Brazilian Journal of Biometrics (2025), **43**, 1–15
doi:10.28951/bjb.v43i3.787

BRAZILIAN JOURNAL OF BIOMETRICS
ISSN:2764-5290

ARTICLE

Impact and interpretation of different measures of genetic diversity in inbred lines applied to crop breeding

María Gabriela Cendoya,*[,1] Martín Grondona,[2] and Andrés Zambelli[3]

[1]Facultad de Ciencias Agrarias, Universidad Nacional de Mar del Plata, Balcarce, Argentina.
[2]Advanta Seeds Biotech Center, College Station, Texas, USA.
[3]Facultad de Agronomía, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina.
*Corresponding author. Email: mcendoya@mdp.edu.ar

(Received: October 25, 2024; Revised: February 18, 2025; Accepted: March 6, 2025; Published: June 2, 2025)

Abstract

The application of molecular genetics in crop breeding has grown significantly, largely due to the success of molecular breeding, which utilizes genotype-based approaches to achieve substantial genetic improvements with favorable cost-effectiveness. A successful molecular breeding strategy involves thorough genotyping, enabling detailed genetic characterization of target germplasm, including genetic diversity analysis, relationships, and population structure. Genotype-based methods, favored for their stability and independence from environmental factors, are preferred over phenotype-based approaches. Genetic diversity is assessed by comparing individual genotypes within and across populations, using statistical methods to calculate genetic distances or similarities. This study focuses on establishing a unified framework to compare and evaluate common similarity measures and their relation to distance metrics, specifically in diploid inbred lines genotyped with biallelic SNPs, for use in genetic improvement efforts.

Keywords: Genetic similarity; Genetic distance; Deep genotyping; Molecular breeding.

1. Introduction

The application of molecular genetics in crop breeding has expanded considerably, driven by the success of "molecular breeding" genotypebased approaches that offer substantial genetic gains with a highly favorable cost-benefit ratio. A successful molecular breeding strategy begins with comprehensive genotyping, allowing for indepth genetic characterization of the target germplasm. This includes the evaluation of genetic diversity, the definition of genetic relationships, and the assessment of population structure (Zambelli, 2023).

Genetic diversity within and between plant populations can be analyzed through morphological traits, biochemical markers such as allozymes, and genotyping. Among these, genotype based methods are preferred over phenotype based approaches due to their stability and the fact that they can be detected in all tissues, irrespective of developmental stage and environmental conditions (Mondini *et al.,* 2009).

To measure genetic diversity, individual genotypes within and across populations are compared. This process begins with constructing a genotype matrix through pairwise comparisons of samples, followed by the calculation of genetic distances and similarities using various statistical methods (Mohammadi & Prasanna, 2003; Patterson *et al.,* 2006; Reif *et al.,* 2005; Warburton & Crossa, 2002).

The inverse relationship between genetic distance and similarity is intuitive, but explaining it in practical applications can be challenging. Selecting the appropriate statistical method requires a nuanced understanding of the distinctions between these approaches and how they influence the final outcomes.

This study aims to establish a cohesive framework for comparing and evaluating widely used similarity and distance measures. It emphasizes the importance of accurately understanding the relationships between these measures, particularly in the context of germplasm comprising diploid inbred lines genotyped with biallelic SNPs.

2. Similarity and Distance Measures for Inbred Lines

Genetic diversity in an inbred line population genotyped with biallelic SNPs can be analyzed using multivariate methodologies based on various similarity or distance measures, where one of the two alleles at each locus (SNP) is arbitrarily designated as the reference allele. Consequently, the resulting SNP information matrix will be:

$$X = [x_{ik}] \tag{1}$$

Where X is a matrix of n rows and m columns, n represents the number of inbred lines (rows) and m represents the number of SNPs (columns). Each entry x_{ik} takes the value 0, 1, or 2 depending on whether line i (L_i) has 0, 1, or 2 copies of the reference allele at SNP k. For homozygous inbred lines, however, x_{ik} can only take the values of 0 or 2. This property of the X matrix simplifies and influences certain overlaps between different similarity and distance indices.

Table 1. Classification of SNPs based on the genotypes of lines L_i and L_j, where m represents the total number of SNPs used for the genotype characterization of the germplasm

		L_i		
		Reference allele	Non-reference allele	Total
L_j	Reference allele	a	b	$a + b$
	Non-reference allele	c	d	$c + d$
	Total	$a + c$	$b + d$	m

Given any two rows of matrix X (Equation 1), representing two homozygous lines L_i and L_j, let "a" denote the number of SNPs where both lines have the reference allele, "d" represent the number of SNPs where both lines have the non-reference allele, "c" denote the number of SNPs where L_i shows the reference allele and L_j does not, and "b" refer the number of SNPs where L_j has the reference allele and L_i does not, and "d" refer the number of SNPs where L_j has the reference allele and L_i does not (as show in Table 1). Based on this SNP classification and the structure of matrix X, different similarity and distance indices between lines can be defined.

2.1 Similarity Indexes for Inbred Lines

2.1.1 Simple matching coefficient or proportion of shared alleles

$$SM(L_i; L_j) = (a + d)/(a + b + c + d) = (a + d)/m \quad (2)$$

For inbred lines, *SM* represents the proportion of shared alleles between lines calculated as $((2a + 2d)/2m)$, irrespective of the reference allele. *SM* ranges from 0, when no alleles are shared, to 1, when all alleles are shared. Each shared allele contributes with $1/m$ to the overall similarity score.

2.1.2 Jaccard coefficient

$$J(L_i; L_j) = a/(a + b + c)$$

The *J* value will vary according to changes in the reference allele. The "*d*" SNPs, where both lines share the non-reference allele, hold the same weight as the "*a*" SNPs, where the lines share the reference allele (Table 1). Ignoring these "*d*" SNPs negatively impacts confidence in the coefficient.

