Dissimilarity between a pair of draws from discrete probability vectors on finite sets of objects

Zarif Ahsan, Xiran Liu, Noah Rosenberg

September 9, 2023

Abstract

Motivated by a problem in population genetics, we examine the combinatorics of dissimilarity between random, unordered draws with replacement from a collection of distinct objects, with potentially different probability distributions for the two draws. Consider two draws of size K taken from a population of I objects. We define an equivalence relation for pairs of draws and enumerate all possible equivalence classes under this relation. The enumeration relies on a series of actions by permutation groups. Given our initial probabilities, we compute the probability of each class and find an expectation for a dissimilarity measure. With these expectations, we determine when the expected dissimilarity between two draws with the same probability distributions is greater than between two draws taken with different probability distributions.

1 Problem Setting

Motivation Consider the following problem from population genetics. For a sexually reproducing diploid population, an individual inherits one copy of a given gene from each of its parents. Each copy can be one of many different variants (alleles), and it is assumed that mating is random so that the probability of inheriting an allele from a parent is determined by the frequency of the allele in the overall population. The combination of these two inheritances constitutes a genotype, where should an individual inherit allele A from one parent and allele B from another, we can equivalently present its genotype as AB or BA. Thus, a genotype consists of an unordered draws of alleles from the population pool.

Given this manner of inheritance, how can we define a notion of genetic dissimilarity? This problem arises particularly as we examine multiple genes (loci) at once, where taking the mean of genetic dissimilarities at each locus gives us a measure of how genetically alike two individuals are. For genes with localized alleles, this can reveal information about shared ancestry or geographic origin of two individuals. We can define a dissimilarity measure $\mathcal{D}$ [Mountain and Ramakrishnan, 2005] whose definition is given later.

In the case of diploid individuals, the expected value for this dissimilarity when comparing individuals from the same population and different populations have already been computed, as well as analytic bounds on both values on the space of all possible allele frequencies [Liu et al., Liu, 2023]. However, not all populations are diploid; more than half of plant species, in particular, are polyploid, with some species tetraploid or even 96-ploid [Heslop-Harrison, 2017]. We define this ploidy number as the number of alleles inherited from both parents. To determine similar expected values and bounds as in the diploid case for arbitrary ploidy, we generalize the problem, which has applications in broader combinatorial settings.

General Problem Consider two random, unordered draws of K items with replacement from I objects. The draws potentially come from populations with different frequencies of the I objects. Each i-th object has a probability of p_i of being drawn from one population and q_i from the other, where $\sum_{i=1}^{I} p_i = \sum_{i=1}^{I} q_i = 1$. We find the number of distinct cases, up to relabeling, of these two draws and the probability of each case. Using allele-sharing dissimilarity measures from population genetics, we can equip two dissimilarity metrics to measure the difference between two such draws, where one measure is a distance metric proper. In doing so, we find the expected value of the dissimilarity between two such draws.

Let $\mathcal{A}_I$ be any set of I objects, which we label A_i for $i \in \{1, \ldots, I\}$. For the Cartesian product $\mathcal{A}_I^K$, we have a group action $S_K \circlearrowright \mathcal{A}_I^K$ given by permuting the order of elements for each $(A_{i_1}, ..., A_{i_K}) \in \mathcal{A}_I$. We can then define the space of unordered draws of size K from $\mathcal{A}_I$, denoted by $\mathcal{G}_I^K$, as a quotient by this group action, i.e. $\mathcal{G}_I^K = \mathcal{A}_I^K / S_K$. In general, we distinguish ordered draws with the letter X and unordered draws with the letter G.

Two ordered draws $X_1 = (X_1^1, \ldots, X_1^K)$ and $X_2 = (X_2^1, \ldots, X_2^K)$ correspond to the same unordered draw if they represent the same equivalence class in $\mathcal{G}_I^K$ (i.e. both are contained in the other's orbit by the S_K action). Each class $G_1 \in \mathcal{G}_I^K$ can also be uniquely represented by a vector $\mathbf{g}_1 = (g_1^1, \ldots, g_1^I)$, where g_1^i is the count of A_i in G_1 and $\sum_{i=1}^{I} g_1^i = K$.

In our population genetic example, $\mathcal{A}_K^I$ denotes the set of alleles present in both populations, A_i the label for each distinct allele, p_i and q_i the frequency of A_i in each population, and each $G \in \mathcal{G}_I^K$ a possible genotype.

We define a dissimilarity measure $\mathcal{D} : \mathcal{G}_I^K \times \mathcal{G}_I^K \to [0, 1]$ on the space of unordered draws. Given two draws G_1 and G_2,

$$\mathcal{D}(G_1, G_2) = 1 - \frac{1}{K^2} \sum_{1 \leqslant i,j \leqslant K} \mathbb{1}_{G_1^i = G_2^j}. \tag{1}$$

which is well-defined under any ordering of G_1 and G_2 we choose (and therefore is a function on $\mathcal{G}_I^K \times \mathcal{G}_I^K$).

Equivalently, as a function of the vector representations $\mathbf{g}_1$ and $\mathbf{g}_2$, this is

$$\mathcal{D}(\mathbf{g}_1, \mathbf{g}_2) = 1 - \frac{1}{K^2}\langle \mathbf{g}_1, \mathbf{g}_2 \rangle \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product. This is not a distance metric, as for any K and I, we can take $\mathbf{g} = (1, \ldots, 1, 0, \ldots, 0)$ for which $\mathcal{D}(\mathbf{g}, \mathbf{g}) = 1 - \frac{1}{K} \neq 0$. The minimum value of $\mathcal{D}$ is reached when only one object is drawn among the two draws, and the maximum value is reached when two draws share no common items.

We list some example draws with $K = 2$ and $I = 4$ that give the extreme values in Table 1.

