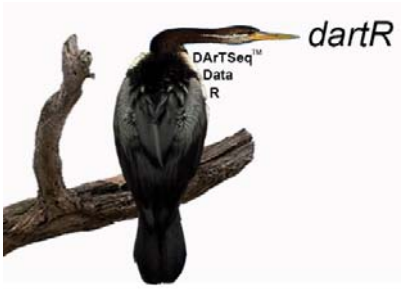
Wildlife Genomics@UC
wildlifegenomics.biomatix.org

UNIVERSITY OF CANBERRA

SNP Analysis using dartR



TechNote: Distance and Visualization in Population Genetics

The Institute for Applied Ecology
University of Canberra ACT 2601
Australia

Email: georges@aerg.canberra.edu.au

dartR is a collaboration between the University of Canberra, CSIRO and Diversity Arrays Technology, and is supported with funding from the ACT Priority Investment Program, CSIRO and the University of Canberra.

TECHNICAL NOTE

Distance and Visualization in Population Genetics

Table of Contents

Introduction

A basic concept in population genetics is the Wright-Fisher model named after Sewall Wright and Ronald Fisher, two individuals who set much of the foundation in theoretical population genetics. In its simplest form, this model consists of a single locus with two alleles in a population of constant finite size and does not incorporate selection or mutation. Over a specified period (e.g. one season, one generation), all individuals randomly mate and then die, leaving only the new, nonoverlapping generation. Through the process of random mating, one allele may become present in greater frequency than another in the offspring of the new generation owing to genetic drift. If the population is finite, this means that one allele must decrease in abundance to accommodate the increase in the abundance of the other. Through time one allele eventually might become extinct resulting in the system moving to a state of fixation, where only one allele exists.

As such, the fundamental processes of population genetics that lead to divergence between populations arise from the interplay of drift, gene flow and selection for one allele over another. The outcome of these processes over time can be captured as genetic distance between populations.

Genetic distances between populations and between individuals are intertwined because one often examines structure by considering genetic difference and similarity among individuals with no preconceived notion of how they might aggregate into natural groupings. Genetic distance between individuals is influenced by the processes of random assortment of alleles (or not) during sexual reproduction, whether the parents were more related in some way than individuals in the general population, other influences on non-random mating, the source of the individuals in the context of barriers to gene flow, etc.

Whether considering populations or individuals, genetic distance is an important concept in population genetics.

There is a daunting plethora of distance and dissimilarity measures in the literature[1]. Fortunately, SNP data have characteristics that limit options substantially. SNP genotypes are comprised of biallelic markers, scored as the frequency of the alternate allele – 0 for homozygous reference allele, 2 for homozygous alternate allele, and 1 for heterozygotes (in diploid organisms). The attributes (SNP loci) are thus all measured as genotypes on the same scale (0, 1 or 2). Thus standardization or normalization is not required as would be under some circumstances in multiallelic systems. The values of 0 and 2 carry equal weight, because the choice of reference allele and alternate allele is arbitrary. The value of 0 is not a true zero. Any distance measures that give differential weight to 0 will yield results that depend on the arbitrary choice of which allele is reference and which is alternate, and this cannot be the case. Such distance measures can be eliminated from consideration.

For SNP data that comprises presence or absence of the amplified sequence tag – 0 for absence and 1 for presence – there exists an array of binary distance measures that are appropriate, but the field of options is manageable.

These notes introduce the concept of distance, the importance of metric properties and the most familiar of distances, Euclidean distance. The rationale for the related techniques of Principal Components Analysis (PCA)[2–4] and Principal Coordinates Analysis (PCoA)[5] is presented, followed by the rationale for genetic distances of relevance to SNP datasets. Application of these techniques using dartR is described to make it clear exactly what algorithms are applied.

We use the concept of distance loosely to encompass the notions of measures of dissimilarity through to metric distances and rigid Euclidean distances. Where the distinction is necessary, a distance is referred to as a non-metric distance, a metric distance, or a Euclidean distance. We refer to individuals or samples or specimens as entities, the SNP loci that are scored for each entity as attributes, and the scores themselves as states.

The Concept of Distance

Possibly the best way to introduce the concept of distance is with a familiar example. Euclidean distance is a common-sense notion derived as an abstraction of physical distance. The distance between two points in our physical space can be measured, as the shortest distance between them, to any desired level of accuracy with a ruler or gauge.

