Using mathematical constraints to explain narrow ranges for allele-sharing dissimilarities

Xiran Liu*
Zarif Ahsan†
Noah A. Rosenberg†

Abstract. Allele-sharing dissimilarity (ASD) statistics are measures of genetic differentiation for pairs of individuals or populations. Given the allele-frequency distributions of two populations—possibly the same population—the expected value of an ASD statistic is computed by evaluating the expectation of the pairwise dissimilarity between two individuals drawn at random, each from its associated allele-frequency distribution. For each of two ASD statistics, which we term $\mathcal{D}_1$ and $\mathcal{D}_2$, we investigate the extent to which the expected ASD is constrained by allele frequencies in the two populations; in other words, how is the magnitude of the measure bounded as a function of the frequency of the most frequent allelic type? We first consider dissimilarity of a population with itself, obtaining bounds on expected ASD in terms of the frequency of the most frequent allelic type in the population. We then examine pairs of populations that might or might not possess the same most frequent allelic type. Across the unit interval for the frequency of the most frequent allelic type, the expected allele-sharing dissimilarity has a range that is more restricted than the $[0, 1]$ interval. The mathematical constraints on expected ASD assist in explaining a pattern observed empirically in human populations, namely that when averaging across loci, allele-sharing dissimilarities between pairs of individuals often tend to vary only within a relatively narrow range.

Key words: allele frequencies, genetic dissimilarity, population genetics

1 Introduction

Statistics based on concepts of allele-sharing dissimilarity (ASD) [Mountain and Cavalli-Sforza, 1997, Mountain and Ramakrishnan, 2005, Gao and Martin, 2009] are important tools in population-genetic data analysis. Beginning with the alleles of two diploid individuals at a genetic locus, a function of the four alleles is computed, producing a value ranging from 0 for the minimum dissimilarity to 1 for the maximum. Among genetic dissimilarity measures, ASD-based statistics are relatively easy to describe and compute. They are meaningful for pairs of individuals, or—if many individuals are considered—pairs of populations, or an individual and a population. Hence, features of allele-sharing dissimilarities are often used for understanding genetic variation within and among populations [Mountain and Cavalli-Sforza, 1997, Mountain and Ramakrishnan, 2005, Witherspoon et al., 2007, Rosenberg, 2011].

Studies of population-genetic statistics that consider dissimilarities across individuals suggest that ranges of observed numerical values of dissimilarity statistics—notably those based on the classic statistic F_{ST}—depend in predictable ways on allele-frequency distributions [Jakobsson et al., 2013, Edge and Rosenberg, 2014, Alcala and Rosenberg, 2017, 2019, 2022]. Consider two populations, each with an allele-frequency distribution at a locus, and consider a bounded statistic that measures the dissimilarity of the two populations as a function of the allele frequencies. Although the statistic is bounded, typically in $[0, 1]$, tighter constraints might exist on the dissimilarity in terms of the separate frequency distributions in the two populations. A value such as 0.55 or 0.7 might then be appropriate to interpret not in relation to the entire unit interval, but in relation to a shorter interval suited to its allele frequencies. Such interpretations have been used to explain unexpected numerical patterns in F_{ST}—such as a low F_{ST} value among high-diversity African populations [Jakobsson et al., 2013], and a high F_{ST} among chimpanzee populations relative to its value between chimpanzees and humans [Alcala and Rosenberg, 2022].

Empirical findings suggest that allele-sharing dissimilarities are also constrained by allele frequencies. For example, the values of allele-sharing dissimilarities have been seen to be quite similar across many computations. Consider, for example, the computations of allele-sharing dissimilarities as averages across many loci for pairs of

*Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305.
†Department of Biology, Stanford University, Stanford, CA 94305.

individuals in Figure 5A of Rosenberg [2011]. In those computations, which consider genome-wide loci in diverse human populations, most pairs of individuals possess allele-sharing dissimilarities between 0.55 and 0.7. Does the narrow range arise from mathematical constraints on ASD measures in relation to the allele frequencies?

We have recently investigated the mathematical properties of two formulations of population-level ASD measures, exploring mathematical properties of the expected genetic dissimilarity between pairs of individuals sampled within and between populations [Liu et al., 2023]. Here, building upon our previous mathematical results, we derive bounds on the two expectations, both within and between populations, in two scenarios: first, when the number of allelic types at a given locus is fixed at I and the allele-frequency distributions within a population can be arbitrary among these I alleles; second, when both I and the frequency M of the most frequent allelic type within a population are held constant. In both scenarios, we focus on the upper and lower bounds on the genetic dissimilarities in terms of the frequency of the most frequent allelic type. We find that indeed, ASD values are mathematically constrained to subintervals of $[0, 1]$, and that the constraints can assist in explaining features of ASD in human populations.

2 Preliminaries

2.1 Definitions

Following Liu et al. [2023], we consider two variants of the ASD concept, which we denote by $\mathcal{D}_1$ and $\mathcal{D}_2$. For $\mathcal{D}_1$, "allele-sharing" for two diploid individuals is interpreted as the number of shared elements in their sets of alleles. $\mathcal{D}_1$ then uses 1 minus the normalized count of the shared alleles as the dissimilarity. Consider a locus with four distinct alleles, A, B, C, and D, the minimum number required so that all possible cases for diploid genotypes exist. Two individuals both with genotype AB have 2 alleles shared, and $\mathcal{D}_1 = 0$. An individual with genotype AB and an individual with genotype AC have 1 shared allele, namely A, and $\mathcal{D}_1 = \frac{1}{2}$.

$\mathcal{D}_2$ instead considers alleles individually, evaluating the fraction of pairs of alleles, one from the first individual and one from the second, that are distinct. For two individuals with genotype AB, $\mathcal{D}_2 = \frac{1}{2}$: among the four possible pairs of alleles — (A,A), (A,B), (B,A), and (B,B), where the first entry in the pair represents an allele from the first individual and the second entry is an allele from the second — two of four contain distinct alleles. An individual with genotype AB and an individual with genotype AC have $\mathcal{D}_2 = \frac{3}{4}$.

Consider a locus with $I \geqslant 2$ allelic types, and suppose the allele frequencies in a population are $\mathbf{p} = (p_1, p_2, \ldots, p_I)$, where p_i represents the frequency of allele i. The frequencies satisfy $0 \leqslant p_i \leqslant 1$ for all i, and $\sum_{i=1}^{I} p_i = 1$. Without loss of generality, let $p_1 = M$ represent the largest entry in the allele-frequency vector $(p_1, p_2, \ldots, p_I)$. When we consider allele-frequency vectors in two populations, we let population 2 have allele frequencies $\mathbf{q} = (q_1, q_2, \ldots, q_I)$, satisfying $0 \leqslant q_i \leqslant 1$ for all i, and $\sum_{i=1}^{I} q_i = 1$. We define

$$\sigma_t = \sum_{i=1}^{I} p_i^t, \quad \tau_t = \sum_{i=1}^{I} q_i^t,$$

for $t = 1, 2, 3, 4$, where $\sigma_1 = \tau_1 = 1$. We also define

$$\rho_{tu} = \sum_{i=1}^{I} p_i^t q_i^u,$$

where (t, u) is equal to $(1, 1)$, $(1, 2)$, $(2, 1)$, or $(2, 2)$.

We denote the dissimilarity $\mathcal{D}$ between two individuals within the same population with allele-frequency vector $\mathbf{p}$ by $\mathcal{D}^w(\mathbf{p})$; here, $\mathcal{D}$ is understood to refer to one of the two dissimilarities, $\mathcal{D}_1$ or $\mathcal{D}_2$. We denote the corresponding dissimilarity between two individuals from different populations with allele-frequency vectors $\mathbf{p}$ and $\mathbf{q}$ by $\mathcal{D}^b(\mathbf{p}, \mathbf{q})$. We often drop the arguments for convenience.

2.2 Review of ASD mathematical results

In Liu et al. [2023], we studied a probabilistic model in which individuals are randomly sampled from allele-frequency distributions and $\mathcal{D}_1$ and $\mathcal{D}_2$ are computed. The expected value of $\mathcal{D}_1^w(\mathbf{p})$ satisfies

$$\mathbb{E}[\mathcal{D}_1^w(\mathbf{p})] = 1 - 2\sigma_2 + 2\sigma_3 - \sigma_4. \tag{1}$$

For $I = 2$, substituting $p_2 = 1 - p_1$ so that $\sigma_t = p_1^t + (1 - p_1)^t$, Eq. 1 becomes $\mathbb{E}[\mathcal{D}_1^w(\mathbf{p})] = 2p_1 - 4p_1^2 + 4p_1^3 - 2p_1^4$. We also have

$$\mathbb{E}[\mathcal{D}_2^w(\mathbf{p})] = 1 - \sigma_2, \tag{2}$$

and for the $I = 2$ case, Eq. 2 simplifies to $\mathbb{E}[\mathcal{D}_2^w(\mathbf{p})] = 2p_1(1 - p_1)$.

For the between-population dissimilarity $\mathcal{D}^b(\mathbf{p})$, we obtain

$$\mathbb{E}[\mathcal{D}_1^b(\mathbf{p}, \mathbf{q})] = 1 - 2\rho_{11} + \rho_{21} + \rho_{12} - \rho_{22}, \tag{3}$$

$$\mathbb{E}[\mathcal{D}_2^b(\mathbf{p}, \mathbf{q})] = 1 - \rho_{11}. \tag{4}$$

Eqs. 1, 2, 3, and 4 correspond to Eqs. 3, 9, 16, and 22 of Liu et al. [2023].

2.3 Review of majorization theory

We recall some results from majorization theory that will assist in finding bounds on ASD statistics. Majorization describes partial orderings on vectors with a shared sum.

Definition 2.1 (Majorization, 1.A.1 of Marshall et al. [2010])**.** Vector $\mathbf{x} \in \mathbb{R}^n$ is said to *majorize* vector $\mathbf{y} \in \mathbb{R}^n$ if, when the components of $\mathbf{x}$ and $\mathbf{y}$ are each rearranged in non-increasing order, (i) $\sum_{i=1}^k x_i \geqslant \sum_{i=1}^k y_i$ for all $k = 1, 2, \ldots, n-1$; and (ii) $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. Equivalently, $\mathbf{y}$ is said to be *majorized* by $\mathbf{x}$.

If $\mathbf{x}$ majorizes $\mathbf{y}$, then we write $\mathbf{x} \succ \mathbf{y}$. Functions that preserve majorization order are said to be Schur-convex.

Definition 2.2 (Schur-convexity, 3.A.1 of Marshall et al. [2010])**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *Schur-convex* if $\mathbf{x} \succ \mathbf{y}$ implies $f(\mathbf{x}) \geqslant f(\mathbf{y})$. The function is *strictly Schur-convex* if $\mathbf{x} \succ \mathbf{y}$ and $\mathbf{x} \neq \mathbf{y}$ implies $f(\mathbf{x}) > f(\mathbf{y})$. A function f is *Schur-concave* if $-f$ is Schur-convex and *strictly Schur-concave* if $-f$ is strictly Schur-concave.

Theorem 2.3 (Schur convexity condition, 3.A.4 of Marshall et al. [2010])**.** *Let $\mathcal{I} \subset \mathbb{R}$ be an open interval and let $f : \mathcal{I}^n \to \mathbb{R}$ be a continuously differentiable function. Function f is Schur-convex if and only if f is symmetric in its n arguments and for all (i, j) with $1 \leqslant i, j \leqslant n$,*

$$(x_i - x_j)\left(\frac{\partial f}{\partial x_i} - \frac{\partial f}{\partial x_j}\right) \geqslant 0.$$

Further, if equality requires $x_i = x_j$, then f is strictly Schur-convex.

Similarly, f is Schur-concave if and only if f is symmetric in its n arguments and for all (i, j) with $1 \leqslant i, j \leqslant n$,

$$(x_i - x_j)\left(\frac{\partial f}{\partial x_i} - \frac{\partial f}{\partial x_j}\right) \leqslant 0.$$

If equality requires $x_i = x_j$, then f is strictly Schur-concave.

Denote the unit $(I-1)$-simplex by Δ^{I-1}:

$$\Delta^{I-1} = \left\{(p_1, p_2, \ldots, p_I) \in \mathbb{R}^I \middle| \sum_{i=1}^I p_i = 1,\ p_i \geqslant 0 \text{ for all } i\right\}.$$

Proposition 2.4 (Majorization inequality for a unit simplex, Section 2.2 of Aw and Rosenberg [2018])**.** *For all vectors $\mathbf{p}$ in the unit $(I-1)$-simplex Δ^{I-1},*

$$\left(\frac{1}{I}, \frac{1}{I}, \ldots, \frac{1}{I}\right) \prec \mathbf{p} \prec (1, 0, \ldots, 0).$$

Proposition 2.5 (Majorization inequality for vectors in the simplex, with a specified value of the largest entry; see the proof of Theorem 3.9 of Aw and Rosenberg [2018])**.** *Let $\mathbf{p}$ be a vector of length I chosen within the simplex, with largest largest entry equal to M; that is, $\mathbf{p} \in \Delta^{\tilde{I}-1}$, where*

$$\Delta^{\tilde{I}-1} = \left\{(p_1, p_2, \ldots, p_I) \in \Delta^{I-1} \middle| p_1 \geqslant p_2 \geqslant \ldots \geqslant p_I,\ p_1 = M\right\}.$$

Then

$$\left(M, \frac{1-M}{I-1}, \frac{1-M}{I-1}, \ldots, \frac{1-M}{I-1}\right) \prec \mathbf{p} \prec \left(M, M, \ldots, M, 1 - (\lceil M^{-1} \rceil - 1)M, 0, 0, \ldots, 0\right).$$

Here, the left-hand vector has $I-1$ entries equal to $\frac{1-M}{I-1}$. The right-hand vector has $\lceil M^{-1}\rceil - 1$ entries equal to M followed by an entry of $1-(\lceil M^{-1}\rceil - 1)M$ and zeroes for the remaining entries. For convenience, we write $\mathbf{p}_{\min} = (M,\ldots,M,1-(\lceil M^{-1}\rceil-1)M,0,\ldots,0)$ and $\mathbf{p}_{\max} = (M,\frac{1-M}{I-1},\ldots,\frac{1-M}{I-1})$, noting that because we will be working with Schur-concave functions, the most majorized vector becomes the "maximum."

Theorem 2.6 (Rearrangement inequality, Theorem 368 of Hardy et al. [1952])**.** *Consider two sets of I real numbers $a_1 \geqslant a_2 \geqslant \ldots \geqslant a_I$ and $b_1 \geqslant b_2 \geqslant \ldots \geqslant b_I$. For each permutation $b_{\sigma(1)}, b_{\sigma(2)}, \ldots, b_{\sigma(I)}$ of $b_1, b_2, \ldots, b_I$,*

$$a_1 b_I + a_2 b_{I-1} + \ldots + a_I b_1 \leqslant a_1 b_{\sigma(1)} + a_2 b_{\sigma(2)} + \ldots + a_I b_{\sigma(I)} \leqslant a_1 b_1 + a_2 b_2 + \ldots + a_I b_I.$$

3 Mathematical constraints on within-population dissimilarity

Using Eqs. 1 and 2, we consider two sets of mathematical constraints on the within-population dissimilarity measures $\mathbb{E}[\mathcal{D}_1^w]$ and $\mathbb{E}[\mathcal{D}_2^w]$. First, fixing the number of allelic types I but permitting the allele-frequency distribution to be arbitrary, we consider general bounds as functions of I. Second, because the largest allele frequency M might impose further restrictions on the allele-frequency distribution, we consider the bounds when fixing both I and M.

3.1 Bounds on $\mathbb{E}[\mathcal{D}_1^w]$ when the number of allelic types I is fixed

Let $\mathbb{E}[\mathcal{D}_1^w]$ be a function of $\mathbf{p} = (p_1, p_2, \ldots, p_I)$ following Eq. 1, where $\mathbf{p} \in \Delta^{I-1}$, the standard $(I-1)$-simplex. Denote $\mathbb{E}[\mathcal{D}_1^w]$ by $f(\mathbf{p})$.

Lemma 3.1. *$f(\mathbf{p}) = \mathbb{E}[\mathcal{D}_1^w]$, as a function of $\mathbf{p} \in \Delta^{I-1}$, is strictly Schur-concave.*

The proof of the lemma appears in Appendix A. Using the strict Schur-concavity of the function $f(\mathbf{p}) = \mathbb{E}[\mathcal{D}_1^w]$ from Lemma 3.1, we arrive at the following theorem.

Theorem 3.2. *Suppose without loss of generality that $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_I$ for $\mathbf{p} = (p_1, p_2, \ldots, p_I)$. Then*

$$0 \leqslant \mathbb{E}[\mathcal{D}_1^w] \leqslant 1 - \frac{2}{I} + \frac{2}{I^2} - \frac{1}{I^3}.$$

Equality at the lower bound is reached if and only if $p_1 = 1$ and $p_i = 0$ for $2 \leqslant i \leqslant I$. Equality at the upper bound is reached if and only if $p_i = \frac{1}{I}$ for $1 \leqslant i \leqslant I$.

Proof. By Proposition 2.4, $\mathbf{p}_{\max} = (\frac{1}{I}, \frac{1}{I}, \ldots, \frac{1}{I})$ is majorized by all $\mathbf{p} \in \Delta^{I-1}$, and $\mathbf{p}_{\min} = (1, 0, \ldots, 0)$ majorizes all $\mathbf{p} \in \Delta^{I-1}$. Because $f(\mathbf{p}) = \mathbb{E}[\mathcal{D}_1^w]$ is strictly Schur-concave by Lemma 3.1, by definition of strict Schur-concavity (Definition 2.2), $f(\mathbf{p}_{\max}) \geqslant f(\mathbf{p})$ for all $\mathbf{p} \in \Delta^{I-1}$ and $f(\mathbf{p}_{\min}) \leqslant f(\mathbf{p})$ for all $\mathbf{p} \in \Delta^{I-1}$. Therefore,

$$\max \mathbb{E}[\mathcal{D}_1^w] = f(\mathbf{p}_{\max}) = 1 - \frac{2}{I} + \frac{2}{I^2} - \frac{1}{I^3},$$
$$\min \mathbb{E}[\mathcal{D}_1^w] = f(\mathbf{p}_{\min}) = 0,$$

with equality if and only if $\mathbf{p}$ lies at the specified points. □

In the simplest case of $I = 2$, we find that $\mathbb{E}[\mathcal{D}_1^w]$ is maximized for $(p_1, p_2) = (\frac{1}{2}, \frac{1}{2})$, with $\mathbb{E}[\mathcal{D}_1^w] = \frac{3}{8}$. It is minimized for $(p_1, p_2) = (1, 0)$, at which $\mathbb{E}[\mathcal{D}_1^w] = 0$.

The relationship between the upper bound on $\mathbb{E}[\mathcal{D}_1^w]$ and I is shown in Figure 1. The figure shows a strictly increasing sequence, as is clear by noting that the derivative of the upper bound of $\mathbb{E}[\mathcal{D}_1^w]$ as a function of I is $(2I^2 - 4I + 3)/I^4$, a strictly positive function for $I \geqslant 2$. As $I \to \infty$, this upper bound approaches 1.

3.2 Bounds on $\mathbb{E}[\mathcal{D}_1^w]$ when the largest allele frequency is fixed

If the largest allele frequency M is fixed, then a tighter constraint is imposed on the range of values that $\mathbb{E}[\mathcal{D}_1^w]$ can take. To derive this constraint, we use Proposition 2.5.

