



Xiran Liu*, Zarif Ahsan, Tarun K. Martheswaran and Noah A. Rosenberg

When is the allele-sharing dissimilarity between two populations exceeded by the allele-sharing dissimilarity of a population with itself?

<https://doi.org/10.1515/sagmb-2023-0004>

Received January 5, 2023; accepted November 10, 2023; published online December 12, 2023

Abstract: Allele-sharing statistics for a genetic locus measure the dissimilarity between two populations as a mean of the dissimilarity between random pairs of individuals, one from each population. Owing to within-population variation in genotype, allele-sharing dissimilarities can have the property that they have a nonzero value when computed between a population and itself. We consider the mathematical properties of allele-sharing dissimilarities in a pair of populations, treating the allele frequencies in the two populations parametrically. Examining two formulations of allele-sharing dissimilarity, we obtain the distributions of within-population and between-population dissimilarities for pairs of individuals. We then mathematically explore the scenarios in which, for certain allele-frequency distributions, the within-population dissimilarity – the mean dissimilarity between randomly chosen members of a population – can exceed the dissimilarity between two populations. Such scenarios assist in explaining observations in population-genetic data that members of a population can be empirically more genetically dissimilar from each other on average than they are from members of another population. For a population pair, however, the mathematical analysis finds that at least one of the two populations always possesses smaller within-population dissimilarity than the value of the between-population dissimilarity. We illustrate the mathematical results with an application to human population-genetic data.

Keywords: allele-sharing; genetic dissimilarity; population genetics

1 Introduction

Statistics that measure the genetic dissimilarity between pairs of populations are widely used for interpreting population-genetic data (Bowcock et al. 1994; Chakraborty and Jin 1993; Gao and Martin 2009; Mountain and Cavalli-Sforza 1997; Mountain and Ramakrishnan 2005; Rosenberg 2011; Tal 2013; Witherspoon et al. 2007). Patterns in numerical values of the statistics appear in calculations of the relative similarity and dissimilarity of different human groups (Mountain and Ramakrishnan 2005; Rosenberg 2011; Witherspoon et al. 2007). Further, genetic dissimilarity statistics, often termed “genetic distances,” underlie frequently applied tools for data analysis and visualization, including methods such as evolutionary tree construction (Bowcock et al. 1994) and multidimensional scaling (Gao and Martin 2009).

Population-level genetic dissimilarity statistics computed at a single genetic locus often proceed by considering pairs of vectors, \mathbf{p} and \mathbf{q} , representing the allele frequencies of two populations. Each vector consists of

*Corresponding author: Xiran Liu, Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, USA, E-mail: xiranliu@stanford.edu

Zarif Ahsan, Tarun K. Martheswaran and Noah A. Rosenberg, Department of Biology, Stanford University, Stanford, CA 94305, USA, E-mail: zarahsan@stanford.edu (Z. Ahsan), tarunkm@stanford.edu (T.K. Martheswaran), noahr@stanford.edu (N.A. Rosenberg)

nonnegative entries that sum to 1. Hence, for a locus with I distinct alleles, such a genetic dissimilarity statistic has domain $\Delta^{I-1} \times \Delta^{I-1}$, where Δ^{I-1} is the simplex $\{p_1, p_2, \dots, p_I: \sum_{i=1}^I p_i = 1 \text{ and } p_i \geq 0 \text{ for all } i\}$.

Among the many genetic dissimilarity statistics that are available (Jorde 1985; Nei 1987), those known as *allele-sharing dissimilarities* form a distinctive subset. Such statistics view a dissimilarity between two populations as the mean of a dissimilarity between pairs of individuals, one from one population and one from the other. With this perspective, they have a simple interpretation as a population-level generalization of an individual-level statistic. They also have a natural connection to a fundamental computation in human population genetics – the apportionment of genetic diversity among different levels of genetic structure (Edge et al. 2022; Lewontin 1972) – which can be viewed in terms of various mean pairwise dissimilarities across certain subsets of individuals (Rosenberg 2011).

Unlike most dissimilarity statistics – such as those based on such principles as the Euclidean distance between functions of allele frequency vectors (Cavalli-Sforza and Edwards 1967) or the dot product of these vectors (Nei 1972) – because they emerge from inter-individual computations among non-identical individuals, allele-sharing dissimilarities can produce nonzero values for the dissimilarity between a polymorphic population and itself. This feature assists in understanding a property of genetic variation in structured populations: the extent to which genetic dissimilarity of individuals from the same population ever exceeds genetic dissimilarity of individuals from different populations, if at all.

Because individuals in a population generally possess a larger number of recent shared ancestors than individuals from different populations, a perspective focused on population-genetic descent predicts that individuals from the same population will be genetically more similar than individuals from different populations. Indeed, in human population genetics, studies of allele-sharing dissimilarity find that the mean dissimilarity across pairs of individuals from different populations does exceed the mean dissimilarity for pairs from the same populations (Mountain and Ramakrishnan 2005; Rosenberg 2011; Tal 2013; Witherspoon et al. 2007). However, such studies also find a perhaps unexpected result that the allele-sharing dissimilarity for *some* pairs of individuals from the same population can exceed the dissimilarity for *some* pairs from different populations.

Here, we seek to explain the properties of allele-sharing dissimilarities within and between populations. We study mathematical properties of population-level allele-sharing dissimilarities under the assumption that individuals in a population represent random draws from the vector of allele frequencies in the population. We consider mean allele-sharing dissimilarities for pairs of individuals from the same population and for pairs of individuals from different populations, evaluating the conditions on allele-frequency vectors under which the allele-sharing dissimilarity for a population to itself can exceed the allele-sharing dissimilarity between two populations. We interpret the results in relation to ongoing efforts to understand human genetic similarity and difference.

2 Methods

2.1 Allele-sharing dissimilarities

An allele-sharing dissimilarity (ASD) is a type of dissimilarity that is based on counting the number of alleles shared at a locus between two diploid individuals. We consider two different versions of the ASD concept.

In one ASD variant, which we denote by D_1 , “allele-sharing” for two diploid individuals is interpreted as the number of shared elements in their multisets of alleles. Consider a locus with four distinct alleles, the minimum number required so that all possible cases exist. Call these alleles A , B , C , and D . For D_1 , two individuals both with genotype AB have 2 alleles shared, as the sets $\{A, B\}$ and $\{A, B\}$ have 2 identical elements. An individual with genotype AB and an individual with genotype AC have 1 allele shared, as the sets $\{A, B\}$ and $\{A, C\}$ have 1 element shared between them, namely A . Two individuals with genotype AA have 2 alleles shared, as multisets $\{A, A\}$ and $\{A, A\}$ have 2 shared elements, A and A . The dissimilarity D_1 then uses 1 minus half the number of the shared alleles as the dissimilarity; the normalization ensures that D_1 lies in $[0,1]$ (Gao and Martin 2009; Mountain and Cavalli-Sforza 1997). With 0, 1, and 2 shared alleles, the dissimilarity equals 1 , $\frac{1}{2}$, and 0 , respectively.

Another variant of ASD, which we denote by 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 (Mountain and Ramakrishnan 2005). For two individuals with

Table 1: Two variants of allele-sharing dissimilarity. All possible pairs of unordered genotypes are shown, along with their values of D_1 and D_2 .

Case	Genotypes	D_1	D_2
1	AA, AA	0	0
2	AA, AB	$\frac{1}{2}$	$\frac{1}{2}$
3	AA, BB	1	1
4	AA, BC	1	1
5	AB, AB	0	$\frac{1}{2}$
6	AB, AC	$\frac{1}{2}$	$\frac{3}{4}$
7	AB, CD	1	1

genotype AB , D_2 is equal to $\frac{1}{2}$, because 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 individual – two of four contain distinct alleles.

Table 1 shows all seven possible pairs of unordered diploid genotypes for two individuals and their corresponding dissimilarities measured by D_1 and D_2 . In only two of seven cases do the two dissimilarities differ.

2.2 Notation

Consider a locus with I distinct alleles. We consider allele-frequency vectors in each of two populations. In Population 1, the allele frequencies are $\mathbf{p} = (p_1, p_2, \dots, p_I)$, where p_i represents the frequency of allele i . In Population 2, they are $\mathbf{q} = (q_1, q_2, \dots, q_I)$. The frequencies satisfy $0 \leq p_i, q_i \leq 1$ for all i , and $\sum_{i=1}^I p_i = \sum_{i=1}^I q_i = 1$.

We are interested in mathematical properties of the distribution of ASD measure D , for pairs of populations – possibly the same population – where D can refer to D_1 or D_2 . We denote the dissimilarity D between two randomly chosen individuals within the same population with allele-frequency vector \mathbf{p} by $D^w(\mathbf{p})$, and the corresponding dissimilarity between two randomly chosen individuals from different populations with allele-frequency vectors \mathbf{p} and \mathbf{q} by $D^b(\mathbf{p}, \mathbf{q})$. We often drop the arguments for convenience.

We will have occasion to use various symmetric sums involving allele frequencies. For $t = 1, 2, 3, 4$, for expressions in the separate populations, we use the notation

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

where $\sigma_1 = \tau_1 = 1$.

For expressions involving both populations, we use

$$\rho_{tu} = \sum_{i=1}^I p_i^t q_i^u, \quad (2)$$

where (t, u) is equal to $(1,1)$, $(1,2)$, $(2,1)$, or $(2,2)$. Note that each of these sums can be viewed as an inner product.

2.3 Assumptions

We seek to perform ASD computations under the assumption that individuals are sampled at random from allele-frequency distributions. With this perspective, for a random pair of individuals, an ASD measure is a random variable that depends on the allele-frequency vectors of two populations of interest, treated as parameters.

At a given locus, we assume that the two alleles of an individual are sampled independently, so that diploid genotypes in a population are assumed to follow Hardy–Weinberg proportions. In other words, the probabilities of diploid genotypes in a population with allele-frequency vector \mathbf{p} equal p_i^2 for homozygous genotypes and $2p_i p_j$ for heterozygous unordered genotypes, with $i \neq j$.

3 Distribution of D^w

We first compute allele-sharing dissimilarities between random pairs of individuals sampled from the same population, evaluating the properties of random variables D_1^w and D_2^w .

3.1 Distribution of D_1^w

D_1^w is a random variable that takes on values 0, $\frac{1}{2}$, and 1. We compute its probability distribution, and we then evaluate its mean and variance.

$$\mathbb{P}[D_1^w = d].$$

We obtain the probability for each possible genotype combination in Table 1. These probabilities appear in Table 2, both as sums and as simplified polynomials.

With the probabilities of all genotype combinations obtained, we can sum across genotype combinations to compute probabilities for $D_1^w(\mathbf{p})$ to equal 0, $\frac{1}{2}$, and 1. The resulting probabilities appear in Table 3.

$$\mathbb{E}[D_1^w].$$

The expected value of $D_1^w(\mathbf{p})$ can be computed from the full probability distribution, via

$$\mathbb{E}[D_1^w(\mathbf{p})] = \sum_{d \in \{0, \frac{1}{2}, 1\}} d \mathbb{P}[D_1^w(\mathbf{p}) = d].$$

Table 2: Probabilities of genotype combinations for pairs of individuals sampled from the same population. For each case, the probability is written as a sum, which is then simplified using Eq. (1).

Case	Genotypes	Probability	Simplified probability
1	AA, AA	$\sum_{i=1}^I p_i^4$	σ_4
2	AA, AB	$4 \sum_{i=1}^I p_i^3 \sum_{\substack{j=1 \\ j \neq i}}^I p_j$	$4\sigma_3 - 4\sigma_4$
3	AA, BB	$\sum_{i=1}^I p_i^2 \sum_{\substack{j=1 \\ j \neq i}}^I p_j^2$	$\sigma_2^2 - \sigma_4$
4	AA, BC	$2 \sum_{i=1}^I p_i^2 \sum_{\substack{j=1 \\ j \neq i}}^I p_j \sum_{\substack{k=1 \\ k \neq i, j}}^I p_k$	$2\sigma_2 - 4\sigma_3 - 2\sigma_2^2 + 4\sigma_4$
5	AB, AB	$2 \sum_{i=1}^I p_i^2 \sum_{\substack{j=1 \\ j \neq i}}^I p_j^2$	$2\sigma_2^2 - 2\sigma_4$
6	AB, AC	$4 \sum_{i=1}^I p_i^2 \sum_{\substack{j=1 \\ j \neq i}}^I p_j \sum_{\substack{k=1 \\ k \neq i, j}}^I p_k$	$4\sigma_2 - 8\sigma_3 - 4\sigma_2^2 + 8\sigma_4$
7	AB, CD	$\sum_{i=1}^I p_i \sum_{\substack{j=1 \\ j \neq i}}^I p_j \sum_{\substack{k=1 \\ k \neq i, j}}^I p_k \sum_{\substack{\ell=1 \\ \ell \neq i, j, k}}^I p_\ell$	$1 - 6\sigma_2 + 8\sigma_3 + 3\sigma_2^2 - 6\sigma_4$

Table 3: Probability distribution of $D_1^w(\mathbf{p})$, the allele-sharing dissimilarity D_1^w for a pair of individuals sampled at random from a population with allele-frequency vector \mathbf{p} . The table is obtained by summing entries in Table 2.