It is possible also to calculate the distance between two points from their coordinates in a space defined by orthogonal Cartesian axes (Figure 1).

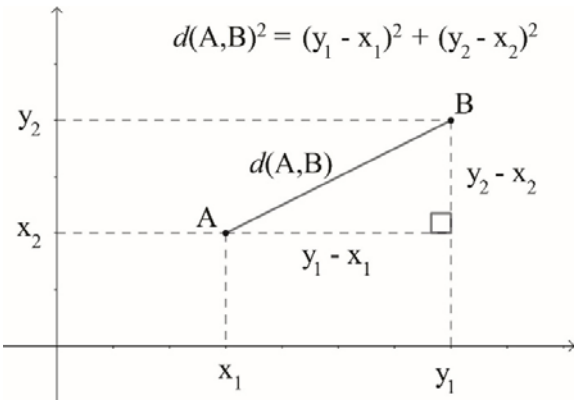


Figure 1. Distance between two points A and B represented in two-dimensional space can be calculated from their Cartesian coordinates using Pythagoras' rule – the square of the hypotenuse of a right-angled triangle is equal to the sum of the squares of the two adjacent sides.

The distance between two points in space is calculated by applying Pythagoras' rule to their projection onto the Cartesian axes (Figure 1).

$$d(A,B)^2 = (y_1 - x_1)^2 + (y_2 - x_2)^2$$

and so the distance between two points A and B can be represented algebraically by

$$d(A,B) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2}$$

This calculation can be generalized to 3 dimensions

$$d(A,B) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + (y_3 - x_3)^2}$$

and beyond to L dimensions

$$d(A,B) = \sqrt{\sum_{i=1}^{L} (x_i - y_i)^2}$$

Metric Distances

Euclidean distance is just one of many distance measures. The concept of distance more generally can be distilled down to three basic properties. For a metric distance:

d(A,B) = 0 if and only if A = B

d(A,B) = d(B,A)

d(A,B) ≤ d(A,C) + d(B,C)

The first condition asserts that indiscernible entities are one and the same. The second condition asserts symmetry. The last condition is referred to as the triangle inequality which enforces the notion that the distance between two points is the shortest. From these properties we can conclude that metric distances must be non-negative.

d(A,B) ≥ 0

In essence, metric distances are well behaved distances.

Graphically, the metric property makes complete sense for a distance (Figure 2). Given three points defined by the distances between them, the position of each of them is uniquely defined. This is necessary (though not sufficient) if we are to draw an analogy between our distances and a representation in a linear physical space.

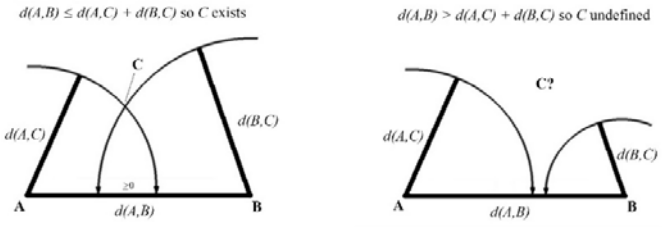


Figure 2. If three distances between A, B and C satisfy the metric properties, then the position of each of A, B and C in space is well defined.

Euclidean Distance is a metric distance.

Euclidean Distance and SNP Genotypes

A genetic distance between two individuals (= samples, specimens) or between two populations is a measure of their genetic dissimilarity. If two individuals or populations have very different genetic profiles, then the measure of dissimilarity will be large. If they have similar genetic profiles, then the measure of dissimilarity will be low. If they have identical genetic profiles, the measure of dissimilarity will be zero. This is a common-sense notion that fits in well with the concept of distance.

Consider a bivariate plot where the axes are defined by the loci and the value taken on each axis is represented by the genotype, which in the case of SNP data is essentially the frequency of the alternate allele (Figure 3). For each individual at a particular locus, the genotype is scored 0 for an individual that is homozygous reference allele (count of zero for the alternate allele); it is scored 1 for an individual that is heterozygous at that locus (count of one for the alternate allele); and it is scored 2 for an individual that is homozygous for the alternate allele (count of 2 for the alternate allele).

In Figure 3, we plot two individuals in a space defined by two loci. Individual A is heterozygous for both Locus 1 and Locus 2 and individual B is homozygous for the alternate allele at each of the two loci.