Theorem 3.3. *Suppose $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_I$, and suppose $p_1 = M$ is fixed, $\frac{1}{I} \leqslant M \leqslant 1$. Let $L = \frac{1-M}{I-1}$ and $R = \lceil M^{-1}\rceil - 1$. $f(\mathbf{p}) = \mathbb{E}[\mathcal{D}_1^w] = 1 - 2\sigma_2 + 2\sigma_3 - \sigma_4$ is bounded by*

$$f(\mathbf{p}_{\min}) \leqslant \mathbb{E}[\mathcal{D}_1^w] \leqslant f(\mathbf{p}_{\max}).$$

Equality with the lower bound is achieved if and only if $\mathbf{p} = \mathbf{p}_{\min} = (M,\ldots,M,1-(\lceil M^{-1}\rceil-1)M,0,\ldots,0)$, producing $f(\mathbf{p}_{\min}) = 1-R(2M^2-2M^3+M^4)-2(1-RM)^2+2(1-RM)^3-(1-RM)^4$. Equality with the upper bound is achieved if and only if $\mathbf{p} = \mathbf{p}_{\max} = (M, \frac{1-M}{I-1}, \ldots, \frac{1-M}{I-1})$, producing $f(\mathbf{p}_{\max}) = 1-2M^2+2M^3-M^4-(I-1)(2L^2-2L^3+L^4)$.

The proof is straightforward: by Proposition 2.5, $\mathbf{p}_{\min} \succ \mathbf{p} \succ \mathbf{p}_{\max}$ for all $\mathbf{p} \in \Delta^{\tilde{I}-1}$. Because f is strictly Schur-concave (Lemma 3.1), by the definition of Schur-concavity (Definition 2.2), $f(\mathbf{p}_{\min}) \leqslant f(\mathbf{p}) \leqslant f(\mathbf{p}_{\max})$ for all $\mathbf{p} \in \Delta^{\tilde{I}-1}$, with the appropriate equality conditions.

In the $I = 2$ case, there is a single choice for $\mathbf{p}$, and the two bounds coincide. For each I from 2 to 9, Figure 2 plots the region specified by the theorem, illustrating that as I increases, the size of the permissible region grows.

The vector $\mathbf{p}$ that produces equality of the lower bound of $\mathbb{E}[\mathcal{D}_1^w]$ given M is exactly the same as the one that minimizes the heterozygosity given I and M; similarly, the vector that produces equality of the upper bound of $\mathbb{E}[\mathcal{D}_1^w]$ given M is the vector tht maximizes heterozygosity given I and M [Reddy and Rosenberg, 2012].

Proposition 3.4. *With fixed I, the region bounded by the upper and lower bounds on $\mathbb{E}[\mathcal{D}_1^w]$ as a function of M has area*

$$S_{\mathbb{E}[\mathcal{D}_1^w]}(I) = \frac{19}{30} - \frac{31}{30I} + \frac{4}{5I^2} - \frac{2}{5I^3} - \sum_{i=2}^{I}\left[\frac{11}{30(i-1)^2} - \frac{3}{10(i-1)^3} + \frac{1}{5(i-1)^4}\right]. \tag{5}$$

The proof appears in Appendix B. Letting $I \to \infty$ in Eq. 5, noting that the Riemann zeta function satisfies $\zeta(2) = \sum_{i=1}^{\infty} 1/i^2 = \pi^2/6$, $\zeta(3) = \sum_{i=1}^{\infty} 1/i^3 \approx 1.202057$, and $\zeta(4) = \sum_{i=1}^{\infty} 1/i^4 = \pi^4/90$, the area approaches

$$\begin{aligned} S_{\mathbb{E}[\mathcal{D}_1^w]}(\infty) &= \frac{19}{30} - \frac{11}{30}\zeta(2) + \frac{3}{10}\zeta(3) - \frac{1}{5}\zeta(4) \\ &= \frac{19}{30} - \frac{11\pi^2}{180} - \frac{\pi^4}{450} + \frac{3}{10}\zeta(3) \approx 0.174343. \end{aligned} \tag{6}$$

The area of the region is plotted as a function of I in Figure 3.

3.3 Bounds on $\mathbb{E}[\mathcal{D}_2^w]$ when the number of allelic types I is fixed

Bounds on $\mathbb{E}[\mathcal{D}_2^w]$ can be obtained similarly to those on $\mathbb{E}[\mathcal{D}_1^w]$. We write $\mathbb{E}[\mathcal{D}_2^w]$ as a function $g(\mathbf{p})$ following Eq. 2, where $\mathbf{p} \in \Delta^{I-1}$. The functional form $\mathbb{E}[\mathcal{D}_2^w] = 1 - \sigma_2$ admits known results for the homozygosity σ_2.

Lemma 3.5. *$g(\mathbf{p}) = \mathbb{E}[\mathcal{D}_2^w]$, as a function of $\mathbf{p} \in \Delta^{I-1}$, is strictly Schur-concave.*

The proof of the lemma follows directly from the strict Schur-convexity of homozygosity σ_2 [Aw and Rosenberg, 2018, p. 720]. As a function of $\mathbf{p}$, $g(\mathbf{p}) = 1 - \sigma_2$, so that $g(\mathbf{p})$ is strictly Schur-concave by Definition 2.2.

Theorem 3.6. *Suppose without loss of generality that $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_I$ for $\mathbf{p} = (p_1, p_2, \ldots, p_I)$. Then*

$$0 \leqslant \mathbb{E}[\mathcal{D}_2^w] \leqslant 1 - \frac{1}{I}.$$

Equality at the lower bound is reached if and only if $p_1 = 1$ and $p_i = 0$ for $2 \leqslant i \leqslant I$. Equality at the upper bound is reached if and only if $p_i = \frac{1}{I}$ for $1 \leqslant i \leqslant I$.

Proof. Using the strict Schur-concavity of function g, the proof follows that of Theorem 3.2. We obtain

$$\begin{aligned} \max \mathbb{E}[\mathcal{D}_2^w] &= g(\mathbf{p}_{\max}) = 1 - \frac{1}{I}, \\ \min \mathbb{E}[\mathcal{D}_2^w] &= g(\mathbf{p}_{\min}) = 0, \end{aligned}$$

with equality if and only if $\mathbf{p}$ lies at the specified points. □

In the $I = 2$ case, we have a maximum value of $\mathbb{E}[\mathcal{D}_2^w] = \frac{1}{2}$ if and only if $(p_1, p_2) = (\frac{1}{2}, \frac{1}{2})$ and a minimum value of $\mathbb{E}[\mathcal{D}_2^w] = 0$ if and only if $(p_1, p_2) = (1, 0)$.

The relationship between the upper bound on $\mathbb{E}[\mathcal{D}_2^w]$ and I appears in Figure 1. It is a strictly increasing sequence, and as $I \to \infty$, the upper bound approaches 1. Note that for $I \geqslant 2$, the upper bound on $\mathbb{E}[\mathcal{D}_2^w]$, $1 - \frac{1}{I}$, strictly exceeds the upper bound on $\mathbb{E}[\mathcal{D}_1^w]$, $1 - \frac{2}{I} + \frac{2}{I^2} - \frac{1}{I^3}$, as $(1 - \frac{1}{I}) - (1 - \frac{2}{I} + \frac{2}{I^2} - \frac{1}{I^3}) = (I-1)^2/I^3 > 0$.

3.4 Bounds on $\mathbb{E}[\mathcal{D}_2^w]$ when the largest allele frequency is fixed

The bounds on $\mathbb{E}[\mathcal{D}_2^w] = 1 - \sigma_2$ when $p_1 = M$ restate known bounds on σ_2.

Theorem 3.7. *Suppose $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_I$, and suppose $p_1 = M$ is fixed, $\frac{1}{I} \leqslant M \leqslant 1$. Let $L = \frac{1-M}{I-1}$ and $R = \lceil M^{-1} \rceil - 1$. $g(\mathbf{p}) = \mathbb{E}[\mathcal{D}_2^w] = 1 - \sigma_2$ is bounded by*

$$g(\mathbf{p}_{\min}) \leqslant \mathbb{E}[\mathcal{D}_2^w] \leqslant g(\mathbf{p}_{\max}).$$

Equality with the lower bound is achieved if and only if $\mathbf{p} = \mathbf{p}_{\min} = (M, \ldots, M, 1 - (\lceil M^{-1} \rceil - 1)M, 0, \ldots, 0)$, producing $g(\mathbf{p}_{\min}) = 1 - RM^2 - (1 - RM)^2$. Equality with the upper bound is achieved if and only if $\mathbf{p} = \mathbf{p}_{\max} = (M, \frac{1-M}{I-1}, \ldots, \frac{1-M}{I-1})$, producing $g(\mathbf{p}_{\max}) = 1 - M^2 - (I-1)L^2$.

The theorem is a restatement of Theorem 2 in Reddy and Rosenberg [2012], which provided the bounds on σ_2 for fixed I and M. In the $I = 2$ case, the upper and lower bounds coincide. The upper and lower bounds are achieved at precisely the same allele-frequency vectors that achieve the upper and lower bounds on $\mathbb{E}[\mathcal{D}_1^w]$.

The relationships between the lower and upper bounds of $\mathbb{E}[\mathcal{D}_1^w]$ and $\mathbb{E}[\mathcal{D}_2^w]$ and the frequency M appear in Figure 2. Both bounds of $\mathbb{E}[\mathcal{D}_2^w]$ exceed than those of $\mathbb{E}[\mathcal{D}_1^w]$, as is clear by noting that $1 - 2\sigma_2 + 2\sigma_3 - \sigma_4 = \mathbb{E}[\mathcal{D}_1^w] = \mathbb{E}[\mathcal{D}_2^w] - \sum_{i=1}^I p_i^2(1-p_i)^2 \leqslant \mathbb{E}[\mathcal{D}_2^w] = 1 - \sigma_2$.

Proposition 3.8. *With fixed I, the region bounded by the upper and lower bounds of $\mathbb{E}[\mathcal{D}_2^w]$ as a function of M has area*

$$S_{\mathbb{E}[\mathcal{D}_2^w]}(I) = \frac{2}{3} - \frac{2}{3I} - \sum_{i=2}^{I} \frac{1}{3(i-1)^2}. \tag{7}$$

As $I \to \infty$, the area approaches

$$S_{\mathbb{E}[\mathcal{D}_2^w]}(\infty) = \frac{2}{3} - \frac{1}{3}\zeta(2) = \frac{2}{3} - \frac{\pi^2}{18} \approx 0.118355. \tag{8}$$

Proposition 3.8 and the calculation of $S_{\mathbb{E}[\mathcal{D}_2^w]}(\infty)$ restate Proposition 16 of Reddy and Rosenberg [2012] for the bounds on σ_2. The area of the region is plotted as a function of I in Figure 3.

4 Mathematical constraints on between-population dissimilarity

Next, we look at the bounds for the between-population dissimilarities, which involve the allele-frequency vectors of two populations, $\mathbf{p}$ and $\mathbf{q}$.

4.1 Bounds on $\mathbb{E}[\mathcal{D}_1^b]$ when the number of allelic types is fixed

If the number of distinct alleles is fixed and the allele-frequency distributions can be arbitrary, then the bounds on both between-population dissimilarity measures are trivial.

Proposition 4.1. *Suppose without loss of generality that $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_I$ for $\mathbf{p} = (p_1, p_2, \ldots, p_I)$, and no constraints are placed on $\mathbf{q}$. Then*

$$0 \leqslant \mathbb{E}[\mathcal{D}_1^b] \leqslant 1.$$

Equality at the lower bound is reached if and only if $p_1 = q_1 = 1$ and $p_i = q_i = 0$ for $2 \leqslant i \leqslant I$. Equality at the upper bound is reached if and only if $\mathbf{p}$ and $\mathbf{q}$ satisfy $p_i q_i = 0$ for $1 \leqslant i \leqslant I$.

Proof. For a pair of individuals, one from population 1 and one from population 2, $\mathcal{D}_1^b \in \{0, \frac{1}{2}, 1\}$. Hence, as a function of allele-frequency vectors, we know $0 \leqslant \mathbb{E}[\mathcal{D}_1^b] \leqslant 1$.

We consider the equality condition $\mathbb{E}[\mathcal{D}_1^b] = 0$. For a pair of individuals, $\mathcal{D}_1^b = 0$ if and only if both individuals have exactly the same diploid genotype. $\mathbb{E}[\mathcal{D}_1^b] = 0$ requires that all pairs of individuals, one from one population and one from the other, possess the same diploid genotype. If either population possesses at least two distinct alleles with nonzero frequency, then the probability is positive that $\mathcal{D}_1^b > 0$ for a pair of random individuals, one from one population and one from the other, so that $\mathbb{E}[\mathcal{D}_1^b] > 0$. We conclude that if $\mathbb{E}[\mathcal{D}_1^b] = 0$, then two populations must have the same single allele. With $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_I$, it follows that $p_1 = q_1 = 1$ and $p_i = q_i = 0$ for $2 \leqslant i \leqslant I$.

For the equality condition $\mathbb{E}[\mathcal{D}_1^b] = 1$, writing $\mathbb{E}[\mathcal{D}_1^b] = 1 - \sum_{i=1}^I p_i q_i - \sum_{i=1}^I p_i q_i (1-p_i)(1-q_i)$, we find that $\mathbb{E}[\mathcal{D}_1^b] = 1$ implies $\sum_{i=1}^I p_i q_i + \sum_{i=1}^I p_i q_i (1-p_i)(1-q_i) = 0$, from which we conclude that the non-negative p_i and q_i bounded above by 1 must satisfy $p_i q_i = 0$ for all i. □

In the case of $I = 2$, $\mathbb{E}[\mathcal{D}_1^b]$ is minimized for $\mathbf{p} = \mathbf{q} = (1, 0)$, and maximized for $\mathbf{p} = (1, 0)$, $\mathbf{q} = (0, 1)$.

4.2 Bounds on $\mathbb{E}[\mathcal{D}_1^b]$ when the largest allele frequencies are fixed

We next consider scenarios in which the largest allele frequency is fixed in both populations. We first investigate the bounds on $\mathbb{E}[\mathcal{D}_1^b]$ in the case that the same allelic type has the largest frequency in both populations.

Theorem 4.2. *Suppose* $p_1 = \max\{p_1, p_2, \ldots, p_I\}$ *and* $q_1 = \max\{q_1, q_2, \ldots, q_I\}$. *Suppose without loss of generality that* $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_I$. *(i) If* $p_1 = M_1$ *and* $q_1 = M_2$ *are fixed, then*

$$M_1 + M_2 - 4M_1M_2 + 2M_1^2M_2 + 2M_1M_2^2 - 2M_1^2M_2^2 \leqslant \mathbb{E}[\mathcal{D}_1^b] \leqslant 1 - 2M_1M_2 + M_1^2M_2 + M_1M_2^2 - M_1^2M_2^2.$$

(ii) Equality with the upper bound is achieved if and only if for all i, $2 \leqslant i \leqslant I$, $p_i = 0$ *or* $q_i = 0$. *(iii) Equality with the lower bound is achieved if and only if* $M_1 \geqslant \frac{1}{2}$, $M_2 \geqslant \frac{1}{2}$, *and* $(p_2, q_2) = (1 - M_1, 1 - M_2)$.

The proof appears in Appendix C. Note that if $M_1 < \frac{1}{2}$ or $M_2 < \frac{1}{2}$, we have not obtained strict inequalities; the bounds in the theorem hold, but the lower bound is not the strictest possible inequality. The bounds in the theorem are depicted in Figure 4A-F.

For a general scenario in which two populations might have different allelic types for their most frequent allele, we also obtain loose inequalities. We proceed by first bounding $\mathbb{E}[\mathcal{D}_1^b]$ in relation to $\mathbb{E}[\mathcal{D}_2^b]$, then deriving bounds for $\mathbb{E}[\mathcal{D}_1^b]$ based on the bounds we obtain for $\mathbb{E}[\mathcal{D}_2^b]$ in Section 4.4.

Theorem 4.3. *Suppose* $\max\{p_1, p_2, \ldots, p_I\} = M_1$ *and* $\max\{q_1, q_2, \ldots, q_I\} = M_2$. *Suppose without loss of generality that* $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_I$. *Then*

$$\mathbb{E}[\mathcal{D}_2^b] - IM_1(1 - M_1)M_2(1 - M_2) \leqslant \mathbb{E}[\mathcal{D}_1^b] \leqslant \mathbb{E}[\mathcal{D}_2^b].$$

The proof appears in Appendix D. The bounds in the theorem are depicted in Figure 5A-F. Together with the lower and upper bounds for $\mathbb{E}[\mathcal{D}_2^b]$ in Section 4.4, we are able to obtain a lower bound and upper bound for $\mathbb{E}[\mathcal{D}_1^b]$.

4.3 Bounds on $\mathbb{E}[\mathcal{D}_2^b]$ when the number of allelic types is fixed

As we observed with $\mathbb{E}[\mathcal{D}_1^b]$, the bounds on $\mathbb{E}[\mathcal{D}_2^b]$ are trivial when the number of distinct alleles is fixed and the allele-frequency distribution can be arbitrary.

Proposition 4.4. *Suppose without loss of generality that* $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_I$ *for* $\mathbf{p} = (p_1, p_2, \ldots, p_I)$, *and no constraints are placed on* $\mathbf{q}$. *Then*

$$0 \leqslant \mathbb{E}[\mathcal{D}_2^b] \leqslant 1.$$

Equality at the lower bound is reached if and only if $p_1 = q_1 = 1$ *and* $p_i = q_i = 0$ *for* $2 \leqslant i \leqslant I$. *Equality at the upper bound is reached if and only if* $\mathbf{p}$ *and* $\mathbf{q}$ *satisfy* $p_iq_i = 0$ *for* $1 \leqslant i \leqslant I$.

Proof. We have $\mathbb{E}[\mathcal{D}_2^b] = 1 - \rho_{11}$, and $0 \leqslant \rho_{11} = \sum_{i=1}^I p_iq_i \leqslant \sum_{i=1}^I p_i = 1$, so that $0 \leqslant \mathbb{E}[\mathcal{D}_2^b] \leqslant 1$.

The equality condition $E[\mathcal{D}_2^b] = 0$ holds if and only if $\rho_{11} = 1$; $\rho_{11} = 1$ implies $\sum_{i=1}^I p_iq_i = \sum_{i=1}^I p_i$, from which $\sum_{i=1}^I p_i(1 - q_i) = 0$ and $q_i = 1$ for each i for which $p_i > 0$. Symmetrically, $\sum_{i=1}^I (1 - p_i)q_i = 0$ and $p_i = 1$ for each i for which $q_i > 0$. We conclude that for all i, $(p_i, q_i) = (1, 1)$ or $(0, 0)$. Because $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_I$, we have $(p_1, q_1) = (1, 1)$ and $(p_i, q_i) = (0, 0)$ for $2 \leqslant i \leqslant I$.

The equality condition $E[\mathcal{D}_2^b] = 1$ holds if and only if $\rho_{11} = 0$, so that $p_iq_i = 0$ for all i. □

For $I = 2$, $\mathbb{E}[\mathcal{D}_2^b]$ has the same minimum and maximum as $\mathbb{E}[\mathcal{D}_1^b]$: $\mathbf{p} = \mathbf{q} = (1, 0)$ for the minimum and $\mathbf{p} = (1, 0)$, $\mathbf{q} = (0, 1)$ for the maximum.

4.4 Bounds on $\mathbb{E}[\mathcal{D}_2^b]$ when the largest allele frequencies are fixed

Unlike for $\mathcal{D}_1^b$, we obtain strict bounds $\mathbb{E}[\mathcal{D}_2^b]$ in the general scenario in which two populations may have different allelic types for the most frequent allele, with frequencies M_1 and M_2 respectively. We recall the rearrangement inequality (Theorem 2.6) and use a related Lemma E.3 in Appendix E.

