 + H_2)} \tag{24}$$

$$H_{\mathrm{adm}}(\gamma^*) = \frac{C_{12}^2 - H_1 H_2}{2(C_{12} - H_S)} = \frac{H_1 + H_2}{2(1 - F_{12})} + \frac{(H_1 - H_2)^2}{8\frac{F_{12}}{1 - F_{12}}(H_1 + H_2)}. \tag{25}$$

Another formulation uses the heterozygosity of a population formed by equal admixture of populations 1 and 2, or H_T. Because $F_{12} = 1 - H_S/H_T$ by eq. 1, $F_{12}/(1 - F_{12}) = (H_T - H_S)/H_S$. Using this relationship in eqs. 24 and 25,

$$\gamma_1^* = \frac{1}{2} + \frac{H_1 - H_2}{8(H_T - H_S)}$$

$$H_{\mathrm{adm}}(\gamma^*) = H_T + \frac{(H_1 - H_2)^2}{16(H_T - H_S)}.$$

□

Proof of Corollary 8. Suppose $H_1 \geqslant H_2$. If case (i) from Proposition 7 applies, then because $H_T > H_S$, $\gamma_1^* \geqslant \frac{1}{2}$. Case (ii) cannot apply because $H_1 < C_{12}$, $H_2 \geqslant C_{12}$, and $H_1 \geqslant H_2$ cannot hold simultaneously. In case (iii), $\gamma_1^* = 1 \geqslant \frac{1}{2}$. For the reverse direction, if $H_1 < H_2$ and case (i) or case (ii) applies, then $\gamma_1^* < \frac{1}{2}$. Case (iii) cannot apply because $H_1 \geqslant C_{12}$, $H_2 < C_{12}$, and $H_1 < H_2$ cannot hold simultaneously. □

Proof of Corollary 9. First, we see that $H_{\mathrm{adm}}(\gamma_1^*) \geqslant H_T$ in case (i) of Proposition 7. In case (ii), $H_2 > H_T = (H_1 + H_2 + 2C_{12})/4$ because $H_2 > H_1$ and $H_2 \geqslant C_{12}$. In case (iii), $H_1 > H_T$

because $H_1 > H_2$ and $H_1 \geqslant C_{12}$. Note that if $H_1 = H_2$, then case (i) applies, producing $H_{\text{adm}}(\gamma_1^*) = H_T$. □

Proof of Corollary 10. We restate the condition $0 < (C_{12} - H_2)/[2(C_{12} - H_S)] < 1$ as

$$0 < \frac{1}{2} + \frac{\left(\frac{H_1 - H_2}{2}\right)}{2\frac{F_{12}}{1 - F_{12}}(H_1 + H_2)} < 1.$$

Subtracting $\frac{1}{2}$ from both sides and multiplying by 2, an equivalent condition is

$$-1 < \frac{(H_1 - H_2)}{2\frac{F_{12}}{1 - F_{12}}(H_1 + H_2)} < 1,$$

or, equivalently, $|H_1 - H_2|/\left[2\frac{F_{12}}{1 - F_{12}}(H_1 + H_2)\right] < 1$. We rearrange this last expression to obtain the desired result. □

Proof of Proposition 11. We apply Proposition 7 with $J = 2$. Substituting $p_{12} = 1 - p_{11}$ and $p_{22} = 1 - p_{21}$ in eqs. 15 and 16, we obtain $C_{12} - H_2 = (p_{11} - p_{21})(1 - 2p_{21})$, $C_{12} - H_1 = (p_{21} - p_{11})(1 - 2p_{11})$, $C_{12} - H_S = (p_{11} - p_{21})^2$, and $C_{12}^2 - H_1H_2 = (p_{11} - p_{21})^2$. Thus, because $p_{11} = p_{21}$ is not permitted, the quantities in eqs. 15 and 16 reduce to those of eqs. 18 and 19, respectively.

To complete the application of Proposition 7 to $K = 2$, note that case (i) of Proposition 7 occurs when $(p_{11} - p_{21})(1 - 2p_{21}) > 0$ and $(p_{21} - p_{11})(1 - 2p_{11}) > 0$. The first of this pair of inequalities requires both $p_{11} - p_{21} > 0$ and $1 - 2p_{21} > 0$, so that $p_{11} > p_{21}$ and $\frac{1}{2} > p_{21}$, or both $p_{11} - p_{21} < 0$ and $1 - 2p_{21} < 0$, so that $p_{11} < p_{21}$ and $\frac{1}{2} < p_{21}$. The second inequality requires both $p_{21} - p_{11} > 0$ and $1 - 2p_{11} > 0$, so that $p_{21} > p_{11}$ and $\frac{1}{2} > p_{11}$, or both $p_{21} - p_{11} < 0$ and $1 - 2p_{11} < 0$, so that $p_{21} < p_{11}$ and $\frac{1}{2} < p_{11}$. Thus, the conditions of case (i) of Proposition 7 obtain if and only if $p_{11} > \frac{1}{2} > p_{21}$ or $p_{21} > \frac{1}{2} > p_{11}$.

Similarly, using the expressions for H_1, H_2, and C_{12} when $K = 2$, the conditions of case (ii) of Proposition 7 are equivalent to $\frac{1}{2} \geqslant p_{21} > p_{11}$ or $p_{11} > p_{21} \geqslant \frac{1}{2}$. The conditions of case (iii) are equivalent to $\frac{1}{2} \geqslant p_{11} > p_{21}$ or $p_{21} > p_{11} \geqslant \frac{1}{2}$. □

Appendix 3: Dirichlet model for allele frequencies

We first provide results concerning H_{adm} in the case that the K source populations have independently and identically distributed (IID) allele frequency vectors. Next, we specify these IID vectors to be Dirichlet distributions.