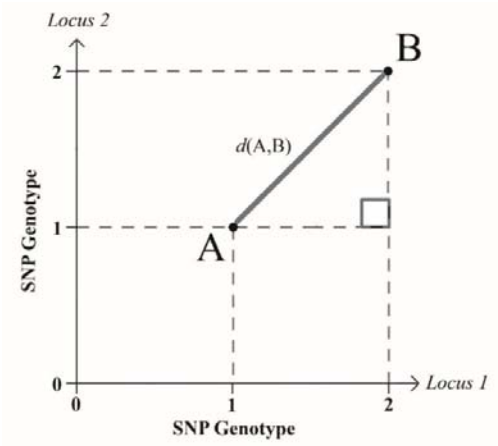


Figure 3. Two individuals, A and B, plotted in a space defined by their genotypes at Locus 1 and Locus 2. The values taken by an individual at a locus are 0, homozygous reference allele; 1, heterozygous; 2, homozygous alternate allele. In effect, each axis represents the frequency of the alternate allele at the corresponding locus.

Euclidean distance is defined by:

$$d(A,B) = \sqrt{\sum_{i=1}^{L}(x_i - y_i)^2}$$

where x_i and y_i are the counts of the alternate allele at locus i for individual A and B respectively, and L is the number of loci. In this case of two loci

$$d(A,B) = \sqrt{(2-1)^2 + (2-1)^2} = \sqrt{2}$$

Note that the contribution of a locus to the distance between A and B is invariant under the choice of which allele is assigned reference and which is assigned alternate. To demonstrate this, reassign the reference allele to alternate and the alternate allele to reference at a locus, that is, let

$$x' = 2 - x$$

$$y' = 2 - y$$

then

$$(x'_i - y'_i)^2 = [(2 - x_i) - (2 - y_i)]^2 = (x_i - y_i)^2$$

so the Euclidean Distance is invariant under our choice of which allele is to be considered reference and which alternate. This result is really important because we cannot have the value of our genetic distance depending on an arbitrary choice of which SNP allele is the reference allele and which is the alternate allele.

An Example of Euclidean Genetic Distance

Consider data for two individuals scored for six loci. The score 0 represents homozygous reference, 2 represents homozygous alternate, and 1 represents the heterozygous state.

	Loc01	Loc02	Loc03	Loc04	Loc05	Loc06
IndA	0	0	2	2	2	1
IndB	0	1	0	1	2	1

Applying the above formula for *d(A,B)* yields

$$d(A,B)^2 = 0 + 1 + 4 + 1 + 0 + 0 = 6$$

$$d(A,B) = \sqrt{6} = 2.45$$

which you can easily confirm for yourself.

Principal Components Analysis (PCA)

Rationale

Principal Components Analysis (PCA)[2–4] is a method of visualizing structure in a multivariable dataset, that is, where the entities have attributes that exceed in

number the 2 or 3 that can be plotted. At first glance, PCA does not appear to be a distance analysis, but hopefully you will appreciate the connection when reading the section on a related technique, Principal Coordinates Analysis (PCoA).

		ATTRIBUTES										
		Locus1	Locus2	Locus3	Locus4	Locus5	Locus6	Locus7	Locus8	Locus9	Locus10	Locus11
ENTITIES	Ind01	0	0	1	0	0	0	1	0	0	0	1
	Ind02	0	0	2	1	0	0	2	1	0	0	2
	Ind03	0	0	2	0	0	0	2	0	0	0	2
	Ind04	0	0	2	1	0	0	2	1	0	0	2
	Ind05	1	0	2	2	NA	0	2	2	0	0	2
	Ind06	0	0	0	0	1	0	0	0	0	0	0
	Ind07	0	1	0	0	0	1	0	0	0	1	0
	Ind08	0	0	0	0	0	0	0	0	0	0	0
	Ind09	0	0	0	0	0	0	0	0	0	0	0
	Ind10	NA	0	0	0	0	0	0	0	0	0	0

Table 1. A SNP genotype dataset comprises the entities (individuals) scored across the attributes (loci) where the data within the individual by locus matrix are the states (genotypes: 0 for homozygous reference allele, 1 for heterozygous, 2 for homozygous reference allele and NA for missed calls).