Theorem 4.5. *Suppose* $\max\{p_1, p_2, \ldots, p_I\} = M_1$ *and* $\max\{q_1, q_2, \ldots, q_I\} = M_2$. *Suppose without loss of generality that* $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_I$.

(i) $\mathbb{E}[\mathcal{D}_2^b]$ *as a function of* $\mathbf{p}$ *and* $\mathbf{q}$, *denoted by* $\ell(\mathbf{p}, \mathbf{q})$, *is bounded by*

$$\ell(\mathbf{p}_*, \mathbf{q}_*) \leqslant \mathbb{E}[\mathcal{D}_2^b] \leqslant \ell(\mathbf{p}^*, \mathbf{q}^*),$$

for particular vectors $\mathbf{p}_*$, $\mathbf{q}_*$, $\mathbf{p}^*$, *and* $\mathbf{q}^*$.

(ii) Equality at the lower bound is reached if

$$p_{*i} = \begin{cases} M_1, & \text{for } i = 1, \dots, \lceil M_1^{-1} \rceil - 1, \\ 1 - (\lceil M_1^{-1} \rceil - 1)M_1, & \text{for } i = \lceil M_1^{-1} \rceil, \\ 0, & \text{for } i = \lceil M_1^{-1} \rceil + 1, \dots, I, \end{cases} \tag{9}$$

and

$$q_{*i} = \begin{cases} M_2, & \text{for } i = 1, \dots, \lceil M_2^{-1} \rceil - 1, \\ 1 - (\lceil M_2^{-1} \rceil - 1)M_2, & \text{for } i = \lceil M_2^{-1} \rceil, \\ 0, & \text{for } i = \lceil M_2^{-1} \rceil + 1, \dots, I. \end{cases} \tag{10}$$

The minimum value is

$$\ell(\mathbf{p}_*, \mathbf{q}_*) = \begin{cases} 1 - (\lceil M_1^{-1} \rceil - 1)M_1M_2 - aM_2, & \text{if } \lceil M_1^{-1} \rceil < \lceil M_2^{-1} \rceil, \\ 1 - (\lceil M_1^{-1} \rceil - 1)M_1M_2 - ab, & \text{if } \lceil M_1^{-1} \rceil = \lceil M_2^{-1} \rceil, \\ 1 - (\lceil M_2^{-1} \rceil - 1)M_1M_2 - M_1b, & \text{if } \lceil M_1^{-1} \rceil > \lceil M_2^{-1} \rceil, \end{cases} \tag{11}$$

where $a = 1 - (\lceil M_1^{-1} \rceil - 1)M_1$ *and* $b = 1 - (\lceil M_2^{-1} \rceil - 1)M_2$.

(iii) Equality at the upper bound is reached if $\mathbf{p}^* = \mathbf{p}_*$ *(Eq. 9), and*

$$q_i^* = \begin{cases} 0, & \text{for } i = 1, \dots, I - \lceil M_2^{-1} \rceil, \\ 1 - (\lceil M_2^{-1} \rceil - 1)M_2, & \text{for } i = I - \lceil M_2^{-1} \rceil + 1, \\ M_2, & \text{for } i = I - \lceil M_2^{-1} \rceil + 2, \dots, I. \end{cases} \tag{12}$$

The maximum value is

$$\ell(\mathbf{p}^*, \mathbf{q}^*) = \begin{cases} 1, & \text{if } \lceil M_1^{-1} \rceil + \lceil M_2^{-1} \rceil \leqslant I, \\ 1 - ab, & \text{if } \lceil M_1^{-1} \rceil + \lceil M_2^{-1} \rceil = I + 1, \\ 1 - aM_2 - M_1b, & \text{if } \lceil M_1^{-1} \rceil + \lceil M_2^{-1} \rceil = I + 2, \\ 1 - (\lceil M_1^{-1} \rceil + \lceil M_2^{-1} \rceil - I - 2)M_1M_2 - aM_2 - M_1b, & \text{if } \lceil M_1^{-1} \rceil + \lceil M_2^{-1} \rceil > I + 2. \end{cases}$$

The proof of the theorem appears in Appendix E. The bounds in the theorem are depicted in Figure 5G-L. The bounds also appear in the loose bounds for $\mathbb{E}[\mathcal{D}_1^b]$ in Theorem 4.3 in the same setting, with $\mathbb{E}[\mathcal{D}_1^b]$ bounded above by the upper bound on $\mathbb{E}[\mathcal{D}_2^b]$ and below by the lower bound on $\mathbb{E}[\mathcal{D}_2^b]$ minus some additional terms.

As a corollary, we obtain the bounds for the specific scenario in which the same allelic type is most frequent in the two populations.

Corollary 4.6. *Suppose* $p_1 = \max\{p_1, p_2, \dots, p_I\} = M_1$ *and* $q_1 = \max\{q_1, q_2, \dots, q_I\} = M_2$. *Suppose without loss of generality that* $p_1 \geqslant p_2 \geqslant \dots \geqslant p_I$.

(i) $\mathbb{E}[\mathcal{D}_2^b]$ *as a function of* $\mathbf{p}$ *and* $\mathbf{q}$, *denoted by* $\ell(\mathbf{p}, \mathbf{q})$, *is bounded by*

$$\ell(\mathbf{p}_*, \mathbf{q}_*) \leqslant \mathbb{E}[\mathcal{D}_2^b] \leqslant \ell(\mathbf{p}^*, \mathbf{q}^*),$$

for particular vectors $\mathbf{p}_*$, $\mathbf{q}_*$, $\mathbf{p}^*$, *and* $\mathbf{q}^*$.

(ii) Equality at the lower bound is reached if $\mathbf{p}_*$ *follows Eq. 9 and* $\mathbf{q}_*$ *follows Eq. 10. The minimum value follows Eq. 11.*

(iii) Equality at the upper bound is reached if $\mathbf{p}^* = \mathbf{p}_*$ *(Eq. 9) and*

$$q_i^* = \begin{cases} M_2, & \text{for } i = 1, \\ 0, & \text{for } i = 2, \dots, I - \lceil M_2^{-1} \rceil + 1, \\ 1 - (\lceil M_2^{-1} \rceil - 1)M_2, & \text{for } i = I - \lceil M_2^{-1} \rceil + 2, \\ M_2, & \text{for } i = I - \lceil M_2^{-1} \rceil + 3, \dots, I. \end{cases} \tag{13}$$

The maximum value is

$$\ell(\mathbf{p}^*, \mathbf{q}^*) = \begin{cases} 1 - M_1M_2, & \text{if } \lceil M_1^{-1} \rceil + \lceil M_2^{-1} \rceil \leqslant I + 1, \\ 1 - M_1M_2 - ab, & \text{if } \lceil M_1^{-1} \rceil + \lceil M_2^{-1} \rceil = I + 2, \\ 1 - M_1M_2 - aM_2 - M_1b, & \text{if } \lceil M_1^{-1} \rceil + \lceil M_2^{-1} \rceil = I + 3, \\ 1 - (\lceil M_1^{-1} \rceil + \lceil M_2^{-1} \rceil - I - 2)M_1M_2 - aM_2 - M_1b, & \text{if } \lceil M_1^{-1} \rceil + \lceil M_2^{-1} \rceil > I + 3. \end{cases}$$

The corollary is proven in Appendix F. The bounds in the theorem are depicted in Figure 4G-L.

5 Data analysis

To investigate how the mathematical bounds with respect to the largest allele frequency affect the values of allele-sharing dissimilarity measures in an empirical setting, we compute the dissimilarities in a dataset of multiallelic loci in human populations.

5.1 Data

We analyze microsatellite genotypes in the H1048 subset of the Human Genome Diversity Project (HGDP-CEPH panel), considering 1048 individuals in 53 populations, typed at 783 microsatellite loci [Rosenberg et al., 2005, Rosenberg, 2006]. For some analyses, we restrict attention to 30 populations with sample size strictly larger than 15, considering a total of 813 individuals. For each locus, individuals with missing data are removed prior to the calculation of genetic dissimilarities for the locus. The dataset is the same as in the analysis of Liu et al. [2023].

5.2 Within-population dissimilarities

For each population and each locus, we compute both the theoretical and the empirical expectations $\mathbb{E}[\mathcal{D}_1^w]$ and $\mathbb{E}[\mathcal{D}_2^w]$. The number of population–locus combinations is $30 \times 783 = 23,490$. The theoretical expectation of $\mathcal{D}_1^w$ is computed by first calculating the allele frequencies of a population and then applying Eq. 1. The empirical expectation of $\mathcal{D}_1^w$ is computed by enumerating all pairs of individuals in the population, calculating their $\mathcal{D}_1^w$ dissimilarity, and averaging over all pairs. The calculation for $\mathcal{D}_2^w$ follows the same process, with Eq. 2. The calculation of theoretical and empirical expectations follows Liu et al. [2023].

We classify each locus by the number of allelic types; considering I from 4 to 14, the number of population–locus combinations is $630 \times 30 = 18,900$, with a minimum count of $1 \times 30 = 30$ for $I = 4$ and a maximum count of $119 \times 30 = 3,570$ for $I = 10$. $I = 14$ has a count of $47 \times 30 = 1,410$. The $153 \times 30 = 4,590$ combinations with a large number of distinct alleles ($I > 14$) are not shown. The theoretical values of $\mathbb{E}[\mathcal{D}_1^w]$ calculated by Eq. 1 from the allele frequencies in the data are visualized in violin plots alongside the theoretical bounds from Figures 1 in Figure 6A. Violin plots for the theoretical $\mathbb{E}[\mathcal{D}_2^w]$ (Eq. 2), empirical $\mathbb{E}[\mathcal{D}_1^w]$, and empirical $\mathbb{E}[\mathcal{D}_2^w]$ are presented in the remaining panels in a similar manner.

The theoretical $\mathbb{E}[\mathcal{D}_1^w]$ values of populations in the dataset strictly adhere to the mathematical bounds we derived for each I (Figure 6A). Similarly, the theoretical $\mathbb{E}[\mathcal{D}_2^w]$ values also adhere to the mathematical bounds (Figure 6B). Data points are concentrated toward the upper bound, a value that can lie substantially below 1. For the empirical $\mathbb{E}[\mathcal{D}_1^w]$ and $\mathbb{E}[\mathcal{D}_2^w]$, computed from empirical pairwise comparisons of diploid individuals rather than from allele frequencies, the plots are similar (Figure 6C and D). For the empirical values, it is not required that a population–locus computation produce a dissimilarity that lies below the upper bound; nevertheless, nearly all data points do lie below the upper bound ($18,896/18,900$ for $\mathbb{E}[\mathcal{D}_1^w]$, $18,900/18,900$ for $\mathbb{E}[\mathcal{D}_2^w]$).

With the largest allele frequency M held fixed, we illustrate the theoretical and empirical dissimilarities in relation to M for the case of $I = 6$ (300 population-locus combinations) in Figure 7. The theoretical dissimilarities strictly reside within the permissible region, tending to fill the space toward the upper bound (Figure 7A and B). The empirical dissimilarities generally lie within the permissible region, sometimes extending beyond it (Figure 7C and D). Data points for other values of I follow similar patterns.

5.3 Between-population dissimilarities

We next calculate the theoretical and empirical $\mathbb{E}[\mathcal{D}_1^b]$ and $\mathbb{E}[\mathcal{D}_2^b]$ for pairs of populations. Among the $\binom{30}{2} \times 783 = 340,605$ combinations of population pairs and loci, two populations share the most frequent allelic type in $169,970$ (49.9%). We consider these pairs, visualizing the bounds from Theorem 4.2 and Corollary 4.6, which provide the bounds in the scenario in which the two populations in a pair share the same most frequent allelic type at a locus.

The theoretical $\mathbb{E}[\mathcal{D}_1^b]$ is calculated by determining the allele frequencies of each population and then applying Eq. 3. For the empirical $\mathbb{E}[\mathcal{D}_1^b]$, we tabulate pairs of individuals, one from each population. We then compute their $\mathcal{D}_1^b$ dissimilarities and take an average across all pairs. The process for $\mathbb{E}[\mathcal{D}_2^b]$ is similar, using Eq. 4. The outcomes for all $\binom{30}{2} \times 4 = 1,740$ combinations for the 4 loci with $I = 5$ are considered in Figure 8 in a three-dimensional space, showing the 1,092 for which the most frequent allelic type is the same in the two populations and ordering population pairs so that $M_1 \geqslant M_2$. Comparable patterns are observed for other values of I.

As seen in the data analysis for within-population dissimilarities, the theoretical $\mathbb{E}[\mathcal{D}_1^b]$ and $\mathbb{E}[\mathcal{D}_2^b]$ lie strictly within the space bounded by the upper and lower bounds (Figure 8A and B). Note that because we only have loose bounds for $\mathbb{E}[\mathcal{D}_1^b]$, more space exists between the data points representing the theoretical $\mathbb{E}[\mathcal{D}_1^b]$ values and

the mathematical bounds. $\mathbb{E}[\mathcal{D}_2^b]$ is bounded more tightly. For the empirical $\mathbb{E}[\mathcal{D}_1^b]$ and $\mathbb{E}[\mathcal{D}_2^b]$, some points fall outside the space demarcated by the bounds (Figure 8C and D).

For the more general case, in which two populations need not have the same allelic type for the most frequent allele, we illustrate the bounds obtained in Theorems 4.3 and 4.5 for all 340,605 combinations of a population pair and locus. The theoretical and empirical $\mathbb{E}[\mathcal{D}_1^b]$ are computed as before. Results for all $\binom{30}{2} \times 4 = 1,740$ combinations with $I = 5$ appear in Figure 9.

Most of the theoretical dissimilarities congregate within the central area of the permissible region (Figure 9A and B). The permissible region is generally larger than in the case in which the most frequent allelic type is the same for a pair of populations, as seen in Figure 8. In the case of $I = 5$, the empirical mean dissimilarities all fall within the permissible range (Figure 9C and D).

5.4 Allele-sharing dissimilarity and heterozygosity

A notable property of $\mathcal{D}_2^w$ is that the expression for its expectation is exactly identical to the expression $1-\sigma_2$ for the heterozygosity of a population, as computed from its allele frequencies. We compare the theoretical and empirical allele-sharing dissimilarities to heterozygosity in two ways. First, we compute the theoretical heterozygosity for each of the 7 geographic regions; this quantity is precisely $\mathbb{E}[\mathcal{D}_2^w]$ for those regions. Next, we compute the theoretical heterozygosity for each of the 53 sampled populations, the value of $\mathbb{E}[\mathcal{D}_2^w]$ for the populations.

Figure 10A plots the theoretical $\mathbb{E}[\mathcal{D}_1^w]$ and $\mathbb{E}[\mathcal{D}_2^w]$ in relation to the theoretical heterozygosity $\mathbb{E}[\mathcal{D}_2^w]$ at the regional level, showing $7 \times 783 = 5,481$ points. The values of $\mathbb{E}[\mathcal{D}_2^w]$ follow the $y = x$ line, as x and y values are equal. The values of $\mathbb{E}[\mathcal{D}_1^w]$ lie below the $y = x$ line, in accord with Theorem 4.3, which—by specifying that two populations have identical frequencies—can be seen to demonstrate that the theoretical $\mathbb{E}[\mathcal{D}_2^w]$ provides an upper bound for the theoretical $\mathbb{E}[\mathcal{D}_1^w]$. Similar results are obtained in Figure 10B for the $53 \times 783 = 41,499$ data points at the population level.

Next, we examine the empirical $\mathbb{E}[\mathcal{D}_1^w]$ and $\mathbb{E}[\mathcal{D}_2^w]$ in relation to the theoretical heterozygosity $\mathbb{E}[\mathcal{D}_2^w]$, computing allele-sharing dissimilarity by considering pairs of individuals in a region or population. Figure 10C plots the $7 \times 783 = 5,481$ data points at the regional level, and Figure 10D plots the $53 \times 783 = 41,499$ data points at the population level. In both panels, the empirical $\mathbb{E}[\mathcal{D}_1^w]$ values are more variable than the theoretical $\mathbb{E}[\mathcal{D}_1^w]$ values in Figure 10A and B. The empirical $\mathbb{E}[\mathcal{D}_2^w]$ values do not precisely equal to the theoretical $\mathbb{E}[\mathcal{D}_2^w]$ values, though the empirical and theoretical values are quite similar.

5.5 Mathematical bounds in empirical allele-sharing dissimilarities

Visualizations of distributions of pairwise genetic dissimilarities between individuals have been important for understanding empirical genetic differences, notably in human populations [Mountain and Ramakrishnan, 2005, Rosenberg, 2011]. In Figure 5 of Rosenberg [2011], distributions of pairwise genetic dissimilarities between individuals, as computed by $\mathcal{D}_1$, are presented in various computations within regions and within geographic regions.

We reproduce Figure 5B and C of Rosenberg [2011], illustrating how the distributions of empirical genetic dissimilarities are informed by mathematical bounds. The calculation uses all 1,048 individuals and 53 populations in the data. In Figure 11A, we show the empirical distribution of allele-sharing dissimilarity between pairs of individuals within regions, averaging across all 783 loci and replotting Figure 5B of Rosenberg [2011]. In Figure 11B-H, we show the empirical distributions within single regions, plotting them alongside mathematical bounds on $\mathbb{E}[\mathcal{D}_1^w]$ for the region. The bounds are calculated from the region-wise allele frequencies for a locus according to Theorem 3.3, then averaged across all loci to obtain the mean lower and upper bounds.

In Figure 12A, we similarly show the empirical distribution of allele-sharing dissimilarity between pairs of individuals within populations, averaging across all 783 loci and replotting Figure 5C of Rosenberg [2011]. In Figure 12B-H, we show the empirical distributions of within-population dissimilarities grouped by region, plotting them alongside mathematical bounds on $\mathbb{E}[\mathcal{D}_1^w]$ for single regions. The bounds are calculated from population-wise allele frequencies for a locus via Theorem 3.3, then averaged across populations within a region and then across all loci to obtain the mean bounds for a region.

Both in Figure 11 and in Figure 12, the theorem specifies a relatively narrow range for values of $\mathbb{E}[\mathcal{D}_1^w]$, dependent on the particular values of the frequency M of the most frequent allelic type in the empirical data. Most of the probability mass lies between the lower and upper bounds. Some empirical dissimilarity values lie outside the range specified by the bounds; it is not required that an empirical dissimilarity lie between the bounds, as the bounds are obtained from an average of theoretical values across loci, whereas the empirical values are obtained for pairs of individuals. Nevertheless, the plots suggest that the mathematical bounds specify informal constraints on the distribution of empirical values of the allele-sharing dissimilarity in population-genetic data.

6 Discussion

Allele-sharing dissimilarities, computed theoretically as expectations based on allele-frequency distributions or empirically based on pairs of individuals, have often been used for studying genetic variation in populations. We have shown that as a function of properties of allele-frequency distributions, the range for expected allele-sharing dissimilarities is substantially narrower than the unit interval. Specifically, considering dissimilarities $\mathcal{D}_1$ and $\mathcal{D}_2$, we have obtained mathematical expressions for constraints on expected ASD within a population when the number of allelic types is fixed (Theorems 3.2 and 3.6), as well as when the frequency of the most frequent allelic type is also fixed (Theorems 3.3 and 3.7). Additional mathematical results concern the area of the region bounded between the smallest and largest within-population ASD values as a function of number of distinct alleles. This region increases in size with an increasing number of allelic types, converging to a value well below 1 (Propositions 3.4 and 3.8). We have also obtained corresponding expressions in between-population scenarios with the number of allelic types fixed (Propositions 4.1 and 4.4) and additionally with fixed frequencies for the most frequent allelic type (Theorems 4.2, 4.3, and 4.5, and Corollary 4.6).