2.1.3 Dice coefficient

$$Di(L_i; L_j) = 2a/(2a + b + c)$$

In this scenario, greater weight is assigned to the SNPs for which both lines share the reference allele. Consequently, *Di* shares the same limitation as *J* in that it does not account for the "*d*" SNPs; as the reference allele changes, the coefficient value also changes.

It is important to note that *J* and *Di* coefficients can be utilized based on a table that categorizes each allele by the presence or absence in the two lines, although this approach is not applicable within the framework we are currently presenting.

2.1.4 Genomic similarity

The *SM* similarity coefficient (Eq. 2) assigns equal weight to all alleles, irrespective of their population frequencies. However, this simplification may overlook the significance of low-frequency alleles, which can offer more valuable insights into the genetic similarity between lines. Sharing a rare allele might indicate a closer relationship than sharing a common, high-frequency allele. For the k^{th} column of matrix *X* (Eq. 1) , the mean is given by $2p_k$, where p_k represents the proportion of lines in which SNP_k has the reference allele. When considering the genotyped lines as the population under study, p_k reflects the reference allele frequency, while $(1 - p_k)$ denotes the non-reference allele frequency. The column-centered matrix *X* is defined as follows:

$$X_c = [x_{ik} - 2p_k] \quad (3)$$

In a population that is in Hardy-Weinberg equilibrium, *X* (Eq. 1) admit 0, 1 or 2 copies of the reference allele and the expected variance of column k^{th} is $2p_k(1 - p_k)$, but since in inbred lines all loci are expected to be homozygous, the variance of the k^{th} column is $4p_k(1-p_k)$, then standardizing *X* by columns gives:

$$X_s = [x_{ik}^s] = \left[\left(\frac{x_{ik} - 2p_k}{\sqrt{4p_k(1 - p_k)}} \right) \right] \quad (4)$$

Patterson *et al.* (2006) made significant contributions to the study of population structure by introducing a normalized version of X, which is proportional to X_s, demonstrating that this scale method is more sensitive in detecting population structures compared to previous methods.

Based on matrix X_c (Eq. 3), a genomic similarity matrix is defined as: $G_1 = \frac{1}{m} X_c X_c^t$, then the similarity between two lines is defined by the entries of G_1:

$$\left[G_1\left(L_i; L_j\right)\right] = \left[\frac{1}{m}\sum_{k=1}^{m}\left(x_{ik} - 2p_k\right)\left(x_{jk} - 2p_k\right)\right] \tag{5}$$

In traditional similarity measures, the similarity of a line with itself is 1 (maximum similarity). However, the genomic similarity matrix G_1 ranges from –1 to 4, deviating from this rule due to the incorporation of allele frequencies. The contributions of different SNPs to this similarity measure, based on Table (1), are defined as follows: "a" SNPs (where $x_{ik} = x_{jk} = 2$) contribute with a value of $\frac{4}{m}(1 - p_k)^2$, "d" SNPs (where$x_{ik} = x_{jk} = 0$) contribute with a value of $\frac{4}{m}(p_k)^2$, "$(b + c)$" SNPs (where $x_{ik} = 2$ and $x_{jk} = 0$, or vice versa) contribute penalizing de similarity with a value of $-\frac{4}{m}p_k(1 - p_k)$.

While the similarity of a line with itself does not equal 1, it is always greater than or equal to its similarity with any other line. This can be expressed mathematically as::

$$G_1\left(L_i; L_i\right) \geq G_1\left(L_i; L_j\right) \quad \forall\, L_i\,,\ L_j.$$

This condition implies that the similarity between any two lines cannot exceed the similarity of each line with itself, which will be an important aspect in subsequent analyses.

In addition to G_1, another genomic similarity index can be derived from the matrix X_s (Eq. 4). This alternative index is represented by the matrix $G_2 = \frac{1}{m} X_s X_s^t$,

$$\left[G_2\left(L_i; L_j\right)\right] = \left[\frac{1}{m}\sum_{k=1}^{m}\frac{\left(x_{ik} - 2p_k\right)\left(x_{jk} - 2p_k\right)}{4p_k(1 - p_k)}\right] \tag{6}$$

G_2 ranges from –1 to ∞ and considers allele frequencies in a different manner compared to G_1: "a" SNPs contribute with $\frac{1}{m}\frac{(1-p_k)}{p_k}$, "d" SNPs contribute with $\frac{1}{m}\frac{p_k}{(1-p_k)}$, "$(b + c)$" SNPs contribution is penalized with a constant value of $-\frac{1}{m}$, regardless of the allele frequencies.

The product $p_k(1 - p_k)$ reach its maximum value of 0.25 when $p_k = 0.5$ and decreases symmetrically as p_k approaches to 0 or 1. By standardizing X columns, the impact of sharing low-frequency alleles on the similarity increased (Figure 1). The theoretical range of G_2 is $(-1; \infty)$, but it is influenced by the SNP filtering criteria. When excluding SNPs with minor allele frequency (MAF) less than 0.05, the maximum value of G_2 is lower than 19 . If only SNP with MAF less than 0.01 are excluded, the maximum value drops below 99. Like G_1, the G_2 similarity also satisfies the inequality $G_2\left(L_i; L_i\right) \geq G_2\left(L_i; L_j\right)$ for all L_i, L_j. This means that the similarity of a line with itself is always greater than or equal to its similarity with any other line.

Figure 1 illustrates that for both G_1 (Eq. 5) and G_2 (Eq. 6), the lower the frequency of the shared allele, the greater its contribution to the similarity, while *SM* is not affected. This effect is more pronounced for G_2. While the impact of a shared allele in G_1 is proportional to the squared probability of the absent allele, in G_2, it is proportional to the odds ratio of the absent allele.