IID allele frequency vectors

We begin by examining the expected values of H_k and H_{adm}.

Proposition 13. Suppose the allele frequency vectors $\underline{p_k}$ are independently and identically distributed for $1 \leqslant k \leqslant K$. Then $\mathbb{E}[H_{\mathrm{adm}}] = \mathbb{E}[H_1] + \left(1 - \sum_{k=1}^{K} \gamma_k^2\right)\left(\sum_{j=1}^{J} \mathrm{Var}\left[p_{1j}\right]\right)$.

Proof. We use eq. 8:

$$\mathbb{E}[H_{\mathrm{adm}}] = 1 - \sum_{k=1}^{K} \gamma_k^2 \left(\sum_{j=1}^{J} \mathbb{E}\left[p_{kj}^2\right]\right) - 2 \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \gamma_k \gamma_\ell \left(\sum_{j=1}^{J} \mathbb{E}\left[p_{kj} p_{lj}\right]\right).$$

Using the IID assumption and simplifying by noting that $1 = (\sum_{k=1}^{K} \gamma_k)^2 = (\sum_{k=1}^{K} \gamma_k^2) + (2 \sum_{k=1}^{K-1} \sum_{\ell=k+1}^{K} \gamma_k \gamma_\ell)$, we have

$$\begin{aligned} \mathbb{E}[H_{\mathrm{adm}}] &= 1 - \left(\sum_{k=1}^{K} \gamma_k^2\right)\left(\sum_{j=1}^{J} \mathbb{E}\left[p_{1j}^2\right]\right) - 2 \sum_{k=1}^{K-1} \sum_{\ell=k+1}^{K} \gamma_k \gamma_\ell \left[\sum_{j=1}^{J} \left(\mathbb{E}\left[p_{1j}\right]\right)^2\right] \\ &= 1 - \sum_{j=1}^{J} \mathbb{E}\left[p_{1j}^2\right] + \sum_{j=1}^{J} \mathbb{E}\left[p_{1j}^2\right]\left(1 - \sum_{k=1}^{K} \gamma_k^2\right) - \sum_{j=1}^{J} \left(\mathbb{E}\left[p_{1j}\right]\right)^2 \left(1 - \sum_{k=1}^{K} \gamma_k^2\right), \end{aligned}$$

from which the result follows. □

An immediate corollary of Proposition 13 is that H_{adm} has expectation greater than or equal to the expectation of the heterozygosity of each of the source populations.

Corollary 14. Suppose the allele frequency vectors $\underline{p_k}$ are independently and identically distributed for $1 \leqslant k \leqslant K$. Then $\mathbb{E}[H_{\mathrm{adm}}] \geqslant \mathbb{E}[H_k]$.

A second corollary results from the Cauchy-Schwarz inequality, by which $\sum_{k=1}^{K} \gamma_k^2 \geqslant \frac{1}{K}$, with equality if and only if $(\gamma_1, \gamma_2, \ldots, \gamma_K) = \left(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}\right)$.

Corollary 15. Suppose the allele frequency vectors $\underline{p_k}$ are independently and identically distributed for $1 \leqslant k \leqslant K$. Considering all admixture vectors $\underline{\gamma} \in \Delta^{K-1}$, $\mathbb{E}[H_{\mathrm{adm}}]$ is maximized at $\underline{\gamma} = \left(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}\right)$, and has maximal value $\mathbb{E}[H_1] + \left(1 - \frac{1}{K}\right)\sum_{j=1}^{J} \mathrm{Var}\left[p_{1j}\right]$.

IID allele frequency vectors from a symmetric Dirichlet distribution

We now further assume that the independently and identically distributed allele frequency vectors follow a symmetric multivariate Dirichlet distribution. This distribution is frequently used for allele frequency distributions (Balding & Nichols, 1995; Pritchard *et al.*, 2000; Huelsenbeck & Andolfatto, 2007), and it is a natural probability distribution to assume for allelic types with the same marginal distributions.

The J-dimensional Dirichlet-(α_1, α_2, …, α_J) distribution is defined over the open unit ($J-$ 1)-simplex Δ^{J-1} and has concentration parameters $\alpha_j > 0$. The means and variances for the individual allele frequencies are (Lange, 1997; Kotz *et al.*, 2000, chapter 49):

$$\begin{aligned} \mathbb{E}\left[p_{kj}\right] &= \frac{\alpha_j}{J\bar{\alpha}} \\ \mathrm{Var}\left[p_{kj}\right] &= \frac{\alpha_j\left(J\bar{\alpha}-\alpha_j\right)}{J^2\bar{\alpha}^2(J\bar{\alpha}+1)}, \end{aligned}$$

where $\bar{\alpha} = \frac{1}{J}\sum_{j=1}^{J}\alpha_j$.

The symmetric Dirichlet distribution assumes $\alpha_1 = \alpha_2 = \ldots = \alpha_J = \bar{\alpha}$, leading to:

$$\begin{aligned} \mathbb{E}\left[p_{kj}\right] &= \frac{1}{J} \\ \mathrm{Var}\left[p_{kj}\right] &= \frac{J-1}{J^2(J\bar{\alpha}+1)}. \end{aligned}$$

Making these substitutions in Proposition 13, we obtain the expectation of H_{adm} under the assumption that the allele frequency vectors follow independent Dirichlet distributions.