G_1	G_2	$\mathbf{g}_1$	$\mathbf{g}_2$	$\mathcal{D}$
(A_1, A_1)	(A_1, A_1)	$(2, 0, 0, 0)$	$(2, 0, 0, 0)$	0
(A_1, A_2)	(A_1, A_2)	$(1, 1, 0, 0)$	$(1, 1, 0, 0)$	0.5
(A_1, A_1)	(A_2, A_2)	$(2, 0, 0, 0)$	$(0, 2, 0, 0)$	1
(A_1, A_2)	(A_3, A_4)	$(1, 1, 0, 0)$	$(0, 0, 1, 1)$	1

Table 1: Pairs of example draws (G_1, G_2), their vector representations $(\mathbf{g}_1, \mathbf{g}_2)$, and the dissimilarities $(\mathcal{D})$ between them for $K = 2$.

Given the probabilities of drawing each object, we obtain the probability of having each of the distinct combinations of two draws and correspondingly the dissimilarity measures we expect for such two draws. We first consider this in the general case in which each draw is taken from different probability distributions and then consider the special case where these two probability distributions are taken to be the same. In the population genetic context, the former corresponds to comparing genotypes from different populations and the latter to comparing genotypes from the same population.

2 Probabilities of All Possible Combinations

2.1 How to enumerate the unique combinations

We identify all the distinct combinations of two draws, up to relabeling, through a series of group actions. Recall that $\mathcal{G}_I^K$, the space of unordered draws of size K from a pool of I objects, is defined as $\mathcal{G}_I^K = \mathcal{A}_I^K / S_K$ where $\mathcal{A}_I^K$ is the space of ordered draws.

Note that $\mathcal{G}_I^K \times \mathcal{G}_I^K$ corresponds to *ordered* pairs of unordered draws. We can uniquely represent an element $(G_1, G_2) \in \mathcal{G}_I^K \times \mathcal{G}_I^K$ via concatenating the two vectors $\mathbf{g}_1$ and $\mathbf{g}_2$ to be a $2 \times I$ matrix, denoted as

$$\hat{\mathbf{g}} = \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{pmatrix} = \begin{pmatrix} g_1^1 & g_1^2 & \cdots & g_1^I \\ g_2^1 & g_2^2 & \cdots & g_2^I \end{pmatrix}, \tag{3}$$

Due to the symmetry of our dissimilarity functions, the order of the two draws does not matter. To account for this symmetry, we define a group action

$S_2 \curvearrowright \mathcal{G}_I^K \times \mathcal{G}_I^K$ defined by the swapping the order of the two draws. The quotient $(\mathcal{G}_I^K \times \mathcal{G}_I^K)/S_2$ thus yields us the space of unordered pairs of unordered draws; we denote this space $\mathcal{P}_I^K$, and an element within it as $[G_1, G_2]_\sim$ or $[\hat{\mathbf{g}}]_\sim$.

We can then define a group action $S_I \curvearrowright \mathcal{P}_I^K$ that corresponds to the effect of relabeling the I objects; in particular, for $\tau \in S_I$, $\tau : m \mapsto n$ for $m, n \in \{1, \ldots, I\}$ if and only if $\tau([G_1, G_2]_\sim)$ replaces all A_m in $[G_1, G_2]_\sim$ with A_n where $[G_1, G_2]_\sim \in \mathcal{P}_I^K$. For example, when $K = 4$ and $I = 8$, the element $(12)(35)(78) \in S_8$ maps the unordered pair of unordered draws $[A_1A_1A_3A_5, A_7A_7A_7A_8]_\sim$ as

$$\begin{aligned}[A_1A_1A_3A_5, A_7A_7A_7A_8]_\sim &\mapsto [A_2A_2A_5A_3, A_8A_8A_8A_7]_\sim \\ &= [A_2A_2A_3A_5, A_7A_8A_8A_8]_\sim.\end{aligned}$$

Note that if we represent (G_1, G_2) in its matrix form $\hat{\mathbf{g}}$, this action is equivalent to permuting the columns of the matrix: for $\tau \in S_I$, $\tau : m \mapsto n$ for $m, n \in \{1, \ldots, I\}$ if and only if $\tau.[\hat{\mathbf{g}}]_\sim$ changes the nth column of $\hat{\mathbf{g}}$ to that of the mth column (this is a well-defined operation when considering the matrix up to a row swap i.e. the S_2 action). Considering the same example, $(12)(35)(78)$ maps the matrix representation of $[A_1A_1A_3A_5, A_7A_7A_7A_8]_\sim$ as

$$\begin{bmatrix} 2 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 & 1 \end{bmatrix}_\sim \mapsto \begin{bmatrix} 0 & 2 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 3 \end{bmatrix}_\sim.$$

We denote the quotient by this action, which corresponds to all unordered pairs of unordered draws up to relabeling, as $\mathcal{C}_I^K = \mathcal{P}_I^K/S_I$. We denote the corresponding case of $\hat{\mathbf{g}} \in \mathcal{G}_I^K \times \mathcal{G}_I^K$ in $\mathcal{C}_I^K$ as $[\hat{\mathbf{g}}]$.

We are therefore counting the number of $[\hat{\mathbf{g}}] \in \mathcal{C}_I^K$, which is the number of $2 \times I$ nonnegative integer matrices whose rows sum to K, up to permutation of rows and columns. This allows for an enumeration of our cases independent of I, moreover, by the following proposition.

Proposition 2.1. *For $I \geqslant 2K$, the elements of $\mathcal{C}_I^K$ can be enumerated independent of I.*

Proof. Take a matrix $\hat{\mathbf{g}} \in \mathcal{G}_I^K \times \mathcal{G}_I^K$ for $I \geqslant 2K$. Since $\hat{\mathbf{g}}$ must have nonnegative entries and its rows $\mathbf{g}_1$ and $\mathbf{g}_2$ must each sum to K, each row of $\hat{\mathbf{g}}$ can have at most K nonzero entries (occurring when each nonzero entry is 1). Therefore, $\hat{\mathbf{g}}$ can have at most $2K$ nonzero columns, which occurs when

$$\hat{\mathbf{g}} = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \end{pmatrix}.$$