In illustrations of the mathematical results using data from human populations, we have found that empirical mean ASD values reflect the theoretical expectations computed from allele-frequency distributions (Figures 6-10). The mathematical bounds on ASD values in relation to the frequency of the most frequent allelic type suggest that ASD values are expected to vary in relatively narrow ranges within the unit interval; indeed, empirical distributions of $\mathcal{D}_1$ are quite constrained (Figures 11 and 12). The mathematical results assist in explaining the relatively narrow ranges for ASD values computed in worldwide human populations, as the frequency of the most frequent allelic type constrains the between-population ASD values.

The bounds are meaningful beyond these computations. In particular, in between-population analyses, a larger range between the bounds permits more variability in the dissimilarity across pairs of populations. Such variability can be relevant in applications that rely on distinguishing the ASD values for different pairs of groups, as greater variability indicates a greater potential to distinguish values for different pairs. Theoretical properties of methods such as neighbor-joining tree construction and multidimensional scaling that rely on dissimilarity matrices, and the effects on these methods of the range between the bounds, can be explored more specifically.

We have considered two ASD measures, $\mathcal{D}_1$, which was used in the data example mimicking the analysis of Rosenberg [2011] (Figures 11 and 12), and $\mathcal{D}_2$, which provides a generalization of heterozygosity (Figure 10). For within-population computations, bounds are provided for both dissimilarities. For between-population computations, however, for $\mathcal{D}_1$, mathematical analysis is more limited. Owing to simpler mathematical expressions, tight bounds can be obtained for $\mathbb{E}[\mathcal{D}_2^b]$ in the between-population case. For $\mathcal{D}_1$, mathematical bounds in Theorem 4.2 are loose in the case that $M_1 < \frac{1}{2}$ or $M_2 < \frac{1}{2}$.

Limitations of the study include the fact that the constraints on the expected allele-sharing dissimilarity consider only on the most frequent allelic type. The frequencies of subsequent allelic types might impose constraints that might be of interest for future investigation, as occurs in various other contexts [Garud and Rosenberg, 2015, Morrison and Rosenberg, 2023]. We also note that in our empirical analysis, we average across all pairs of individuals, either within or between populations, to obtain the empirical $\mathbb{E}[\mathcal{D}]$. The reuse of each individual in multiple pairs violates the assumption that pairs are independent draws from the allele-frequency distributions, so that the empirical results do not quite mimic the computation performed theoretically. The theoretical results assume that pairs of alleles within an individual are independently drawn from the allele-frequency distribution—but empirically, the two alleles can be dependent due to inbreeding. The violation of the assumptions can contribute to deviations of the empirical observations from the theoretical values.

Additionally, our mathematical expressions are for dissimilarity values computed based on a single genetic locus. In empirical studies such as Rosenberg [2011], however, measures are typically calculated on multiple loci and averaged together. An explicitly multilocus analysis that considers the constraints at multiple loci could provide further insight into the behavior of an empirical mean across many loci.

Acknowledgments. We acknowledge National Institutes of Health grant R01 HG005855 for support.

References

N. Alcala and N. A. Rosenberg. Mathematical constraints on F_{ST}: biallelic markers in arbitrarily many populations. *Genetics*, 206:1581–1600, 2017.

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

N. Alcala and N. A. Rosenberg. Mathematical constraints on F_{ST}: multiallelic markers in arbitrarily many populations. *Philosophical Transactions of the Royal Society B*, 377:20200414, 2022.

A. J. Aw and N. A. Rosenberg. Bounding measures of genetic similarity and diversity using majorization. *Journal of Mathematical Biology*, 77:711–737, 2018.

M. D. Edge and N. A. Rosenberg. Upper bounds on F_{ST} in terms of the frequency of the most frequent allele and total homozygosity: the case of a specified number of alleles. *Theoretical Population Biology*, 97:20–34, 2014.

X. Gao and E. R. Martin. Using allele sharing distance for detecting human population stratification. *Human Heredity*, 68: 182–191, 2009.

N. R. Garud and N. A. Rosenberg. Enhancing the mathematical properties of new haplotype homozygosity statistics for the detection of selective sweeps. *Theoretical Population Biology*, 102:94–101, 2015.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, Cambridge, 1952.

M. Jakobsson, M. D. Edge, and N. A. Rosenberg. The relationship between F_{ST} and the frequency of the most frequent allele. *Genetics*, 193:515–528, 2013.

X. Liu, Z. Ahsan, T. K. Martheswaran, and N. A. Rosenberg. When is the allele-sharing dissimilarity between two populations exceeded by the allele-sharing dissimilarity of a population with itself? *Statistical Applications in Genetics and Molecular Biology*, 22:20230004, 2023.

A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, New York, 2nd edition, 2010.

M. L. Morrison and N. A. Rosenberg. Mathematical bounds on Shannon entropy given the abundance of the ith most abundant taxon. *Journal of Mathematical Biology*, 87:76, 2023.

J. L. Mountain and L. L. Cavalli-Sforza. Multilocus genotypes, a tree of individuals, and human evolutionary history. *American Journal of Human Genetics*, 61:705–718, 1997.

J. L. Mountain and U. Ramakrishnan. Impact of human population history on distributions of individual-level genetic distance. *Human Genomics*, 2:1–16, 2005.

S. B. Reddy and N. A. Rosenberg. Refining the relationship between homozygosity and the frequency of the most frequent allele. *Journal of Mathematical Biology*, 64:87–108, 2012.

N. A. Rosenberg. Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of Human Genetics*, 70:841–847, 2006.

N. A. Rosenberg. A population-genetic perspective on the similarities and differences among worldwide human populations. *Human Biology*, 83:659–684, 2011.

N. A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, 1:e70, 2005.

D. J. Witherspoon, S. Wooding, A. R. Rogers, E. E. Marchani, W. S. Watkins, M. A. Batzer, and L. B. Jorde. Genetic similarities within and between human populations. *Genetics*, 176:351–359, 2007.

A Proof of Lemma 3.1

$f(\mathbf{p}) = 1 - 2\sigma_2 + 2\sigma_3 - \sigma_4$ is symmetric in the p_i by construction, and all first partial derivatives $\frac{\partial f}{\partial p_i}$ exist. By Theorem 2.3, to show that f is Schur-concave, it suffices to show that $(p_1 - p_2)(\frac{\partial f}{\partial p_1} - \frac{\partial f}{\partial p_2}) \leqslant 0$ for all $\mathbf{p} \in \Delta^{I-1}$.

We have $\frac{\partial \sigma_2}{\partial p_1} = 2p_1$, $\frac{\partial \sigma_3}{\partial p_1} = 3p_1^2$, and $\frac{\partial \sigma_4}{\partial p_1} = 4p_1^3$, so that

$$\begin{aligned}(p_1 - p_2)\left(\frac{\partial f(\mathbf{p})}{\partial p_1} - \frac{\partial f(\mathbf{p})}{\partial p_2}\right) &= (p_1 - p_2)\big[-4(p_1 - p_2) + 6(p_1^2 - p_2^2) - 4(p_1^3 - p_2^3)\big] \\ &= -(p_1 - p_2)^2\left[4 - 6(p_1 + p_2) + 2(p_1 + p_2)^2 + 2(p_1^2 + p_2^2)\right].\end{aligned}$$

For $0 \leqslant x \leqslant 1$, $4 - 6x + 2x^2 \geqslant 0$ with equality if and only if $x = 1$. Hence $4 - 6(p_1 + p_2) + 2(p_1 + p_2)^2 \geqslant 0$ always holds for $0 \leqslant p_1 + p_2 \leqslant 1$. We then have $(p_1 - p_2)(\frac{\partial f}{\partial p_1} - \frac{\partial f}{\partial p_2}) \leqslant 0$. By Theorem 2.3, f is Schur-concave.

To verify strict Schur-concavity, note that $4 - 6(p_1 + p_2) + 2(p_1 + p_2)^2 = 0$ requires $p_1 + p_2 = 1$, so that $4-6(p_1+p_2)+2(p_1+p_2)^2+2(p_1^2+p_2^2) > 0$ for all permissible (p_1, p_2): either $p_1+p_2 \neq 1$ and $4-6(p_1+p_2)+2(p_1+p_2)^2 > 0$, or $p_1 + p_2 = 1$ and $2(p_1 + p_2)^2 > 0$. We conclude that $(p_1 - p_2)(\frac{\partial f}{\partial p_1} - \frac{\partial f}{\partial p_2}) = 0$ implies $p_1 = p_2$.

B Proof of Proposition 3.4

The desired area is calculated by considering M in segments. For $M \in [\frac{1}{i}, \frac{1}{i-1})$, $\lceil M^{-1} \rceil = i$. The area then equals

$$S_{\mathbb{E}[\mathcal{D}_1^w]}(I) = \int_{M=\frac{1}{I}}^{1} f(\mathbf{p}_{\max})\, dM - \sum_{i=2}^{I} \int_{M=\frac{1}{i}}^{\frac{1}{i-1}} f(\mathbf{p}_{\min})\, dM. \tag{14}$$

The first term is

$$\begin{aligned}\int_{M=\frac{1}{I}}^{1} f(\mathbf{p}_{\max})\, dM &= \int_{M=\frac{1}{I}}^{1} \left[1 - 2M^2 + 2M^3 - M^4 - \frac{2(1-M)^2}{I-1} + \frac{2(1-M)^3}{(I-1)^2} - \frac{(1-M)^4}{(I-1)^3}\right] dM \\ &= \frac{19}{30} - \frac{5}{3I} + \frac{11}{6I^2} - \frac{6}{5I^3} + \frac{2}{5I^4}.\end{aligned} \tag{15}$$

The second term is

$$\begin{aligned}\sum_{i=2}^{I} \int_{M=\frac{1}{i}}^{\frac{1}{i-1}} f(\mathbf{p}_{\min})\, dM &= \sum_{i=2}^{I} \int_{M=\frac{1}{i}}^{\frac{1}{i-1}} \big[1 - (i-1)(2M^2 - 2M^3 + M^4) \\ &\qquad -2[1-(i-1)M]^2 + 2[1-(i-1)M]^3 - [1-(i-1)M]^4\big] \\ &= \sum_{i=2}^{I} \frac{12 - 72i + 199i^2 - 335i^3 + 310i^4 - 150i^5 + 30i^6}{30i^4(i-1)^4} \\ &= -\frac{19}{30I} + \frac{31}{30I^2} - \frac{4}{5I^3} + \frac{2}{5I^4} + \sum_{i=2}^{I}\left[\frac{11}{30(i-1)^2} - \frac{3}{10(i-1)^3} + \frac{1}{5(i-1)^4}\right].\end{aligned} \tag{16}$$

Subtracting Eq. 16 from Eq. 15 in Eq. 14, we obtain the quantity in Eq. 5.

C Proof of Theorem 4.2

First, for the upper bound, because $\sum_{i=2}^{I} p_i = 1 - M_1$ and $\sum_{i=2}^{I} q_i = 1 - M_2$, Eq. 3 can be written

$$\begin{aligned}\mathbb{E}[\mathcal{D}_1^b] &= \big(1 - 2M_1M_2 + M_1^2M_2 + M_1M_2^2 - M_1^2M_2^2\big) + \left(-2\sum_{i=2}^{I} p_iq_i + \sum_{i=2}^{I} p_i^2q_i + \sum_{i=2}^{I} p_iq_i^2 - \sum_{i=2}^{I} p_i^2q_i^2\right) \\ &= \big(1 - 2M_1M_2 + M_1^2M_2 + M_1M_2^2 - M_1^2M_2^2\big) - \sum_{i=2}^{I} p_iq_i\,(2 - p_i - q_i + p_iq_i) \\ &\leqslant 1 - 2M_1M_2 + M_1^2M_2 + M_1M_2^2 - M_1^2M_2^2.\end{aligned} \tag{17}$$

The last inequality holds because $0 \leqslant p_i < 1$ and $0 \leqslant q_i < 1$ for all $i = 2, 3, \ldots, I$, so that $2 - p_i - q_i + p_i q_i > 0$, $p_i q_i \geqslant 0$, and $\sum_{i=2}^{I} p_i q_i (2 - p_i - q_i + p_i q_i) \geqslant 0$. Equality with the upper bound requires that for all i, $2 \leqslant i \leqslant I$, $p_i q_i = 0$. That is, for all $i = 2, 3, \ldots, I$, $p_i = 0$ or $q_i = 0$, so that allele 1 is the only allele shared between populations.

Next, for the lower bound,

$$
\begin{aligned}
&-2(1-M_1)(1-M_2) + (1-M_1)^2(1-M_2) + (1-M_1)(1-M_2)^2 - (1-M_1)^2(1-M_2)^2 \\
&\quad - \left(-2 \sum_{i=2}^{I} p_i q_i + \sum_{i=2}^{I} p_i^2 q_i + \sum_{i=2}^{I} p_i q_i^2 - \sum_{i=2}^{I} p_i^2 q_i^2 \right) \\
&= \sum_{i=2}^{I} p_i q_i (1 - p_i) + \sum_{i=2}^{I} p_i q_i (1 - q_i) - (1 - M_1)(1 - M_2)(M_1 + M_2) + \sum_{i=2}^{I} p_i^2 q_i^2 - \left(\sum_{i=2}^{I} p_i \right)^2 \left(\sum_{i=2}^{I} q_i \right)^2 \\
&\leqslant \sum_{i=2}^{I} p_i q_i (1 - p_i) + \sum_{i=2}^{I} p_i q_i (1 - q_i) - (1 - M_1)(1 - M_2)(M_1 + M_2) + \left(\sum_{i=2}^{I} p_i \right)^2 \left(\sum_{i=2}^{I} q_i \right)^2 - \left(\sum_{i=2}^{I} p_i \right)^2 \left(\sum_{i=2}^{I} q_i \right)^2 \\
&= \sum_{i=2}^{I} p_i \big[q_i (1 - q_i) - M_2 (1 - M_2) \big] + \sum_{i=2}^{I} q_i \big[p_i (1 - p_i) - M_1 (1 - M_1) \big] \qquad (18) \\
&\leqslant 0.
\end{aligned}
$$

The first inequality uses the fact that the p_i and q_i are all non-negative, so that $(\sum_{i=2}^{I} p_i)^2 (\sum_{i=2}^{I} q_i)^2 \geqslant \sum_{i=2}^{I} p_i^2 q_i^2$. The last inequality uses the fact that $p_i \leqslant M_1$, $p_i \leqslant 1 - M_1$, $q_i \leqslant M_2$, and $q_i \leqslant 1 - M_2$. The function $f(x) = x(1-x)$ is nondecreasing for $x \in [0, \frac{1}{2})$, and one of M_1 and $1 - M_1$ must lie in $[0, \frac{1}{2}]$, so $p_i \leqslant M_1$ and $p_i \leqslant 1 - M_1$ implies $f(p_i) \leqslant f(M_1)$; analogously, $f(q_i) \leqslant f(M_2)$.

Applying Eq. 17, we therefore have

$$
\begin{aligned}
\mathbb{E}[\mathcal{D}_1^b] &\geqslant \big(1 - 2M_1M_2 + M_1^2M_2 + M_1M_2^2 - M_1^2M_2^2\big) \\
&\qquad - 2(1-M_1)(1-M_2) + (1-M_1)^2(1-M_2) + (1-M_1)(1-M_2)^2 - (1-M_1)^2(1-M_2)^2 \\
&= M_1 + M_2 - 4M_1M_2 + 2M_1^2M_2 + 2M_1M_2^2 - 2M_1^2M_2^2.
\end{aligned}
$$

Equality with the lower bound requires $\big(\sum_{i=2}^{I} p_i\big)^2 \big(\sum_{i=2}^{I} q_i\big)^2 = \sum_{i=2}^{I} p_i^2 q_i^2$. Because $p_2 \geqslant p_3 \geqslant \ldots \geqslant p_I$, this condition requires $p_2 = 1 - p_1 = 1 - M_1$ and hence $q_2 = 1 - q_1 = 1 - M_2$, making use of assumptions $M_1 \geqslant \frac{1}{2}$ and $M_2 \geqslant \frac{1}{2}$. Equality with the lower bound also requires that the expression in Eq. 18 equal 0; allele-frequency distributions $(p_1, p_2, p_3, \ldots, p_I) = (M_1, 1 - M_1, 0, \ldots, 0)$ and $(q_1, q_2, q_3, \ldots, q_I) = (M_2, 1 - M_2, 0, \ldots, 0)$ produce a value of 0 in Eq. 18.

D Proof of Theorem 4.3

We write $\mathbb{E}[\mathcal{D}_1^b]$ in the form

$$
\mathbb{E}[\mathcal{D}_1^b] = 1 - 2 \sum_{i=1}^{I} p_i q_i + \sum_{i=1}^{I} p_i^2 q_i + \sum_{i=1}^{I} p_i q_i^2 - \sum_{i=1}^{I} p_i^2 q_i^2 = 1 - \sum_{i=1}^{I} p_i q_i - \sum_{i=1}^{I} p_i (1 - p_i) q_i (1 - q_i).
$$

For the upper bound, because $\mathbb{E}[\mathcal{D}_1^b] \leqslant 1 - \sum_{i=1}^{I} p_i q_i = \mathbb{E}[\mathcal{D}_2^b]$, the upper bound of $\mathbb{E}[\mathcal{D}_2^b]$ can also serve as a (loose) upper bound for $\mathbb{E}[\mathcal{D}_1^b]$.

To obtain a loose lower bound, we must bound from above the quantity $\sum_{i=1}^{I} p_i (1 - p_i) q_i (1 - q_i)$ given $\max\{p_1, p_2, \ldots, p_I\} = M_1$ and $\max\{q_1, q_2, \ldots, q_I\} = M_2$. First, note that for $p_1, p_2, \ldots, p_I$ with $p_1 \geqslant p_2 \geqslant \ldots p_I \geqslant 0$ and $\sum_{i=1}^{I} p_i = 1$, we have $p_i(1-p_i) \geqslant p_j(1-p_j)$ for $i < j$. This result follows because $f(x) = x(1-x)$ is maximized at $x = \frac{1}{2}$, declining symmetrically around the maximum, and p_j lies farther from $\frac{1}{2}$ than does p_i; the claim is verified in two cases, $p_i \geqslant \frac{1}{2}$, for which $p_i - \frac{1}{2} \leqslant \frac{1}{2} - p_j$, and $p_i < \frac{1}{2}$, for which $\frac{1}{2} - p_i \leqslant \frac{1}{2} - p_j$.

It follows that for each i, $p_i(1-p_i) \leqslant M_1(1-M_1)$ and $q_i(1-q_i) \leqslant M_2(1-M_2)$, so that $\sum_{i=1}^{I} p_i(1-p_i)q_i(1-q_i) \leqslant I M_1(1-M_1)M_2(1-M_2)$.

E Proof of Theorem 4.5

Theorem 4.5 states the bounds on $\mathbb{E}[\mathcal{D}_2^b]$ and gives sufficient conditions on the p_i and q_i at which the bounds are reached—given an upper bound M_1 on the p_i and an upper bound M_2 on the q_i (also assuming $\sum_{i=1}^I p_i = \sum_{i=1}^I q_i = 1$). The upper bounds need not occur at the same allele.