Value of the dissimilarity (d)	$\mathbb{P}[D_1^w(\mathbf{p}) = d]$
0	$2\sigma_2^2 - \sigma_4$
$\frac{1}{2}$	$4\sigma_2 - 4\sigma_3 - 4\sigma_2^2 + 4\sigma_4$
1	$1 - 4\sigma_2 + 4\sigma_3 + 2\sigma_2^2 - 3\sigma_4$

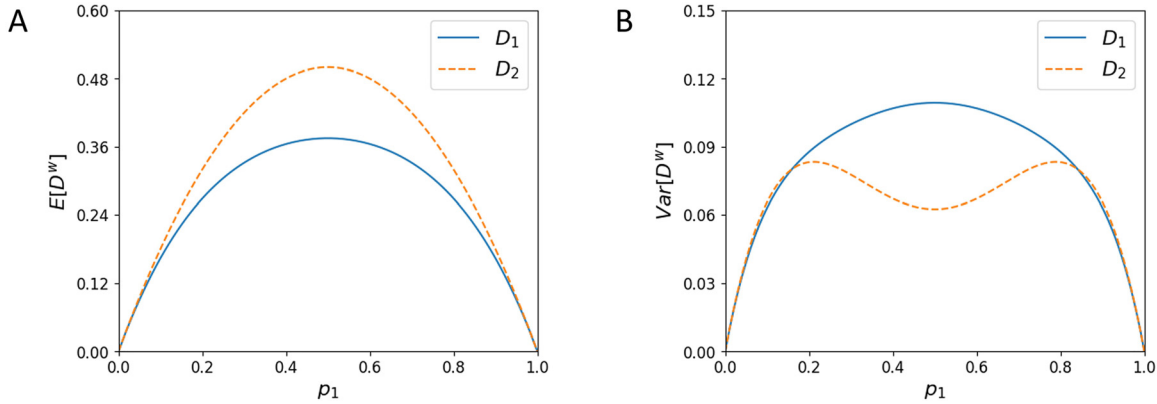


Figure 1: Mean and variance of the within-population dissimilarities D_1^w and D_2^w for $I = 2$ alleles as functions of the frequency p_1 of one of the alleles. (A) Mean, Eqs. (4) and (10). (B) Variance, Eqs. (8) and (14).

Using the probabilities in Table 3, the result is

$$\mathbb{E}[D_1^w(\mathbf{p})] = 1 - 2\sigma_2 + 2\sigma_3 - \sigma_4. \quad (3)$$

In the $I = 2$ case, using $p_2 = 1 - p_1$ so that $\sigma_t = p_1^t + (1 - p_1)^t$, Eq. (3) becomes:

$$\mathbb{E}[D_1^w(\mathbf{p})] = 2p_1 - 4p_1^2 + 4p_1^3 - 2p_1^4. \quad (4)$$

Figure 1A plots Eq. (4) as a function of p_1 . In the figure, we can observe that the mean value of the dissimilarity increases from a value of 0 at $p_1 = 0$, when the population is monomorphic, to a peak of $\frac{3}{8}$ at $p_1 = \frac{1}{2}$. It then decreases symmetrically to 0 at $p_1 = 1$.

$\text{Var}[D_1^w]$.

To obtain the variance of the distribution of $D_1^w(\mathbf{p})$, we first calculate

$$\begin{aligned} \mathbb{E}[D_1^w(\mathbf{p})^2] &= \sum_{d \in \{0, \frac{1}{2}, 1\}} d^2 \mathbb{P}[D_1^w(\mathbf{p}) = d] \\ &= 1 - 3\sigma_2 + 3\sigma_3 + \sigma_2^2 - 2\sigma_4. \end{aligned} \quad (5)$$

The variance can then be obtained from Eqs. (3) and (5) by $\text{Var}[D_1^w(\mathbf{p})] = \mathbb{E}[D_1^w(\mathbf{p})^2] - \mathbb{E}[D_1^w(\mathbf{p})]^2$:

$$\text{Var}[D_1^w(\mathbf{p})] = \sigma_2 - \sigma_3 - 3\sigma_2^2 + 8\sigma_2\sigma_3 - 4\sigma_2\sigma_4 - 4\sigma_3^2 + 4\sigma_3\sigma_4 - \sigma_4^2. \quad (6)$$

For the $I = 2$ case, we once again use that $p_2 = 1 - p_1$:

$$\mathbb{E}[D_1^w(\mathbf{p})^2] = p_1 - p_1^2 \quad (7)$$

$$\text{Var}[D_1^w(\mathbf{p})] = p_1 - 5p_1^2 + 16p_1^3 - 32p_1^4 + 40p_1^5 - 32p_1^6 + 16p_1^7 - 4p_1^8. \quad (8)$$

Figure 1B plots Eq. (8) as a function of p_1 . Like the mean, the variance of the dissimilarity increases from 0 at $p_1 = 0$ to a peak at $p_1 = \frac{1}{2}$, decreasing symmetrically to 0 at $p_1 = 1$. The maximal variance is $\frac{7}{64}$.

3.2 Distribution of D_2^w

We compute the distribution of random variable D_2^w . This computation uses the same probabilities for genotype pairs as those used for D_1^w in Table 2.

$$\mathbb{P}[D_2^w = d].$$

We compute the probability for each of the possible values of D_2^w by summing probabilities in Table 2. The resulting probabilities appear in Table 4.

Table 4: Probability distribution of $D_2^w(\mathbf{p})$, the allele-sharing dissimilarity D_2^w for a pair of individuals sampled at random from a population with allele-frequency vector \mathbf{p} . The table is obtained by summing entries in Table 2.

Value of the dissimilarity (d)	$\mathbb{P}[D_2^w(\mathbf{p}) = d]$
0	σ_4
$\frac{1}{2}$	$4\sigma_3 + 2\sigma_2^2 - 6\sigma_4$
$\frac{3}{4}$	$4\sigma_2 - 8\sigma_3 - 4\sigma_2^2 + 8\sigma_4$
1	$1 - 4\sigma_2 + 4\sigma_3 + 2\sigma_2^2 - 3\sigma_4$

$$\mathbb{E}[D_2^w].$$

Summing across the possible values for the dissimilarity,

$$\mathbb{E}[D_2^w(\mathbf{p})] = \sum_{d \in \{0, \frac{1}{2}, \frac{3}{4}, 1\}} d \mathbb{P}[D_2^w(\mathbf{p}) = d],$$

yielding the result

$$\mathbb{E}[D_2^w(\mathbf{p})] = 1 - \sigma_2. \quad (9)$$

Note that Eq. (9) gives the “expected heterozygosity,” the probability that two draws from the allele-frequency distribution produce distinct alleles.

For the $I = 2$ case, we have $\sigma_2 = p_1^2 + (1 - p_1)^2 = 1 - 2p_1 + 2p_1^2$, so Eq. (9) simplifies to

$$\mathbb{E}[D_2^w(\mathbf{p})] = 2p_1 - 2p_1^2 = 2p_1(1 - p_1). \quad (10)$$

Figure 1A plots Eq. (10) as a function of p_1 . The mean value of the dissimilarity is symmetric around a peak at $(\frac{1}{2}, \frac{1}{2})$, equaling 0 at $p_1 = 0$ and $p_1 = 1$.

$$\text{Var}[D_2^w].$$

The variance of the distribution of D_2^w is obtained using $\text{Var}[D_2^w] = \mathbb{E}[D_2^w(\mathbf{p})^2] - \mathbb{E}[D_2^w(\mathbf{p})]^2$. We first find

$$\begin{aligned} \mathbb{E}[D_2^w(\mathbf{p})^2] &= \sum_{d \in \{0, \frac{1}{2}, \frac{3}{4}, 1\}} d^2 \mathbb{P}[D_2^w(\mathbf{p}) = d] \\ &= 1 - \frac{7}{4}\sigma_2 + \frac{1}{2}\sigma_3 + \frac{1}{4}\sigma_2^2. \end{aligned} \quad (11)$$

Therefore,

$$\text{Var}[D_2^w(\mathbf{p})] = \frac{1}{4}\sigma_2 + \frac{1}{2}\sigma_3 - \frac{3}{4}\sigma_2^2. \quad (12)$$

For the $I = 2$ case, we use $p_2 = 1 - p_1$ to obtain

$$\mathbb{E}[D_2^w(\mathbf{p})^2] = p_1 - 2p_1^3 + p_1^4 \quad (13)$$

$$\text{Var}[D_2^w(\mathbf{p})] = p_1 - 4p_1^2 + 6p_1^3 - 3p_1^4. \quad (14)$$

Figure 1B plots Eq. (14). The variance has peaks at $(\frac{3-\sqrt{3}}{6}, \frac{1}{12})$ and $(\frac{3+\sqrt{3}}{6}, \frac{1}{12})$, between which it has a local minimum at $(\frac{1}{2}, \frac{1}{16})$. It equals 0 at $p_1 = 0$ and $p_1 = 1$.

3.3 Comparison of D_1^w and D_2^w

Comparing $\mathbb{E}[D_1^w]$ (Eq. (3)) and $\mathbb{E}[D_2^w]$ (Eq. (9)), we quickly observe that if $p_i \neq 1$ for all i , then

$$\mathbb{E}[D_1^w] < \mathbb{E}[D_2^w]. \quad (15)$$

The result follows by noting $(1 - p_i)^2 > 0$ for all i , so that $\sum_{i=1}^I p_i^2(2p_i) < \sum_{i=1}^I p_i^2(1 + p_i^2)$ and $2\sigma_3 < \sigma_2 + \sigma_4$, from which we obtain Eq. (15). In fact, Eq. (15) follows from Table 1: for all possible genotype combinations, $D_1^w \leq D_2^w$,

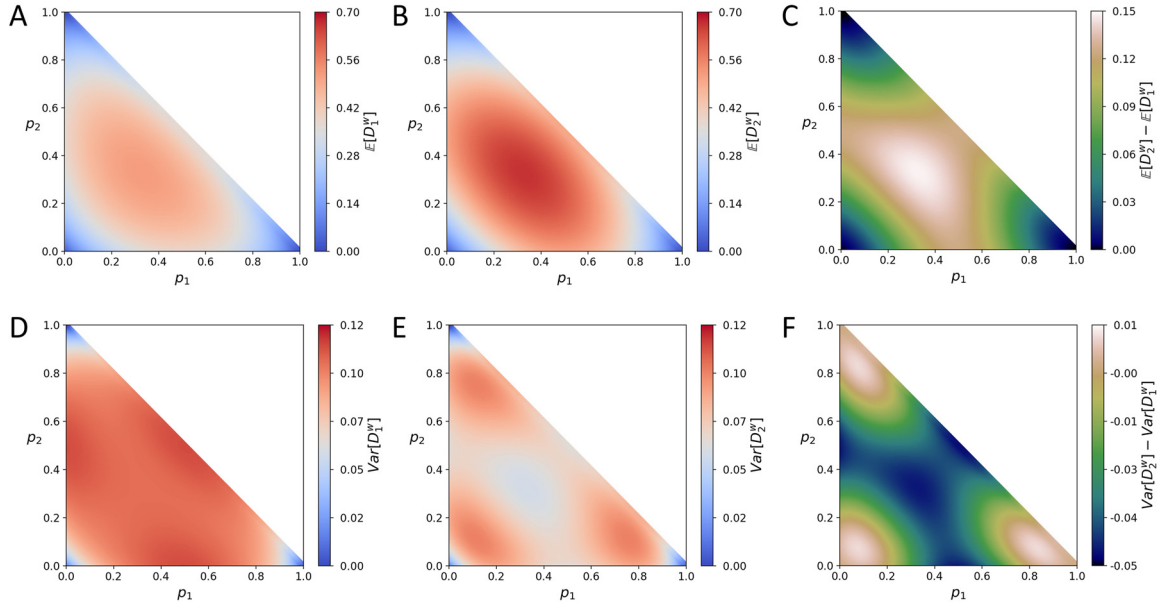


Figure 2: Mean and variance of the within-population dissimilarities D_1^w and D_2^w for $I = 3$ alleles as functions of the frequencies p_1 and p_2 of two of the alleles. (A) Mean of D_1^w , Eq. (3). (B) Mean of D_2^w , Eq. (9). (C) $\mathbb{E}[D_2^w] - \mathbb{E}[D_1^w]$. (D) Variance of D_1^w , Eq. (6). (E) Variance of D_2^w , Eq. (12). (F) $\text{Var}[D_2^w] - \text{Var}[D_1^w]$.

and the inequality is strict in two of seven cases, at least one of which must have nonzero probability if $p_i \neq 1$ for all i .