Therefore, by permutation of the columns, we can have a representative of $[\hat{\mathbf{g}}]$, denoted $\hat{\mathbf{h}}$, such that the last $I - 2K$ columns are zero, so the left $2 \times 2K$ entries of $\hat{\mathbf{h}}$ are a $2 \times 2K$ nonnegative-integer matrix with rows summing to K, i.e. an element of $\mathcal{G}_{2K}^K \times \mathcal{G}_{2K}^K$. This yields a bijection between $\mathcal{C}_I^K$ and $\mathcal{C}_{2K}^K$, namely

$$\begin{aligned}\mathcal{C}_{2K}^K &\longleftrightarrow \mathcal{C}_I^K \\ [\hat{\mathbf{h}}] &\longmapsto [\begin{pmatrix} \hat{\mathbf{h}} & 0 & \cdots & 0 \\ & 0 & \cdots & 0 \end{pmatrix}].\end{aligned}$$

Therefore, we can enumerate the elements in $\mathcal{C}_I^K$ via the elements in $\mathcal{C}_{2K}^K$, so the enumeration of $\mathcal{C}_I^K$ is independent of I for $I \geqslant 2K$. □

This result is trivial, considering that $\mathcal{C}_I^K$ is the space of unordered pairs of unordered draws up to relabelling, where each unordered pair has at most $2K$ distinct objects and therefore can be relabelled to have nothing beyond the *first* $2K$ distinct objects.

Note that in the case that $I < 2K$, we can still use this enumeration, but some cases (i.e. those with more than I distinct objects) will be zero probability. We can account for this later, however, by introducing $2K - I$ objects to our initial pool $\mathcal{A}_I^K$ with 0 probability.

Varying K, $\left|\mathcal{C}_{2K}^K\right|$ is equivalent to the integer sequence A331722 in the On-Line Encyclopedia of Integer Sequences:

$$1, 2, 7, 21, 66, 192, 565, 1579, 4348, 11582, 30205, 76736, \ldots \quad .$$

This sequence uses the 2-column matrix instead of the 2-row one and allows any number of nonzero columns [OEIS Foundation Inc., 2023].

Table 2 displays an example of this enumeration for the case of $K = 2$, providing representative draws and their matrix for each element in $\mathcal{C}_I^2$ as well as their $\mathcal{D}$ value.

Table 2: Enumeration of elements in $\mathcal{C}_I^K$ for $K = 2$,$I \geqslant 2$, each given by its representative draws and matrix.

Draws	$\hat{\mathbf{g}}$	$\mathcal{D}$
A_1A_1, A_1A_1	$\begin{pmatrix} 2 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \end{pmatrix}$	0
A_1A_1, A_1A_2	$\begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$	$\frac{1}{2}$
A_1A_1, A_2A_2	$\begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{pmatrix}$	1
A_1A_1, A_2A_3	$\begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$	1
A_1A_2, A_1A_2	$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$	$\frac{1}{2}$
A_1A_2, A_1A_3	$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}$	$\frac{3}{4}$
A_1A_2, A_3A_4	$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$	1

2.2 Probability of each combination

We assume the first draw is taken with replacement from $\mathcal{A}_I^K$ with probability p_i for drawing the ith object A_i, and the second from $\mathcal{A}_I^K$ with probability q_i for drawing A_i.

2.2.1 Probability of a Pair of Unordered Draws

Let the first draw be represented by $\mathbf{g}_1$ and the second by $\mathbf{g}_2$. We have that the probability of each possible ordering of $\mathbf{g}_1$, X_1, and each possible ordering of $\mathbf{g}_2$, X_2, are given by

$$\mathbb{P}(X_1) = \prod_{i=1}^{I} p_i^{g_1^i}, \tag{4}$$

$$\mathbb{P}(X_2) = \prod_{i=1}^{I} q_i^{g_2^i}. \tag{5}$$

The probability of the unordered draw g_1 comes by summing across all possible orderings, which are distinct permutations of the vector g_1. Since each ordering has the same probability, we have

$$\mathbb{P}(\mathbf{g}_1) = \sum_{\text{reorderings of } X_1} \mathbb{P}(X_1) \tag{6}$$

$$= \binom{K}{g_1^1, \ldots, g_1^I} \prod_{i=1}^{I} p_i^{g_1^i} \tag{7}$$

where $\binom{K}{g_1^1,\ldots,g_1^I} = K! \prod_{i=1}^{I} \frac{1}{g_1^i!}$ is a multinomial coefficient. Likewise, for $\mathbf{g}_2$,

$$\mathbb{P}(\mathbf{g}_2) = \sum_{\text{reorderings of } X_2} \mathbb{P}(X_2) \tag{8}$$

$$= \binom{K}{g_2^1, \ldots, g_2^I} \prod_{i=1}^{I} q_i^{g_2^i}. \tag{9}$$

Therefore, the probability of obtaining an ordered pair consisting of these draws is given by

$$\mathbb{P}(\hat{\mathbf{g}}) = \mathbb{P}(\mathbf{g}_1)\mathbb{P}(\mathbf{g}_2) \tag{10}$$

$$= \binom{K}{g_1^1, \ldots, g_1^I} \binom{K}{g_2^1, \ldots, g_2^I} \prod_{i=1}^{I} p_i^{g_1^i} q_i^{g_2^i} \tag{11}$$

where we assume $\mathbf{g}_1$ is taken from the probability distribution given by $p_1, \ldots, p_I$ and $\mathbf{g}_2$ is taken from the probability distribution given by $q_1, \ldots, p_I$.

We can also consider the reversed case, where $\mathbf{g}_1$ is taken from the distribution given by $q_1, \ldots, q_I$ and $\mathbf{g}_2$ from that of $p_1, \ldots, p_I$. If $\mathbf{g}_1 = \mathbf{g}_2$, this event is

identical to the previous one. If not, however, we combine the original case and the reversed case into a single event to compute the probability of the unordered pair of unordered draws:

$$\mathbb{P}([\hat{\mathbf{g}}]_\sim) = \frac{1}{1+\mathbb{1}_{\mathbf{g}_1=\mathbf{g}_2}}\Big[\mathbb{P}\binom{\mathbf{g}_1}{\mathbf{g}_2} + \mathbb{P}\binom{\mathbf{g}_2}{\mathbf{g}_1}\Big] \tag{12}$$

$$= \frac{1}{1+\mathbb{1}_{\mathbf{g}_1=\mathbf{g}_2}}\binom{K}{g_1^1,\ldots,g_1^I}\binom{K}{g_2^1,\ldots,g_2^I}\Big(\prod_{i=1}^{I} p_i^{g_1^i} q_i^{g_2^i} + \prod_{i=1}^{I} p_i^{g_2^i} q_i^{g_1^i}\Big). \tag{13}$$

Here we sum across each of the two possible ordering of the draws, dividing by $1+\mathbb{1}_{\mathbf{g}_1=\mathbf{g}_2}$ to account for when our two draws are the same.