Corollary 16. Suppose the allele frequency vectors $\underline{p_k}$ are independently and identically distributed for $1 \leqslant k \leqslant K$, all with symmetric multivariate Dirichlet distributions with concentration parameter $\bar{\alpha}$. Then

$$\begin{aligned} \mathbb{E}[H_k] &= \left(1-\frac{1}{J}\right)\left(1-\frac{1}{J\bar{\alpha}+1}\right), \\ \mathbb{E}[H_{\mathrm{adm}}] &= \left(1-\frac{1}{J}\right)\left(1-\frac{1}{J\bar{\alpha}+1}\sum_{k=1}^{K}\gamma_k^2\right). \end{aligned}$$

This corollary implies that both $\mathbb{E}[H_k]$ and $\mathbb{E}[H_{\mathrm{adm}}]$ are increasing functions of J and $\bar{\alpha}$.

The next proposition considers the special case of $K = 2$ and $J = 2$, further specifying a uniform distribution for γ_1.

Proposition 17. Consider $K = 2$ and $J = 2$. Suppose that the values of p_{11} and p_{21} are independently chosen from a uniform-[0,1] distribution. Suppose also that γ_1 is also chosen from a uniform-[0, 1] distribution. Then $\mathbb{P}[H_{\mathrm{adm}}(\gamma_1) > \max\{H_1, H_2\}] = 1 - \log 2 \approx 0.307$.

Proof. Using Proposition 11, we identify the regions of the unit square for (p_{11}, p_{21}) in which $\max_{\gamma_1 \in (0,1)} H_{\mathrm{adm}}(\gamma_1) > \max\{H_1, H_2\}$. These regions are

$\left\{(p_{11}, p_{21}) \mid \frac{1}{2} < p_{11} < 1, 0 < p_{21} < \frac{1}{2}\right\}$ and $\left\{(p_{11}, p_{21}) \mid 0 < p_{11} < \frac{1}{2}, \frac{1}{2} < p_{21} < 1\right\}$.

Within those regions, we must determine the portion of the unit interval for γ_1 in which $H_{\mathrm{adm}}(\gamma_1) > \max\{H_1, H_2\}$. $H_{\mathrm{adm}}(\gamma_1)$ is a quadratic function of γ_1. We ignore the set of zero volume with $H_1 = H_2$. In the regions for (p_{11}, p_{21}) in which

$\max_{\gamma_1 \in (0,1)} H_{\mathrm{adm}}(\gamma_1) > \max\{H_1, H_2\}$ and $H_2 > H_1$, the interval for γ_1 in which $H_{\mathrm{adm}}(\gamma_1) > H_1$ is $\left(0, \frac{1 - 2p_{21}}{p_{11} - p_{21}}\right)$. In the regions for (p_{11}, p_{21}) in which $max_{\gamma_1 \in (0,1)} H_{\mathrm{adm}}(\gamma_1) > \max\{H_1, H_2\}$ and $H_1 > H_2$, the interval for γ_1 in which $H_{\mathrm{adm}}(\gamma_1) > H_1$ is $\left(\frac{p_{21} - 1 + p_{11}}{p_{21} - p_{11}}, 1\right)$.

The desired probability is the volume within the unit cube for $(p_{11}, p_{21}, \gamma_1)$ of the regions in which $H_{\mathrm{adm}}(\gamma_1) > \max\{H_1, H_2\}$. The volume is

$$\int_{1/2}^{1}\int_{1 - p_{11}}^{1/2}\int_{0}^{\frac{1 - 2p_{21}}{p_{11} - p_{21}}} 1 \mathrm{d}\gamma_1 \mathrm{d}p_{21} \mathrm{d}p_{11} + \int_{1/2}^{1}\int_{0}^{1 - p_{11}}\int_{\frac{p_{21} - 1 + p_{11}}{p_{21} - p_{11}}}^{1} 1 \mathrm{d}\gamma_1 \mathrm{d}p_{21} \mathrm{d}p_{11}$$
$$\int_{0}^{1/2}\int_{1 - p_{11}}^{1}\int_{\frac{p_{21} - 1 + p_{11}}{p_{21} - p_{11}}}^{1} 1 \mathrm{d}\gamma_1 \mathrm{d}p_{21} \mathrm{d}p_{11} + \int_{0}^{1/2}\int_{1/2}^{1 - p_{11}}\int_{0}^{\frac{1 - 2p_{21}}{p_{11} - p_{21}}} 1 \mathrm{d}\gamma_1 \mathrm{d}p_{21} \mathrm{d}p_{11}$$
$$= 4\frac{1 - \log 2}{4}.$$

□

References

Alcala N and Rosenberg NA 2017 Mathematical constraints on F_{ST}: biallelic markers in arbitrarily many populations, Genetics 206, 1581–1600. [PubMed: 28476869]

Alcala N and Rosenberg NA 2019 G'_{ST}, Jost's D, and F_{ST} are similarly constrained by allele frequencies: a mathematical, simulation, and empirical study, Mol. Ecol 28, 1624–1636. [PubMed: 30589985]

Balding DJ and Nichols RA 1995 A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity, Genetics 96, 3–12.

Boca SM and Rosenberg NA 2011 Mathematical properties of F_{st} between admixed populations and their parental source populations, Theor. Pop. Biol 80, 208–216. [PubMed: 21640742]

Buerkle CA and Lexer C 2008 Admixture as the basis for genetic mapping, Trends Ecol. Evol 23, 686–694. [PubMed: 18845358]

Chakraborty R 1986 Gene admixture in human populations: Models and predictions, Yrbk. Phys. Anthropol 29, 1–43.