For $I = 2$, Eq. (15) can be observed in Figure 1A, as it can be seen that the curve for $\mathbb{E}[D_2^w]$ exceeds that for $\mathbb{E}[D_1^w]$. The largest excess occurs at $p_1 = p_2 = \frac{1}{2}$. Figure 2C plots the difference $\mathbb{E}[D_2^w] - \mathbb{E}[D_1^w]$ for the case of $I = 3$, and the maximal difference in the figure also occurs when alleles have the same frequency, $(p_1, p_2, p_3) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

For the variances, Figure 1B finds that for $I = 2$, $\text{Var}[D_1^w] > \text{Var}[D_2^w]$ for intermediate p_1 , and that the two variances are comparable for p_1 near 0 or 1, with some p_1 values producing $\text{Var}[D_1^w] < \text{Var}[D_2^w]$. Figure 2F illustrates a similar result for $I = 3$. For both $I = 2$ and $I = 3$, at intermediate allele frequencies, $\text{Var}[D_1^w] > \text{Var}[D_2^w]$; at extreme allele frequencies, the two variances are comparable, sometimes with $\text{Var}[D_1^w] < \text{Var}[D_2^w]$.

4 Distribution of D^b

We now examine allele-sharing dissimilarities between pairs of individuals from different populations. Let \mathbf{p} be the allele frequency vector for the population from which the first individual is sampled, and let \mathbf{q} be the corresponding vector for the population of the second individual; the special case of $\mathbf{q} = \mathbf{p}$ follows Section 3. We evaluate the properties of the random variables D_1^b and D_2^b .

4.1 Distribution of D_1^b

$\mathbb{P}[D_1^b = \mathbf{a}]$. We obtain the probability for each possible genotype combination for a pair of individuals from different populations. For this computation, we use the polynomials in Eqs. (1) and (2). The resulting probabilities appear in Table 5.

We sum across genotype combinations to obtain probabilities for D_1^b to equal particular values. Table 6 provides these probabilities.

Table 5: Probability of genotype combinations for pairs of individuals sampled from two populations. For each case, the probability is written as a sum, which is then simplified using Eqs. (1) and (2).

Case	Genotypes	Probability	Simplified probability
1	AA, AA	$\sum_{i=1}^I p_i^2 q_i^2$	ρ_{22}
2	AA, AB	$2 \sum_{i=1}^I p_i^2 q_i \sum_{j=1, j \neq i}^I q_j + 2 \sum_{i=1}^I p_i q_i^2 \sum_{j=1, j \neq i}^I p_j$	$2\rho_{21} + 2\rho_{12} - 4\rho_{22}$
3	AA, BB	$\sum_{i=1}^I p_i^2 \sum_{j=1, j \neq i}^I q_j^2$	$\sigma_2 \tau_2 - \rho_{22}$
4	AA, BC	$\sum_{i=1}^I p_i^2 \sum_{j=1, j \neq i}^I q_j \sum_{k=1, k \neq i, j}^I q_k + \sum_{i=1}^I q_i^2 \sum_{j=1, j \neq i}^I p_j \sum_{k=1, k \neq i, j}^I p_k$	$\sigma_2 + \tau_2 - 2\sigma_2 \tau_2 - 2\rho_{21} - 2\rho_{12} + 4\rho_{22}$
5	AB, AB	$2 \sum_{i=1}^I p_i q_i \sum_{j=1, j \neq i}^I p_j q_j$	$2\rho_{11}^2 - 2\rho_{22}$
6	AB, AC	$4 \sum_{i=1}^I p_i q_i \sum_{j=1, j \neq i}^I p_j \sum_{k=1, k \neq i, j}^I q_k$	$4\rho_{11} - 4\rho_{21} - 4\rho_{12} - 4\rho_{11}^2 + 8\rho_{22}$
7	AB, CD	$\sum_{i=1}^I p_i \sum_{j=1, j \neq i}^I p_j \sum_{k=1, k \neq i, j}^I q_k \sum_{\ell=1, \ell \neq i, j, k}^I q_\ell$	$1 - \sigma_2 - \tau_2 + \sigma_2 \tau_2 - 4\rho_{11} + 4\rho_{21} + 4\rho_{12} - 6\rho_{22} + 2\rho_{11}^2$

Table 6: Probability distribution of $D_1^b(\mathbf{p}, \mathbf{q})$, the allele-sharing dissimilarity D_1^b for a pair of individuals sampled at random from two populations with allele-frequency vectors \mathbf{p} and \mathbf{q} . The table is obtained by summing entries in Table 5.

Value of the dissimilarity (d)	$\mathbb{P}[D_1^b(\mathbf{p}, \mathbf{q}) = d]$
0	$2\rho_{11}^2 - \rho_{22}$
$\frac{1}{2}$	$4\rho_{11} - 2\rho_{21} - 2\rho_{12} - 4\rho_{11}^2 + 4\rho_{22}$
1	$1 - 4\rho_{11} + 2\rho_{21} + 2\rho_{12} + 2\rho_{11}^2 - 3\rho_{22}$

$$\mathbb{E}[D_1^b].$$

As we did for the within-population dissimilarity $D_1^w(\mathbf{p})$, we compute the expected value of the distribution of the between-population dissimilarity $D_1^b(\mathbf{p}, \mathbf{q})$ as

$$\mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})] = \sum_{d \in \{0, \frac{1}{2}, 1\}} d \mathbb{P}[D_1^b(\mathbf{p}, \mathbf{q}) = d].$$

Using the values in Table 6, we obtain

$$\mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})] = 1 - 2\rho_{11} + \rho_{21} + \rho_{12} - \rho_{22}. \quad (16)$$

For the $I = 2$ case, with $p_2 = 1 - p_1$ and $q_2 = 1 - q_1$, Eq. (16) simplifies to

$$\mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})] = p_1 + q_1 - 4p_1q_1 + 2p_1^2q_1 + 2p_1q_1^2 - 2p_1^2q_1^2. \quad (17)$$

Figure 3A plots Eq. (17). The figure has maxima of 1 at $(p_1, q_1) = (1, 0)$ and $(0, 1)$, when the two populations have the greatest difference in allele frequency, and equals 0 at $(0, 0)$ and $(1, 1)$. It has a saddle surface with a value of $\frac{3}{8}$ at saddle point $(p_1, q_1) = (\frac{1}{2}, \frac{1}{2})$.

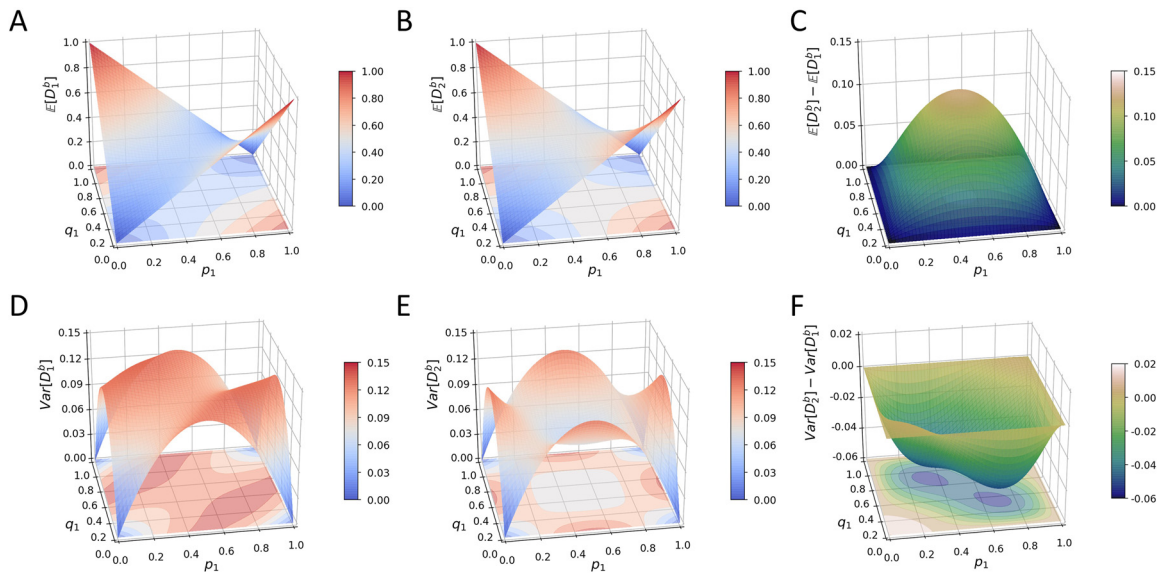


Figure 3: Mean and variance of the between-population dissimilarities D_1^b and D_2^b for $I = 2$ alleles as functions of the frequencies (p_1, q_1) of one of the alleles. (A) Mean of D_1^b , Eq. (17). (B) Mean of D_2^b , Eq. (23). (C) $\mathbb{E}[D_2^b] - \mathbb{E}[D_1^b]$. (D) Variance of D_1^b , Eq. (21). (E) Variance of D_2^b , Eq. (27). (F) $\text{Var}[D_2^b] - \text{Var}[D_1^b]$.

$\text{Var}[D_1^b]$. We first compute

$$\begin{aligned}\mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})^2] &= \sum_{d \in \{0, \frac{1}{2}, 1\}} d^2 \mathbb{P}[D_1^b(\mathbf{p}, \mathbf{q}) = d] \\ &= 1 - 3\rho_{11} + \frac{3}{2}\rho_{21} + \frac{3}{2}\rho_{12} - 2\rho_{22} + \rho_{11}^2.\end{aligned}\quad (18)$$

Using $\text{Var}[D_1^b(\mathbf{p}, \mathbf{q})] = \mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})^2] - \mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})]^2$, the variance is thus

$$\begin{aligned}\text{Var}[D_1^b(\mathbf{p}, \mathbf{q})] &= \rho_{11} - \frac{1}{2}\rho_{21} - \frac{1}{2}\rho_{12} - 3\rho_{11}^2 + 4\rho_{11}\rho_{21} + 4\rho_{11}\rho_{12} - 4\rho_{11}\rho_{22} - \rho_{21}^2 - \rho_{12}^2 - 2\rho_{12}\rho_{21} \\ &\quad + 2\rho_{12}\rho_{22} + 2\rho_{21}\rho_{22} - \rho_{22}^2.\end{aligned}\quad (19)$$

For the $I = 2$ case, we have $p_1 = 1 - p_2$ and $q_1 = 1 - q_2$. Equations (18) and (19) simplify to

$$\mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})^2] = \frac{1}{2}p_1 + \frac{1}{2}q_1 - 2p_1q_1 + \frac{1}{2}p_1^2 + \frac{1}{2}q_1^2 \quad (20)$$

$$\begin{aligned}\text{Var}[D_1^b(\mathbf{p}, \mathbf{q})] &= \frac{1}{2}p_1 + \frac{1}{2}q_1 - 4p_1q_1 - \frac{1}{2}p_1^2 - \frac{1}{2}q_1^2 + 8p_1^2q_1 + 8p_1q_1^2 - 4p_1^3q_1 - 24p_1^2q_1^2 - 4p_1q_1^3 \\ &\quad + 20p_1^3q_1^2 + 20p_1^2q_1^3 - 4p_1^4q_1^2 - 24p_1^3q_1^3 - 4p_1^2q_1^4 + 8p_1^4q_1^3 + 8p_1^3q_1^4 - 4p_1^4q_1^4.\end{aligned}\quad (21)$$

Figure 3D shows that the variance has higher values away from the four corners (0,0), (1,0), (0,1), and (1,1) for (p_1, q_1) , equaling 0 in each of these corners.

4.2 Distribution of D_2^b

$\mathbb{P}[D_2^b = d]$. We use Table 5 to obtain the probabilities of particular values of D_2^b . The resulting probabilities appear in Table 7.