2.2.2 Probability up to Relabelling

To find the probability of $[\hat{\mathbf{g}}] \in \mathcal{C}_I^K$, we have to sum over all distinct relabellings of our draws, which corresponds to summed probability of the orbit of $[\hat{\mathbf{g}}]_\sim$ under S_I.

Let $N(\hat{\mathbf{g}})$ be the number of distinct objects in $\hat{\mathbf{g}}$ (i.e. the number of non-zero columns). Without loss of generality, we can assume the first $N(\hat{\mathbf{g}})$ columns of $\hat{\mathbf{g}}$ are nonzero (as we can always find such a matrix in $[\hat{\mathbf{g}}]$ via a permutation of the columns). Note that a relabelling of the objects A_i corresponds to a relabelling of their frequencies p_i and q_i, so to find the probability of $\hat{\mathbf{g}}$, we sum across all distinct reassignments of our p_i and q_i present in our case, yielding

$$P([\hat{\mathbf{g}}]) = C(\hat{\mathbf{g}}) \sum_{i_1 \neq \cdots \neq i_{N(\hat{\mathbf{g}})}} \Big(\prod_{j=1}^{N(\hat{\mathbf{g}})} p_{i_j}^{g_1^j} q_{i_j}^{g_2^j} + \prod_{i=1}^{N(\hat{\mathbf{g}})} p_{i_j}^{g_2^j} q_{i_j}^{g_1^j}\Big) \tag{14}$$

where

$$C(\hat{\mathbf{g}}) = \frac{1}{(1+\mathbb{1}_{\mathbf{g}_1=\mathbf{g}_2})|\mathrm{Stab}_{S_{N(\hat{\mathbf{g}})}}[\hat{\mathbf{g}}]_\sim|}\binom{K}{g_1^1,\ldots,g_1^I}\binom{K}{g_2^1,\ldots,g_2^I}. \tag{15}$$

We can pull out the factor of $1+\mathbb{1}_{\mathbf{g}_1=\mathbf{g}_2}$ since the relation $\mathbf{g}_1 = \mathbf{g}_2$ is preserved under any permutation of the columns of $\hat{\mathbf{g}}$.

Depending on the values in $\hat{\mathbf{g}}$, note that certain relabellings of the objects does not change the pair of draws; for instance, when $K = I = 4$ the relabelling given by $(13)(24) \in S_I$ takes the case $[(A_1, A_1, A_2, A_2), (A_3, A_3, A_4, A_4)]_\sim$ to $[(A_3, A_3, A_4, A_4), (A_1, A_1, A_2, A_2)]_\sim$, which is equivalent. Thus to prevent double counting within the sum due to this, we divide by the number of relabellings of the p_i that preserve $[\hat{\mathbf{g}}]_\sim$, given by the stabilizer subgroup

$$\mathrm{Stab}_{S_{N(\hat{\mathbf{g}})}}[\hat{\mathbf{g}}]_\sim = \{\sigma \in S_{N(\hat{\mathbf{g}})} \subset S_I : \sigma([\hat{\mathbf{g}}]_\sim) = [\hat{\mathbf{g}}]_\sim\}. \tag{16}$$

We can compute the size of this subgroup, moreover, observing the matrix representation in the next section.

2.2.3 Using the matrix representation to compute the stabilizer

Any element of $S_{N(\hat{\mathbf{g}})}$ acts on $[\hat{\mathbf{g}}]_\sim$ by permuting the nonzero columns of $\hat{\mathbf{g}}$ (where, again, we assume the first $N(\hat{\mathbf{g}})$ columns of $\hat{\mathbf{g}}$ are nonzero). Therefore, to find $|\text{Stab}_{S_{N(\hat{\mathbf{g}})}}[\hat{\mathbf{g}}]_\sim|$, we count all permutations of the nonzero columns of $\hat{\mathbf{g}}$ that yield the same matrix up to a row swap.

Denote the unordered list of non-zero entries in the first row of the matrix as r_1, and that of the second row as r_2. The order of the items does not matter, so $r_1 = r_2$ if they have exactly the same items. Suppose there are L unique columns in $\hat{\mathbf{g}}$ that are nonzero: $c_1, \ldots, c_L$. Use $\{c_\ell\}$ to denote all columns in matrix $\hat{\mathbf{g}}$ that equal to c_ℓ.

The denominator $|\text{Stab}_{S_{N(\hat{\mathbf{g}})}}[\hat{\mathbf{g}}]_\sim|$ in $C([\hat{\mathbf{g}}])$ can be computed as

$$|\text{Stab}_{S_{N(\hat{\mathbf{g}})}}[\hat{\mathbf{g}}]_\sim| = \prod_{\ell=1}^{L} |\{c_\ell\}|! \times \left[1 + (\mathbb{1}_{r_1=r_2} - \mathbb{1}_{\mathbf{g}_1=\mathbf{g}_2})\right], \tag{17}$$

where $\mathbb{1}_\cdot$ is the indicator variable .

The first component $\prod_{\ell=1}^{L} |\{c_\ell\}|!$ in the expression counts all the possible ways to rearrange all nonzero columns in the matrix $\hat{\mathbf{g}}$ while keeping the combination scenario equivalent when we fix the order of the two draws. It counts all the possible ways to rearrange each group of the same nonzero columns by counting all the permutations of $|\{c_\ell\}|$ objects, then multiplies the numbers of permutations together to get the total count.