The proof proceeds by a series of lemmas. Informally, Lemma E.1 shows that for fixed a_i, we can reduce the sum of products $\sum_{i=1}^I a_i b_i$ by a particular choice of the value of a specific frequency b_ℓ (if it is not already optimized).

Lemma E.1. *Suppose a collection of $I \geqslant 2$ fixed non-negative values $a_1, a_2, \ldots, a_I$ is given, with $a_1 \geqslant a_2 \geqslant \ldots \geqslant a_I$. Suppose $b_1, b_2, \ldots, b_I$ are non-negative values satisfying three conditions:*
(1) monotonicity, $b_1 \leqslant b_2 \leqslant \ldots \leqslant b_I$;
(2) fixed total sum, $\sum_{i=1}^I b_i = B$; and
(3) boundedness from above, $b_i \leqslant b^*$ for all $i = 1, 2, \ldots, I$, where $\frac{B}{I} \leqslant b^* < B$.
Consider ℓ with $2 \leqslant \ell \leqslant I$ and $(I-\ell)b^* < B$. Suppose $b_i = b^*$ for each i with $\ell < i \leqslant I$, and suppose $b_\ell < \min\left(b^*, B-(I-\ell)b^*\right)$. Then there exists a set of values $b'_1, b'_2, \ldots, b'_I$ with $b'_i = b^*$ for each i with $\ell < i \leqslant I$, satisfying conditions (1), (2), and (3), such that $b'_\ell = \min\left(b^*, B-(I-\ell)b^*\right)$, and

$$\sum_{i=1}^I a_i b'_i \leqslant \sum_{i=1}^I a_i b_i. \tag{19}$$

Proof. For convenience, write $s = \min\left(b^*, B-(I-\ell)b^*\right)$, so that $b'_\ell = s > 0$. Let $b_\ell < s = b'_\ell$. We have $b_i = b'_i = b^*$ for each i with $\ell < i \leqslant I$. Let $x = b'_\ell - b_\ell$, a positive quantity representing the difference between the value we will place in the ℓth entry in our new sequence and the value in the current sequence. Because $\sum_{i=1}^I b_i = \sum_{i=1}^I b'_i = B$,

$$x = b'_\ell - b_\ell = \sum_{i=1}^{\ell-1}(b_i - b'_i) > 0.$$

Let k be the unique index that satisfies $\sum_{i=1}^{k-1} b_i \leqslant x$ and $\sum_{i=1}^{k} b_i > x$. We set the values of b'_i so that $b'_i = 0$ for each i with $1 \leqslant i \leqslant k-1$, $b'_k = b_k - (x - \sum_{i=1}^{k-1} b_i)$, $b'_i = b_i$ for $k+1 \leqslant i \leqslant \ell-1$, and $b'_i = b^*$ for $\ell \leqslant i \leqslant I$.

Note that $k \leqslant \ell$ always holds. For contradiction, suppose $k > \ell$. Then $\sum_{i=1}^{k-1} b_i = \sum_{i=1}^{\ell-1} b_i + \sum_{i=\ell}^{k-1} b_i = \left[(\sum_{i=1}^{\ell-1} b'_i) + x\right] + \sum_{i=\ell}^{k-1} b_i$. We have $\sum_{i=\ell+1}^I b_i = \sum_{i=\ell+1}^I b^* = (I-\ell)b^* < B$; because $\sum_{i=1}^I b_i = B$, it follows that $\sum_{i=1}^{\ell} b_i = B - \sum_{i=\ell+1}^I b_i > 0$. Next, because $b_\ell \geqslant b_i$ for each i with $1 \leqslant i \leqslant \ell-1$, we have $b_\ell > 0$. As a result, $\sum_{i=1}^{k-1} b_i = \left[(\sum_{i=1}^{\ell-1} b'_i) + x\right] + \sum_{i=\ell}^{k-1} b_i \geqslant x + b_\ell > x$, contradicting the condition $\sum_{i=1}^{k-1} b_i \leqslant x$ in the definition of k.

We have constructed a sequence of values b'_i that continues to satisfy the monotonicity, fixed-total-sum, and boundedness-from-above conditions. (1) For monotonicity, $b'_i = 0$ for $1 \leqslant i \leqslant k-1$, $b'_k \leqslant b_k \leqslant b_i = b'_i$ for $k+1 \leqslant i \leqslant \ell-1$, $b'_{\ell-1} = b_{\ell-1} \leqslant b_\ell < b'_\ell$, and $b'_\ell = b^* = b'_i$ for $\ell+1 \leqslant i \leqslant I$. (2) For fixed total sum, $\sum_{i=1}^I b'_i = (\sum_{i=1}^{k-1} b'_i) + b'_k + (\sum_{i=k+1}^{\ell-1} b'_i) + b'_\ell + (\sum_{i=\ell+1}^I b'_i) = 0 + [b_k - (x - \sum_{i=1}^{k-1} b_i)] + (\sum_{i=k+1}^{\ell-1} b_i) + (b_\ell + x) + (\sum_{i=\ell+1}^I b_i) = \sum_{i=1}^I b_i = B$. (3) For boundedness from above, $b'_i = 0 < b^*$ for $1 \leqslant i \leqslant k-1$, $b'_k \leqslant b_k < b^*$, $b'_i = b_i \leqslant b_\ell < b^*$ for $k+1 \leqslant i \leqslant \ell-1$, and $b'_i = b^*$ for $\ell \leqslant i \leqslant I$.

It remains to show that Eq. 19 holds. We have

$$\begin{aligned}
\sum_{i=1}^{I} a_i b'_i - \sum_{i=1}^{I} a_i b_i &= \left(\sum_{i=1}^{\ell-1} a_i b'_i\right) + a_\ell b'_\ell - \left(\sum_{i=1}^{\ell-1} a_i b_i\right) - a_\ell b_\ell \\
&= \left[\sum_{i=1}^{\ell-1} a_i (b'_i - b_i)\right] + a_\ell (b'_\ell - b_\ell) \\
&= \left[\sum_{i=1}^{k-1} a_i (0 - b_i)\right] + a_k \left[\left[b_k - \left(x - \sum_{i=1}^{k-1} b_i\right)\right] - b_k\right] + a_\ell x \\
&= a_\ell x - \left(\sum_{i=1}^{k-1} a_i b_i\right) - a_k \left(x - \sum_{i=1}^{k-1} b_i\right) \\
&= a_\ell \left[\sum_{i=1}^{k-1} b_i + \left(x - \sum_{i=1}^{k-1} b_i\right)\right] - \left(\sum_{i=1}^{k-1} a_i b_i\right) - a_k \left(x - \sum_{i=1}^{k-1} b_i\right) \\
&= \left[\sum_{i=1}^{k-1} (a_\ell - a_i) b_i\right] + (a_\ell - a_k)\left(x - \sum_{i=1}^{k-1} b_i\right) \\
&\leqslant 0.
\end{aligned}$$

In the last step, the inequality holds because $k \leqslant \ell$ and the a_i are monotonically decreasing, so that $a_\ell \leqslant a_i$ for all i, $1 \leqslant i \leqslant \ell$. □

Lemma E.2 is similar to Lemma E.1, but in the reverse direction. It shows that for fixed a_i, we can *increase* $\sum_{i=1}^{I} a_i b_i$ by a particular choice of the value of a specific frequency b_ℓ (if it is not already optimized).

Lemma E.2. *Suppose a collection of $I \geqslant 2$ fixed non-negative values $a_1, a_2, \ldots, a_I$ is given, with $a_1 \geqslant a_2 \geqslant \ldots \geqslant a_I$. Suppose $b_1, b_2, \ldots, b_I$ are non-negative values satisfying three conditions:*
(1) monotonicity, $b_1 \geqslant b_2 \geqslant \ldots \geqslant b_I$;
(2) fixed total sum, $\sum_{i=1}^{I} b_i = B$; and
(3) boundedness from above, $b_i \leqslant b^*$ for all $i = 1, 2, \ldots, I$, where $\frac{B}{I} \leqslant b^* < B$.
Consider ℓ with $1 \leqslant \ell \leqslant I - 1$ and $(\ell - 1)b^* < B$. Suppose $b_i = b^*$ for each i with $1 \leqslant i < \ell$, and suppose $b_\ell < \min\left(b^*, B - (\ell - 1)b^*\right)$. Then there exists a set of values $b'_1, b'_2, \ldots, b'_I$ with $b'_i = b^*$ for each i with $1 \leqslant i < \ell$, satisfying conditions (1), (2), and (3), such that $b'_\ell = \min\left(b^*, B - (\ell - 1)b^*\right)$, and

$$\sum_{i=1}^{I} a_i b'_i \geqslant \sum_{i=1}^{I} a_i b_i. \tag{20}$$

Proof. The proof is similar to that of Lemma E.1. Write $s = \min\left(b^*, B - (\ell - 1)b^*\right)$, so that $b'_\ell = s > 0$. Let $b_\ell < s = b'_\ell$. We now have $b_i = b'_i = b^*$ for each i with $1 \leqslant i < \ell$. Let $x = b'_\ell - b_\ell$, a positive quantity representing the difference between the value we place in the ℓth entry in our new sequence and the value in the current sequence. Because $\sum_{i=1}^{I} b_i = \sum_{i=1}^{I} b'_i = B$,

$$x = b'_\ell - b_\ell = \sum_{i=\ell+1}^{I} (b_i - b'_i) > 0.$$

Let k be the unique index that satisfies $\sum_{i=k+1}^{I} b_i \leqslant x$ and $\sum_{i=k}^{I} b_i > x$. We set the values of b'_i so that $b'_i = b^*$ for $1 \leqslant i \leqslant \ell$, $b'_i = b_i$ for $\ell + 1 \leqslant i \leqslant k - 1$, $b'_k = b_k - (x - \sum_{i=k+1}^{I} b_i)$, and $b'_i = 0$ for each i with $k + 1 \leqslant i \leqslant I$.

We show $k \geqslant \ell$. For contradiction, suppose $k < \ell$. Then $\sum_{i=k+1}^{I} b_i = \sum_{i=k+1}^{\ell} b_i + \sum_{i=\ell+1}^{I} b_i = \sum_{i=k+1}^{\ell} b_i + \left[(\sum_{i=\ell+1}^{I} b'_i) + x\right]$. We have $\sum_{i=1}^{\ell-1} b_i = \sum_{i=1}^{\ell-1} b^* = (\ell - 1)b^* < B$; because $\sum_{i=1}^{I} b_i = B$, it follows that $\sum_{i=\ell}^{I} b_i = B - \sum_{i=1}^{\ell-1} b_i > 0$. Next, because $b_\ell \geqslant b_i$ for each i with $\ell + 1 \leqslant i \leqslant I$, we have $b_\ell > 0$. As a result, $\sum_{i=k+1}^{I} b_i = \sum_{i=k+1}^{\ell} b_i + \left[(\sum_{i=\ell+1}^{I} b'_i) + x\right] \geqslant b_\ell + x > x$, contradicting the condition $\sum_{i=k+1}^{I} b_i \leqslant x$ in the definition of k.

The constructed sequence of values b'_i continues to satisfy the monotonicity, fixed-total-sum, and boundedness-from-above conditions. (1) For monotonicity, $b'_\ell = b^* = b'_i$ for $1 \leqslant i \leqslant \ell - 1$, $b'_\ell > b_\ell \geqslant b_{\ell+1} = b'_{\ell+1}$, $b'_i = b_i \geqslant b_k \geqslant b'_k$ for $\ell + 1 \leqslant i \leqslant k - 1$, and $b'_i = 0$ for $k + 1 \leqslant i \leqslant I$. (2) For fixed total sum, $\sum_{i=1}^{I} b'_i = (\sum_{i=1}^{\ell-1} b'_i) + b'_\ell + (\sum_{i=\ell+1}^{k-1} b'_i) +$

$b'_k + (\sum_{i=k+1}^{I} b'_i) = (\sum_{i=1}^{\ell-1} b_i) + (b_\ell + x) + (\sum_{i=\ell+1}^{k-1} b_i) + [b_k - (x - \sum_{i=k+1}^{I} b_i)] + 0 = B$. (3) For boundedness from above, $b'_i = b^*$ for $1 \leqslant i \leqslant \ell$, $b'_i = b_i \leqslant b_\ell < b^*$ for $\ell + 1 \leqslant i \leqslant k-1$, $b'_k \leqslant b_k < b^*$, and $b'_i = 0 < b^*$ for $k+1 \leqslant i \leqslant I$.

It remains to show that Eq. 20 holds. We have

$$\begin{aligned}
\sum_{i=1}^{I} a_i b'_i - \sum_{i=1}^{I} a_i b_i &= a_\ell b'_\ell + \left(\sum_{i=\ell+1}^{I} a_i b'_i \right) - a_\ell b_\ell - \left(\sum_{i=\ell+1}^{I} a_i b_i \right) \\
&= a_\ell (b'_\ell - b_\ell) + \left[\sum_{i=\ell+1}^{I} a_i (b'_i - b_i) \right] \\
&= a_\ell x + a_k \left[\left[b_k - (x - \sum_{i=k+1}^{I} b_i) \right] - b_k \right] + \left[\sum_{i=k+1}^{I} a_i (0 - b_i) \right] \\
&= a_\ell x - a_k x + \left(a_k \sum_{i=k+1}^{I} b_i \right) - \left(\sum_{i=k+1}^{I} a_i b_i \right) \\
&= a_\ell \left[\sum_{i=k+1}^{I} b_i + \left(x - \sum_{i=k+1}^{I} b_i \right) \right] - \left(\sum_{i=k+1}^{I} a_i b_i \right) - a_k \left(x - \sum_{i=k+1}^{I} b_i \right) \\
&= \left[\sum_{i=k+1}^{I} (a_\ell - a_i) b_i \right] + (a_\ell - a_k) \left(x - \sum_{i=k+1}^{I} b_i \right) \\
&\geqslant 0.
\end{aligned}$$

In the last step, the inequality holds because $k \geqslant \ell$ and the a_i are monotonically decreasing, so that $a_\ell \geqslant a_i$ for all i, $\ell \leqslant i \leqslant I$. □

Lemma E.3 now uses Lemmas E.1 and E.2 to find the minimum and maximum of the sum of products $\sum_{i=1}^{I} a_i b_i$, allowing both a_i and b_i to vary.

Lemma E.3. *Consider all possible sets of non-negative real numbers $\{a_1, a_2, \ldots, a_I\}$ and $\{b_1, b_2, \ldots, b_I\}$ with fixed sums $\sum_{i=1}^{I} a_i = A$ and $\sum_{i=1}^{I} b_i = B$, where $I \geqslant 2$, $A > 0$, and $B > 0$. Suppose that the a_i are non-decreasing, with $a_1 \geqslant a_2 \geqslant \ldots \geqslant a_I$ and that the b_i are monotonic, with $b_1 \geqslant b_2 \geqslant \ldots \geqslant b_I$ or $b_1 \leqslant b_2 \leqslant \ldots \leqslant b_I$. Suppose also that $a_i \leqslant a^*$ and $b_i \leqslant b^*$ for all i, with $0 < a^* < A$ and $0 < b^* < B$. Let $\alpha = \lceil A/a^* \rceil$ and $\beta = \lceil B/b^* \rceil$. The values of I, A, B, a^*, and b^* are fixed and given. Consider the following conditions:*

1. *$a_i = a^*$ for $1 \leqslant i \leqslant \alpha - 1$, $a_\alpha = A - (\alpha - 1)a^*$, and $a_i = 0$ for $\alpha + 1 \leqslant i \leqslant I$.*
2. *$b_i = 0$ for $1 \leqslant i \leqslant I - \beta$, $b_{I-\beta+1} = B - (\beta - 1)b^*$, and $b_i = b^*$ for $I - \beta + 2 \leqslant i \leqslant I$.*
3. *$b_i = b^*$ for $1 \leqslant i \leqslant \beta - 1$, $b_\beta = B - (\beta - 1)b^*$, and $b_i = 0$ for $\beta + 1 \leqslant i \leqslant I$.*

Then (i) $\sum_{i=1}^{I} a_i b_i$ achieves its maximal value if Conditions 1 and 3 hold. (ii) $\sum_{i=1}^{I} a_i b_i$ achieves its minimal value if Conditions 1 and 2 hold.

Proof. (i) For the upper bound, by the rearrangement inequality (Theorem 2.6), if the a_i are fixed with $a_1 \geqslant a_2 \geqslant \ldots \geqslant a_I$ and the b_i are free to vary subject to $b_1 \geqslant b_2 \geqslant \ldots \geqslant b_I$ (and $0 \leqslant b_i \leqslant b^*$, $\sum_{i=1}^{I} b_i = B$), then for each permutation σ of $(1, 2, \ldots, I)$,

$$\sum_{i=1}^{I} a_i b_i \geqslant \sum_{i=1}^{I} a_i b_{\sigma(i)}.$$

In other words, to maximize $\sum_{i=1}^{I} a_i b_i$, it suffices to proceed by assuming that $b_1 \geqslant b_2 \geqslant \ldots \geqslant b_I$.

We apply Lemma E.2 with $\ell = 1$. We conclude that the maximal value of $\sum_{i=1}^{I} a_i b_i$ is achieved with $b_1 = \min(b^*, B) = b^*$. Fixing $b_1 = b^*$, we next apply Lemma E.2 with $\ell = 2$. We find that the maximal value of $\sum_{i=1}^{I} a_i b_i$ achieved at $b_2 = \min(b^*, B - b^*)$.

We proceed by fixing $b_\ell = b^*$ for each $\ell = 3, 4, \ldots, \lceil B/b^* \rceil - 1$, repeatedly applying Lemma E.2 provided $\ell b^* < B$—that is, while $\ell < B/b^*$, or $\ell \leqslant \lceil B/b^* \rceil - 1$, and continuing to assign $b_\ell = b^*$. The next value of ℓ is $\ell = \lceil B/b^* \rceil$. If $\lceil B/b^* \rceil \leqslant I - 1$, then Lemma E.2 yields $b_\ell = B - \big(\lceil B/b^* \rceil - 1\big)b^*$ and $b_i = 0$ for all $i > \ell$. If $\lceil B/b^* \rceil = I$, then we have reached a trivial case in which $b_I = B - (I - 1)b^*$ and $b_i = b^*$ for all i, $1 \leqslant i \leqslant I - 1$.

We arrive at Condition 3: with the a_i in non-increasing order held constant, the b_i that satisfy Condition 3 produce the maximum of $\sum_{i=1}^{I} a_i b_i$. By symmetry, we fix the b_i as in Condition 3 and apply Lemma E.2 with the roles of the a_i and b_i interchanged. We find analogously that the a_i follow Condition 1.

(ii) For the lower bound, by the rearrangement inequality (Theorem 2.6), if the a_i are fixed with $a_1 \geqslant a_2 \geqslant \ldots \geqslant a_I$ and the b_i are free to vary subject to $b_1 \leqslant b_2 \leqslant \ldots \leqslant b_I$ (and $0 \leqslant b_i \leqslant b^*$, $\sum_{i=1}^{I} b_i = B$), then for each permutation σ of $(1, 2, \ldots, I)$,

$$\sum_{i=1}^{I} a_i b_i \leqslant \sum_{i=1}^{I} a_i b_{\sigma(i)}.$$

In other words, to minimize $\sum_{i=1}^{I} a_i b_i$, it suffices to proceed by assuming that $b_1 \leqslant b_2 \leqslant \ldots \leqslant b_I$.

We apply Lemma E.1 with $\ell = I$. We conclude that the minimal value of $\sum_{i=1}^{I} a_i b_i$ is achieved with $b_I = \min(b^*, B) = b^*$. Fixing $b_I = b^*$, we next apply Lemma E.1 with $\ell = I - 1$. We find that the minimal value of $\sum_{i=1}^{I} a_i b_i$ is achieved at $b_{I-1} = \min(b^*, B - b^*)$.