$\mathbb{E}[D_2^b]$. For D_2^b , we substitute the values from Table 7 into

$$\mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})] = \sum_{d \in \{0, \frac{1}{2}, \frac{3}{4}, 1\}} d \mathbb{P}[D_2^b(\mathbf{p}, \mathbf{q}) = d].$$

We obtain

$$\mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})] = 1 - \rho_{11}. \quad (22)$$

This quantity is the between-population analogue of expected heterozygosity, the probability that two random draws, one from the allele-frequency distribution of a locus in one population and one from the corresponding distribution in a second population, represent the same allele.

For the $I = 2$ case, Eq. (22) simplifies to

$$\mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})] = p_1 + q_1 - 2p_1q_1. \quad (23)$$

Table 7: Probability distribution of $D_2^b(\mathbf{p}, \mathbf{q})$, the allele-sharing dissimilarity D_2^b for a pair of individuals sampled at random from two populations with allele-frequency vectors \mathbf{p} and \mathbf{q} . The table is obtained by summing entries in Table 5.

Value of the dissimilarity (d)	$\mathbb{P}[D_2^b(\mathbf{p}, \mathbf{q}) = d]$
0	ρ_{22}
$\frac{1}{2}$	$2\rho_{21} + 2\rho_{12} + 2\rho_{11}^2 - 6\rho_{22}$
$\frac{3}{4}$	$4\rho_{11} - 4\rho_{21} - 4\rho_{12} - 4\rho_{11}^2 + 8\rho_{22}$
1	$1 - 4\rho_{11} + 2\rho_{21} + 2\rho_{12} + 2\rho_{11}^2 - 3\rho_{22}$

Figure 3B plots Eq. (23). The figure has maxima of 1 at $(p_1, q_1) = (1, 0)$ and $(0, 1)$ and equals 0 at $(0, 0)$ and $(1, 1)$. It has a saddle surface with a value of $\frac{1}{2}$ at saddle point $(p_1, q_1) = (\frac{1}{2}, \frac{1}{2})$.

$\text{Var}[D_2^b]$.

We find that

$$\begin{aligned} \mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})^2] &= \sum_{d \in \{0, \frac{1}{2}, 1\}} d^2 \mathbb{P}[D_2^b(\mathbf{p}, \mathbf{q}) = d] \\ &= 1 - \frac{7}{4}\rho_{11} + \frac{1}{4}\rho_{21} + \frac{1}{4}\rho_{12} + \frac{1}{4}\rho_{11}^2. \end{aligned} \quad (24)$$

Therefore, by $\text{Var}[D_2^b(\mathbf{p}, \mathbf{q})] = \mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})^2] - \mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})]^2$,

$$\text{Var}[D_2^b(\mathbf{p}, \mathbf{q})] = \frac{1}{4}\rho_{11} + \frac{1}{4}\rho_{21} + \frac{1}{4}\rho_{12} - \frac{3}{4}\rho_{11}^2. \quad (25)$$

For the $I = 2$ case, Eqs. (24) and (25) simplify to

$$\mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})^2] = \frac{1}{2}p_1 + \frac{1}{2}q_1 + \frac{1}{2}p_1^2 + \frac{1}{2}q_1^2 - p_1q_1 - p_1^2q_1 - p_1q_1^2 + p_1^2q_1^2 \quad (26)$$

$$\text{Var}[D_2^b(\mathbf{p}, \mathbf{q})] = \frac{1}{2}p_1 + \frac{1}{2}q_1 - 3p_1q_1 - \frac{1}{2}p_1^2 - \frac{1}{2}q_1^2 + 3p_1^2q_1 + 3p_1q_1^2 - 3p_1^2q_1^2. \quad (27)$$

Figure 3E plots Eq. (27). The variance is greatest at $(p_1, q_1) = (\frac{1}{2}, 0)$, $(\frac{1}{2}, 1)$, $(0, \frac{1}{2})$, and $(1, \frac{1}{2})$ and equals 0 at $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$. It has a local minimum at $(p_1, q_1) = (\frac{1}{2}, \frac{1}{2})$.

4.3 Comparison of D_1^b and D_2^b

The two measures for the between-population dissimilarity have the same expected value, $\mathbb{E}[D_1^b] = \mathbb{E}[D_2^b]$, if for all i , at least one of p_i , $1 - p_i$, q_i , and $1 - q_i$ is zero. The condition for equality can be seen from $\mathbb{E}[D_2^b] - \mathbb{E}[D_1^b] = \rho_{11} - \rho_{21} - \rho_{12} + \rho_{22} = \sum_{i=1}^I p_i(1 - p_i)q_i(1 - q_i)$. Excluding these equality cases, we have

$$\mathbb{E}[D_1^b] < \mathbb{E}[D_2^b]. \quad (28)$$

Note that $D_1^b \leq D_2^b$ for all possible genotype combinations in Table 1.

The inequality in Eq. (28) can be observed for the $I = 2$ case in Figure 3C, where the surface plot of $\mathbb{E}[D_2^b] - \mathbb{E}[D_1^b]$ remains greater than or equal to 0, with equality only on the boundary. The largest difference occurs at $p_1 = q_1 = \frac{1}{2}$.

Figure 3F compares the variances of D_1^b and D_2^b for the case of $I = 2$. Across most of the parameter space, $\text{Var}[D_1^b] > \text{Var}[D_2^b]$. The excess is greatest at points $(p_1, q_1) = (\frac{1}{3}, \frac{2}{3})$ and $(\frac{2}{3}, \frac{1}{3})$.

5 The relative magnitudes of $\mathbb{E}[D^w]$ and $\mathbb{E}[D^b]$

We now examine the relative magnitudes of the expectations $\mathbb{E}[D^w]$ and $\mathbb{E}[D^b]$. We determine the conditions under which the expectation of a within-population dissimilarity exceeds that of a between-population dissimilarity.

5.1 Inequality relationship between $\mathbb{E}[D_1^w(\mathbf{p})]$ and $\mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})]$

For arbitrary I , using Eqs. (3) and (16), the expression $\mathbb{E}[D_1^w(\mathbf{p})] > \mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})]$ is equivalent to

$$-2\sigma_2 + 2\sigma_3 - \sigma_4 + 2\rho_{11} - \rho_{21} - \rho_{12} + \rho_{22} > 0. \quad (29)$$

This condition can be written with vector notation. Let $\tilde{\mathbf{p}} = (p_1^2, p_2^2, \dots, p_I^2)$ and $\tilde{\mathbf{q}} = (q_1^2, q_2^2, \dots, q_I^2)$, treating \mathbf{p} , \mathbf{q} , $\tilde{\mathbf{p}}$, and $\tilde{\mathbf{q}}$ as row vectors. We have the identities $\sigma_2 = \mathbf{p}\mathbf{p}^T$, $\sigma_3 = \mathbf{p}\tilde{\mathbf{p}}^T = \tilde{\mathbf{p}}\mathbf{p}^T$, $\sigma_4 = \tilde{\mathbf{p}}\tilde{\mathbf{p}}^T$, $\rho_{11} = \mathbf{p}\mathbf{q}^T$, $\rho_{12} = \mathbf{p}\tilde{\mathbf{q}}^T$, $\rho_{21} = \tilde{\mathbf{p}}\mathbf{q}^T$, and $\rho_{22} = \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T$.

Equation (29) thus becomes

$$-2\mathbf{p}\mathbf{p}^T + 2\mathbf{p}\tilde{\mathbf{p}}^T - \tilde{\mathbf{p}}\tilde{\mathbf{p}}^T + 2\mathbf{p}\mathbf{q}^T - \tilde{\mathbf{p}}\mathbf{q}^T - \mathbf{p}\tilde{\mathbf{q}}^T + \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T > 0, \quad (30)$$

which simplifies to

$$\begin{pmatrix} \mathbf{p} & \mathbf{p} - \tilde{\mathbf{p}} \end{pmatrix} \begin{pmatrix} (\mathbf{p} - \mathbf{q})^T \\ [(\mathbf{p} - \tilde{\mathbf{p}}) - (\mathbf{q} - \tilde{\mathbf{q}})]^T \end{pmatrix} < 0. \quad (31)$$

For $I = 2$, we can further simplify this condition on p_1 and q_1 , noting $p_2 = 1 - p_1$ and $q_2 = 1 - q_1$.

Theorem 1. Consider a locus with $I = 2$ distinct alleles. For individuals sampled from two populations with allele frequency vectors $\mathbf{p} = (p_1, 1 - p_1)$ and $\mathbf{q} = (q_1, 1 - q_1)$, $\mathbb{E}[D_1^w(\mathbf{p})] > \mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})]$ holds if and only if

$$\begin{cases} 0 < q_1 < p_1 & \text{if } 0 < p_1 \leq a, \\ g(p_1) < q_1 < p_1 & \text{if } a \leq p_1 < \frac{1}{2}, \\ p_1 < q_1 < g(p_1) & \text{if } \frac{1}{2} < p_1 \leq 1 - a, \\ p_1 < q_1 < 1 & \text{if } 1 - a \leq p_1 < 1, \end{cases} \quad (32)$$

where

$$g(x) = \frac{2x^3 - 4x^2 + 4x - 1}{2x(1 - x)},$$

and

$$a = \frac{1}{3} \left(\frac{\sqrt[3]{3\sqrt{33} - 13}}{2^{2/3}} - \frac{2^{5/3}}{\sqrt[3]{3\sqrt{33} - 13}} + 2 \right) \approx 0.3522$$

is the unique real root of $2x^3 - 4x^2 + 4x - 1$.

Proof. We simplify Eq. (29) noting $p_2 = 1 - p_1$ and $q_2 = 1 - q_1$. To find the region where $\mathbb{E}[D_1^w(\mathbf{p})] > \mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})]$, we solve the polynomial inequality

$$p_1 - q_1 - 4p_1^2 + 4p_1q_1 + 4p_1^3 - 2p_1^2q_1 - 2p_1q_1^2 - 2p_1^4 + 2p_1^2q_1^2 > 0, \quad (33)$$

with $0 \leq p_1 \leq 1$ and $0 \leq q_1 \leq 1$. Solving for q_1 in terms of p_1 , we find that the expression in Eq. (33) is 0 at $q_1 = p_1$ and at $q_1 = g(p_1)$, and for fixed p_1 , it is positive when q_1 lies between the two roots. The unique real root for $g(x) = x$ is at $x = \frac{1}{2}$, so that $g(p_1) < p_1$ for $p_1 < \frac{1}{2}$ and $g(p_1) > p_1$ for $p_1 > \frac{1}{2}$.

For $0 \leq p_1 < \frac{1}{2}$, $g(p_1) < 0$ for $p_1 < a$, so that for $0 \leq p_1 \leq a$, the region where the expression in Eq. (33) is positive includes the full interval $(0, p_1)$ for q_1 . For $a \leq p_1 \leq \frac{1}{2}$, it is positive only in interval $(g(p_1), p_1)$ for q_1 .

For $\frac{1}{2} < p_1 < 1$, $g(p_1) = 1$ for $p_1 = 1 - a$, with $g(p_1) < 1$ for p_1 in $[\frac{1}{2}, 1 - a)$ and $g(p_1) > 1$ for p_1 in $(1 - a, 1]$. Hence, for p_1 in $[\frac{1}{2}, 1 - a]$, the expression in Eq. (33) is positive for q_1 in $(p_1, g(p_1))$, and for p_1 in $[1 - a, 1]$, it is positive for q_1 in $(p_1, 1)$. \square

Figure 4A plots the region identified in Theorem 1. That a nonempty region exists indicates that sometimes, allele frequencies for a biallelic locus produce a within-population dissimilarity that exceeds the between-population dissimilarity. Note that because the choice of which allele is labeled 1 and which is labeled 2 is arbitrary, (p_1, q_1) is included in the region if and only if $(1 - p_1, 1 - q_1)$ is also included.