The second component, $[1 + (\mathbb{1}_{r_1=r_2} - \mathbb{1}_{\mathbf{g}_1=\mathbf{g}_2})]$, is a multiplier used to account for possibly equivalent scenarios when swapping the two draws (two rows) because the order of draws does not matter. If the two draws have different item distributions (two rows having different unordered lists of entries), then swapping them would make the combination of the two draws a different scenario, so we would not multiply the first component by two. If two draws have the same pattern in their item distribution, that is, two rows have the same unordered list of non-zero entries, then we can swap them and still keep the combination scenario the same, so we need to multiply the first component by two. However, if the two draws are identical, then we should avoid double-counting, therefore, we only multiply the first component by one in this case. Recall that relabeling of the objects is not considered at this point.

2.2.4 Final probability expression

Note that if $\mathbb{1}_{\mathbf{g}_1=\mathbf{g}_2} = 1$, then $\mathbb{1}_{r_1=r_2}$. Therefore,

$$(1 + \mathbb{1}_{\mathbf{g}_1=\mathbf{g}_2})(1 + (\mathbb{1}_{r_1=r_2} - \mathbb{1}_{\mathbf{g}_1=\mathbf{g}_2})) = \mathbb{1}_{r_1=r_2}, \tag{18}$$

and substituting our result from Eq. 17 into our probability expression from Eq. 14, we get

$$P([\hat{\mathbf{g}}]) = C(\hat{\mathbf{g}}) \sum_{i_1 \neq \cdots \neq i_{N(\hat{\mathbf{g}})}} \left(\prod_{j=1}^{N(\hat{\mathbf{g}})} p_{i_j}^{g_1^j} q_{i_j}^{g_2^j} + \prod_{i=1}^{N(\hat{\mathbf{g}})} p_{i_j}^{g_2^j} q_{i_j}^{g_1^j} \right) \tag{19}$$

where

$$C(\hat{\mathbf{g}}) = \frac{1}{(1 + \mathbb{1}_{r_1=r_2})(\prod_{\ell=1}^{L} |\{c_\ell\}|!)} \binom{K}{g_1^1, \ldots, g_1^I} \binom{K}{g_2^1, \ldots, g_2^I}. \tag{20}$$

This yields our probability expression when each unordered draw is taken with different probability distributions. When $\mathbf{p} = \mathbf{q}$,

$$P([\hat{\mathbf{g}}]) = 2C(\hat{\mathbf{g}}) \sum_{i_1 \neq \cdots \neq i_{N(\hat{\mathbf{g}})}} \prod_{j=1}^{N(\hat{\mathbf{g}})} p_{i_j}^{g_1^j + g_2^j}. \tag{21}$$

which is the probability of $[\hat{\mathbf{g}}]$ when both unordered draws are taken with the same probability distributions.

With these expressions, we can compute the probability of each possible case given an enumeration of all elements in $\mathcal{C}_I^K$ as we do in Tables 3 and 4, where sequential lettering is used to denote separate objects (i.e. A is A_1, B is A_2, C is A_3, and so on).

Table 3: Combinations of two draws with $K = 2$, draws taken with different probability distributions

	G1	G2	$\mathcal{D}$	Probability
1	AA	AA	0	$\sum_{i_1}^{I} p_{i_1}^2 q_{i_1}^2$
2	AA	AB	$\frac{1}{2}$	$4\sum_{i_1\neq i_2}^{I} p_{i_1}^2 q_{i_1} q_{i_2} + p_{i_1} p_{i_2} q_{i_1}^2$
3	AA	BB	1	$\sum_{i_1\neq i_2}^{I} p_{i_1}^2 q_{i_2}^2$
4	AA	BC	1	$\sum_{i_1\neq i_2\neq i_3}^{I} p_{i_1}^2 q_{i_2} q_{i_3} + p_{i_2} p_{i_3} q_{i_1}^2$
5	AB	AB	$\frac{1}{2}$	$2\sum_{i_1\neq i_2}^{I} p_{i_1} p_{i_2} q_{i_1} q_{i_2}$
6	AB	AC	$\frac{3}{4}$	$4\sum_{i_1\neq i_2\neq i_3}^{I} p_{i_1} p_{i_2} q_{i_1} q_{i_3}$
7	AB	CD	1	$\sum_{i_1\neq i_2\neq i_3\neq i_4}^{I} p_{i_1} p_{i_2} q_{i_3} q_{i_4}$

Table 4: Combinations of two draws with $K = 3$, draws taken with different probability distributions