We proceed by fixing $b_\ell = b^*$ for each $\ell = I - 2, I - 3, \ldots, I - \lceil B/b^* \rceil + 2$, repeatedly applying Lemma E.1 provided $\sum_{i=\ell}^{I} b_i < B$, that is, while $(I - \ell + 1)b^* < B$, or $\ell \geqslant I - \lceil B/b^* \rceil + 2$, and continuing to assign $b_\ell = b^*$. The next value of ℓ is $\ell = I - \lceil B/b^* \rceil + 1$. If $I - \lceil B/b^* \rceil + 1 \geqslant 2$, then Lemma E.1 yields $b_\ell = B - (\lceil B/b^* \rceil - 1)b^*$, and $b_i = 0$ for all $i < \ell$. If $I - \lceil B/b^* \rceil + 1 = 1$, then we have reached a trivial case in which $b_1 = B - (I - 1)b^*$ and $b_i = b^*$ for all i, $2 \leqslant i \leqslant I$.

We arrive at Condition 2: with the a_i in non-increasing order held constant, the b_i that satisfy Condition 2 produce the minimum of $\sum_{i=1}^{I} a_i b_i$. By symmetry, if we fix the b_i as in Condition 2, and write the b_i in reverse order, with $c_i = b_{I+1-i}$, then we can apply Lemma E.1 with the c_i in the role of the a_i and the reversed a_i, or $d_i = a_{I+1-i}$, in the role of the b_i. We obtain that the d_i follow Condition 2, and consequently, that the $a_i = d_{I+1-i}$ follow Condition 1. □

Proof of Theorem 4.5 The function $\ell(\mathbf{p}, \mathbf{q}) = 1 - \rho_{11} = 1 - \sum_{i=1}^{I} p_i q_i$, with $\sum_{i=1}^{I} p_i = 1$ and $\sum_{i=1}^{I} q_i = 1$, is minimized when $\sum_{i=1}^{I} p_i q_i$ is maximized, and maximized when $\sum_{i=1}^{I} p_i q_i$ is minimized.

(i) Via Lemma E.3i, $\sum_{i=1}^{I} p_i q_i$ reaches its upper bound if $\mathbf{p}_* = (M_1, M_1, \ldots, M_1, 1 - (\lceil M_1^{-1} \rceil - 1)M_1, 0, \ldots, 0)$ and $\mathbf{q}_* = (M_2, \ldots, M_2, 1 - (\lceil M_2^{-1} \rceil - 1)M_2, 0, \ldots, 0)$, producing the lower bound for $\ell(\mathbf{p}, \mathbf{q})$.

(ii) Via Lemma E.3ii, $\sum_{i=1}^{I} p_i q_i$ reaches its lower bound if $\mathbf{p}^* = (M_1, M_1, \ldots, M_1, 1 - (\lceil M_1^{-1} \rceil - 1)M_1, 0, \ldots, 0)$ and $\mathbf{q}^* = (0, \ldots, 0, 1 - (\lceil M_2^{-1} \rceil - 1)M_2, M_2, \ldots, M_2)$, producing the upper bound for $\ell(\mathbf{p}, \mathbf{q})$.

The values of $\ell(\mathbf{p}_*, \mathbf{q}_*)$ and $\ell(\mathbf{p}^*, \mathbf{q}^*)$ can be obtained by computing $\ell(\mathbf{p}, \mathbf{q})$ with the vectors specified.

F Proof of Corollary 4.6

This proof follows that of Theorem 4.5. With the additional requirement that M_1 and M_2 are the frequencies for the same allele in both populations, we can write $\ell(\mathbf{p}, \mathbf{q})$ as,

$$\ell(\mathbf{p}, \mathbf{q}) = 1 - \rho_{11} = 1 - M_1 M_2 - \sum_{i=2}^{I} p_i q_i.$$

We must find $(p_2, p_3, \ldots, p_I)$ and $(q_2, q_3, \ldots, q_I)$ that give the upper and lower bounds for $\sum_{i=2}^{I} p_i q_i$.

With $\sum_{i=2}^{I} p_i = 1 - M_1$ and $\sum_{i=2}^{I} q_i = 1 - M_2$, by Lemma E.3, the minimum of $\sum_{i=2}^{I} p_i q_i$ is reached at

$$\mathbf{p}^* = (M_1, M_1, \ldots, M_1, 1 - (\lceil M_1^{-1} \rceil - 1)M_1, 0, \ldots, 0),$$

with $\lceil M_1^{-1} \rceil - 1$ entries of M_1 followed by an entry of $1 - (\lceil M_1^{-1} \rceil - 1)M_1$ and 0 for the rest, and

$$\mathbf{q}^* = (M_2, 0, \ldots, 0, 1 - (\lceil M_2^{-1} \rceil - 1)M_2, M_2, \ldots, M_2),$$

with one entry of M_2, $I - \lceil M_2^{-1} \rceil$ entries of 0, followed by an entry of $1 - (\lceil M_2^{-1} \rceil - 1)M_2$ and $\lceil M_2^{-1} \rceil - 2$ entries of M_2. These values minimize $\sum_{i=2}^{I} p_i q_i$, thereby maximizing $\ell(\mathbf{p}, \mathbf{q}) = \mathbb{E}[\mathcal{D}_2^b]$.

Similarly, by Lemma E.3, the maximum of $\sum_{i=2}^{I} p_i q_i$ is reached at

$$\mathbf{p}_* = (M_1, M_1, \ldots, M_1, 1 - (\lceil M_1^{-1} \rceil - 1)M_1, 0, \ldots, 0),$$
$$\mathbf{q}_* = (M_2, M_2, \ldots, M_2, 1 - (\lceil M_2^{-1} \rceil - 1)M_2, 0, \ldots, 0).$$

These values maximize $\sum_{i=2}^{I} p_i q_i$, thus minimizing $\ell(\mathbf{p}, \mathbf{q}) = \mathbb{E}[\mathcal{D}_2^b]$.

The values of $\ell(\mathbf{p}^*, \mathbf{q}^*)$ and $\ell(\mathbf{p}_*, \mathbf{q}_*)$ can be obtained accordingly.

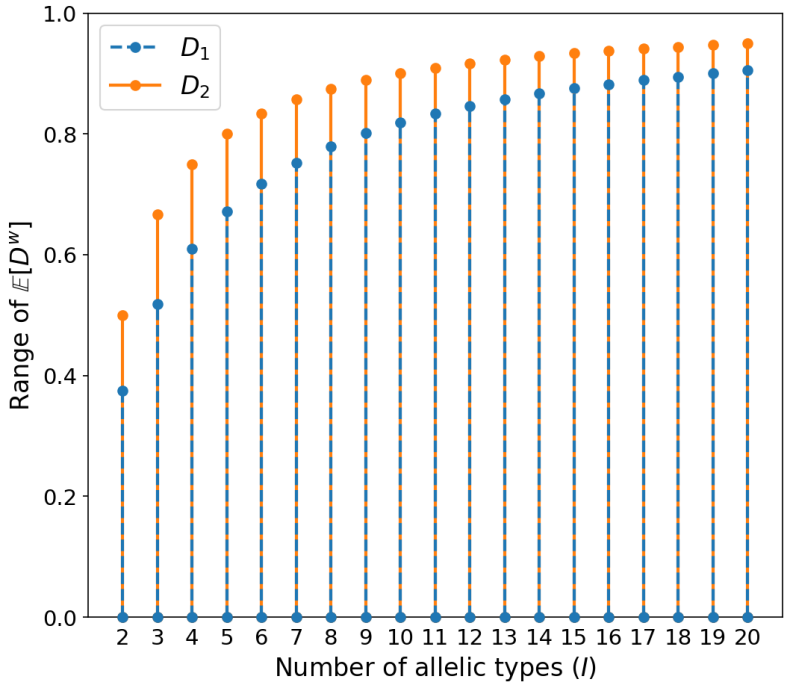

Figure 1: Range of $\mathbb{E}[\mathcal{D}_1^w]$ and $\mathbb{E}[\mathcal{D}_2^w]$ as functions of the number of allelic types I when the allele-frequency vector is permitted to be arbitrary, as stated in Theorems 3.2 and 3.6.

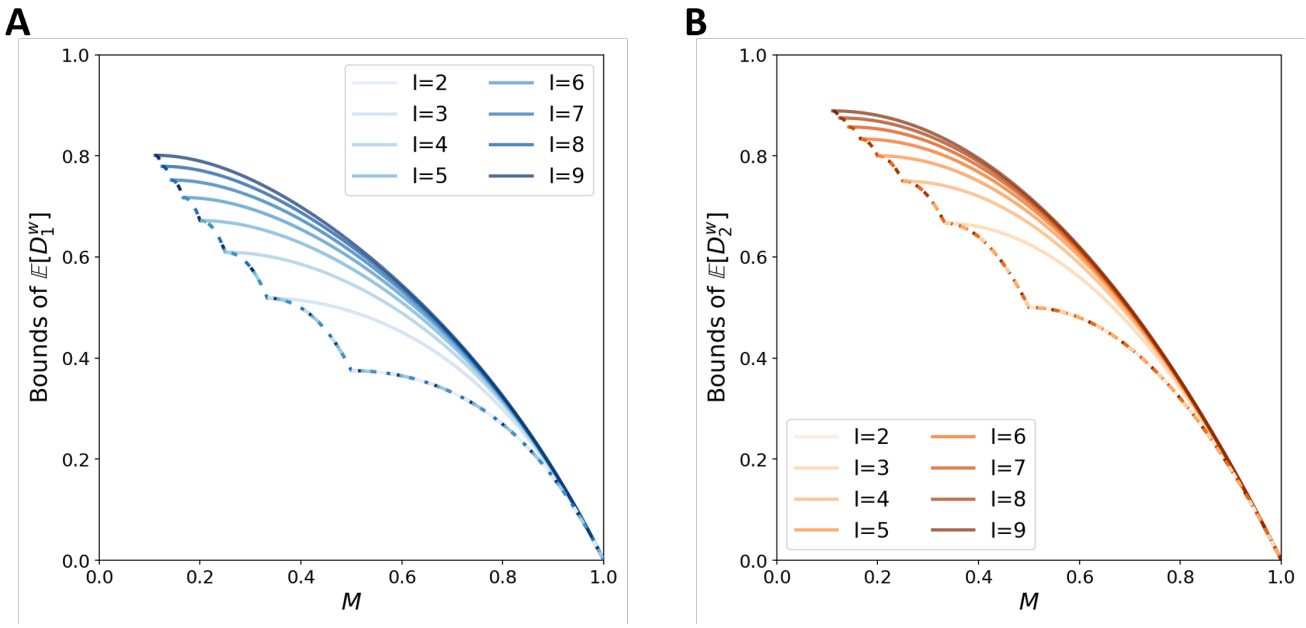

Figure 2: Bounds of expected dissimilarities for $I = 2$ to 9 allelic types when the largest allele frequency is fixed to be M, as stated in Theorems 3.3 and 3.7. (A) $\mathbb{E}[\mathcal{D}_1^w]$. (B) $\mathbb{E}[\mathcal{D}_2^w]$. The solid line corresponds to the upper bound, and the dashed line corresponds to the lower bound, with lower bounds for different I values overlapping.

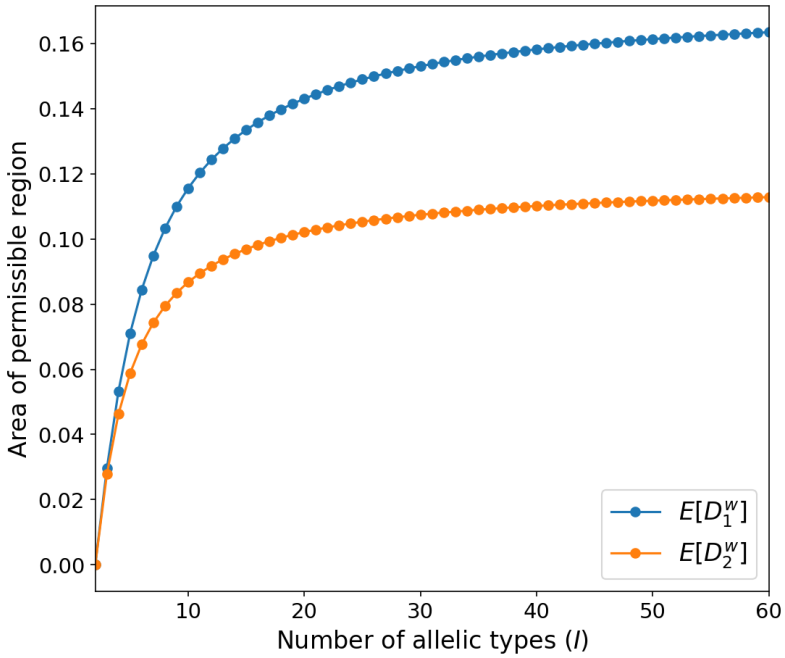

Figure 3: Area of the permissible region for expected dissimilarities for $I = 2$ to 60 when the largest allele frequency is fixed to be M and M ranges from $\frac{1}{I}$ to 1, as stated in Propositions 3.4 and 3.8.

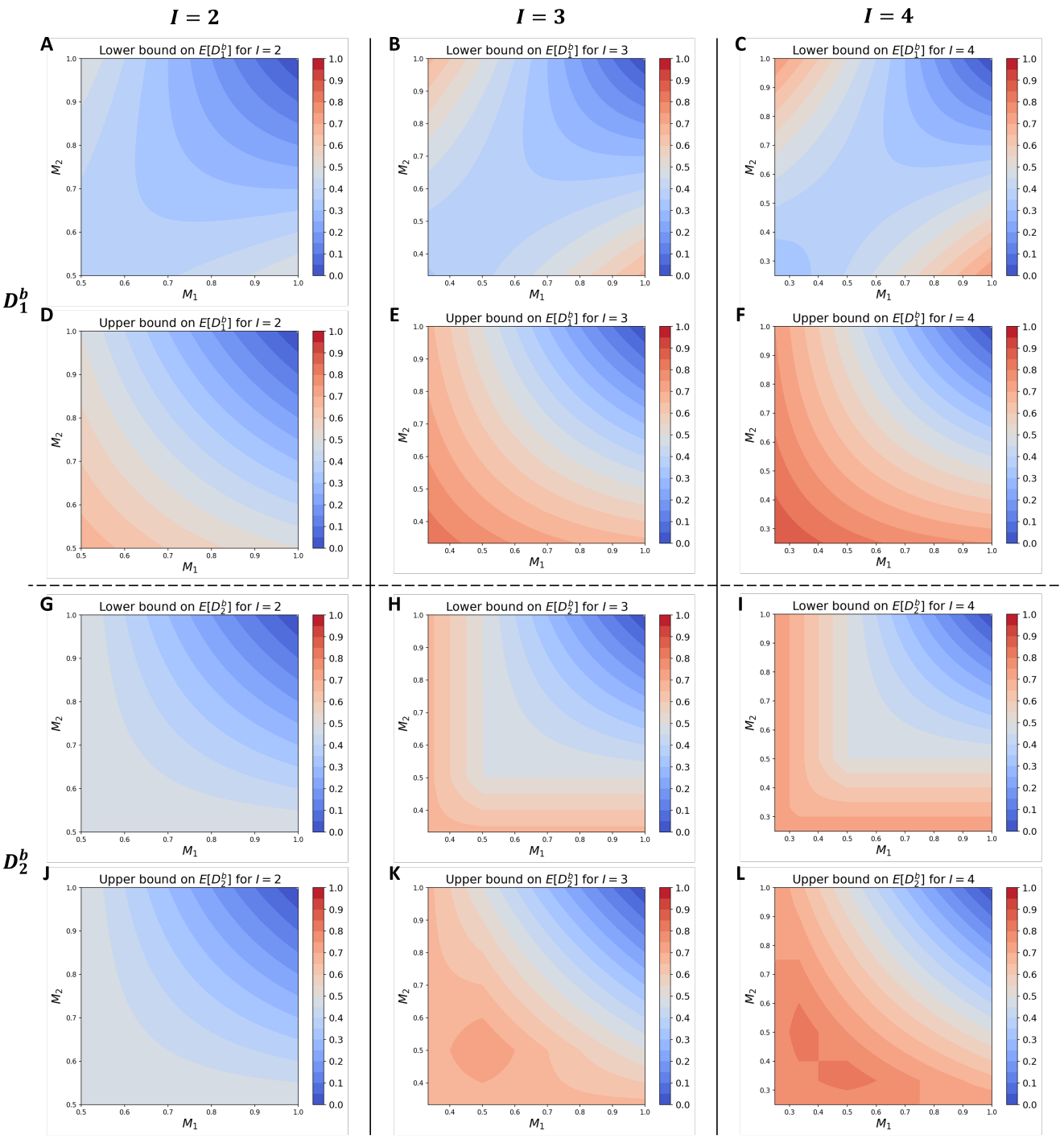

Figure 4: Bounds for $\mathbb{E}[\mathcal{D}_1^b]$ and $\mathbb{E}[\mathcal{D}_2^b]$ for $I = 2$ to 4 when two populations share the most frequent allelic type. The frequencies of the most frequent allelic type are fixed to be M_1 and M_2 in the two populations. (A) $\mathbb{E}[\mathcal{D}_1^b]$, lower bound for $I = 2$. (B) $\mathbb{E}[\mathcal{D}_1^b]$, lower bound for $I = 3$. (C) $\mathbb{E}[\mathcal{D}_1^b]$, lower bound for $I = 4$. (D) $\mathbb{E}[\mathcal{D}_1^b]$, upper bound for $I = 2$. (E) $\mathbb{E}[\mathcal{D}_1^b]$, upper bound for $I = 3$. (F) $\mathbb{E}[\mathcal{D}_1^b]$, upper bound for $I = 4$. (G) $\mathbb{E}[\mathcal{D}_2^b]$, lower bound for $I = 2$. (H) $\mathbb{E}[\mathcal{D}_2^b]$, lower bound for $I = 3$. (I) $\mathbb{E}[\mathcal{D}_2^b]$, lower bound for $I = 4$. (J) $\mathbb{E}[\mathcal{D}_2^b]$, upper bound for $I = 2$. (K) $\mathbb{E}[\mathcal{D}_2^b]$, upper bound for $I = 3$. (L) $\mathbb{E}[\mathcal{D}_2^b]$, upper bound for $I = 4$. Bounds for $\mathbb{E}[\mathcal{D}_1^b]$ follow Theorem 4.2. Bounds for $\mathbb{E}[\mathcal{D}_1^b]$ follow Corollary 4.6. The lower bound for $\mathbb{E}[\mathcal{D}_1^b]$ is loose if $M_1 < \frac{1}{2}$ or $M_2 < \frac{1}{2}$. The lower bound for $\mathbb{E}[\mathcal{D}_1^b]$ if $M_1 \geqslant \frac{1}{2}$ and $M_2 \geqslant \frac{1}{2}$, the upper bound for $\mathbb{E}[\mathcal{D}_1^b]$, and the bounds for $\mathbb{E}[\mathcal{D}_2^b]$ are strict.