We can calculate the area of the region in the unit square representing the probability $\mathbb{P}(\mathbb{E}[D_1^w] > \mathbb{E}[D_1^b])$ under the assumption that p_1 and q_1 are independently and identically distributed with uniform-[0,1] distribution:

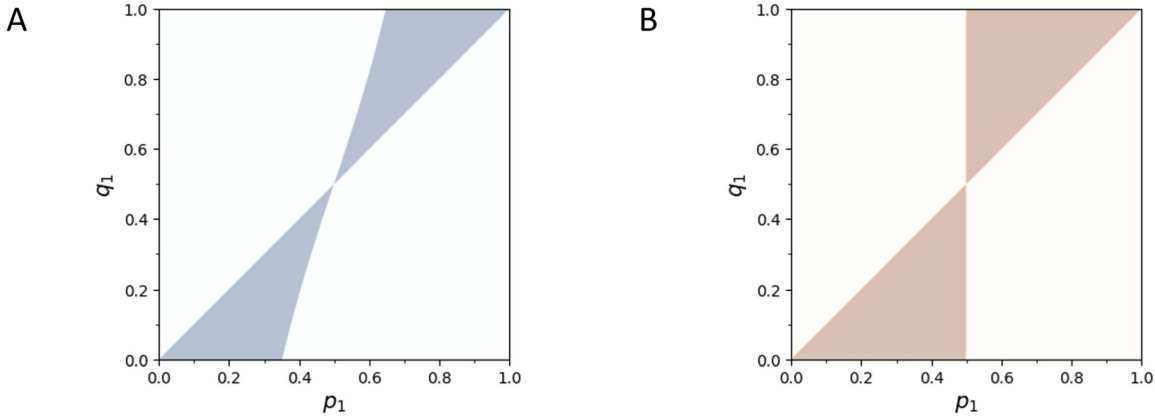


Figure 4: Values of (p_1, q_1) for which $\mathbb{E}[D_1^w] > \mathbb{E}[D_1^b]$ in the case of $I = 2$ alleles, shaded in color. (A) D_1 , Theorem 1. (B) D_2 , Theorem 2.

$$\begin{aligned}
& \mathbb{P}(\mathbb{E}[D_1^w] > \mathbb{E}[D_1^b]) \\
&= \int_{p_1=0}^a \int_{q_1=0}^{p_1} 1 dq_1 dp_1 + \int_{p_1=a}^{\frac{1}{2}} \int_{q_1=g(p_1)}^{p_1} 1 dq_1 dp_1 + \int_{p_1=\frac{1}{2}}^{1-a} \int_{q_1=p_1}^{g(p_1)} 1 dq_1 dp_1 + \int_{p_1=1-a}^1 \int_{q_1=p_1}^1 1 dq_1 dp_1 \\
&= 2 \left[\int_{p_1=0}^a p_1 dp_1 + \int_{p_1=a}^{\frac{1}{2}} \frac{-4p_1^3 + 6p_1^2 - 4p_1 + 1}{2p_1(1-p_1)} dp_1 \right] \\
&= -a^2 + 2a - \frac{1}{2} - 2 \log 2 - \log a - \log(1-a) \\
&\approx 0.17179.
\end{aligned} \tag{34}$$

To evaluate $\mathbb{P}(\mathbb{E}[D_1^w] > \mathbb{E}[D_1^b])$ more generally, for each I from 2 to 20, we perform a simulation. In particular, for each I , we consider independently and identically distributed vectors \mathbf{p} and \mathbf{q} from the uniform distribution over the simplex Δ^{I-1} (the Dirichlet-(1, 1, ..., 1) distribution, where the vector of 1's has length I). We sample 100,000 replicate pairs (\mathbf{p}, \mathbf{q}) , and for each pair we evaluate if $\mathbb{E}[D_1^w] > \mathbb{E}[D_1^b]$.

Figure 5A plots the resulting probability. We can observe that for $I = 2$, the simulated $\mathbb{P}(\mathbb{E}[D_1^w] > \mathbb{E}[D_1^b])$ accords with the analytical value in Eq. (34). The probability then decreases with increasing I .

5.2 Inequality relationship between $\mathbb{E}[D_2^w(\mathbf{p})]$ and $\mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})]$

For arbitrary I , via Eqs. (9) and (22), the expression $\mathbb{E}[D_2^w(\mathbf{p})] > \mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})]$ is equivalent to

$$\rho_{11} - \sigma_2 > 0. \tag{35}$$

With $\sigma_2 = \mathbf{p}\mathbf{p}^T$ and $\rho_{11} = \mathbf{p}\mathbf{q}^T$, Eq. (35) thus becomes

$$\mathbf{p}(\mathbf{p} - \mathbf{q})^T < 0. \tag{36}$$

For $I = 2$, Eq. (36) can be simplified to a condition on p_1 and q_1 , again noting $p_2 = 1 - p_1$ and $q_2 = 1 - q_1$.

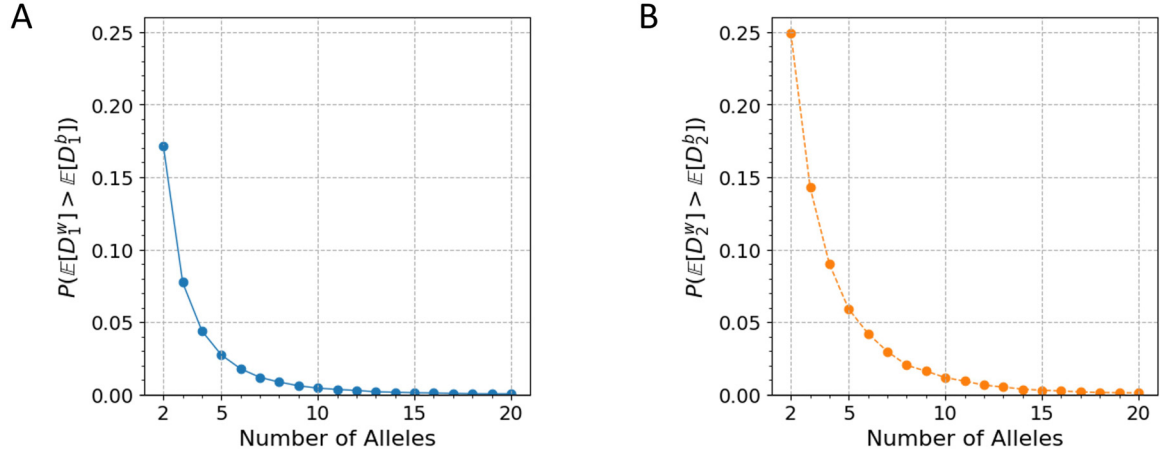


Figure 5: The probability $\mathbb{P}(\mathbb{E}[D^w] > \mathbb{E}[D^b])$ for simulated pairs of allele frequency vectors (\mathbf{p}, \mathbf{q}) with I distinct alleles. (A) D_1 . (B) D_2 . Independent and identical uniform distributions are simulated for each I , $2 \leq I \leq 20$, by drawing uniformly from the simplex Δ^{I-1} (100,000 replicates).

Theorem 2. Consider a locus with $I = 2$ distinct alleles. For individuals sampled from two populations with allele frequency vectors $\mathbf{p} = (p_1, 1 - p_1)$ and $\mathbf{q} = (q_1, 1 - q_1)$, $\mathbb{E}[D_2^w(\mathbf{p})] > \mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})]$ holds if and only if

$$\begin{cases} 0 < q_1 < p_1 & \text{if } 0 < p_1 < \frac{1}{2}, \\ p_1 < q_1 < 1 & \text{if } \frac{1}{2} < p_1 < 1. \end{cases} \quad (37)$$

Proof. With $p_2 = 1 - p_1$ and $q_2 = 1 - q_1$, Eq. (35) simplifies to

$$p_1 - q_1 - 2p_1^2 + 2p_1q_1 > 0.$$

Solving this inequality, we arrive at the result. \square

Figure 4B plots the region identified in Theorem 2. This region describes the locations in which allele frequencies for a biallelic locus produce a within-population dissimilarity that exceeds the between-population dissimilarity. As is true for D_1 , (p_1, q_1) is included in the region if and only if $(1 - p_1, 1 - q_1)$ is also included.

The area of the region in the unit square, representing $\mathbb{P}(\mathbb{E}[D_2^w] > \mathbb{E}[D_2^b])$ under the assumption that p_1 and q_1 are independently and identically distributed with uniform-[0,1] distribution, is straightforward:

$$\begin{aligned} & \mathbb{P}(\mathbb{E}[D_2^w] > \mathbb{E}[D_2^b]) \\ &= \int_{p_1=0}^{\frac{1}{2}} \int_{q_1=0}^{p_1} 1 \, dq_1 \, dp_1 + \int_{p_1=\frac{1}{2}}^1 \int_{q_1=p_1}^1 1 \, dq_1 \, dp_1 \\ &= \frac{1}{4}. \end{aligned} \quad (38)$$

We evaluate $\mathbb{P}(\mathbb{E}[D_2^w] > \mathbb{E}[D_2^b])$ for each I from 2 to 20 by simulation. For each I , we consider independently and identically distributed vectors \mathbf{p} and \mathbf{q} from the uniform distribution over the simplex Δ^{I-1} (the Dirichlet-(1, 1, ..., 1) distribution), sampling 100,000 replicate pairs (\mathbf{p}, \mathbf{q}) , and evaluating the fraction of pairs for which $\mathbb{E}[D_2^w] > \mathbb{E}[D_2^b]$.

Figure 5B plots the resulting probability, illustrating the agreement between the simulated $\mathbb{P}(\mathbb{E}[D_2^w] > \mathbb{E}[D_2^b])$ and the analytical value in Eq. (38) for $I = 2$. The probability then decreases as I increases.

5.3 Comparison of the $\mathbb{E}[D^w] - \mathbb{E}[D^b]$ inequalities for D_1 and D_2

The inequality $\mathbb{E}[D^w] > \mathbb{E}[D^b]$, where the mean dissimilarity between individuals from the same population exceeds that between individuals from different populations, holds under different scenarios for D_1 and D_2 . Comparing Eqs. (34) and (38), we see that for the case of $I = 2$, $\mathbb{E}[D_1^w] > \mathbb{E}[D_1^b]$ holds over a smaller fraction of the parameter space than the corresponding inequality $\mathbb{E}[D_2^w] > \mathbb{E}[D_2^b]$ (Figure 4). Further, if the former inequality holds, then the latter always holds as well.

In Figure 5, we also observe that the probabilities $\mathbb{P}(\mathbb{E}[D^w] > \mathbb{E}[D^b])$ are higher for D_2 than for D_1 in simulations with different numbers of alleles. Hence, use of D_2 rather than D_1 produces a greater probability that the within-population genetic dissimilarity exceeds the between-population dissimilarity.

6 The relative magnitudes of $\overline{\mathbb{E}[D^w]}$ and $\mathbb{E}[D^b]$

We have seen that both for D_1 and for D_2 , it is possible for the expected dissimilarity $\mathbb{E}[D^w]$ of random pairs of individuals within a population to exceed the expected dissimilarity $\mathbb{E}[D^b]$ of random pairs between that population and a second population. However, we will see that for a pair of populations, the mean of their two within-population dissimilarities never exceeds their between-population dissimilarity.

For a pair of populations with allele frequency vectors \mathbf{p} and \mathbf{q} , let $\overline{\mathbb{E}[D_1^w(\mathbf{p}, \mathbf{q})]} = \frac{1}{2}(\mathbb{E}[D_1^w(\mathbf{p})] + \mathbb{E}[D_1^w(\mathbf{q})])$, and let $\overline{\mathbb{E}[D_2^w(\mathbf{p}, \mathbf{q})]} = \frac{1}{2}(\mathbb{E}[D_2^w(\mathbf{p})] + \mathbb{E}[D_2^w(\mathbf{q})])$.

6.1 Inequality relationship between $\overline{\mathbb{E}[D_1^w]}(\mathbf{p}, \mathbf{q})$ and $\mathbb{E}[D_1^b](\mathbf{p}, \mathbf{q})$

Theorem 3. $\overline{\mathbb{E}[D_1^w(\mathbf{p}, \mathbf{q})]} \leq \mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})]$, with equality if and only if $\mathbf{p} = \mathbf{q}$.

Proof. We use Eqs. (3) and (16) to rewrite $\overline{\mathbb{E}[D_1^w(\mathbf{p}, \mathbf{q})]} - \mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})]$, obtaining

$$\begin{aligned} & \frac{\mathbb{E}[D_1^w(\mathbf{p})] + \mathbb{E}[D_1^w(\mathbf{q})]}{2} - \mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})] \\ &= -\sigma_2 + \sigma_3 - \frac{1}{2}\sigma_4 - \tau_2 + \tau_3 - \frac{1}{2}\tau_4 - \rho_{21} - \rho_{12} + \rho_{22} + 2\rho_{11}. \end{aligned}$$

Rewriting in terms of the vectors \mathbf{p} , \mathbf{q} , $\tilde{\mathbf{p}}$, and $\tilde{\mathbf{q}}$, we have

$$\begin{aligned} & \frac{\mathbb{E}[D_1^w(\mathbf{p})] + \mathbb{E}[D_1^w(\mathbf{q})]}{2} - \mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})] \\ &= -(\mathbf{p} - \mathbf{q})(\mathbf{p} - \mathbf{q})^T + (\mathbf{p} - \mathbf{q})(\tilde{\mathbf{p}} - \tilde{\mathbf{q}})^T - \frac{1}{2}(\tilde{\mathbf{p}} - \tilde{\mathbf{q}})(\tilde{\mathbf{p}} - \tilde{\mathbf{q}})^T \\ &= -\frac{1}{2}\|\mathbf{p} - \mathbf{q}\|^2 - \frac{1}{2}\|(\mathbf{p} - \mathbf{q}) - (\tilde{\mathbf{p}} - \tilde{\mathbf{q}})\|^2 \\ &\leq 0. \end{aligned}$$

Equality is reached in the last step if and only if $\mathbf{p} = \mathbf{q}$. □

6.2 Inequality relationship between $\overline{\mathbb{E}[D_2^w(\mathbf{p}, \mathbf{q})]}$ and $\mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})]$

Theorem 4. $\overline{\mathbb{E}[D_2^w(\mathbf{p}, \mathbf{q})]} \leq \mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})]$, with equality if and only if $\mathbf{p} = \mathbf{q}$.