Case	G1	G2	$\mathcal{D}$	Probability
1	AAA	AAA	0	$\sum_{i_1}^{I} p_{i_1}^3 q_{i_1}^3$
2	AAA	AAB	$\frac{1}{3}$	$3\sum_{i_1\neq i_2}^{I} p_{i_1}^3 q_{i_1}^2 q_{i_2} + p_{i_1}^2 p_{i_2} q_{i_1}^3$
3	AAA	ABB	$\frac{2}{3}$	$3\sum_{i_1\neq i_2}^{I} p_{i_1}^3 q_{i_1} q_{i_2}^2 + p_{i_1} p_{i_2}^2 q_{i_1}^3$
4	AAA	ABC	$\frac{2}{3}$	$3\sum_{i_1\neq i_2\neq i_3}^{I} p_{i_1}^3 q_{i_1} q_{i_2} q_{i_3} + p_{i_1} p_{i_2} p_{i_3} q_{i_1}^3$
5	AAA	BBB	1	$\sum_{i_1\neq i_2}^{I} p_{i_1}^3 q_{i_2}^3$
6	AAA	BBC	1	$3\sum_{i_1\neq i_2\neq i_3}^{I} p_{i_1}^3 q_{i_2}^2 q_{i_3} + p_{i_2}^2 p_{i_3} q_{i_1}^3$
7	AAA	BCD	1	$\sum_{i_1\neq i_2\neq i_3\neq i_4}^{I} p_{i_1}^3 q_{i_2} q_{i_3} q_{i_4} + p_{i_2} p_{i_3} p_{i_4} q_{i_1}^3$
8	AAB	AAB	$\frac{4}{9}$	$9\sum_{i_1\neq i_2}^{I} p_{i_1}^2 p_{i_2} q_{i_1}^2 q_{i_2}$
9	AAB	AAC	$\frac{5}{9}$	$9\sum_{i_1\neq i_2\neq i_3}^{I} p_{i_1}^2 p_{i_2} q_{i_1}^2 q_{i_3}$
10	AAB	ABB	$\frac{5}{9}$	$9\sum_{i_1\neq i_2}^{I} p_{i_1}^2 p_{i_2} q_{i_1} q_{i_2}^2$
11	AAB	ABC	$\frac{2}{3}$	$18\sum_{i_1\neq i_2\neq i_3}^{I} p_{i_1}^2 p_{i_2} q_{i_1} q_{i_2} q_{i_3} + p_{i_1} p_{i_2} p_{i_3} q_{i_1}^2 q_{i_2}$
12	AAB	ACC	$\frac{7}{9}$	$9\sum_{i_1\neq i_2\neq i_3}^{I} p_{i_1}^2 p_{i_2} q_{i_1} q_{i_3}^2 + p_{i_1} p_{i_3}^2 q_{i_1}^2 q_{i_2}$
13	AAB	ACD	$\frac{7}{9}$	$9\sum_{i_1\neq i_2\neq i_3\neq i_4}^{I} p_{i_1}^2 p_{i_2} q_{i_2} q_{i_3} q_{i_4} + p_{i_2} p_{i_3} p_{i_4} q_{i_1}^2 q_{i_2}^2$
14	AAB	BCC	$\frac{8}{9}$	$9\sum_{i_1\neq i_2\neq i_3}^{I} p_{i_1}^2 p_{i_2} q_{i_2} q_{i_3}^2$
15	AAB	BCD	$\frac{8}{9}$	$9\sum_{i_1\neq i_2\neq i_3\neq i_4}^{I} p_{i_1}^2 p_{i_2} q_{i_2} q_{i_3} q_{i_4} + p_{i_2} p_{i_3} p_{i_4} q_{i_1}^2 q_{i_2}$
16	AAB	CCD	1	$9\sum_{i_1\neq i_2\neq i_3\neq i_4}^{I} p_{i_1}^2 p_{i_2} q_{i_3}^2 q_{i_4}$
17	AAB	CDE	1	$3\sum_{i_1\neq i_2\neq i_3\neq i_4\neq i_5}^{I} p_{i_1}^2 p_{i_2} q_{i_3} q_{i_4} q_{i_5} + p_{i_3} p_{i_4} p_{i_5} q_{i_1}^2 q_{i_2}$
18	ABC	ABC	$\frac{2}{3}$	$6\sum_{i_1\neq i_2\neq i_3}^{I} p_{i_1} p_{i_2} p_{i_3} q_{i_1} q_{i_2} q_{i_3}$
19	ABC	ABD	$\frac{7}{9}$	$18\sum_{i_1\neq i_2\neq i_3\neq i_4}^{I} p_{i_1} p_{i_2} p_{i_3} q_{i_1} q_{i_2} q_{i_4}$
20	ABC	ADE	$\frac{8}{9}$	$9\sum_{i_1\neq i_2\neq i_3\neq i_4\neq i_5}^{I} p_{i_1} p_{i_2} p_{i_3} q_{i_1} p_{i_4} p_{i_5}$
21	ABC	DEF	1	$\sum_{i_1\neq i_2\neq i_3\neq i_4\neq i_5\neq i_6}^{I} p_{i_1} p_{i_2} p_{i_3} q_{i_4} q_{i_5} q_{i_6}$

Here, $\sum_{i_1\neq i_2}^{I} a_{i_1 i_2} = \sum_{i_1=1}^{I} \sum_{\substack{i_2=1 \\ i_2\neq i_1}}^{I} a_{i_1 i_2}$, $\sum_{i_1\neq i_2\neq i_3}^{I} a_{i_1 i_2 i_3} = \sum_{i_1=1}^{I} \sum_{\substack{i_2=1 \\ i_2\neq i_1}}^{I} \sum_{\substack{i_3=1 \\ i_3\neq i_1 \\ i_3\neq i_2}}^{I} a_{i_1 i_2 i_3}$, and so on.

3 Expected dissimilarity value

Let $\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})]$ be the expected dissimilarity between two random unordered draws with replacement as a function of our drawing probability vectors $\mathbf{p} = (p_1,\ldots,p_I)$ and $\mathbf{q} = (q_1,\ldots,q_I)$. We can compute $\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})]$ for a given K algorithmically via the cases and their corresponding probabilities in the previous section, taking

$$\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})] = \sum_{[\hat{\mathbf{g}}]\in\mathcal{C}_I^K} \mathcal{D}(\hat{\mathbf{g}})P([\hat{\mathbf{g}}]) \tag{22}$$

where each of the probabilities $P([\hat{\mathbf{g}}])$ can be computed via the previous derivations. In the $K = 2$ or $K = 3$ case, this is equivalent to taking the dot product of the $\mathcal{D}$ and Probability columns of Table 3 and 4, respectively, and reducing the resultant polynomials modulo $p_1 + \cdots + p_I - 1 = 0$ and $q_1 + \cdots + q_I - 1 = 0$.