SNP genotype datasets can be organised into a matrix of Individuals (rows) by Loci (columns) (Table 1). Principal Components Analysis begins with the entities (individuals or samples or specimens) each represented as a point in a multivariable space defined by the L independent loci (see Figure 3). It then remaps those points, maintaining their relative positions, in a new ordered set of orthogonal axes derived as linear transformation of the original axes. They are ordered in the amount of information they contain, so that the first few axes tend to contain information on any structure in the data (signal) and later axes tend to contain only noise[6].

We might, for example, represent N individuals in a space defined by distances calculated from their SNP scores for L loci, that is, as a cloud of points in an L-dimensional space (Figure 4a). With the variation among populations spread across L axes, say 10,000 axes, it is impossible to peruse the data to identify any genetic structure.

Instead, the basis defined by the original L coordinate axes can be centred then rotated to form a new set of L coordinate axes without changing the relative proximity of the depicted points. In Principal Coordinates Analysis, the L axes are rotated such that the first new axis is in the direction of maximal variation in the data; the second new axis, orthogonal to the first, lies in the direction of maximal remaining variation; the third new axis, orthogonal to the first two, lies in the direction of maximal remaining variation, and so on (Figure 4b). The new, L ordered axes are then adopted as the basis for our new reference system (Figure 4c).

The final solution is a multivariate space defined by a basis with distinct ordered dimensions equal to the number of entities minus 1 or the number of attributes, whichever is the lesser, after removing redundancy.

The mathematics of PCA draw from either the covariance matrix or the correlation matrix (standardized covariance matrix). Because SNP data are all

measured on the same scale, the covariance matrix is preferred as standardization is not necessary.

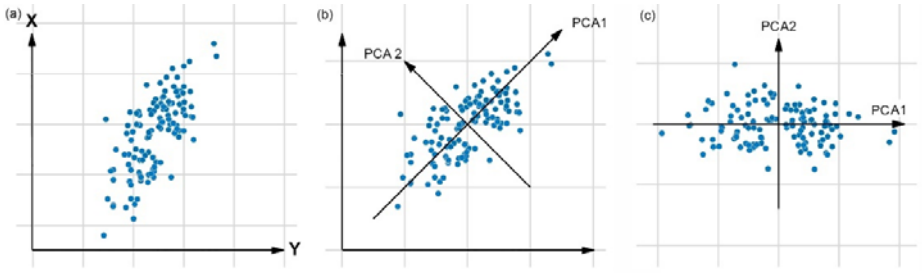


Figure 4. The process of reduced space representation used in Principal Components Analysis illustrated in two dimensions. (a) The entities are represented in Euclidean space defined by Cartesian coordinates X and Y; (b) New axes are selected as linear combinations of the original axes after centering and optionally standardizing the data, with the first (PCA1) in the direction of maximal variation in the data, the second (PCA2) orthogonal to the first in the direction of maximal remaining variation; (c) The new axes are used as a new basis for the spatial representation of the entities.

Using PCA, we can take a SNP data matrix (genotypes or presence-absence data), represent the entities (individuals or samples or specimens) in a space defined by the L loci, represent that space in new orthogonal axes ordered on the contribution of variance in their direction, and examine important patterns of variation among the entities in a relatively few dimensions, preferably 2 or 3. This is a very powerful visual technique.

The analysis yields the following information.

Eigenvalues	eigenvalues give the component of variation in the direction of each dimension of the reduced space representation.
Scores	scores are the coefficients of the linear relationships that define the new PCA axes and are used to plot the entities in the ordinated space.
Loadings	loadings are the correlations between the original variables (the SNP loci) and the principal components. They give an indication of which loci are contributing to variation along particular PCA axes.

Consider the reduced space representation of individuals sampled from across a wide geographic range (Figure 5). Here we have a projection of the multivariable data in only two dimensions, capturing 42.8% of total variation among entities, to reveal the essential structure present.

Note that were the data to have been drawn from a panmictic population (arguably the null proposition), each of the original variables would, on average, be expected to capture the same quantity of variance, and the PCA would fail (the first two axes would each represent only a small percentage of the total variance). The visualization of Figure 5 is informative because the different populations differ in genetic composition. There is structure to uncover.