Proof. We rewrite $\overline{\mathbb{E}[D_2^w(\mathbf{p}, \mathbf{q})]} - \mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})]$ using Eqs. (9) and (22):

$$\begin{aligned} & \frac{\mathbb{E}[D_2^w(\mathbf{p})] + \mathbb{E}[D_2^w(\mathbf{q})]}{2} - \mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})] \\ &= -\frac{\sigma_2}{2} - \frac{\tau_2}{2} + \rho_{11}. \end{aligned}$$

In terms of the vectors \mathbf{p} and \mathbf{q} , we have

$$\begin{aligned} & \frac{\mathbb{E}[D_2^w(\mathbf{p})] + \mathbb{E}[D_2^w(\mathbf{q})]}{2} - \mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})] \\ &= -\frac{1}{2}\mathbf{p}\mathbf{p}^T - \frac{1}{2}\mathbf{q}\mathbf{q}^T + \mathbf{p}\mathbf{q}^T \\ &= -\frac{1}{2}\|\mathbf{p} - \mathbf{q}\|^2 \\ &\leq 0, \end{aligned}$$

with equality if and only if $\mathbf{p} = \mathbf{q}$. □

6.3 Comparison of the $\overline{\mathbb{E}[D^w]} - \mathbb{E}[D^b]$ inequalities for \mathcal{D}_1 and \mathcal{D}_2

The inequality $\overline{\mathbb{E}[D^w(\mathbf{p}, \mathbf{q})]} \leq \mathbb{E}[D^b(\mathbf{p}, \mathbf{q})]$, with equality if and only if $\mathbf{p} = \mathbf{q}$, holds for both \mathcal{D}_1 and \mathcal{D}_2 . Comparing the proofs of Theorems 3 and 4, we see that

$$\overline{\mathbb{E}[D_1^w(\mathbf{p}, \mathbf{q})]} - \mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})] = \overline{\mathbb{E}[D_2^w(\mathbf{p}, \mathbf{q})]} - \mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})] - \frac{1}{2}\|(\mathbf{p} - \mathbf{q}) - (\tilde{\mathbf{p}} - \tilde{\mathbf{q}})\|^2. \quad (39)$$

The extent to which $\overline{\mathbb{E}[D_1^w(\mathbf{p}, \mathbf{q})]} < \mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})]$ for $\mathbf{p} \neq \mathbf{q}$, or $\overline{\mathbb{E}[D_1^w(\mathbf{p}, \mathbf{q})]} - \mathbb{E}[D_1^b(\mathbf{p}, \mathbf{q})]$, has a greater absolute value than the corresponding extent to which $\overline{\mathbb{E}[D_2^w(\mathbf{p}, \mathbf{q})]} < \mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})]$ for $\mathbf{p} \neq \mathbf{q}$, or $\overline{\mathbb{E}[D_2^w(\mathbf{p}, \mathbf{q})]} - \mathbb{E}[D_2^b(\mathbf{p}, \mathbf{q})]$.

7 Data analysis

7.1 Data

Our theoretical analysis predicts features of dissimilarities \mathcal{D}_1 and \mathcal{D}_2 in within-population and between-population computations. To compare to empirical observations, we examine multiallelic microsatellite data from the Human Genome Diversity Project (HGDP-CEPH panel). We consider the 1048 individuals and 783 microsatellite loci from Rosenberg et al. (2005), employing the H1048 subset of the HGDP-CEPH panel (Rosenberg 2006). We follow previous uses of the HGDP-CEPH panel in considering 53 populations and 7 geographic regions. We focus on 30 populations for which the number of sampled individuals is greater than 15. Across these 30 populations, the total number of individuals considered is 813.

7.2 Theoretical computations

For our theoretical calculations, given a population in the data set and a locus, we compute allele frequencies. We then apply our theoretical formulas to the allele frequency vectors. Note that if a locus is missing genotypes

in an individual, then we omit that individual from the calculation of population allele frequencies at the locus, so that we maintain the property that allele frequencies at a locus in a population sum to 1.

7.3 Empirical computations

For empirical calculations, we consider the actual diploid individuals in the HGDP-CEPH data, for within-population computations comparing all pairs of individuals within a population. For between-population computations, we compare all pairs of individuals, one each from two populations. Pairwise dissimilarities between diploid genotypes are obtained according to Table 1. We compute within-population and between-population dissimilarities as the means across relevant pairs, and we compute variances of dissimilarity distributions across pairs of individuals.

For this analysis, we omit individuals with missing data prior to computation of empirical ASD values. In between-population comparisons, all allelic types present in one but not the other population are assigned a frequency of 0 in the population in which they are absent.

We perform the theoretical and empirical calculations for all 783 loci.

7.4 Results of data analysis

Figure 6 compares empirical and theoretical means and variances of within-population dissimilarities across pairs of individuals, considering 100 randomly sampled loci in 30 populations. Figure 6A compares the empirical value of $\mathbb{E}[D_1^w]$ computed by averaging D_1^w values for all pairs of sampled individuals with the theoretical value predicted from the allele frequencies and Eq. (3). The theoretical calculation generally predicts the empirical dissimilarity, with most points clustering along the diagonal ($r = 0.962$). In Figure 6B, a similar plot for $\mathbb{E}[D_2^w]$ using Eq. (9) for the theoretical computation produces closer agreement between the empirical and theoretical values ($r = 0.999$).

Figure 6C and D compare empirical and theoretical variances across pairs of individuals for within-population dissimilarities, using Eqs. (6) and (12) for the theoretical computation. The theoretical variance predicts the empirical variance, but the agreement is not as close as for the mean ($r = 0.676$ for $\text{Var}[D_1^w]$, $r = 0.732$ for $\text{Var}[D_2^w]$).

Figure 7 plots analogous comparisons for between-population dissimilarities, considering a subset of loci from Figure 6. In Figure 7A, we see a close relationship between empirical $\mathbb{E}[D_1^b]$ and theoretical $\mathbb{E}[D_1^b]$ similar to the relationship observed in Figure 6A ($r = 0.943$). As was seen in Figure 6B, in Figure 7B, we see a stronger relationship between the empirical value of $\mathbb{E}[D_2^b]$ and the theoretical value ($r = 1.000$).

Figure 7C and D consider relationships between empirical and theoretical between-population variances for D_1 and D_2 . As was observed in Figure 6C and D, empirical and theoretical variance are correlated ($r = 0.676$ for $\text{Var}[D_1^b]$, $r = 0.731$ for $\text{Var}[D_2^b]$), but the agreement for variances is not as close as for the mean.

Figure 8 empirically examines the inequalities in Theorems 3 and 4 stating that when computed from allele frequencies, the mean of the within-population dissimilarities for two populations is always less than the dissimilarity between them. It shows all population pairs from Figures 6 and 7 with a single random locus.

In Figure 8A, we find that the theoretical values of $\mathbb{E}[D_1^b]$ and $\mathbb{E}[D_1^w]$, computed from allele frequencies alone, follow the predicted inequality, with $\mathbb{E}[D_1^b] > \mathbb{E}[D_1^w]$. However, the theorem does not necessarily apply to dissimilarities computed from actual diploid individuals, and indeed, some exceptions are observed in which the empirical $\mathbb{E}[D_1^b]$ is smaller than $\mathbb{E}[D_1^w]$ (Figure 8C). Similar results hold for $\mathbb{E}[D_2^b]$ and $\mathbb{E}[D_2^w]$ in Figure 8B and D.

Figure 9 tabulates the fraction of loci for which the empirical within-population dissimilarity of a population (denoted Population 1) exceeds the population's empirical between-population dissimilarity with a second population (Population 2), or $\mathbb{E}[D^w] > \mathbb{E}[D^b]$. The populations are arranged geographically, following a general decrease in within-population genetic diversity with migration distance from Africa, as measured by expected

heterozygosity $1 - \sigma_2$ (Prugnolle et al. 2005; Ramachandran et al. 2005). In Figure 9A, for D_1 , if Population 1 is a population with relatively low within-population heterozygosity, such as a Native American population, then its within-population dissimilarity rarely exceeds its between-population dissimilarity with a second population (rightmost columns). The fraction of loci for which $\mathbb{E}[D^w] > \mathbb{E}[D^b]$ is greatest for intermediate-heterozygosity South Asian populations (central columns). If Population 2 is a high-heterozygosity African population, then for all non-African choices of Population 1, the within-population dissimilarity of Population 1 rarely exceeds the between-population dissimilarity with an African Population 2 (bottom rows). Similar patterns are seen in Figure 9B for D_2 , with the additional observation that the within-population dissimilarity of Population 1 often exceeds the between-population dissimilarity when low-heterozygosity Native American populations are placed in the role of Population 2 (top rows).

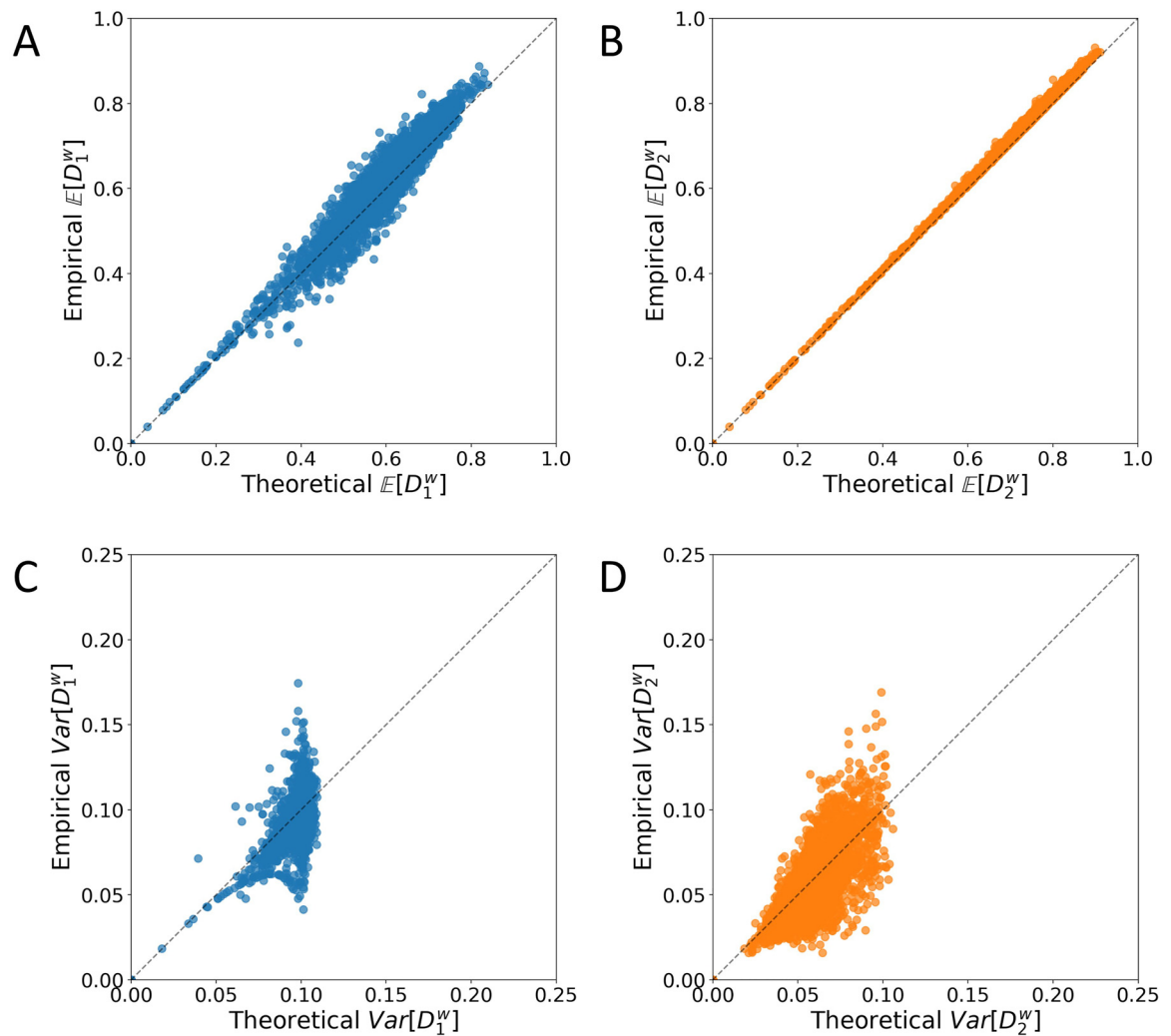


Figure 6: Empirical and theoretical mean and variance of within-population allele-sharing dissimilarities. Each panel considers 100 randomly sampled loci (among 783) in 30 populations with sample size greater than 15 ($100 \times 30 = 3000$ data points in each panel). (A) $\mathbb{E}[D_1^w]$. (B) $\mathbb{E}[D_2^w]$. (C) $\text{Var}[D_1^w]$. (D) $\text{Var}[D_2^w]$. Empirical values rely on dissimilarity calculations according to Table 1 from pairs of diploid individuals, and theoretical values are calculated from allele frequencies according to Eqs. (3), (6), (9) and (12).