Edge MD and Rosenberg NA 2014 Upper bounds on F_{ST} in terms of the frequency of the most frequent allele and total homozygosity: the case of a specified number of alleles, Theor. Pop. Biol 97, 20–34. [PubMed: 25132646]

Gravel S 2012 Population genetics models of local ancestry, Genetics 191, 607–619. [PubMed: 22491189]

Graybill FA 1976 "Theory and application of the linear model", Duxbury, Pacific Grove, CA.

Hedrick PW 1999 Perspective: highly variable loci and their interpretation in evolution and conservation, Evolution 53, 313–318. [PubMed: 28565409]

Hedrick PW 2005 A standardized genetic differentiation measure, Evolution 59, 1633–1638. [PubMed: 16329237]

Horn RA and Johnson CR 2012 "Matrix analysis", Cambridge University Press, New York, NY.

Huelsenbeck JP and Andolfatto P 2007 Inference of population structure under a Dirichlet process model, Genetics 175, 1787–1802. [PubMed: 17237522]

Jakobsson M, Edge MD, and Rosenberg NA 2013 The relationship between F_{ST} and the frequency of the most frequent allele, Genetics 193, 515–528. [PubMed: 23172852]

Kotz S, Balakrishnan N, and Johnson NL 2000 “Continuous Multivariate Distributions. Volume 1: Models and Applications”, Wiley, New York.

Lange K 1997 “Mathematical and Statistical Methods for Genetic Analysis”, Springer, New York.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, and Myers RM 2008 Worldwide human relationships inferred from genome-wide patterns of variation, Science 319, 1100–1104. [PubMed: 18292342]

Long JC 1991 The genetic structure of admixed populations, Genetics 127, 417–428. [PubMed: 2004712]

Long JC and Kittles RA 2003 Human genetic diversity and the nonexistence of biological races, Hum. Biol. 75, 449–471. [PubMed: 14655871]

Magnus JR and Neudecker H 2007 “Matrix differential calculus with applications in statistics and econometrics”, John Wiley & Sons, Chichester, UK, 3rd edition.

Maruki T, Kumar S, and Kim Y 2012 Purifying selection modulates the estimates of population differentiation and confounds genome-wide comparisons across single-nucleotide polymorphisms, Mol. Biol. Evol. 29, 3617–3623. [PubMed: 22826460]

Mehta RS, Feder AF, Boca SM, and Rosenberg NA 2019 The relationship between haplotype-based F_{ST} and haplotype length, Genetics 213, 281–295. [PubMed: 31285255]

Millar RB 1987 Maximum likelihood estimation of mixed stock fishery composition, Can. J. Fish. Aquat. Sci 44, 583–590.

Mooney JA, Huber CD, Service S, Sul JH, Marsden CD, Zhang Z, Sabatti C, Ruiz-Linares A, Bedoya G, Costa Rica/Colombia Consortium for Genetic Investigation of Bipolar Endophenotypes, Freimer N, and Lohmueller KE 2018 Understanding the hidden complexity of Latin American population isolates, Am. J. Hum. Genet 103, 707–726. [PubMed: 30401458]

Nagylaki T 1998 Fixation indices in subdivided populations, Genetics 148, 1325–1332. [PubMed: 9539445]

Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, and Li JZ 2012 Genomic patterns of homozygosity in worldwide human populations, Am. J. Hum. Genet 91, 275–292. [PubMed: 22883143]

Pemberton TJ, DeGiorgio M, and Rosenberg NA 2013 Population structure in a comprehensive genomic data set on human microsatellite variation, G3: Genes, Genomes, Genetics 3, 891–907. [PubMed: 23550135]

Pritchard JK, Stephens M, and Donnelly P 2000 Inference of population structure using multilocus genotype data, Genetics 155, 945–959. [PubMed: 10835412]

Reddy SB and Rosenberg NA 2012 Refining the relationship between homozygosity and the frequency of the most frequent allele, J. Math. Biol 64, 87–108. [PubMed: 21305294]

Risch N, Choudhry S, Via M, Basu A, Sebro R, Eng C, Beckman K, Thyne S, Chapela R, Rodriguez-Santana JR, Rodriguez-Cintron W, Avila PC, Ziv E, and Burchard EG 2009 Ancestry-related assortative mating in Latino populations, Genome Biol. 10, R132. [PubMed: 19930545]

Rosenberg NA and Calabrese PP 2004 Polyploid and multilocus extensions of the Wahlund inequality, Theor. Pop. Biol 66, 381–391. [PubMed: 15560915]

Rosenberg NA, Li LM, Ward R, and Pritchard JK 2003 Informativeness of genetic markers for inference of ancestry, Am. J. Hum. Genet 73, 1402–1422. [PubMed: 14631557]

San Lucas FA, Rosenberg NA, and Scheet P 2012 Haploscope: a tool for the graphical display of haplotype structure in populations, Genet. Epidemiol 35, 17–21.

Schroeder KB, Jakobsson M, Crawford MH, Schurr TG, Boca SM, Conrad DF, Tito RY, Osipova LP, Tarskaia LA, Zhadanov SI, Wall JD, Pritchard JK, Malhi RS, Smith DG, and Rosenberg NA 2009 Haplotypic background of a private allele at high frequency in the Americas, Mol. Biol. Evol 26, 995–1016. [PubMed: 19221006]

Verdu P and Rosenberg NA 2011 A general mechanistic model for admixture histories of hybrid populations, Genetics 189, 1413–1426. [PubMed: 21968194]

Wang S, Ray N, Rojas W, Parra MV, Bedoya G, Gallo C, Poletti G, Mazzotti G, Hill K, Hurtado AM, Camrena B, Nicolini H, Klitz W, Barrantes R, Molina JA, Freimer NB, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Dipierri JE, Alfaro EL, Bailliet G, Bianchi NO, Llop E, Rothhammer