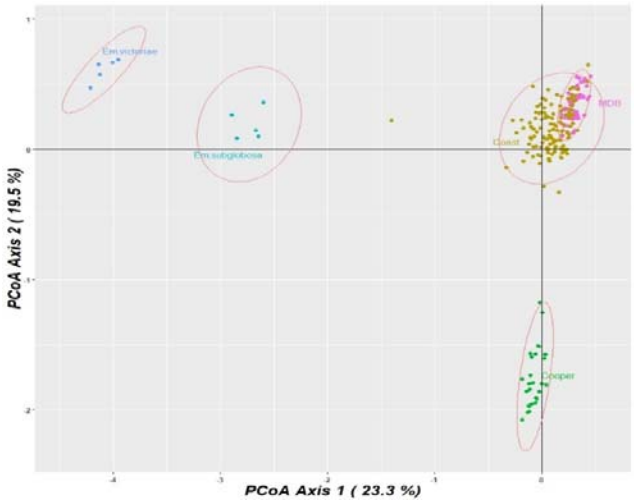


Figure 5. A PCA plot of individuals of turtle species in the genus *Emydura* sampled from across the Australian continent. A total of 42.8% of variation is captured in the first two of the ordered axes. Substantial structure across the landscape is clearly evident.

How many dimensions to retain?

The question now arises as to how many dimensions are required to fully capture the structure. A PCA plot can be misleading if one arbitrarily chooses, for convenience, only two dimensions in which to visualize the solution. Separation in the 2-D plot can be accepted as real, but proximity cannot. For example, in Figure 5, the individuals drawn from the Murray-Darling Basin (MDB) and the coastal rivers of Queensland and NSW (Coastal) are in close proximity in the 2-D plot. But were we to consider a third dimension (% variation explained < 19.5% but nevertheless possibly substantial), the MDB and Coastal individuals might be well separated. Interpreting proximity in a PCA requires careful interpretation.

There are a few techniques that can assist in the decision on how many dimensions to include in the final PCA solution and how to sensibly interpret the structure if more than three such dimensions are indicated.

The most commonly used approach is a scree plot[7], where the eigenvalues (variance explained) associated with each of the new ordered dimensions are plotted (Figure 6).

It is usual only to consider dimensions that have more than the average eigenvalue (the Kaiser-Guttman criterion[8]) because anything less than the average explains only that explained by each of the original variables, on average. Even so, this may leave many dimensions still in consideration (14 in Figure 6). A rule of thumb is to only consider axes that explain more than 10% of total variation, or one can look for a sudden drop in the percentage variation explained as a threshold.

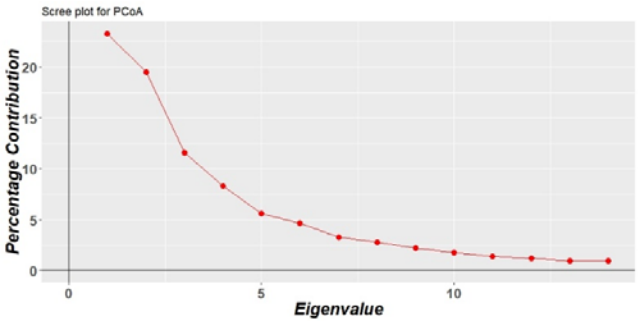


Figure 6. A Scree Plot[7] of the percentage contribution to total variance of each axis in the ordinated space. Eigenvalues give the component of variation in the direction of each dimension of the reduced space representation. Because the dimensions are ordered, the first eigenvalue is the largest, followed by the second, the third and so on for N-1 dimensions. In this case, at least three dimensions (> 10%) are indicated.

Common sense can also prevail. If you have three primary clusters in your reduced space representation, arising from aggregations of individuals that are diverging from other such aggregations independently through drift, then two dimensions will be adequate. Any three points (centroids) can be represented in 2-D space. If you have four primary clusters, diverging from each other independently through drift, then three dimensions will be required – required because, having captured the variation between the three major clusters in two dimensions, it is highly unlikely that the fourth cluster will diverge from the other three in the direction of their common plane.

The number of dimensions can be decided by considering both the scree plot and the gross structure among entities in the PCA. However, to make the decision to use only 2 or 3 dimensions as the final solution, simply because it is easy to present as a graph, is not appropriate.

Visualizing structure in 4 or more dimensions

Final solutions in 2D can be reported in publications, but what do you do if 3 or more dimensions are indicated?