We show, in general, that

Theorem 3.1. *For any* $\mathbf{p}$, $\mathbf{q}$ *and* K, I

$$\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})] = 1 - \langle \mathbf{p},\mathbf{q}\rangle \tag{23}$$

Proof. Let Γ_1, Γ_2 be random variables corresponding to our two unordered draws. Recall that for two instances G_1 of Γ_1 and G_2 of Γ_2

$$\mathcal{D}(G_1,G_2) = 1 - \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} \mathbb{1}_{G_1^i = G_2^j}.$$

We can rewrite this as

$$\begin{aligned}
\mathcal{D}(G_1,G_2) &= \frac{K^2}{K^2} - \frac{1}{K^2}\sum_{1\leqslant i,j\leqslant K} \mathbb{1}_{G_1^i=G_2^j} \\
&= \frac{1}{K^2}\sum_{1\leqslant i,j\leqslant K} 1 - \mathbb{1}_{G_1^i=G_2^j} \\
&= P(\Gamma_1^S \neq \Gamma_2^T | \Gamma_1 = G_1, \Gamma_2 = G_2)
\end{aligned}$$

where S and T are random variables corresponding randomly selected indices $i, j \in [1, K]$. For fixed s and t and note that

$$P(\Gamma_1^S \neq \Gamma_2^T | S = s, T = t) = 1 - \langle \mathbf{p},\mathbf{q}\rangle$$

for random G_1 and G_2 since this is equivalent to checking whether two draws of a singular object are not matching. We also have that

$$\begin{aligned}
P(\Gamma_1^S \neq \Gamma_2^T | S = s, T = t) &= 1 \cdot \mathbb{1}_{\Gamma_1^s \neq \Gamma_2^t} + 0 \cdot \mathbb{1}_{\Gamma_1^s = \Gamma_2^t} \\
&= \mathbb{E}[\mathbb{1}_{\Gamma_1^s \neq \Gamma_2^t}] \\
&= \sum_{(G_1,G_2)\in\mathcal{G}_I^K\times\mathcal{G}_I^K} P(\Gamma_1 = G_1)P(\Gamma_2 = G_2)\mathbb{1}_{G_1^s \neq G_2^t}.
\end{aligned}$$

Note that our random choice of s and t are independent of G_1 and G_2. Therefore,

$$\begin{aligned}
\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})] &= \sum_{(G_1,G_2)\in\mathcal{G}_I^K\times\mathcal{G}_I^K} P(\Gamma_1=G_1,\Gamma_2=G_2)\cdot\mathcal{D}(G_1,G_2)\\
&= \sum_{(G_1,G_2)\in\mathcal{G}_I^K\times\mathcal{G}_I^K} P(\Gamma_1=G_1,\Gamma_2=G_2)\cdot P(\Gamma_1^S\neq\Gamma_2^T|\Gamma_1=G_1,\Gamma_2=G_2)\\
&= \sum_{\mathcal{G}_I^K\times\mathcal{G}_I^K}\sum_{1\leqslant i,j\leqslant K} P(\Gamma_1=G_1,\Gamma_2=G_2)P(S=i,T=j)P(\Gamma_1^S\neq\Gamma_2^T|\Gamma_1=G_1,\Gamma_2=G_2,S=i,T=j).
\end{aligned}$$

We have that $P(\Gamma_1^S\neq\Gamma_2^T|\Gamma_1=G_1,\Gamma_2=G_2,S=i,T=j)=\mathbb{1}_{G_1^i\neq G_2^j}$ and each particular pair of i and j is equally likely, so $P(S=i,T=j)=\frac{1}{K^2}$. Thus, we can simply this expression to

$$\begin{aligned}
\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})] &= \sum_{\mathcal{G}_I^K\times\mathcal{G}_I^K}\sum_{1\leqslant i,j\leqslant K} P(\Gamma_1=G_1)P(\Gamma_2=G_2)\cdot P(S=i,T=j)\cdot\mathbb{1}_{G_1^i\neq G_2^j}\\
&= \sum_{\mathcal{G}_I^K\times\mathcal{G}_I^K}\sum_{1\leqslant i,j\leqslant K}\frac{1}{K^2}P(\Gamma_1=G_1)P(\Gamma_2=G_2)\cdot\mathbb{1}_{G_1^i\neq G_2^j}\\
&= \sum_{\mathcal{G}_I^K\times\mathcal{G}_I^K} P(\Gamma_1=G_1)P(\Gamma_2=G_2)\cdot\mathbb{1}_{G_1^s\neq G_2^t}\\
&= P(\Gamma_1^S\neq\Gamma_2^T|S=s,T=t)\\
&= 1-\langle\mathbf{p},\mathbf{q}\rangle.
\end{aligned}$$

□

With this, we can answer one of our original questions, i.e. whether the expected dissimilarity between draws is the smallest when both are taken with equivalent probability distributions. We find that

Corollary 3.2. $\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{p})]\leqslant\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})]$ *if and only if* $\langle\mathbf{p},\mathbf{q}\rangle\leqslant\langle\mathbf{p},\mathbf{p}\rangle$.

Proof. $\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{p})]\leqslant\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})]$ if and only if $\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{p})]-\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})]\leqslant=0$, where

$$\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{p})]-\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})]=\langle\mathbf{p},\mathbf{q}\rangle-\langle\mathbf{p},\mathbf{p}\rangle.$$

Thus, $\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{p})]\leqslant\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})]$ if and only if $\langle\mathbf{p},\mathbf{q}\rangle\leqslant\langle\mathbf{p},\mathbf{p}\rangle$. □

Therefore, there are probabilities for which the expected dissimilarity between draws is *greater* when taken with the same probability distributions rather than different. For example, when $\mathbf{p}=(0.8,0.2,0,\ldots,0)$ and $\mathbf{q}=(0.9,0.1,0,\ldots,0)$,

$$\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{p})]=1-0.68=0.32\geqslant 0.26=\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})].$$

However, we do find that

Theorem 3.3. *For any K, I and $\mathbf{p}$, $\mathbf{q}$:*

$$\frac{1}{2}\big(\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{p})]+\mathbb{E}[\mathcal{D}(\mathbf{q},\mathbf{q})]\big)\leqslant\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})].$$

Proof. This is true if and only if

$$\frac{1}{2}\big(\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{p})] + \mathbb{E}[\mathcal{D}(\mathbf{q},\mathbf{q})]\big) - \mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})] \leqslant 0.$$

where we have

$$\begin{aligned}\frac{1}{2}\big(\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{p})] + \mathbb{E}[\mathcal{D}(\mathbf{q},\mathbf{q})]\big) - \mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})] &= -\frac{1}{2}\langle \mathbf{p},\mathbf{p}\rangle - \frac{1}{2}\langle \mathbf{q},\mathbf{q}\rangle + \langle \mathbf{p},\mathbf{q}\rangle \\ &= -\frac{1}{2}\big(\langle \mathbf{p},\mathbf{p}\rangle + \langle \mathbf{q},\mathbf{q}\rangle - 2\langle \mathbf{p},\mathbf{q}\rangle\big) \\ &= -\frac{1}{2}\langle \mathbf{p}-\mathbf{q},\mathbf{p}-\mathbf{q}\rangle \\ &\leqslant 0. \end{aligned}$$

□

Therefore, $\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})]$ is always bounded below by at least one of $\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{p})]$ or $\mathbb{E}[\mathcal{D}(\mathbf{q},\mathbf{q})]$.