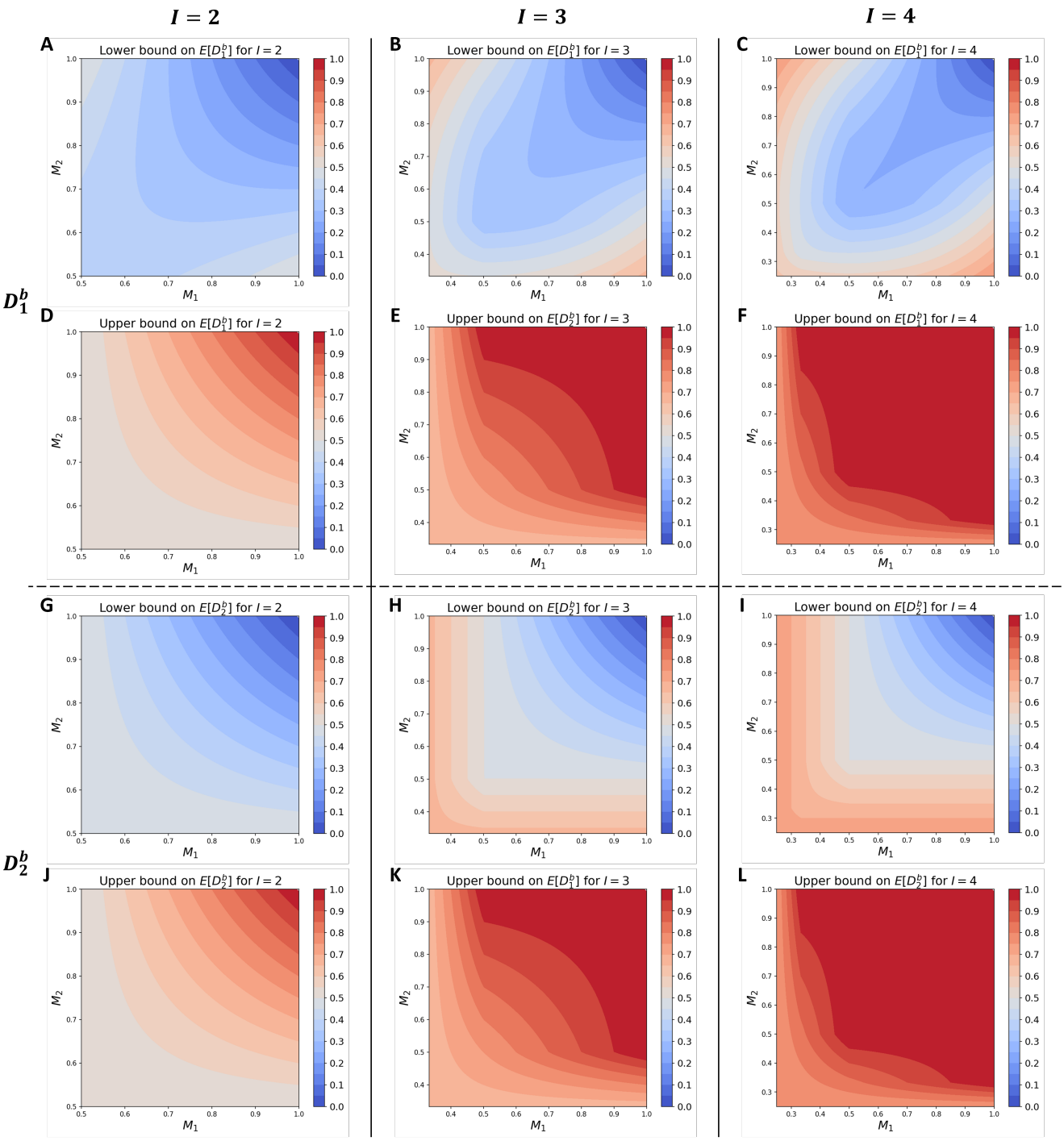

Figure 5: Bounds for $\mathbb{E}[\mathcal{D}_1^b]$ and $\mathbb{E}[\mathcal{D}_2^b]$ for $I = 2$ to 4 when two populations do not necessarily share the most frequent allelic type. The frequencies of the most frequent allelic type are fixed to be M_1 and M_2 in the two populations. (A) $\mathbb{E}[\mathcal{D}_1^b]$, lower bound for $I = 2$. (B) $\mathbb{E}[\mathcal{D}_1^b]$, lower bound for $I = 3$. (C) $\mathbb{E}[\mathcal{D}_1^b]$, lower bound for $I = 4$. (D) $\mathbb{E}[\mathcal{D}_1^b]$, upper bound for $I = 2$. (E) $\mathbb{E}[\mathcal{D}_1^b]$, upper bound for $I = 3$. (F) $\mathbb{E}[\mathcal{D}_1^b]$, upper bound for $I = 4$. (G) $\mathbb{E}[\mathcal{D}_2^b]$, lower bound for $I = 2$. (H) $\mathbb{E}[\mathcal{D}_2^b]$, lower bound for $I = 3$. (I) $\mathbb{E}[\mathcal{D}_2^b]$, lower bound for $I = 4$. (J) $\mathbb{E}[\mathcal{D}_2^b]$, upper bound for $I = 2$. (K) $\mathbb{E}[\mathcal{D}_2^b]$, upper bound for $I = 3$. (L) $\mathbb{E}[\mathcal{D}_2^b]$, upper bound for $I = 4$. Bounds for $\mathbb{E}[\mathcal{D}_1^b]$ follow Theorem 4.3. Bounds for $\mathbb{E}[\mathcal{D}_2^b]$ follow Theorem 4.5. Bounds for $\mathbb{E}[\mathcal{D}_1^b]$ are loose and bounds for $\mathbb{E}[\mathcal{D}_2^b]$ are strict.

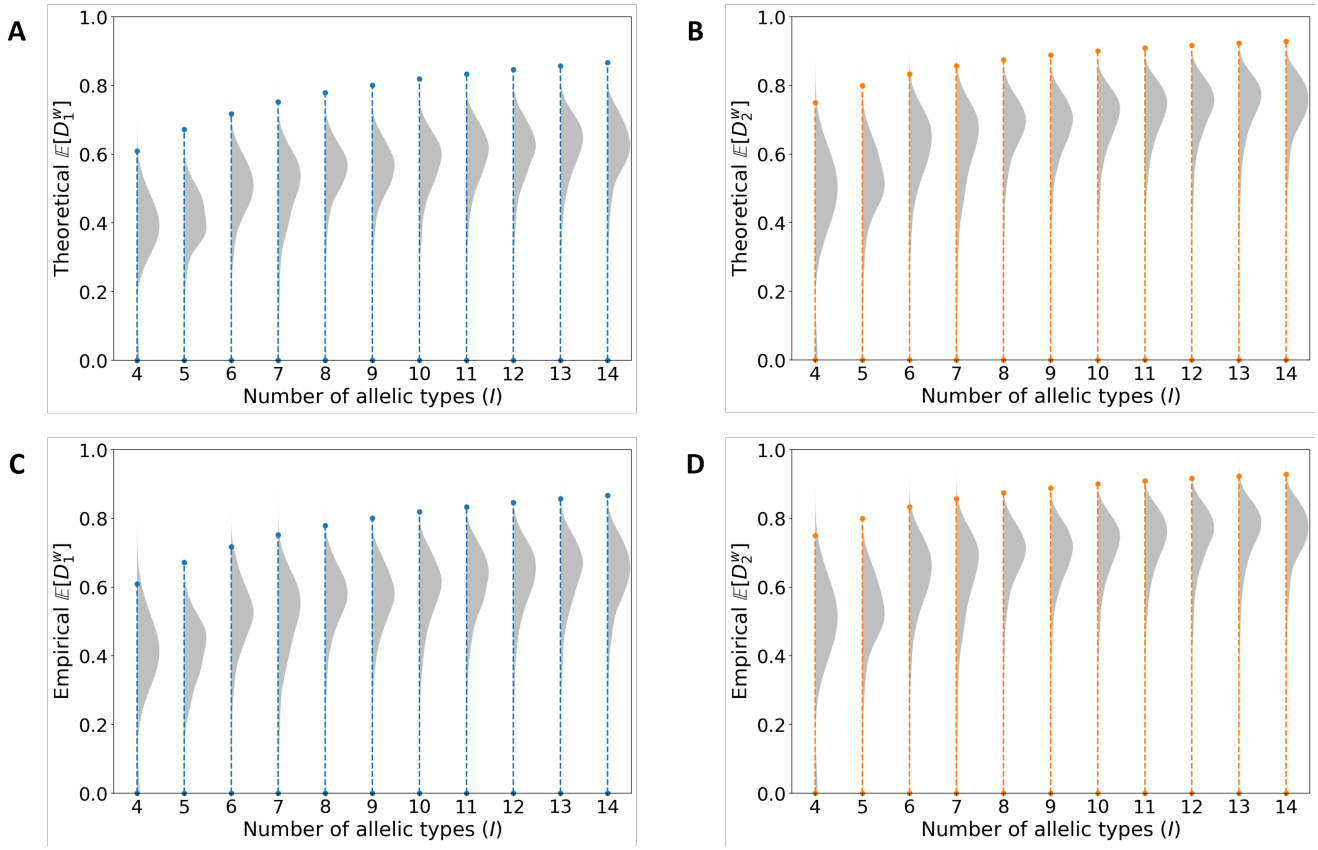

Figure 6: Violin plots of $\mathbb{E}[\mathcal{D}^w]$ in human population-genetic data; 30 populations with sample size larger than 15 are considered at 630 loci, so that each panel contains $30 \times 630 = 18,900$ data points. Population–locus combinations are grouped by values of I; loci with $I > 14$ are not shown. Only half of each violin is depicted. (A) Theoretical $\mathbb{E}[\mathcal{D}_1^w]$ (Eq. 1). (B) Theoretical $\mathbb{E}[\mathcal{D}_2^w]$ (Eq. 2). (C) Empirical $\mathbb{E}[\mathcal{D}_1^w]$. (D) Empirical $\mathbb{E}[\mathcal{D}_2^w]$. "Theoretical" values are calculated based on the allele frequencies in a population, and "empirical" values are obtained by averaging across all pairs of individuals in the population. The permissible regions for $\mathbb{E}[\mathcal{D}_1^w]$ for arbitrary allele frequencies (Theorem 3.2) appear in the background in panels (A) and (C); the permissible regions for $\mathbb{E}[\mathcal{D}_2^w]$ (Theorem 3.6) appear in panels (B) and (D).

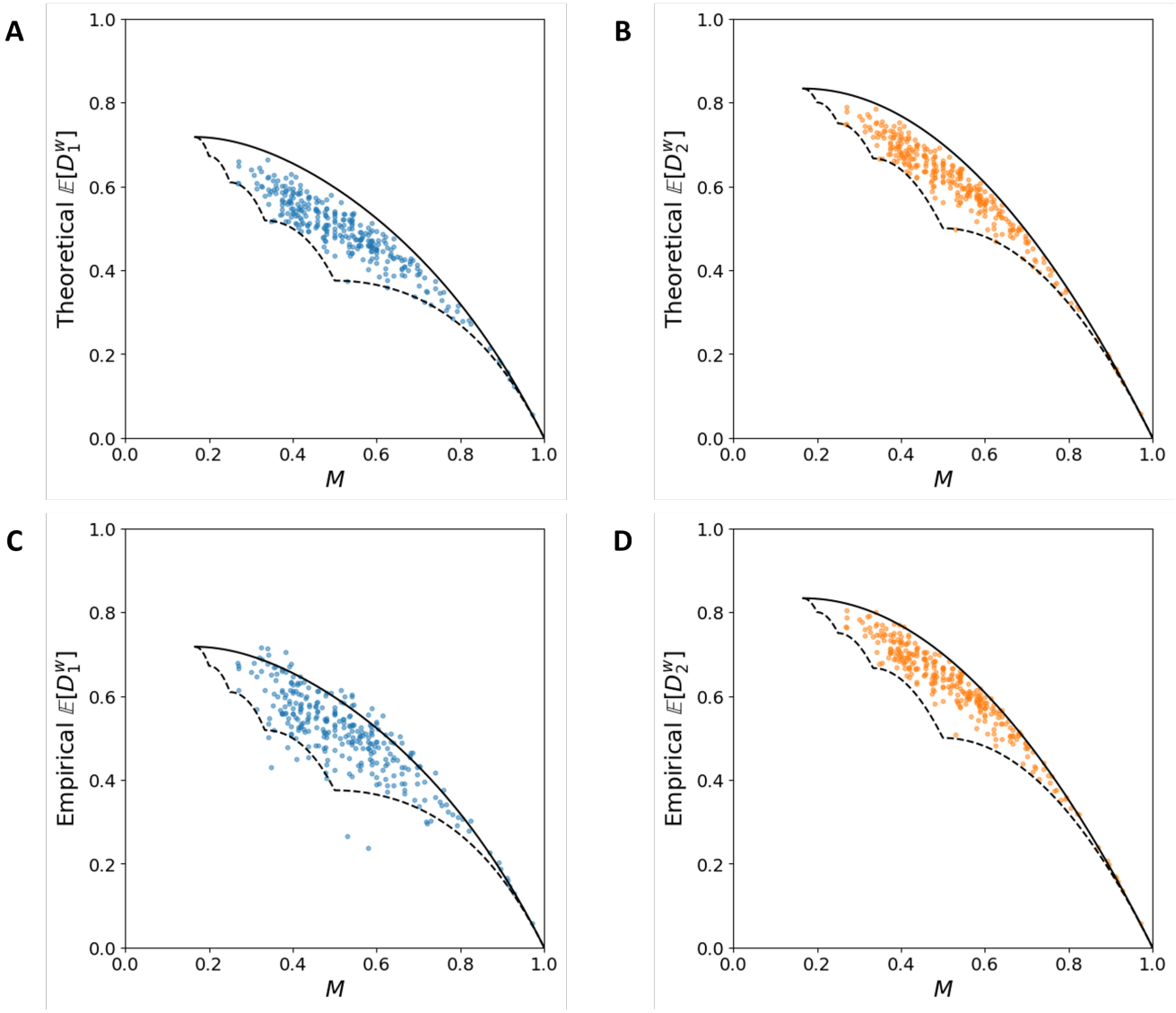

Figure 7: The relationship between the within-population allele-sharing dissimilarity $\mathbb{E}[\mathcal{D}^w]$ and the largest allele frequency M in empirical data. The plot considers 30 populations with sample size larger than 15 and 10 loci with a number of distinct alleles equal to 6, a total of $30 \times 10 = 300$ population–locus combinations. (A) Theoretical $\mathbb{E}[\mathcal{D}_1^w]$ (Eq. 1). (B) Theoretical $\mathbb{E}[\mathcal{D}_2^w]$ (Eq. 2). (C) Empirical $\mathbb{E}[\mathcal{D}_1^w]$. (D) Empirical $\mathbb{E}[\mathcal{D}_2^w]$. "Theoretical" values are calculated based on the allele frequencies in a population, and "empirical" values are obtained by averaging across all pairs of individuals in the population. The permissible region for $\mathbb{E}[\mathcal{D}_1^w]$ in relation to M (Theorem 3.3) appears in panels (A) and (C); the permissible region for $\mathbb{E}[\mathcal{D}_2^w]$ in relation to M (Theorem 3.7) appears in panels (B) and (D).

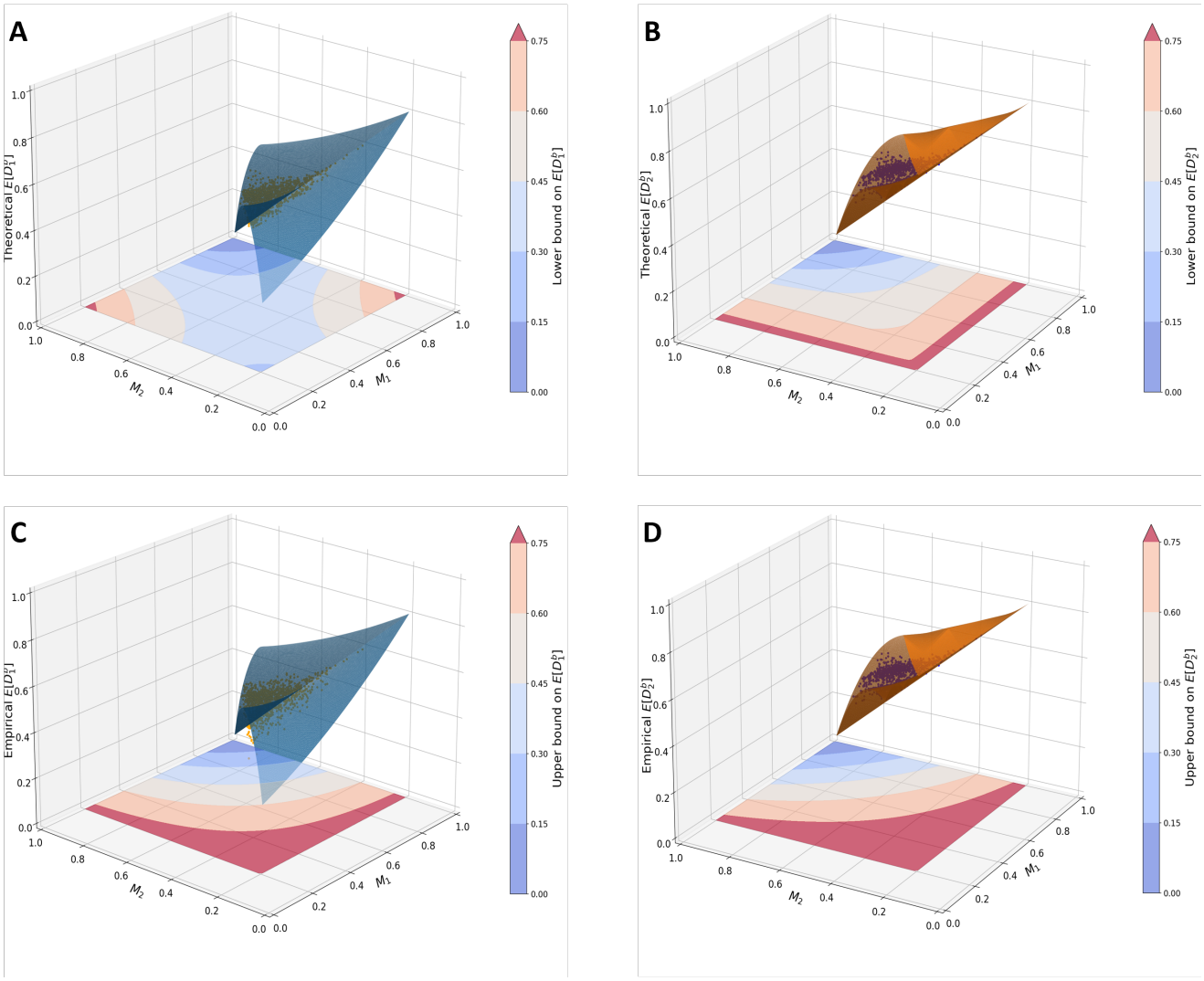

Figure 8: The relationship between the between-population allele-sharing dissimilarity $\mathbb{E}[\mathcal{D}^b]$ and the largest allele frequencies M_1 and M_2 of two populations that have the same most frequent allelic type. The plot considers $\binom{30}{2} = 435$ pairs of populations, both with sample size larger than 15, and 4 loci with number of distinct alleles equal to 5, showing the 1,092 of 1,740 combinations for which the two populations have the same most frequent allelic type. Contour plots of the lower and upper bounds are shown in the xy-plane; contour plots cover the whole plane, but for visual simplicity, only the triangle with $M_1 \geqslant M_2$ is plotted on the z axis. (A) Theoretical $\mathbb{E}[\mathcal{D}_1^b]$ (Eq. 3) and contour plot of the lower bound. (B) Theoretical $\mathbb{E}[\mathcal{D}_2^b]$ (Eq. 4) and contour plot of the lower bound. (C) Empirical $\mathbb{E}[\mathcal{D}_1^b]$ and contour plot of the upper bound. (D) Empirical $\mathbb{E}[\mathcal{D}_2^b]$ and contour plot of the upper bound. "Theoretical" values are calculated based on the allele frequencies in two populations, and "empirical" values are obtained by averaging across all pairs of individuals, one each from two populations. Each pair of populations is ordered such that $M_1 \geqslant M_2$. The permissible region for $\mathbb{E}[\mathcal{D}_1^b]$ in relation to M_1 and M_2 (Theorem 4.2) appears in panels (A) and (C); the permissible region for $\mathbb{E}[\mathcal{D}_2^b]$ in relation to M_1 and M_2 (Corollary 4.6) appears in panels (B) and (D).