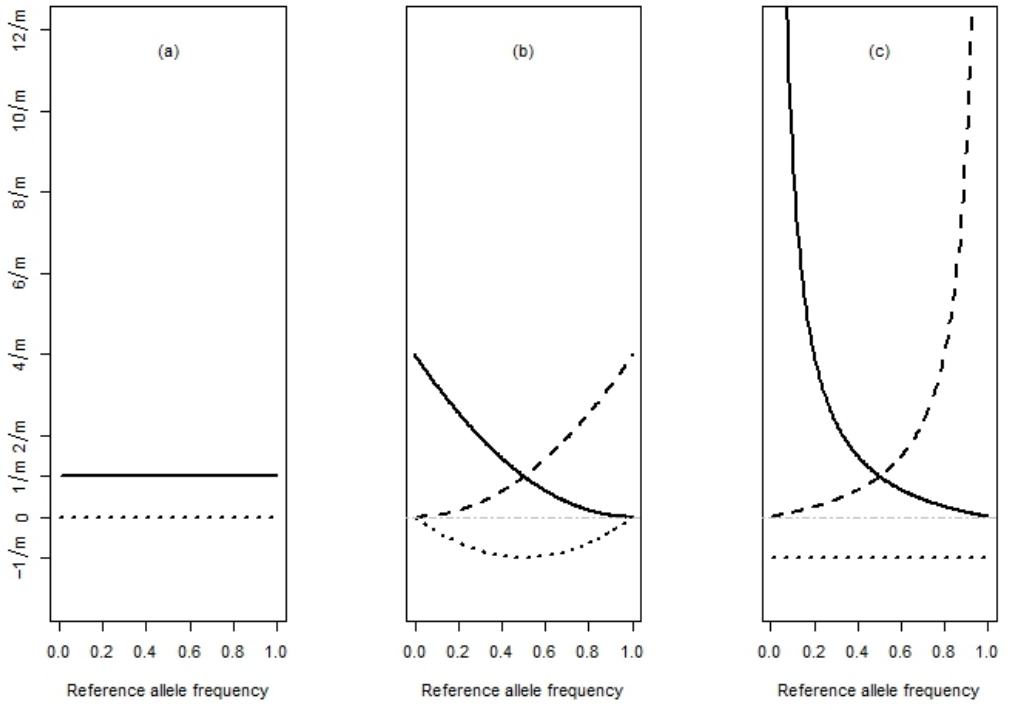

Figure 1. Contribution of SNPs to genomic similarity indexes as a function of their reference allele frequencies. Solid lines represent contribution of shared reference alleles, dashed lines contribution of a non-reference shared alleles, and dotted lines contribution (penalty) of the non-shared alleles.

In the context of genetic prediction (GP) for a given trait, lines in X rows are a random sample from a population, and for the genetic effects the normal distribution assumed is :

$$g \sim N\left(0;\ \sigma_g^2 G\right)$$

where: $g = [g_i]$ is the vector of breeding values for each individual concerning the trait, G is the genomic additive relationship matrix, σ_g^2 represents the genetic variance of the trait within the population captured by the model.

Several formulations for the genomic relationship matrix G have been proposed (Fernando *et al.,* 2017; Meuwissen *et al.,* 2001; VanRaden, 2008; VanRaden, 2007). However, the most referenced matrices are related to G_1 and G_2.

The original proposal is known as the animal model, where $G = A$, A is proportional to the additive relationship matrix, and relies solely on pedigree data to calculate the expected probabilities that pairs of loci are identical by descent (Wright, 1922).

For the matrix G, when genomic information from SNPs it is available, VanRaden (2008) first suggested the following formulation:

$$G_{VR1} = \frac{1}{2\sum_{k=1}^{m} p_k\left(1-p_k\right)} X_c X_c^t \tag{7}$$

In this equation, p_k represents the population frequency of the reference allele for the k^{th} SNP. This approach creates a genomic relationship matrix analogous to the additive genetic relationship matrix (A), in the sense that its expected value ($E(G_{\nu R1})$) is A. Treating the columns of X as random variables, then $E(X_c X_c^t) = \left(2\sum_{k=1}^{m} p_k(1-p_k)\right) A$, hence $E(G_{\nu R1}) = A$ (Habier *et al.,* 2007).

It is important to note that $G_{\nu R1}$ is related to G_1, but while G_1 is based on p_k the frequency of the reference allele for the k^{th} SNP in the groups of lines represented in X's rows, $G_{\nu R1}$ uses the frequencies of those alleles in the population from which the lines in X were selected. These difference is some times not clearly establish in some papers and can lead to confusion. If the lines in X can be considered as a random sample of the population, then p_k values used in G_1 are the estimators of the ones needed for $G_{\nu R1}$.

When there are enough markers, so they are in linkage disequilibrium with all the QTL (Quantitative Trait Loci), $G_{\nu R1}$ provides more information about the covariance between relatives than matrix A, because realized relationships differs from expected relationships (Nejati-Javaremi *et al.,* 1997; VanRaden & Tooker, 2007).

The similarity index G_2 is associated with the second formula proposed by VanRaden (2008) for constructing the genomic relationship matrix, defined as:

$$G_{\nu R2} = \left[\frac{1}{m}\sum_{k=1}^{m}\frac{(x_{ik}-2p_k)\left(x_{jk}-2p_k\right)}{2p_k(1-p_k)}\right] \tag{8}$$

Following the same procedure than (Habier *et al.,* 2007) it can be proved that $E(G_{\nu R2}) = A$. Notice that while G_2 used the reference alleles frequencies in the groups of lines represented in X, $G_{\nu R2}$ needs the populations frequencies.

Patterson *et al.* (2006) employed a matrix proportional to G_2 to investigate population structure using different methodologies. This highlights the versatility and relevance of the G_2 similarity index in population genetics studies.