These results generalize the same inequalities and expressions found in the diploid case of population genetic (i.e. when $K = 2$), showing that as a measure of genetic difference, $\mathcal{D}$ is independent of our ploidy number K.

4 Conclusion

We have examined the problem of measuring dissimilarity between random, unordered draws with replacement, using a dissimilarity measure previously used in population genetic contexts. Formalizing our draws via quotient under a group action, we find all salient cases of pairs of draws for our dissimilarity measure, quotienting up to relabelling to yield a means of enumerating all possible cases independent of I. We find the probability of each of thesee cases, which can thereon be used to algorithmically compute expected values of our dissimilarity measure, as well as other dissimilarity measure we may equip to the space. In general, however, we show that the expected value of our dissimilarity can be expressed with the inner product of the vectors of our probability distributions (where we assume each draw is not necessarily taken with the same probability distributions).

Contrary to what may be expected from an intuitive notion of dissimilarity, it is possible for our expected dissimilarity to decrease when not requiring our two sets of probability distributions to be the same. Nonetheless, we do find that the expected dissimilarity for when our two draws are taken with different probability distributions (represented by vectors $\mathbf{p}$) is bounded below by the one of the expected dissimilarities when we take both draws with probability $\mathbf{p}$ or $\mathbf{q}$, since we find that $\frac{1}{2}(\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{p})] + \mathbb{E}[\mathcal{D}(\mathbf{q},\mathbf{q})]) \leqslant \mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})]$. This result connects to our original population-genetic motivation as $\mathbb{E}[\mathcal{D}(\mathbf{p},\mathbf{q})]$ measures the genetic difference between populations given a vector of its allele frequencies, and intrapopulation genetic variation when $\mathbf{p} = \mathbf{q}$. Therefore, our results lend indsight into the conditions under which a population is more genetically

different from itself than others, generalizing previous work in the $K = 2$ case in population genetics. [Liu et al., Liu, 2023]

Random, unordered draws arise in multiple combinatorial contexts, moreover. For instance, random samples from a large population of objects with sufficiently small number of labels behave similarly to random, unordered draws with replacement, in which case $\mathbb{E}[\mathcal{D}]$ allows us to measure variability among samples. The enumeration of elements in $\mathcal{C}_I^K$ links to multiple other combinatorial problems, as the nonzero columns of elements in $\mathcal{C}_I^K$ correspond to partitions of the vector $\binom{K}{K}$ into integer vectors, up to a swapping of the rows of each partition element.

Our problem can also be understood as a card game played from two infinite decks, with potentially different drawing probabilities. Two players draw K cards from a different deck as their hand, from which we compute the dissimilarity between the two hands via $\mathcal{D}$, which is equivalent to finding the probability that, selecting one card at random from the hand, the players select different cards. We find that finding an expectation of this dissimilarity is the same as if only one card was drawn from each deck and compared. Indeed, multiple different card games can be adapted to follow this problem setting, including infinite-deck versions of War and Anomia. If the game were structured in a way that a player was trying to maximize the probability of a match via rigging the deck, our results show that this can be done without necessitating that the player match their drawing probabilities to the other's.

There exist several open questions within the problem, however. We were unable to determine a generating function for the elements in $\mathcal{C}_I^K$ or an algorithmic means to enumerate its elements (its elements can be found, however, by reducing our space of nonnegative-integer matrices modulo our group actions). Particularly for other dissimilarity measures for which a general expected value cannot be easily computed, this may be beneficial for its algorithmic computation. Moreover, while we found the size of $\mathcal{C}_I^K$ to be equivalent to the OEIS sequence A331722 [OEIS Foundation Inc., 2023], a closed form expression for this sequence may also aid in connecting the combinatorial nature of this problem to other contexts, as well as aid in understanding the computational complexity of our algorithmic means of computing the expected value.

Additionally, while we did find that the expected dissimilarity may be greater for two draws with identical probability distributions than those without, we expect this event to be unlikely and approach 0 as the number of nonzero-frequency objects increases (as seen in the $K = 2$ case in previous work [Liu et al.]). Determining a closed form expression for this probability and its limit, therefore, will also yield insight into the importance of the I parameter in the behavior of $\mathbb{E}[\mathcal{D}]$ as a measure of dissimilarity. As is, however, our current work provides a mathematical basis for a problem which has extensive biological importance, connecting genetic concepts with broader combinatorial notions.

5 Acknowledgments

The authors would like to thank the Stanford University Research Institute in Mathematics (SURIM) 2023 Program and Lernik Asserian for supporting this project, Tarun Martheswaran for his contributions to the predecessor of this project, and the rest of Rosenberg Lab for their mentorship and guidance.

References

J. S. Heslop-Harrison. Polyploidy. In *Reference Module in Life Sciences*. Elsevier, 2017. ISBN 978-0-12-809633-8. doi: https://doi.org/10.1016/B978-0-12-809633-8.06934-X.

X. Liu. *Computational Methods and Mathematical Measures for Population Relationships*. PhD thesis, Stanford University, 2023.

X. Liu, Z. Ahsan, T. Martheswaran, and N. Rosenberg. When is the allele-sharing dissimilarity between two populations exceeded by the allele-sharing dissimilarity of a population with itself? *Statistical Applications in Genetic and Molecular Biology*, in revision.

J. L. Mountain and U. Ramakrishnan. Impact of human population history on distributions of individual-level genetic distance. *Human Genomics*, 2:4–19, 2005.

OEIS Foundation Inc. The On-Line Encyclopedia of Integer Sequences, 2023. Published electronically at http://oeis.org.