There are three approaches to examining structure in more dimensions than two. The first is to construct a 3D plot and rotate it to best display the distinction between clusters. That approach is appropriate if the scree plot indicates that three dimensions are sufficient to represent the structure in the dataset.

Separation between two clusters of entities in two dimensions can be accepted as real, but adjacency might belie separation in deeper dimensions. In Figure 5, above the MDB and the Coast are overlaid, but this might not be sustained on examination of Axes 3 and 4. A second approach is to plot deeper dimensions taken two at a time, say by plotting PCA1 against PCA3 for example. This would show the true relationship between clusters MDB and Coastal that are overlaid in

the 2D solution of Figure 5, for example. The presentation of multiple pairwise plots is common in reporting PCA results.

A third approach is to run a separate PCA on each distinct cluster to examine structure within[9]. This is like treating the deeper dimensions as 2D or 3D 'bubbles' that emanate from a single point (the cluster centroid) in the initial 2D space. An example might be to run a separate PCA on the individuals in the MDB and Coastal cluster of Figure 13.

Impact of missing values

With SNP genotypes, missing data can arise because the read depth is insufficient to yield full coverage of the sequence tags in the genome, or because of mutations at one or both of the restriction enzyme recognition sites in some individuals. In SNP presence-absence data, missing values arise because the read depth is insufficient to be definitive about the absence of a particular sequence tag. SNP datasets typically have substantial numbers of missing values.

Classical PCA will not accept missing values, so if a locus is not scored for a particular individual, either the locus must be deleted or the individual must be deleted. This is clearly very wasteful of information.

The data loss can be managed to an extent by pre-filtering on call rate by locus (say requiring a call rate > 95%) or by individual (say requiring a call rate > 80%), but the remaining missing data need to be accommodated. Numerous ways for handling missing data in PCA have been suggested[10], but the most common method is to replace a missing value with the mean of the allele frequencies for the affected individual (mean-imputed missing data).

Mean-imputed missing data can lead to the individuals (or samples or specimens) with a high proportion of missing data to be drawn out of their natural grouping and toward the origin, which can lead to potential misinterpretation[13]. For example, if individuals in the PCA aggregate into natural clusters, perhaps representing geographic isolates, and these clusters are on either side of the origin, then an individual with a high frequency of missing values will be drawn out of its cluster toward the origin. Its location intermediate to the two clusters might be falsely interpreted as a case of admixture. substantial individuals with missing data corrected by mean-imputation will also distort confidence envelopes that may be applied to clusters.

Perhaps a better way of handling missing values in SNP datasets used for population genetics is to implement a local imputation. Missing values for an individual are replaced with the mean allele frequency for the population from which the individual was drawn. In this way, the individual is displaced toward the centroid of the population from which it was drawn, not the origin of the PCA. Alternatively, a random genotype can be applied to the missing value at a locus by drawing allele frequencies at random from the frequency distribution for the population from which the individual is drawn. The individual is displaced in a

random direction within the bounds of the cluster representing the population from which it was drawn.

Strictly, both of these methods of local imputation should be applied to panmictic populations, which means the group of individuals sampled from the same locality. In practice, it is unlikely to matter too much so long as the imputation is restricted to the aggregation that is appearing as distinct in the PCA.

An alternative approach is to apply pairwise deletion of loci rather than the global deletion dictated by classical PCA[14]. We do this by calculating a matrix of Euclidean distances for individuals taken pairwise, whereby loci with a missing value for one or both individuals are removed. Provided the Euclidean distance is scaled for the number of loci in the pair (as introduced later), a Principal Coordinates Analysis (PCoA, see next section) can be applied to the distance matrix to deliver the ordination. This approach capitalizes on the observation that PCA and PCoA using Euclidean distance yield the same visualizations[16:43-4]. There are cryptic implications of this approach, not least of which is the disruption of the metric and/or Euclidean properties of the distance matrix, so the resultant eigenvalues should be examined for negative values. Negative eigenvalues are unlikely unless the frequency of missing values is extreme.

Implementation of PCA in dartR

When it receives a genlight object, the script gl.pcoa() in package dartR uses the glPCA() function of {adegenet}[11,12] to undertake a PCA with parameters set as the defaults of center=TRUE, scale=FALSE, alleleAsUnit=FALSE. A PCA can be undertaken on either SNP genotype data or SNP presence-absence data.