2.2 Distances for Inbred Lines

Genetic distance between genotypes can be calculated using various statistical approaches (Nei, 1972; Nei *et al.,* 1983; Reif *et al.,* 2005; Weir & Cockerham, 1996). From a mathematical perspective, when individuals are characterized by a vector of multiple quantitative traits, the distance between them can be defined as the "magnitude" of their difference vector. Several common distance metrics can be employed, each measured using different norms: Manhattan (L1 norm), Euclidean (L2 norm) and Standardized Euclidean.

2.2.1 Manhattan distance or Taxi cab

The Manhattan distance, denoted as $d_{Man}\left(L_i; L_j\right)$, is defined mathematically as:

$$d_{Man}\left(L_i; L_j\right) = \sum_{k=1}^{m}\left|x_{ik}-x_{jk}\right| = 2(b+c) \tag{9}$$

This metric measures the distance by summing the absolute differences between corresponding coordinates of two vectors. It is analogous to calculating the distance between two points on a grid, where movement is restricted to parallel paths along the axes. The range of Manhattan distance is

between 0 and $2m$, making it dependent on the total number of markers. This dependence is not desired when assessing genetic distances.

Nei *et al.* (1983) proposed a modified version of the Manhattan distance that addresses this limitation. This rescaled distance, denoted as $d_{Nei83}\left(L_i; L_j\right)$, is defined as:

$$d_{Nei83}\ \left(L_i; L_j\right) = \frac{1}{2m}\sum_{k=1}^{m}\left|x_{ik} - x_{jk}\right| = \frac{b+c}{m} \tag{10}$$

In this formulation, d_{Nei83} ranges from 0 to 1 and is independent of the number of markers. This distance can be interpreted as the proportion of non-shared alleles between two genotypes. The original expression of this distance was in terms of the presence-absence of each allele at a locus. However, when applied to biallelic loci and homozygous lines, it can be simplified as equation (10).

2.2.2 Euclidean distance

The Euclidean distance between two genotypes, denoted as $d_E\left(L_i; L_j\right)$ is defined mathematically as:

$$\left(d_E\left(L_i; L_j\right)\right)^2 = \sum_{k=1}^{m}\left(x_{ik} - x_{jk}\right)^2 \tag{11}$$

This distance is based on the L2 norm and represents the length of the straight line connecting the vector representations of the genotypes in the rows of matrix X.

According to equation (11) and Table (1), there are "a" SNPs (where $x_{ik} = x_{jk} = 2$) and "d" SNPs (where $x_{ik} = x_{jk} = 0$), these SNPs do not contribute to the distance between lines. Only the "$b + c$" SNPs where $x_{ik} = 2$ and $x_{jk} = 0$, or vice versa contribute. Specifically, for these contributing SNPs, the squared difference is given by $(2 - 0)^2 = 4$. Thus, the Euclidean distance can be expressed as:

$$d_E\ \left(L_i; L_j\right) = 2\sqrt{(b+c)}$$

Consequently, the Euclidean distances between lines depend solely on the number of non-shared alleles and range from 0 to $2\sqrt{m}$. Similar to the Manhattan distance, this dependence on the number of markers is undesirable for assessing genetic distances. To address these limitations, Wright (1978) and Goodman & Stuber (1983) proposed a modification to the Euclidean distance known as Modified Rogers' distance, denoted as $d_{MR}\ \left(L_i; L_j\right)$ and defined as:

$$d_{MR}\left(L_i; L_j\right) = \frac{1}{2\sqrt{m}}d_E\ \ \left(L_i; L_j\right) = \frac{1}{2\sqrt{m}}\sqrt{\sum_{k=1}^{m}\left(\ x_{ik} - \ x_{jk}\right)^2} = \sqrt{\frac{b+c}{m}} \tag{12}$$

Modified Rogers' distance ranges from 0 and 1 and is independent of the number of markers. According to Eq. (13), the Modified Rogers' distance d_{MR} considers the alleles frequencies as follow, there are "a" SNPs (where $x_{ik} = x_{jk} = 2$) and "d" SNPs (where $x_{ik} = x_{jk} = 0$), these SNPs do not contribute to the distance between lines; only the "$b + c$" SNPs where $ik = 2$ and $x_{jk} = 0$, or vice versa contribute, precisely each one contributes to the squared d_{MR} by $\frac{1}{m}$.

In its original definition, Modified Rogers' distance considered all alleles at each locus. However, for biallelic markers and homozygous (inbred) lines, this expression can be simplified to align with equation (12).

From the relationship between L1 an L2 norms, it follows that: $d_{MR}\left(L_i;L_j\right) \leq d_{Nei83}\left(L_i;L_j\right)$ for every pair of lines. Particularly when measuring genetic distances, it is found that: $d_{MR}\left(L_i;L_j\right) = \sqrt{d_{Nei83}\left(L_i;L_j\right)}$ This means that when the proportion of shared alleles is high, Modified Rogers' distance magnifies the distances between individuals compared to d_{Nei83}.

Note that both d_{Nei83} and d_{MR} depend solely on the number of non-shared alleles but do not consider their frequencies. This characteristic highlights their utility in measuring genetic distances while also indicating a potential limitation in fully capturing genetic diversity.