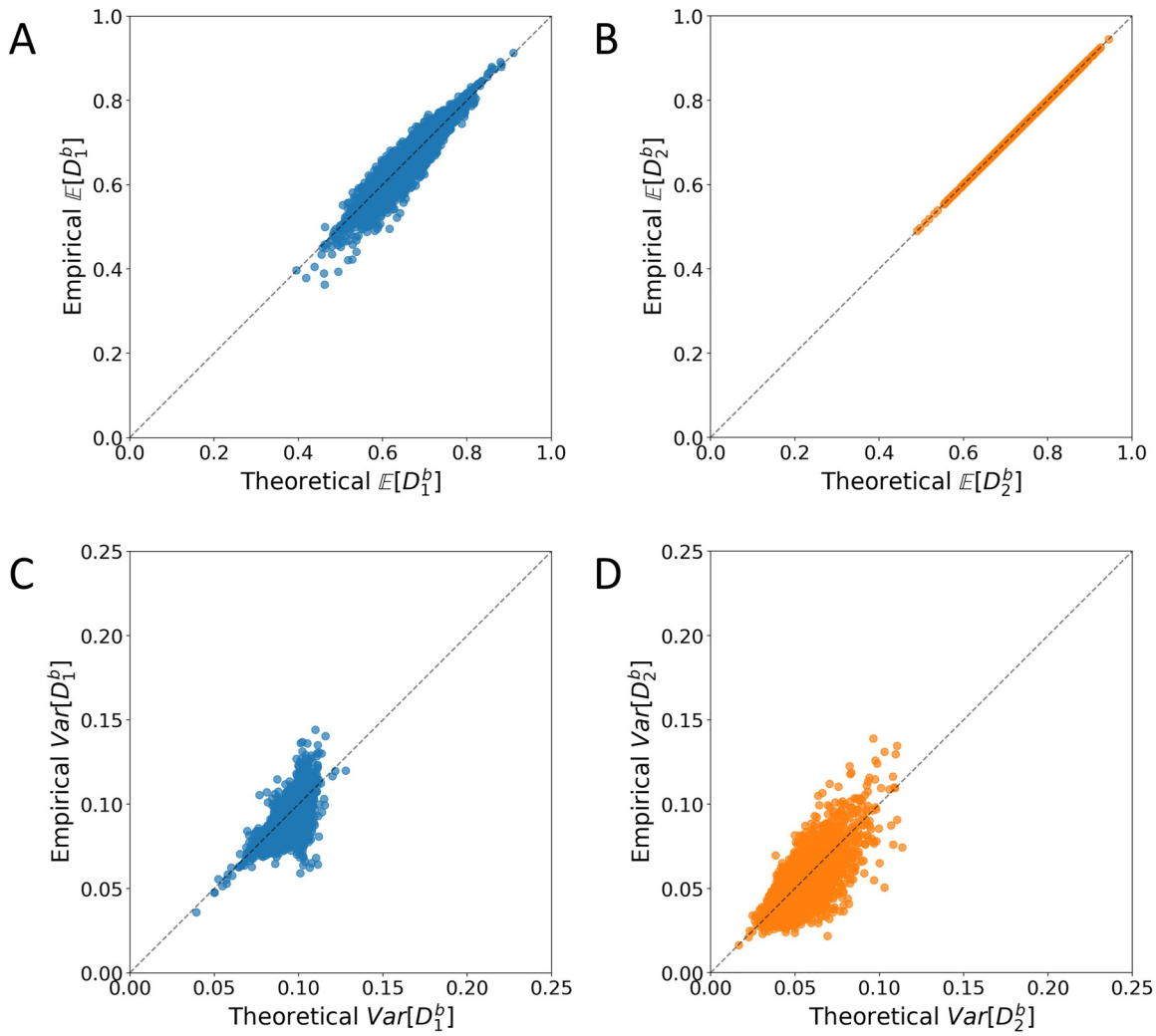


Figure 7: Empirical and theoretical mean and variance of between-population allele-sharing dissimilarities. Each panel considers 10 randomly sampled loci in pairs among the 30 populations with sample size greater than 15 ($10 \times \binom{30}{2} = 4350$ data points in each panel). The 10 loci are taken from among those used in Figure 6. (A) $E[D_1^b]$. (B) $E[D_2^b]$. (C) $Var[D_1^b]$. (D) $Var[D_2^b]$. Empirical values rely on dissimilarity calculations according to Table 1 from pairs of diploid individuals, and theoretical values are calculated from allele frequencies according to Eqs. (16), (19), (22) and (25).

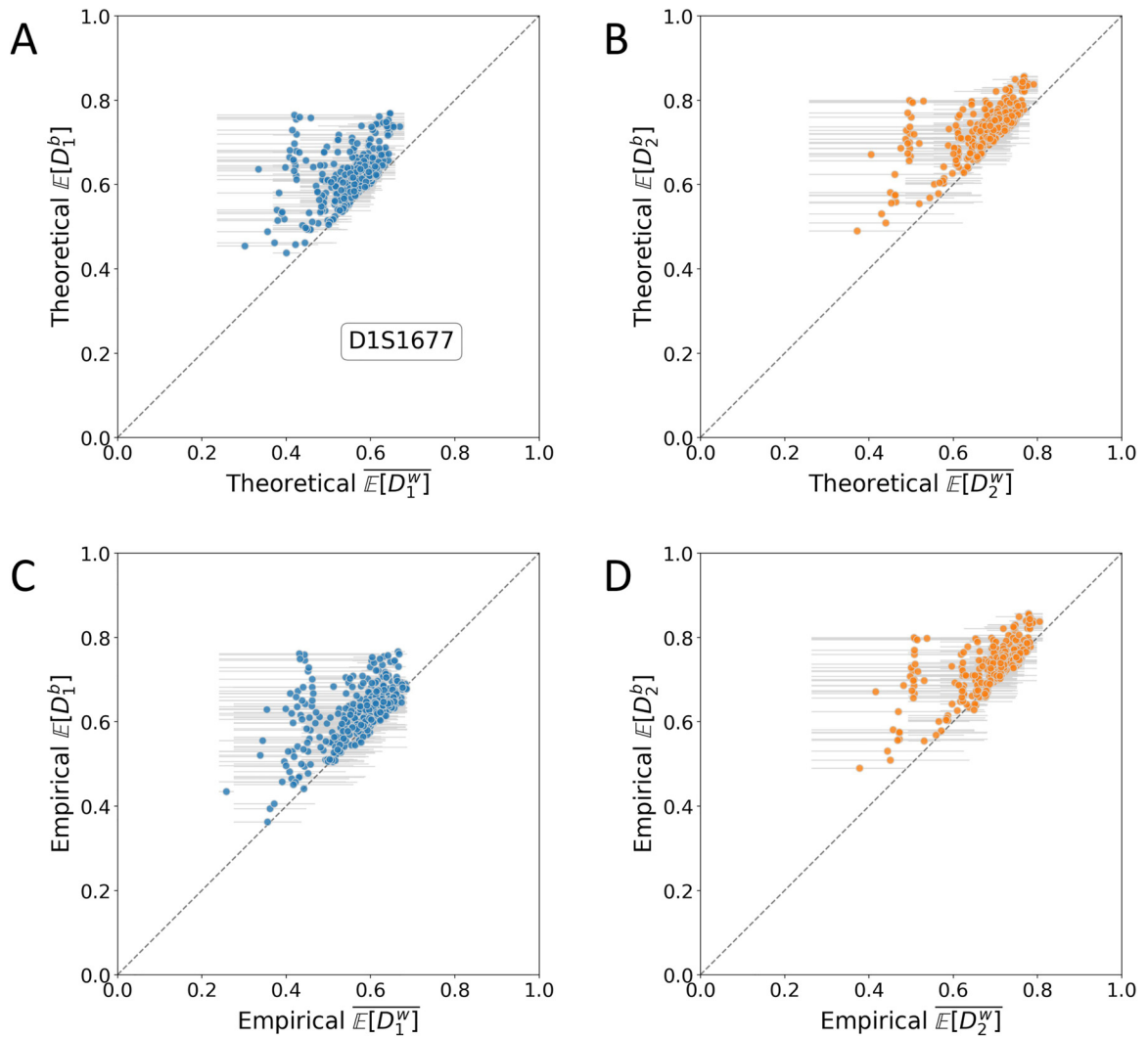


Figure 8: Empirical and theoretical $\mathbb{E}[D^b]$ and $\mathbb{E}[D^w]$. Each panel considers a random locus, D1S1677, in 435 pairs of populations with sample size greater than 15. The locus is among those used in Figures 6 and 7. The upper left triangle is the region in which the between-population dissimilarity of two populations exceeds the mean of the within-population dissimilarities of the two populations, $\mathbb{E}[D^b] > \mathbb{E}[D^w]$, as proven for theoretical dissimilarities (Theorems 3 and 4). The two ends of a horizontal gray line indicate the $\mathbb{E}[D^w]$ values for two populations whose mean within-population dissimilarity is plotted at the midpoint of the line. (A) Theoretical values of D_1 . (B) Theoretical values of D_2 . (C) Empirical values of D_1 . (D) Empirical values of D_2 .

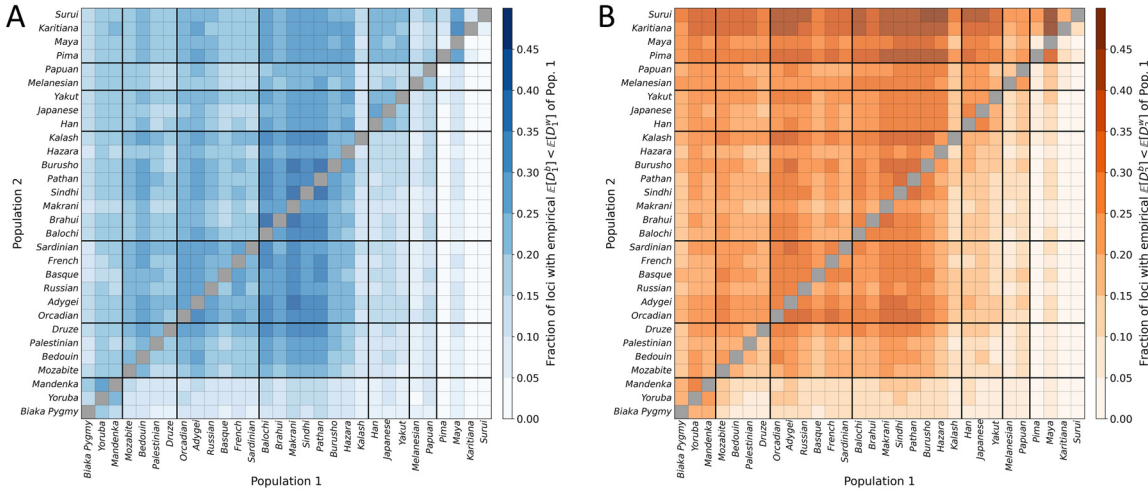


Figure 9: Fraction of loci for which $\mathbb{E}[D^b] < \mathbb{E}[D^w]$. Each panel considers all 783 loci in pairs among the 30 populations with sample size greater than 15. Each cell denotes a pair of populations, with Population 1 considered for the within-population dissimilarity. Geographical regions are separated by bold black lines. (A) D_1 . (B) D_2 .