F, Excoffier L, and Ruiz-Linares A 2008 Geographic patterns of genome admixture in Latin American Mestizos, PLoS Genet. 4, e1000037. [PubMed: 18369456]

Zhu X, Tang H, and Risch N 2008 Admixture mapping and the role of population structure for localizing disease genes, Adv. Genet 60, 547–569. [PubMed: 18358332]

Zou JY, Park DS, Burchard EG, Torgerson DG, Pino-Yanes M, Song YS, Sankararaman S, Halperin E, and Zaitlen N 2015 Genetic and socioeconomic study of mate choice in Latinos reveals novel assortment patterns, Proc. Natl. Acad. Sci. USA 112, 13621–13626. [PubMed: 26483472]

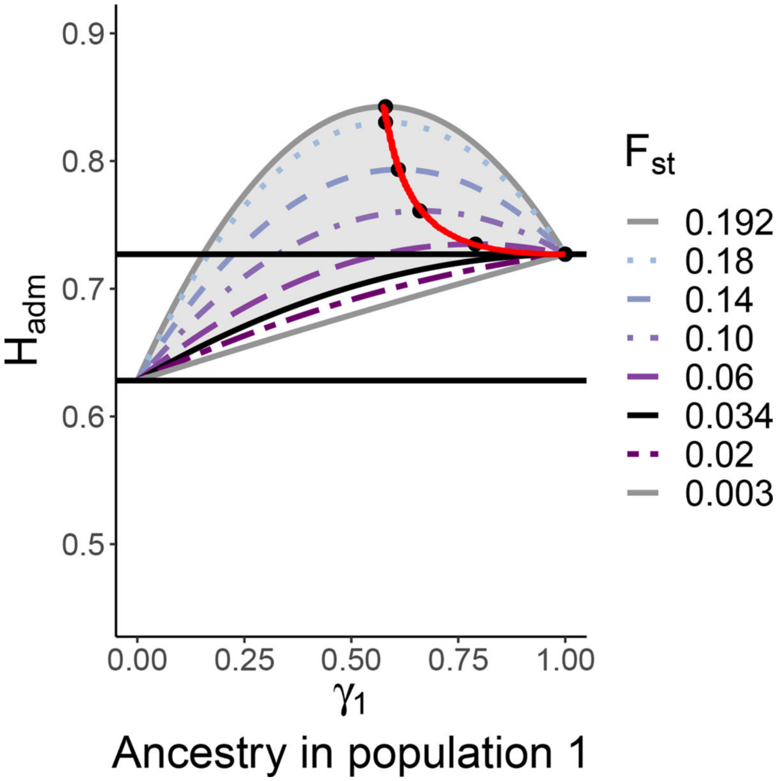

Figure 1:

H_{adm} versus γ_1 for fixed values of H_1 and H_2. We choose $H_1 = 0.727$ and $H_2 = 0.628$; the horizontal lines represent $H_{adm} = H_1$ and $H_{adm} = H_2$. Eq. 13 is plotted for multiple values of F_{12}, considering the allowable range of F_{12} values in [0.003,0.192] as specified by eq. 5. The red curve, which plots (γ_1, H_{adm}) in terms of H_1, H_2, and F_{12} in the form of eqs. 24 and 25, indicates the maxima of H_{adm} as F_{12} varies, with black dots specifying the maxima for the specific plotted values of F_{12}. The shaded region corresponds to the region where $\gamma_1^* \in (0, 1)$, as specified by Corollary 10; the value $F_{12} \approx 0.034$ gives the boundary of this region. The values chosen for H_1 and H_2 are, respectively, the mean heterozygosities across 8 European and 29 Native American populations, based on population-wise estimates in Table S20 of Pemberton *et al.* (2013). The value of γ_1 can be viewed as the fraction of European ancestry in an admixed population and $1 - \gamma_1$ can be considered the fraction of Native American ancestry.

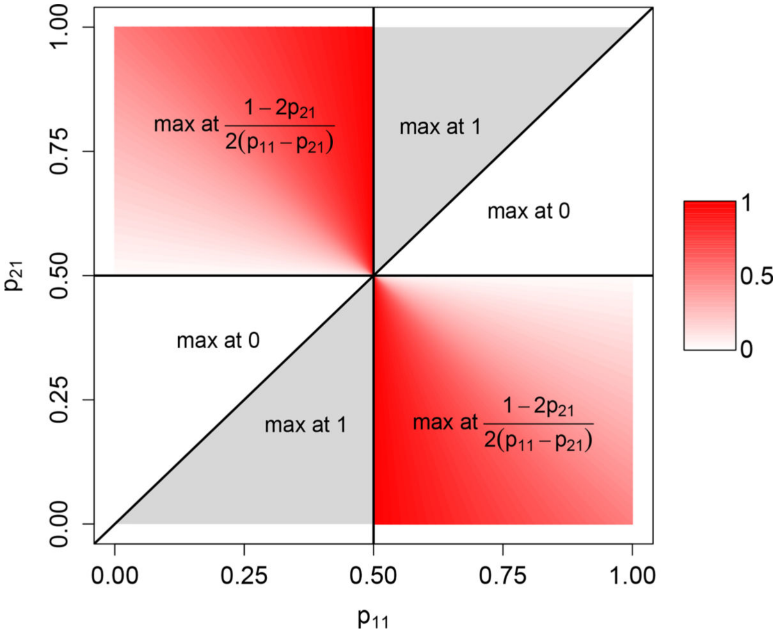

Figure 2:

The admixture coefficient γ_1 that maximizes H_{adm} in the case of $K = 2$ source populations and $J = 2$ allelic types. The plot shows the unit square for (p_{11}, p_{21}). In the red regions, the maximizing value of γ_1 lies in (0,1), whereas in the white and gray regions, it lies on one or the other boundary. The figure depicts the result of Proposition 11.