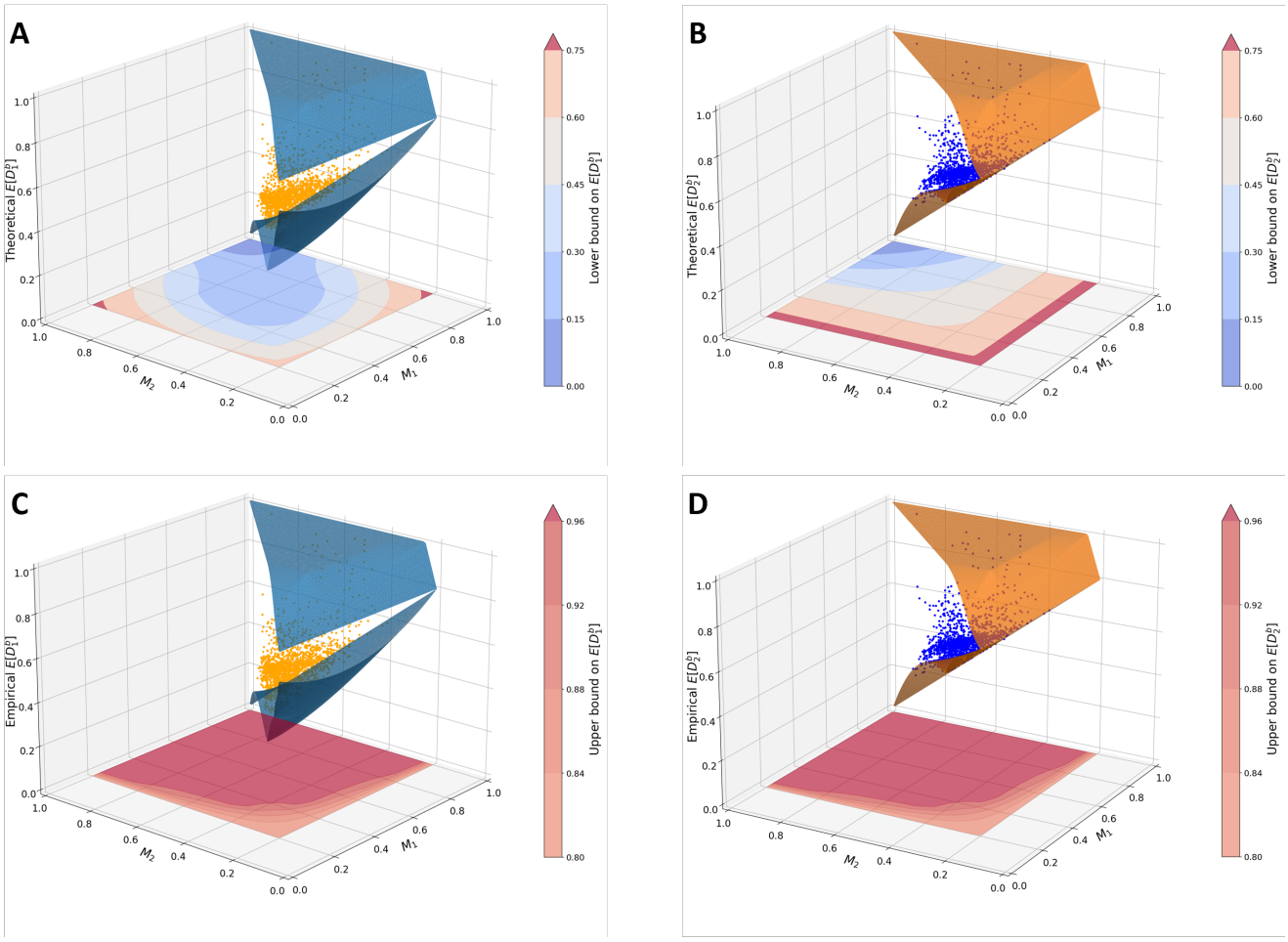

Figure 9: The relationship between the between-population allele-sharing dissimilarity $\mathbb{E}[\mathcal{D}^b]$ and the largest allele frequencies M_1 and M_2 of two populations that do not necessarily have the same most frequent allelic type. The plot considers $\binom{30}{2} = 435$ pairs of populations, both with sample size larger than 15, and 4 loci with number of distinct alelles equal to 5, a total of 1740 combinations. Contour plots of the lower and upper bounds are shown in the xy-plane; contour plots cover the whole plane, but for visual simplicity, only the triangle with $M_1 \geqslant M_2$ is plotted on the z axis. (A) Theoretical $\mathbb{E}[\mathcal{D}_1^b]$ (Eq. 3) and contour plot of the lower bound. (B) Theoretical $\mathbb{E}[\mathcal{D}_2^b]$ (Eq. 4) and contour plot of the lower bound. (C) Empirical $\mathbb{E}[\mathcal{D}_1^b]$ and contour plot of the upper bound. (D) Empirical $\mathbb{E}[\mathcal{D}_2^b]$ and contour plot of the upper bound. "Theoretical" values are calculated based on the allele frequencies in two populations, and "empirical" values are obtained by averaging across all pairs of individuals, one each from two populations. Each pair of populations is ordered such that $M_1 \geqslant M_2$. The permissible region for $\mathbb{E}[\mathcal{D}_1^b]$ in relation to M_1 and M_2 (Theorem 4.3) appears in panels (A) and (C); the permissible region for $\mathbb{E}[\mathcal{D}_2^b]$ in relation to M_1 and M_2 (Theorem 4.5) appears in panels (B) and (D).

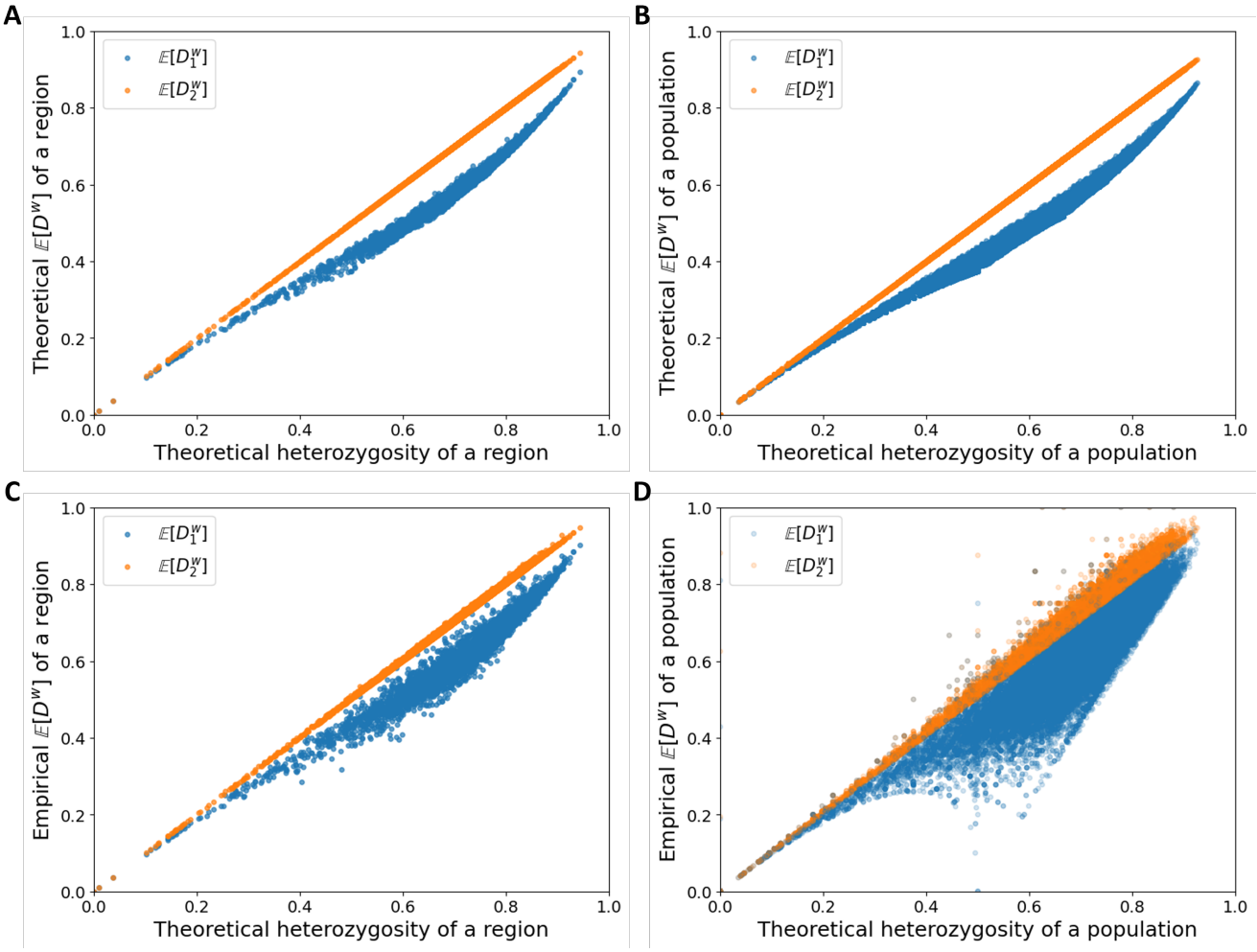

Figure 10: Within-population allele-sharing dissimilarity in relation to theoretical heterozygosity $\mathbb{E}[\mathcal{D}_2^w]$. The plots consider 7 geographic regions and 53 populations, for a total of $7 \times 783 = 5,481$ points in panels (A) and (B), and $53 \times 783 = 41,499$ points in panels (C) and (D) (A) Theoretical $\mathbb{E}[\mathcal{D}_1^w]$ and $\mathbb{E}[\mathcal{D}_2^w]$ for regions (Eqs. 1 and 2). (B) Theoretical $\mathbb{E}[\mathcal{D}_1^w]$ and $\mathbb{E}[\mathcal{D}_2^w]$ for populations (Eqs. 1 and 2). (C) Empirical $\mathbb{E}[\mathcal{D}_1^w]$ and $\mathbb{E}[\mathcal{D}_2^w]$ for regions. (D) Empirical $\mathbb{E}[\mathcal{D}_1^w]$ and $\mathbb{E}[\mathcal{D}_2^w]$ for populations. "Theoretical" values are calculated based on the allele frequencies in a population, and "empirical" values are obtained by averaging across all pairs of individuals in the population.

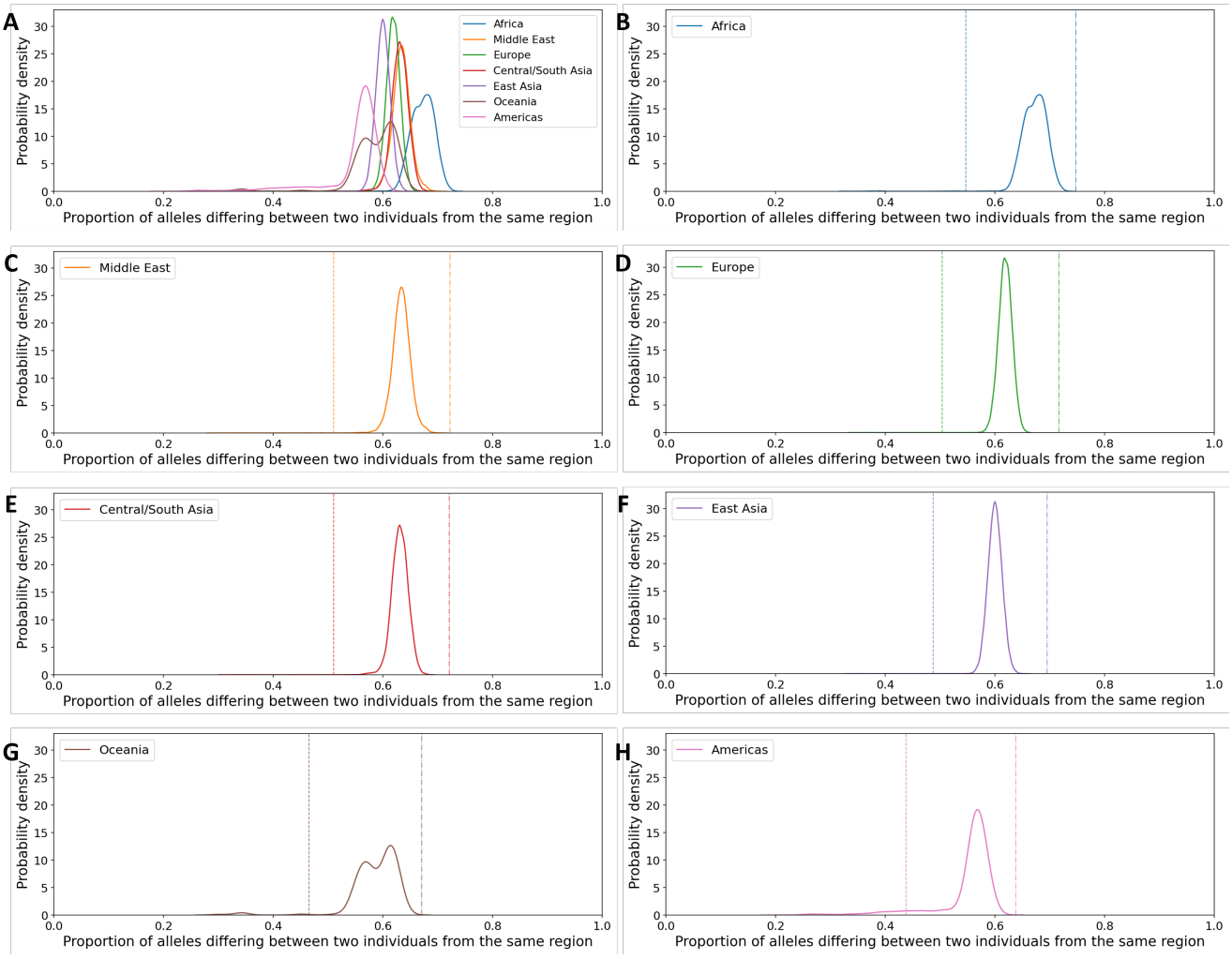

Figure 11: Distributions of empirical values of within-region $\mathcal{D}_1$ averaged across all 783 loci for pairs of individuals in human population-genetic data. For each region, mathematical bounds for $\mathbb{E}[\mathcal{D}_1^w]$ are calculated from allele frequencies within a region according to Theorem 3.3, averaging across loci. (A) Seven regions displayed together. (B) Africa. (C) Middle East. (D) Europe. (E) Central/South Asia. (F) East Asia. (G) Oceania. (H) Americas.

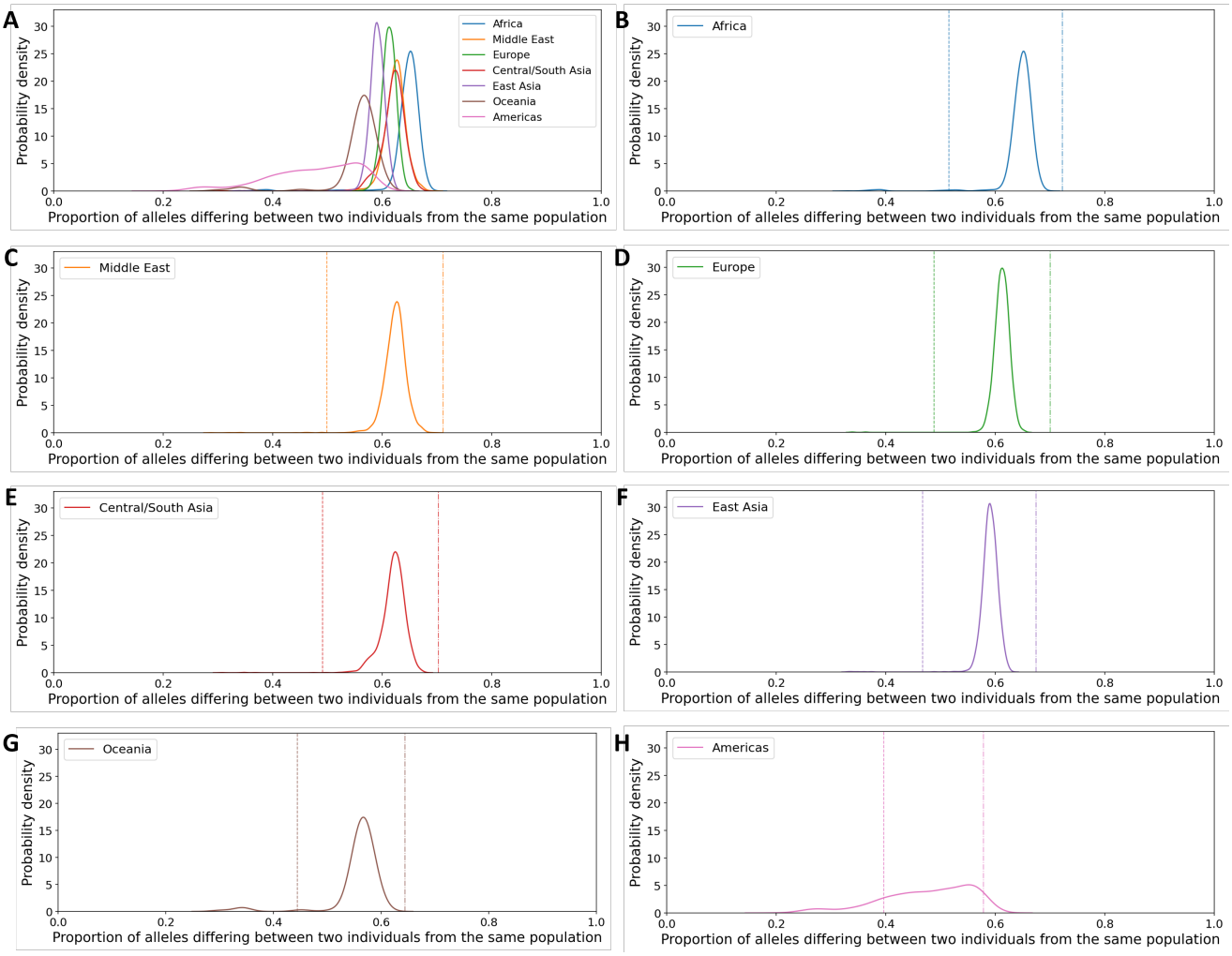

Figure 12: Distributions of empirical values of within-population $\mathcal{D}_1$ averaged across all 783 loci for pairs of individuals in human population-genetic data. For each region, mathematical bounds for $\mathbb{E}[\mathcal{D}_1^w]$ are calculated for each population from allele frequencies within the population according to Theorem 3.3, averaging across loci. Bounds are then averaged across populations within regions. (A) Seven regions displayed together. (B) Africa. (C) Middle East. (D) Europe. (E) Central/South Asia. (F) East Asia. (G) Oceania. (H) Americas.