8 Discussion

Allele-sharing statistics are often used to quantify genetic dissimilarity within and between populations. Because they typically share a larger number of recent ancestors, individuals from the same population might be predicted to possess a lower genetic dissimilarity than those from different populations. We have mathematically explored the circumstances under which this prediction fails, when the genetic dissimilarity within a population exceeds the genetic dissimilarity between two populations. The analysis characterizes the properties of allele frequency vectors that give rise to this counterintuitive scenario, illustrating its occurrence in human population-genetic data.

When does within-population dissimilarity for a population exceed between-population dissimilarity with a second population? The conditions that permit this inequality in the case of $I = 2$ alleles are instructive (Theorems 1 and 2 and Figure 4). In this case, two populations have unbalanced allele frequencies, with Population 2 more unbalanced than Population 1, but the two populations are similar in their frequencies. In Population 1, dissimilarity is generated from comparisons of homozygotes for one allele and homozygotes for the other allele. However, because Population 2 has allele frequencies that are more unbalanced than those of Population 1, fewer comparisons of distinct homozygotes occur in the between-population comparison. This phenomenon results in a within-population dissimilarity in Population 1 that exceeds the between-population dissimilarity. Beyond $I = 2$, such an excess is observed in empirical calculations with $I \geq 2$ alleles (Figure 9), as well as in simulations, though with decreasing probability as I increases (Figure 5).

Although a population can possess greater within-population dissimilarity than its between-population dissimilarity to a second population, we find that for arbitrary numbers of alleles I , it is not possible for *both* populations in a pair to possess greater within-population dissimilarity than the between-population dissimilarity (Theorems 3 and 4). In data, “theoretical” dissimilarities obtained by treating allele frequencies in the data as parametric frequencies of two populations follow this inequality strictly, with greater between-population dissimilarity than at least one of the two within-population dissimilarities (Figure 8A and B). Similarly, the mean of the two within-population dissimilarities is strictly less than the between-population dissimilarity in theoretical calculations (Figure 8A and B); while “empirical” dissimilarities calculated from individual genotypes *can* violate the inequality, we find that these violations are generally mild (Figure 8C and D).

The results can contribute to understanding unexpected phenomena involving allele-sharing dissimilarities in human populations. We have seen that within-population dissimilarities in Population 1 sometimes

exceed between-population dissimilarities, often in comparisons that involve a lower-diversity Population 2 and a higher-diversity Population 1 (Figure 9); in essence, a high-diversity population can possess enough variation that its inter-individual dissimilarity can exceed the dissimilarity between populations. Our theoretical calculations provide a basis for this scenario, and in fact, we saw for $I = 2$ that it is not unlikely in certain parts of the allele frequency space (Figure 4).

Our theoretical analysis deepens a line of inquiry on mathematical effects on allele-sharing. For each of two dissimilarity functions, we have obtained probability distributions of within- and between-population allele-sharing dissimilarities across pairs of individuals as functions of allele frequencies (Tables 3, 4, 6, 7), focusing on the mean and variance of the dissimilarity statistics (Eqs. (3), (6), (9), (12), (16), (19), (22) and (25)). The expressions for these quantities, and inequalities concerning their relationships (Theorems 1–4), augment previous efforts on the mathematics of allele-sharing dissimilarities in terms of allele frequencies (Chakraborty and Jin 1993; Tal 2013).

The two variants of allele-sharing dissimilarity that we studied, D_1 and D_2 , share many features. For $I = 2$ and $I = 3$ alleles, the expected values of D_1^w and D_2^w are maximal when all alleles have the same frequency (Figures 1A and 2A, B). Trends in expectations of D_1^b and D_2^b at $I = 2$ are also similar (Figure 3A and B), as are the regions in which $\mathbb{E}[D^w] > \mathbb{E}[D^b]$ for $I = 2$ (Figure 4), and the simulated probabilities $\mathbb{P}(\mathbb{E}[D^w] > \mathbb{E}[D^b])$ for $I \geq 2$ (Figure 5).

However, some consistent differences between the two dissimilarities are also observed. $D_2 \geq D_1$ for all genotypes (Table 1), and hence, $\mathbb{E}[D_2^w] \geq \mathbb{E}[D_1^w]$ (Figures 1 and 2C and Eq. (15)) and $\mathbb{E}[D_2^b] \geq \mathbb{E}[D_1^b]$ (Figure 3C and Eq. (28)). Although both dissimilarities have $\overline{\mathbb{E}[D^w]} \leq \mathbb{E}[D^b]$ (Theorems 3 and 4), $\overline{\mathbb{E}[D_1^w]} - \mathbb{E}[D_1^b] \leq \overline{\mathbb{E}[D_2^w]} - \mathbb{E}[D_2^b]$ (Eq. (39)), so that the extent to which $\overline{\mathbb{E}[D^w]}$ lies below $\mathbb{E}[D^b]$ has greater magnitude for D_1 .

The within-population variance across pairs of individuals is not uniformly higher for either dissimilarity (Figures 1B and 2F); at $I = 2$, it has different shapes, as $\text{Var}[D_2^w]$ has two maxima, whereas $\text{Var}[D_1^w]$ has only one (Figure 1B). D_2 has larger regions in which $\mathbb{E}[D^w] > \mathbb{E}[D^b]$ for $I = 2$ (Figure 4) and for $I \geq 2$ (Figure 5). In the empirical analysis, D_2 has a closer match between empirical and theoretical mean values of the dissimilarity (Figures 6B and 7B). Its patterns in the fraction of loci for which $\mathbb{E}[D^w] > \mathbb{E}[D^b]$ align more closely with the heterozygosity values of the populations, with the probability of $\mathbb{E}[D^w] > \mathbb{E}[D^b]$ larger when Population 1 is a higher-diversity population and Population 2 is a lower-diversity population (Figure 9B). Notably, expressions for $\mathbb{E}[D_2]$ are closely tied to heterozygosity (Eq. (9)) and its between-population analogue (Eq. (22)), potentially explaining the tighter connection of heterozygosity to its associated observations. Thus, the lesser-used D_2 – which, unlike D_1 , allows the dissimilarity of an individual and itself to be nonzero (Table 1) – does possess a more easily interpreted pattern in the probability that $\mathbb{E}[D^w] > \mathbb{E}[D^b]$.

Does our analysis suggest a preference for D_1 over D_2 , or vice versa? To summarize, D_1 has been used more frequently than D_2 , and it also has the property that the dissimilarity of an individual and itself is zero. The less frequently used D_2 does not have this property, but it produces simpler expressions for its within-population and between-population expectations, with more natural interpretations of those expectations and their consequences. We conclude that although D_1 has a number of desirable properties, D_2 does as well, and it perhaps merits attention commensurate with that given to D_1 .

This work has several possible extensions. We have focused on the first and second moments of allele-sharing dissimilarities across pairs of individuals; the full distributions (Tables 3, 4, 6, 7) could also be further investigated. We examined $I = 2$ in the greatest detail, but special cases that fix a maximal value of I could also be considered. We chose the two most frequently used ASD variants, D_1 and D_2 , but a variant designed for genotypes obtained by observation of band patterns (Chakraborty and Jin 1993) could also be studied.

We have only considered allele-sharing dissimilarity between population pairs at a single locus, and it will be of interest to investigate dissimilarities that average across many loci. Our theoretical calculations focus on dissimilarities between two random individuals chosen from specified allele-frequency distributions at a locus. Although such distributions have nonzero probability only on the discrete values $\{0, \frac{1}{2}, 1\}$ for D_1 and $\{0, \frac{1}{2}, \frac{3}{4}, 1\}$ for D_2 , when an allele-sharing dissimilarity is calculated as an average across L loci, the $2L + 1$ values $\{0, \frac{1}{2L}, \frac{1}{L}, \frac{3}{2L}, \dots, \frac{L-1}{L}, \frac{2L-1}{2L}, 1\}$ become possible values for D_1 (all multiples of $\frac{1}{2L}$ in $[0,1]$), and the $4L$

values $\{0, \frac{1}{2L}, \frac{3}{4L}, \dots, \frac{4L-3}{4L}, \frac{2L-1}{2L}, \frac{4L-1}{4L}, 1\}$ for \mathcal{D}_2 (all multiples of $\frac{1}{4L}$ in $[0,1]$, other than $\frac{1}{4L}$ itself). Thus, the mean allele-sharing dissimilarity of a random pair of individuals across many loci – computed either theoretically or empirically – has many possible numerical values, potentially giving rise to continuous approximations for associated probability distributions.

We note significant caveats in interpreting our empirical analysis in relation to our theoretical computations. The empirical computations make use of all pairs of individuals drawn from specified samples; each sampled individual appears in many pairs, so that the empirical analysis does not follow the assumption of the theoretical analysis that pairs represent independent draws from allele frequency distributions. A second difference of the empirical and theoretical analyses is that the theoretical analysis assumes that pairs of alleles *within* an individual are independent draws from the allele-frequency distribution, whereas inbreeding can induce dependence of these alleles empirically. Such deviations from the assumptions of the theoretical analysis in conducting the empirical analysis could be explored in simulations that do and do not permit inbreeding and reuse of pairs of individuals and in empirical samples large enough to avoid such reuses.

Allele-sharing dissimilarities have long been used in population genetics. The mathematical relationships we have obtained assist both in predicting their properties in relation to allele frequencies and in understanding empirical aspects of their values. When counterintuitive phenomena are obtained with such dissimilarities – such as a greater within-population dissimilarity than the between-population dissimilarity – the mathematical results can potentially provide insight into the unexpected observations.

Research ethics: Not applicable.

Author contributions: Study design: XL, NAR; Mathematical analysis: XL, ZA, TKM, NAR; Data analysis: XL, ZA; Manuscript preparation: XL, ZA, NAR. The authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: The authors state no conflict of interest.

Research funding: We acknowledge NIH grant R01 HG005855 and NSF grant BCS-2116322 for support.

Data availability: Not applicable.

References

- Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R., and Cavalli-Sforza, L.L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368: 455–457.
- Cavalli-Sforza, L.L. and Edwards, A.W.F. (1967). Phylogenetic analysis: models and estimation procedures. *Am. J. Hum. Genet.* 19: 233–257.
- Chakraborty, R. and Jin, L. (1993). A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances. In: Pena, S.D.J., Chakraborty, R., Epplen, J.T., and Jeffreys, A.J. (Eds.), *DNA fingerprinting: state of the science*. Birkhäuser Verlag, Basel, pp. 153–175.
- Edge, M.D., Ramachandran, S., and Rosenberg, N.A. (2022). Celebrating 50 years since Lewontin's apportionment of human diversity. *Phil. Trans. Roy. Soc. Lond. B Biol. Sci.* 377: 20200405.
- Gao, X. and Martin, E.R. (2009). Using allele sharing distance for detecting human population stratification. *Hum. Hered.* 68: 182–191.
- Jorde, L.B. (1985). Human genetic distance studies: present status and future prospects. *Annu. Rev. Anthropol.* 14: 343–373.
- Lewontin, R.C. (1972). The apportionment of human diversity. *Evol. Biol.* 6: 381–398.
- Mountain, J.L. and Cavalli-Sforza, L.L. (1997). Multilocus genotypes, a tree of individuals, and human evolutionary history. *Am. J. Hum. Genet.* 61: 705–718.
- Mountain, J.L. and Ramakrishnan, U. (2005). Impact of human population history on distributions of individual-level genetic distance. *Hum. Genom.* 2: 4–19.
- Nei, M. (1972). Genetic distance between populations. *Am. Nat.* 106: 283–292.
- Nei, M. (1987). *Molecular evolutionary genetics*. Columbia University Press, New York.
- Prugnolle, F., Manica, A., and Balloux, F. (2005). Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* 15: R159–R160.
- Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W., and Cavalli-Sforza, L.L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. USA* 102: 15942–15947.

- Rosenberg, N.A. (2006). Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann. Hum. Genet.* 70: 841–847.
- Rosenberg, N.A. (2011). A population-genetic perspective on the similarities and differences among worldwide human populations. *Hum. Biol.* 83: 659–684.
- Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K., and Feldman, M.W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* 1: e70.
- Tal, O. (2013). Two complementary perspectives on inter-individual genetic distance. *Biosystems* 111: 18–36.
- Witherspoon, D.J., Wooding, S., Rogers, A.R., Marchani, E.E., Watkins, W.S., Batzer, M.A., and Jorde, L.B. (2007). Genetic similarities within and between human populations. *Genetics* 176: 351–359.