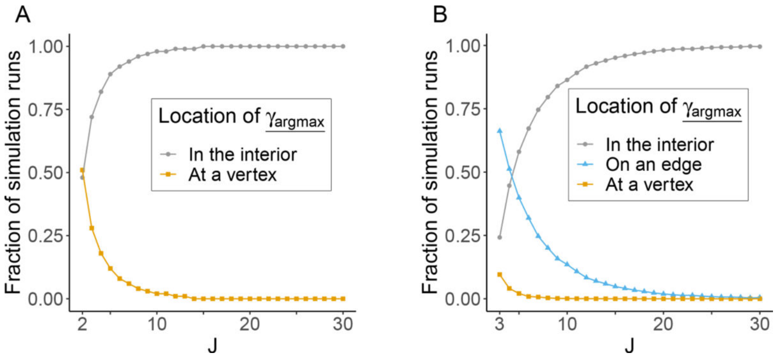

Figure 3:

Location of the maximum of H_{adm} in simulation replicates. (A) $K = 2$. (B) $K = 3$. The location $\underline{\gamma_{\text{arg max}}}$ can be in the interior of the simplex Δ^{K-1}, corresponding to nontrivial admixture of all source groups, or on the boundary of the simplex. For $K = 3$, it can be on an edge, corresponding to admixture of two of three source populations, and for both $K = 2$ and $K = 3$, it can be at a vertex, corresponding to membership in only one source population. For each (K, J), points plotted are based on 10,000 simulations with independently and identically distributed Dirichlet-(1, 1, …, 1) distributions for the allele frequency vectors $\underline{p_k}$ in the K populations.

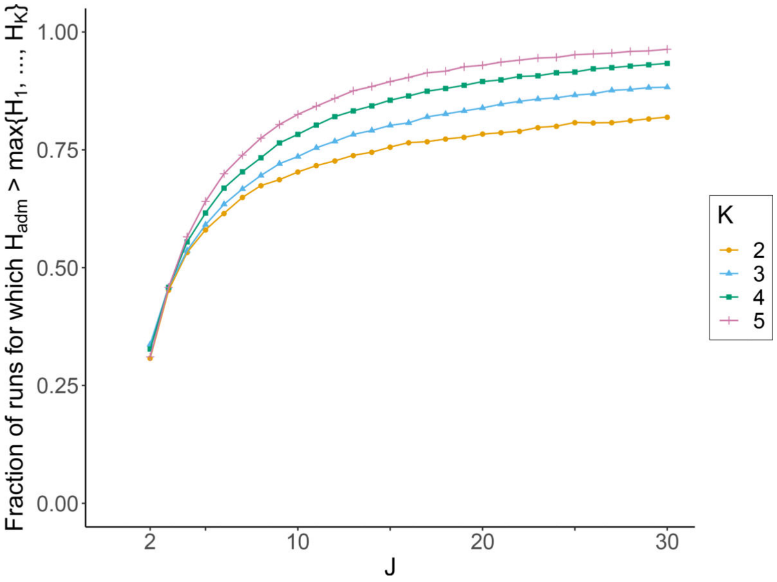

Figure 4:

The fraction of simulation replicates for which $H_{\text{adm}} > \max\{H_1, \ldots, H_K\}$, for various values of K and J. For each (K, J), points plotted are based on 50,000 simulation replicates with independently and identically distributed Dirichlet-(1, 1, …, 1) distribitions for the allele frequency vectors $\underline{p_k}$ in the K populations, and a Dirichlet-(1, 1, …, 1) distribution for the admixture coefficient vector $\underline{\gamma}$.

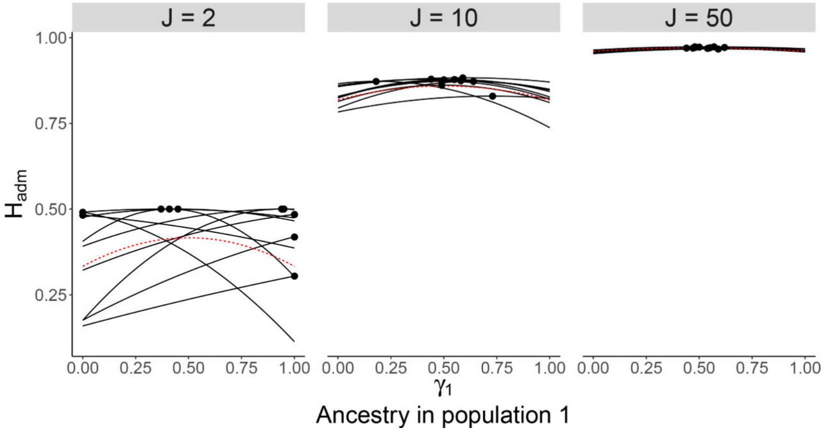

Figure 5:

H_{adm} versus γ_1 for 10 simulation replicates for $K = 2$ source populations, for each of three values of the number of allelic types J. For each replicate, allele frequency vectors $\underline{p_k}$ in the two populations are simulated according to Dirichlet-(1, 1, …, 1) distributions, and H_{adm} is plotted as a function of γ_1 according to eq. 8. The maximum of H_{adm} is indicated by a black circle in each replicate. The red dashed lines represent the expected values of H_{adm} according to Corollary 16 in Appendix 3.