2.2.3 Standardized Euclidean distance

The standardized Euclidean distance between lines L_i and L_j can be defined as the Euclidean distance (based on L2 norm) between the corresponding rows in the standardized matrix X_s (Eq. 4):

$$d_E\left(X_s[i,];X_s[j,]\right) = \sqrt{\sum_{k=1}^{m}\left(\frac{\left(X_{ik}-2p_k\right)}{2\sqrt{p_k(1-p_k)}}-\frac{\left(X_{jk}-2p_k\right)}{2\sqrt{p_k(1-p_k)}}\right)^2}$$

The Euclidean distance d_E depends on the number of markers, since each marker contributes a non-negative value to the total distance. To obtain a distance metric that is independent of the number of markers, denoted as d_{SE}, d_E can be rescale as follows:

$$d_{SE}\left(L_i;L_j\right) = \frac{1}{\sqrt{m}}\sqrt{\sum_{k=1}^{m}\frac{\left(x_{ik}-x_{jk}\right)^2}{4p_k\left(1-p_k\right)}} \tag{13}$$

According to equation (13) and Table (1), the standardized Euclidean distance d_{SE} considers the alleles frequencies as follow, there are "a" SNPs (where $x_{ik} = x_{jk} = 2$) and "d" SNPs (where $x_{ik} = x_{jk} = 0$), these SNPs do not contribute to the distance between lines; only the "$b + c$" SNPs where $x_{ik} = 2$ and $x_{jk} = 0$, or vice versa contribute. Specifically, the contribution of these SNPs squared of d_{SE} is given by $\frac{1}{m}\frac{(2-0)^2}{4p_k(1-p_k)} = \frac{4}{4p_k(1-p_k)} = \frac{1}{p_k(1-p_k)}$.

d_{SE} depends solely on the non-shared alleles, weighting their contributions by the product of the frequencies of both possible alleles for each locus. When the reference allele frequency for a locus is 0.5 and the lines do not share this allele, its contribution to the squared distance reaches its minimum value of $\left(\frac{4}{m}\right)$. Conversely, as the reference allele frequency approaches 0 or 1, its contribution increases significantly, tending toward infinity. Therefore, not sharing alleles with a frequency of 0.5 has less impact than not sharing alleles with frequencies close to 0 or 1. As a result, the range of d_{SE} is $(0,\ \infty)$ (Fig. 2). The maximum value of d_{SE} will vary depending on SNP filtering criteria. For instance, excluding SNPs with MAF less than 0.05 results in a maximum value of less than 4.59. In contrast, if SNPs with MAF less than 0.01 are excluded, the maximum value increases but remain less than 10.06.

Hamming distance is a popular measure in information theory and coding theory, for two strings of equal length, Hamming distance is defined as the number of positions at which the corresponding symbols differ. When applied to the representation of lines in the rows of matrix X there are $b + c$

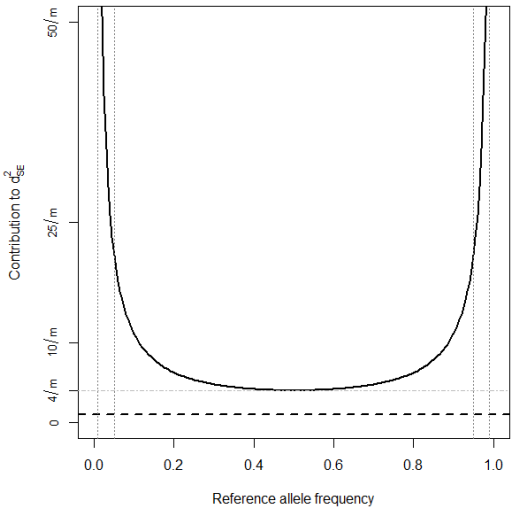

Figure 2. Contribution of the non-shared reference alleles on the squared standardized Euclidean distance d_{SE}. The vertical dashed lines indicate the commonly used thresholds for MAF at 0.01 and 0.05.

SNPs where the lines do not share an allele. Thus, it is related to d_{Man} (Eq. 9) and d_{Nei83} (Eq. 10) as follows:

$$d_{Hamm}\left(L_i;L_j\right) = b + c = \frac{1}{2}d_{Man}\left(L_i;L_j\right) = m\, d_{Nei83}\left(L_i;L_j\right)$$

Table 2. Genetic distance metrics for inbred lines based on allele classifications from Table (1). Standardized Euclidean distance d_{SE} is expressed in terms of the entries of matrix X and the corresponding allele frequencies

Genetic distance	Range	Euclidean	Based on alleles frequencies
$d_{Nei83}\left(L_i;L_j\right) = \frac{b+c}{m}$	$[0,1]$	No	No
$d_{MR}\left(L_i;L_j\right) = \sqrt{\frac{(b+c)}{m}}$	$[0,1]$	Yes	No
$d_{SE}\left(L_i;L_j\right) = \frac{1}{\sqrt{m}}\sqrt{\sum_{k=1}^{m}\frac{\left(x_{ik}-x_{jk}\right)^2}{4p_k(1-p_k)}}$	$[0,\infty]$	Yes	Yes

3. Relationships between similarities and distances for inbreed lines

Similarity and distance are inversely correlated, allowing for the association of a distance metric with each similarity index. For each similarity indexes $K\left(L_i;L_i\right)$ several distances can be associated.

When a similarity index K range between 0 and 1, an associated distance can be expressed as $1-K$. This is the case of $d_{Nei83}\left(L_i;L_j\right) = 1 - SM\left(L_i;L_j\right)$. For similarity indexes that do not fall within the $[0,1]$ range, a kernel distance can still be defined if they satisfy the condition: $K\left(L_i;L_i\right) \geq K\left(L_i;L_j\right)$ for all L_i, L_j. In this case, the squared kernel distance is defined as:

$$\left(d_K\left(L_i;L_j\right)\right)^2 = K\left(L_i;L_i\right) + K\left(L_j;L_j\right) - 2K\left(L_i;L_j\right)$$

The kernel distance d_K satisfices the following properties:

$d_K(L_i; L_i) = 0$ for every L_i, $d_K(L_i; L_j) \geq 0$,
$d_K(L_i; L_j) = d_K(L_j; L_i)$ for all genotypes L_i and L_j.
Here K serves as the kernel of the distance (Aronszajn, 1950; Hein & Bousquet, 2005).
When using the similarity measure SM as the kernel, the corresponding distance is defined as:

$$\begin{aligned}\left(d_{SM}(L_i; L_j)\right)^2 &= SM(L_i; L_i) + SM(L_j; L_j) - 2SM(L_i; L_j) \\ &= 2\left(1 - SM(L_i; L_j)\right) \\ &= 2\, d_{Nei83}(L_i; L_j) \\ &= 2\left(d_{MR}(L_i; L_j)\right)^2\end{aligned}$$

Thus, the kernel distance based on SM similarity is proportional to the Modified Rogers' distance (d_{MR}). Similarly, when using the genomic relationship matrix G_1 as a kernel the resulting distance is also proportional to d_{MR}:

$$\begin{aligned}\left(d_{G1}(L_i; L_j)\right)^2 &= G_1(L_i; L_i) + G_1(L_j; L_j) - 2G_1(L_i; L_j) \\ &= \frac{1}{m}\sum_{k=1}^{m}(x_{ik} - 2p_k)^2 + \frac{1}{m}\sum_{k=1}^{m}\left(x_{jk} - 2p_k\right)^2 - 2\frac{1}{m}\sum_{k=1}^{m}(x_{ik} - 2p_k)\left(x_{jk} - 2p_k\right) \\ &= \frac{1}{m}\sum_{k=1}^{m}\left(x_{ik} - x_{jk}\right)^2 \\ &= \left(2d_{MR}(L_i; L_j)\right)^2 \\ &= 2\left(d_{SM}(L_i; L_j)\right)^2\end{aligned}$$

Although G_1 similarity, in its calculations, account for allele frequencies (in the lines represented in X), the induced kernel distance is proportional to a distance that remains unaffected by these frequencies, with the constant of proportionality depending on the allele frequency across all loci.

In contrast, when using the similarity index from the genomic relationship matrix G_2 (Eq. 6) as a kernel, the distance obtained is $d_{SE}(L_i; L_j)$.

$$\begin{aligned}\left(d_{G2}(L_i; L_j)\right)^2 &= \left(G_2(L_i; L_i) + G_2(L_j; L_j) - 2G_2(L_i; L_j)\right) \\ &= \frac{1}{m}\sum_{k=1}^{m}\frac{(x_{ik} - 2p_k)^2}{4p_k(1-p_k)} + \frac{1}{m}\sum_{k=1}^{m}\frac{\left(x_{jk} - 2p_k\right)^2}{4p_k(1-p_k)} - 2\frac{1}{m}\sum_{k=1}^{m}\frac{(x_{ik} - 2p_k)\left(x_{jk} - 2p_k\right)}{4p_k(1-p_k)} \\ &= \frac{1}{m}\sum_{k=1}^{m}\left(\frac{x_{ik}}{2\sqrt{p_k(1-p_k)}} - \frac{x_{jk}}{2\sqrt{p_k(1-p_k)}}\right)^2 \\ &= \left(d_{SE}(L_i; L_j)\right)^2\end{aligned}$$

Distance d_{SE} is shaped by population allele frequencies at loci where lines do not share the allele. This distinction highlights how different similarity measures can influence genetic distance assessments in diverse ways.

4. Graphical visualization of the distance between inbred lines

Multidimensional scaling (MDS), as described by Cox & Cox (2000), is a powerful multivariate data visualization technique that allows researchers to represent complex data in a lower-dimensional space. Given a distance matrix $d = [d_{ij}]$ that quantifies the dissimilarities between individuals, MDS aims to find a corresponding set of points in a real vector space of dimension p. The goal is to ensure that the Euclidean distance $\left(\widetilde{d}_{ij}\right)$ between points i and j is "as close as possible" to the original dissimilarity measure (d_{ij}). The interpretation of "as close as possible" can vary, leading to different methodologies for MDS. Classical MDS, also known as principal coordinate analysis (PCoA) minimize the stress function defined as:

$$Stress\left(d,\widetilde{d}\right) = \frac{\sum_{i\neq j}\left[d_{ij}-\widetilde{d}_{ij}\right]^2}{\sum_{i\neq j} d_{ij}^2}$$

This approach focuses on preserving the relationships among individuals while reducing dimensionality. Principal component analysis (PCA) is another linear dimensionality reduction technique widely used in exploratory data analysis and visualization. PCA transforms the data into a new coordinate system where the first axis captures the largest variance, followed by subsequent axes that explain the remaining variance. The mathematical foundation for PCA involves singular value decomposition (SVD) applied to the matrix:

$$\frac{1}{\sqrt{m}}X_c = UDV^t$$

Where: the columns of U and V are the left and right singular vector of $\frac{1}{\sqrt{m}}X_c$, respectively. U and V are othonormal matrices with rank r equal to the rank of X_c, D is a diagonal matrix containing the singular values of $\frac{1}{\sqrt{m}}X_c$ in decreasing order: $\lambda_1 > \lambda_2 > \cdots > \lambda_r$.

The SVD of $\frac{1}{\sqrt{m}}X_c$ yields: $\frac{1}{m}X_cX_c^t = U\widetilde{D}U^t$ where $\widetilde{D}$ is a diagonal matrix with the squared singular values of $\frac{1}{\sqrt{m}}X_c$. This expression satisfies the definition of eigenvalue decomposition of $\frac{1}{m}X_cX_c^t$, indicating that the eigenvalues and eigenvectors of $\frac{1}{m}X_cX_c^t$, correspond to the squared singular values and left singular vectors of $\frac{1}{\sqrt{m}}X_c$, respectively. To reduce dimensionality, only the first few singular vectors and values are utilized. For example, if we desire to retain only two dimensions, we can denote:

$U_{1:2}$: The first two columns of U,

$D_{1:2;1:2}$: The diagonal matrix containing just the first two singular values.