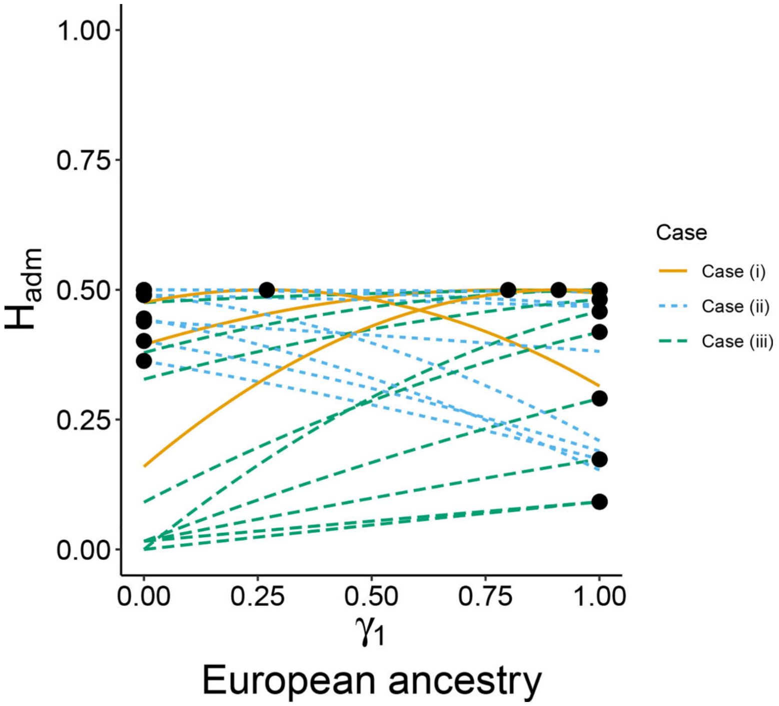

Figure 6:

H_{adm} versus γ_1 for 20 random biallelic loci from Pemberton *et al.* (2012). The two source populations providing the allele frequencies are the European and Native American populations, with γ_1 corresponding to membership in the European population. H_{adm} is plotted according to eq. 8. Circles indicate the location of the maximum along each curve. Different colors and line types correspond to the three cases in Proposition 11 for the location of the maximal H_{adm}.

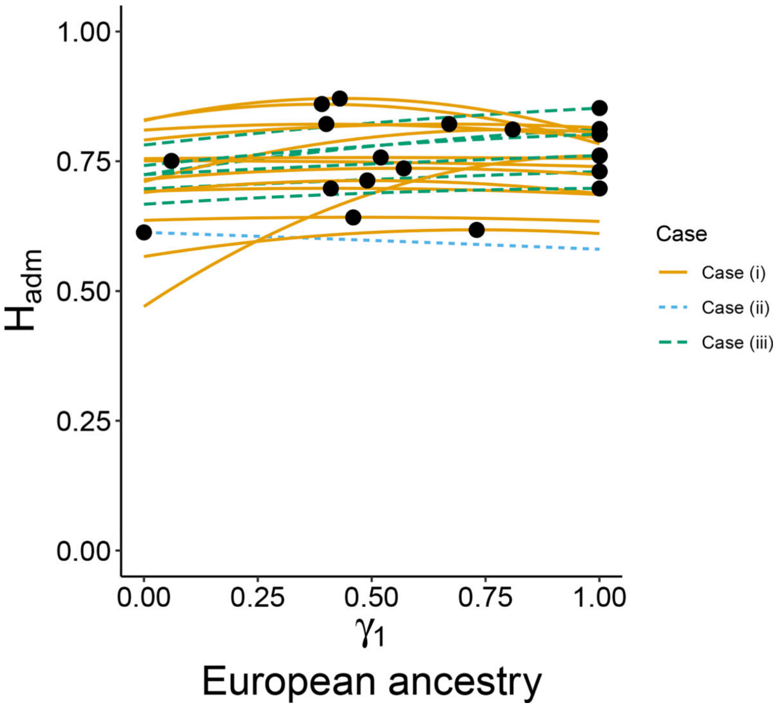

Figure 7:

H_{adm} versus γ_1 for 20 random multiallelic loci from Wang *et al.* (2008). The two source populations providing the allele frequencies are the European and Native American populations, with γ_1 corresponding to membership in the European population. H_{adm} is plotted according to eq. 8. Circles indicate the location of the maximum along each curve. Different colors and line types correspond to the three cases in Proposition 7 for the location of the maximal H_{adm}.

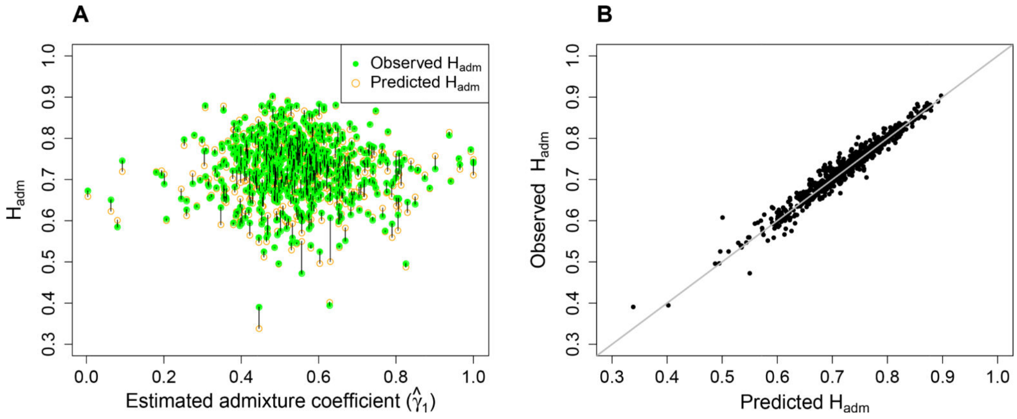

Figure 8:

Predicted and observed H_{adm}. (A) The predicted and observed H_{adm} values for an admixed Mestizo population are plotted against the locus-wise estimated European admixture fraction $\hat{\gamma}_1$ in the Mestizo population, estimated by maximum likelihood. The prediction is based on eq. 8, using European and Native American allele frequencies estimated from Wang *et al.* (2008) as $\underline{p_1}$ and $\underline{p_2}$, respectively, together with the maximum likelihood estimate of γ_1. The observation is based on H_{adm} computed from Definition 1, inserting estimated allele frequencies from Wang *et al.* (2008) for the Mestizo population. (B) The observed H_{adm} value is plotted against the predicted H_{adm} value. The identity line is shown in gray. In both panels, each point represents one of the 678 loci used. The correlation coefficient between the predicted and observed H_{adm} values is 0.978.

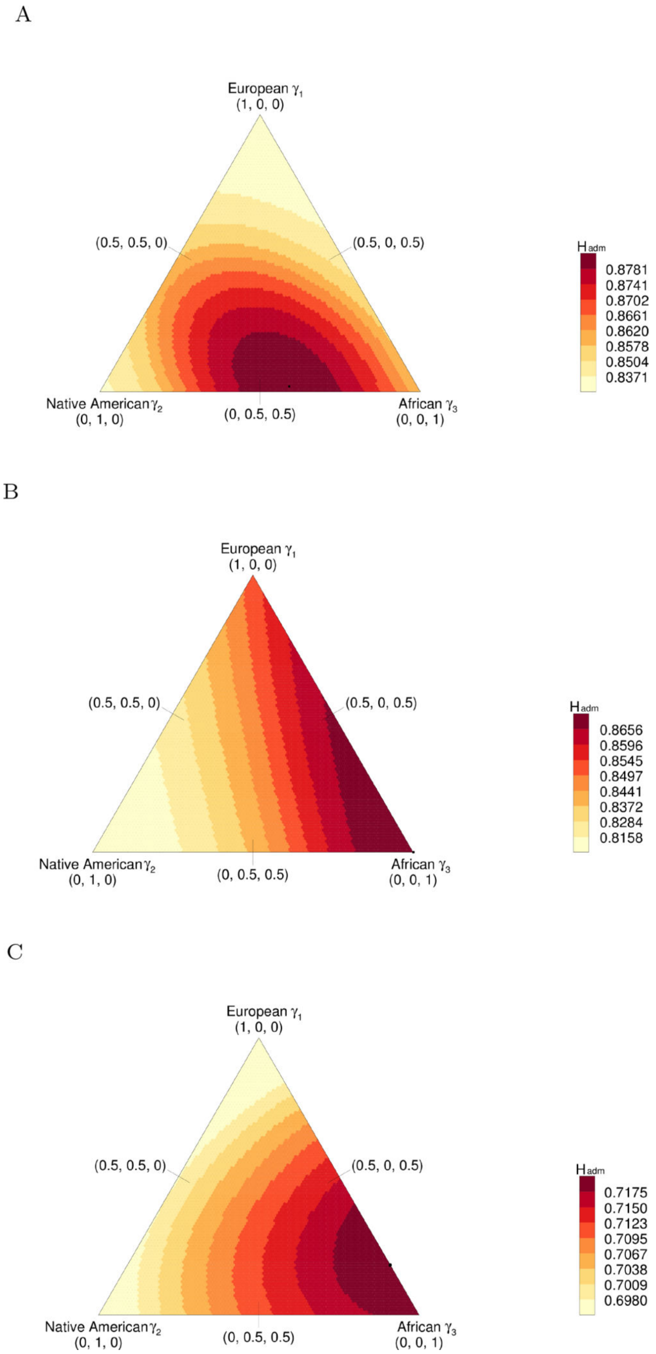

Figure 9:

H_{adm} versus (γ_1, γ_2, γ_3) for three loci. The loci are from Wang *et al.* (2008) and have 14, 14, and 8 distinct alleles, respectively. The value of H_{adm} is computed from eq. 8. Black circles indicate the maximum H_{adm}. (A) Locus D2S1399: the maximum lies in the interior of the region. (B) Locus GATA101G01: the maximum lies at the (0, 0, 1) vertex. (C) Locus GATA146D07: the maximum lies on the $\gamma_2 = 0$ edge.

Table 1:

Notation

Type of quantity	Symbol	Description
Indices	$j = 1,..., J$	Index over alleles
	$k = 1,..., K$	Index over source populations
Allele frequencies	p_{kj}	Frequency of allelic type j in population k
	$\mathbf{p}_k$	$J \times 1$ vector of allele frequencies for population k
	$\mathbf{P}$	$J \times K$ matrix of allele frequencies in the source populations
	$\bar{p}_j$	Frequency of allelic type j in the admixed population
Admixture fractions	γ_k	Admixture fraction for population k
	$\boldsymbol{\gamma}$	$K \times 1$ vector of admixture fractions
Heterozygosities	H_k	Heterozygosity for population k; probability that two alleles drawn from population k differ in type
	H_{adm}	Heterozygosity for the admixed population
	$C_{k\ell}$	Probability that an allele drawn from population k and an allele drawn from population ℓ differ in type
Fixation index	$F_{k\ell}$	Fixation index F_{ST} between populations k and ℓ