The Euclidean distances between rows $\frac{1}{\sqrt{m}}X_c$ are best represented in two dimensions by the Euclidean distances between rows of $U_{1:2}D_{1:2;1:2}$.

Consequently, applying PCA to $\frac{1}{\sqrt{m}}X_c$ is equivalent to applying classical MDS to the Euclidean distance matrix (Venables & Ripley, 2013) . It's important to note that the covariance matrix for the columns of X is given by $\frac{1}{m}X_c^tXc$, which leads to the total variance in the columns of X:

$$Tr\left(\frac{1}{m}X_c^tX_c\right) = Tr\left(\frac{1}{m}X_cX_c^t\right) = Tr\left(D\right) = \sum_{k=1}^{r}\left(\lambda_k\right)^2$$

This relationship explains why reduced–dimensional representations recover a maximum percentage of the total variance in the columns of X .

Given these concepts, it follows that since d_{MR} (Eq. 12) represents the Euclidean distance between rows of $\left(\frac{1}{2\sqrt{m}}X_c\right)$, applying PCA to $\frac{1}{2\sqrt{m}}X_c$ will yield principal coordinates that provide the best representation of d_{MR} among lines in r dimensions. Furthermore, performing PCA on both $\frac{1}{2\sqrt{m}}X_c$ and $\frac{1}{4m}X_c{X_c}^t = \frac{1}{4}G_1$ will yield equivalent representations for these lines, thus also representing d_{MR}. Similarly, applying PCA on $\frac{1}{\sqrt{m}}X_s$, or on G_2, will result in principal coordinates that best represent the standardized Euclidean distanced_{SE} (Eq. 13). Since d_{Nei83} is a non-Euclidean distance metric, classical MDS is required to create plots similar to PCA plots that accurately represent the d_{Nei83} distances between lines in r dimensions.

A critical consideration for accurately representing distances in plots is ensuring that the aspect ratio of the x–axis and y–axis is set to 1. This means that one unit on the x-axis should occupy the same visual space as one unit on the y–axis. Only by adhering to this principle can the straight–line distance between two points be perceived accurately by the human eye, thereby reflecting the true distance that it is intended to represent.

5. Conclusions

The methods for measuring genetic distance and similarity between inbred lines can be categorized based on the L1 and L2 norms utilized. When characterizing the genetic information of a population of inbred lines, it is crucial to determine whether to use Manhattan distance, Euclidean distance based on raw marker information, or Euclidean distance based on standardized marker information to best capture genetic diversity. If the objective of the genetic characterization is to understand the behavior of a specific trait, it becomes essential to consider the underlying gene action associated with that trait. This understanding will influence the importance of accounting for allele frequencies in the population when making descriptions. For example, if two lines do not share an allele at a particular locus, that locus will contribute to their differentiation in trait behavior. However, the question arises: Should this contribution be weighted by allele frequency or treated independently? From a statistical or mathematical standpoint, there is no definitive right or wrong answer to this question; it ultimately depends on the specific context of the study.

For example, if the genotypes under study come from a well-established breeding program, most of the genes are likely already selected to be part of the population. In this case, if low-frequency alleles are suspected not to contribute to the trait of interest, then SM similarity, along with d_{Nei83} or d_{MR}, -which are not influenced by allele frequency- would be preferred over G_1, G_2 and d_{SE}. Conversely, if the genotypes consist of a breeding program where few foreign materials have been recently introduced, the valuable new alleles will likely be at low frequency. In this scenario, G_1, G_2 and d_{SE}. will be recommended because they pay more attention to low-frequency new alleles (Figure: 3). Additionally, when the objective of the study is to estimate genetic variance components using G_2, the additive variance of allele effects may be overestimated if linkage disequilibrium has been induced by selection within the population (Fernando *et al.,* 2017).

Researchers must carefully evaluate their objectives and decide which framework they wish to adopt for their analyses. This thoughtful consideration will enhance the clarity and relevance of their genetic characterizations.

In Table (2) are summarized the key features of different distances used for the genetic comparison of inbred lines. The primary distinction between them lies in whether they induce Euclidean geometry and whether they are influenced by the number of markers used for genotyping. Thus, when genetic distance is proportional to the number of non-shared alleles $(b + c)$, it is characterized by a Manhattan geometry (L1). Conversely, if genetic distance is proportional to the square root of the number of non-shared alleles $\left(\sqrt{b+c}\right)$, it is defined by Euclidean geometry. In both scenarios, allele frequencies within the population of those loci are not considered.

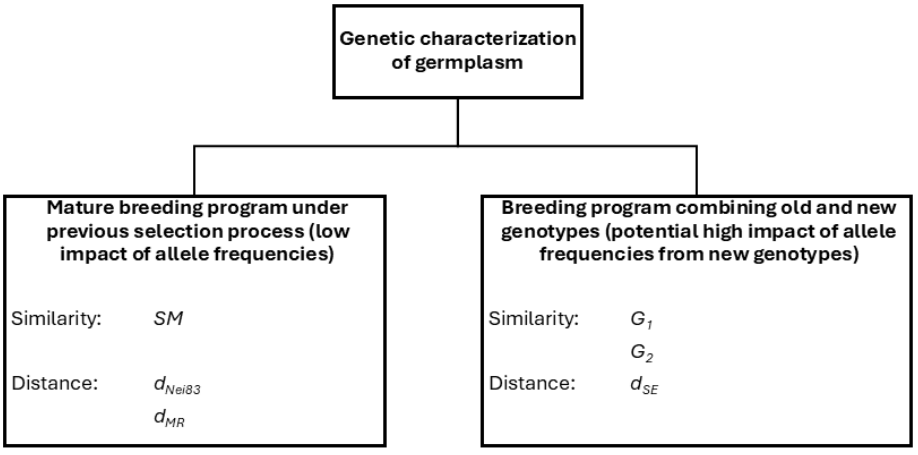

Figure 3. Summary chart outlining key considerations for the genetic characterization of a germplasm base forming a breeding program. This includes the selection process that enhances the frequency of allele combinations associated with improved agronomic performance, as well as the implications of incorporating foreign genotypes into mature breeding program.

The standardized Euclidean distance d_{SE} (Eq. 13) is the only distance metric that takes into account the frequencies of population alleles. Understanding that relationship between distances, similarities, and the interconnections between these measures underscores the importance of their implications in genetic analyses. By recognizing these connections, researchers can make more informed decisions about which metrics to use in their studies.

If the objective of genetic characterization is to estimate the genomic relationship matrix for use in genomic prediction models, then , $G_{\nu R1}$ (Eq. 7) and $G_{\nu R2}$ (Eq. 8) are the most suitable similarity indices. Since both measures are analogous to the additive genetic relationship matrix A (VanRaden, 2008).

Additionally, it is essential to evaluate whether to consider rare alleles to avoid overestimating genetic differences within a specific population. Depending on the trait being predicted, selecting either $G_{\nu R1}$ or $G_{\nu R2}$ may be more appropriate. If the aim is to characterize population structure, PCA analysis on G_2 has been recommended as a useful and sensitive tool (Cavalli-Sforza, 1994; Patterson *et al.,* 2006).

In summary, the selection of the appropriate measure of similarity or distance depends on the specific objectives of genetic characterization for inbred line germplasm. Understanding the different properties of these measures is crucial for choosing the most suitable one for ultimately enhancing the efficiency of a crop breeding strategy.

Acknowledgments

This research is part of the Ph.D. thesis authored by María Gabriela Cendoya, at the Facultad de Ciencias Agrarias, Universidad Nacional de Mar del Plata, Argentina.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Data curation**: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Formal analysis**: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Funding acquisition**: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Investigation**: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Methodology**: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Project administration**:CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Software**: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Resources**: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Supervision**: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Validation**: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Visualization**: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Writing - original draft**: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A. **Writing - review and editing**: CENDOYA, M. G., GRONDONA, M., ZAMBELLI, A.

References

1. Aronszajn, N. Theory of reproducing kernels. *Transactions of the American mathematical society* **68,** 337–404 (1950).
2. Cavalli-Sforza, L. L. *The history and geography of human genes* 428 (1994).
3. Cox, T. F. & Cox, M. A. *Multidimensional scaling* 328 (CRC press, 2000).
4. Fernando, R., Cheng, H, Sun, X & Garrick, D. A comparison of identity-by-descent and identity-by-state matrices that are used for genetic evaluation and estimation of variance components. *Journal of Animal Breeding and Genetics* **134,** 213–223 (2017).
5. Goodman, M. M. & Stuber, C. W. *Races of maize. VI. Isozyme variation among races of maize in Bolivia.* (1983).
6. Habier, D., Fernando, R. L. & Dekkers, J. C. M. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177,** 2389–2397 (2007).
7. Hein, M. & Bousquet, O. *Hilbertian metrics and positive definite kernels on probability measures* in *International Workshop on Artificial Intelligence and Statistics* (2005), 136–143.
8. Meuwissen, T. H., Hayes, B. J. & Goddard, M. Prediction of total genetic value using genome-wide dense marker maps. *genetics* **157,** 1819–1829 (2001).
9. Mohammadi, S. A. & Prasanna, B. Analysis of genetic diversity in crop plants—salient statistical tools and considerations. *Crop science* **43,** 1235–1248 (2003).
10. Mondini, L., Noorani, A. & Pagnotta, M. A. Assessing plant genetic diversity by molecular tools. *Diversity* **1,** 19–35 (2009).
11. Nei, M. Genetic distance between populations. *The American Naturalist* **106,** 283–292 (1972).
12. Nei, M., Tajima, F. & Tateno, Y. Accuracy of estimated phylogenetic trees from molecular data: II. Gene frequency data. *Journal of molecular evolution* **19,** 153–170 (1983).
13. Nejati-Javaremi, A, Smith, C & Gibson, J. Effect of total allelic relationship on accuracy of evaluation and response to selection. *Journal of animal science* **75,** 1738–1745 (1997).
14. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS genetics* **2,** e190 (2006).
15. Reif, J. C., Melchinger, A. E. & Frisch, M. Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop science* **45,** 1–7 (2005).

16. VanRaden, P. & Tooker, M. *Genetic evaluations using combined data from all breeds and crossbred cows* in *Crossbreeding of Dairy Cattle: The Science and the Impact. 4th Biennial WE Petersen Symposium, University of Minnesota, St. Paul. http://www.ansci.umn.edu/petersen_symposium/petersen2007.htm Accessed March* **19** (2007), 23–28.

17. VanRaden, P. M. Efficient methods to compute genomic predictions. *Journal of dairy science* **91,** 4414–4423 (2008).

18. VanRaden, P. Efficient estimation of breeding values from dense genomic data. *J Dairy Sci* **90,** 374–375 (2007).

19. Venables, W. N. & Ripley, B. D. *Modern applied statistics with S-PLUS* 400 (Springer Science & Business Media, 2013).

20. Warburton, M. & Crossa, J. *Data analysis in the CIMMYT applied biotechnology center: for fingerprinting and genetic diversity studies* 2002.

21. Weir, B. S. & Cockerham, C. Genetic data analysis II: Methods for discrete population genetic data. Sinauer Assoc. *Inc., Sunderland, MA, USA* (1996).

22. Wright, S. *Evolution and genetics of populations vol IV.* (The University of Chicago Press, 1978).

23. Wright, S. Coefficients of inbreeding and relationship. *The American Naturalist* **56,** 330–338 (1922).

24. Zambelli, A. The importance of deep genotyping in crop breeding. *Journal of Basic and Applied Genetics* **34,** 47–56 (2023).