Package glPCA handles missing values by substituting them with the mean allele frequency for the associated entity (mean-imputed missing data). As outlined above, entities (individuals/samples/specimens) with a high proportion of missing data will thus be drawn out of their natural grouping and toward the origin, which can lead to potential misinterpretation[13]. The appropriate strategy to avoid this is to

(a) Filter stringently on call rate, using a threshold of at least 95% loci called.

(b) Remove individuals for which call rate is exceptionally low, say <80%.

(c) Impute the remaining missing values on a population-by-population basis, where populations can be considered panmictic.

An example of a script to implement these steps follows:

```
gl <- gl.filter.callrate(gl, method="loc", threshold=0.95)
gl <- gl.filter.callrate(gl, method="ind", threshold=0.80)
gl <- gl.impute(gl, method=random)
pca <- gl.pcoa(gl)
```

To undertake pairwise deletion of missing values, a script along the following lines is appropriate.

```
gl <- gl.filter.callrate(gl, method="loc", threshold=0.95)
```

```
gl <- gl.filter.callrate(gl, method="ind", threshold=0.80)
d <- gl.dist.ind(gl, method="euclidean", scale=TRUE)
pca <- gl.pcoa(d)
```

Principal Co-ordinates Analysis (PCoA)

Rationale

Principal Co-ordinates Analysis[5] is a visualization technique that represents a distance matrix in a Euclidean space defined by an ordered set of orthogonal axes, as does PCA. They are ordered in the amount of information they contain, so that the first few axes tend to contain information on any structure in the data (signal) and later axes tend to contain only noise[6].

The primary difference between PCA and PCoA is that PCA works with the original data and the covariance (or correlation) matrix derived from the original data whereas PCoA works with a distance matrix and does not directly draw upon the original data that was used to generate that matrix. Indeed there are some circumstances where the distances are measured directly, such as immunological distance[15]. Because the mathematics of PCA moves forward from the covariance (or correlation) matrix, the insight attributed to John Gower[5] was to substitute, at this point in the analysis, any distance matrix following a simple transformation. This yields an ordered representation of those distances in multivariable space. akin to PCA

Intuitively, we represent N individuals in a space defined by the distances between them, that is, as a cloud of points in an (N-1)-dimensional space (Figure 7a). It is (N-1) because any two points can be represented with a line, that is in 1-D space, any three points define a plane, that is a 2-D space, etc.

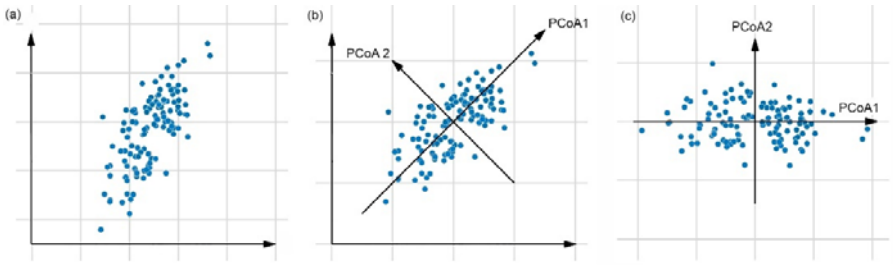


Figure 7. The process of reduced space representation used in Principal Coordinates Analysis illustrated in two dimensions. (a) The distance matrix is represented in Euclidean space defined by Cartesian coordinates X and Y; (b) New axes are selected as linear combinations of the original axes after centering and standardizing the data, with the first (PCoA1) in the direction of maximal variation in the data, the second (PCoA2) orthogonal to the first in the direction of maximal remaining variation; (c) The new axes are used as the new basis for the spatial representation of the distances. Any distance measure can be used, but a metric distance is preferred.

Note the important distinction between Figure 4(a) where the N entities are represented in a space of L dimensions defined by the loci and Figure 7(a) where the entities are represented in an N-1 space with coordinate axes not directly

connected to any raw data based solely on their distances. Apart from that, the mathematics of PCA and PCoA are very similar.

As with PCA, the basis defined by the original N-1 coordinate axes can be rotated to form a new set of N-1 coordinate axes without changing the relative proximity of the depicted points (Figure 7b). The N-1 axes are rotated such that the first new axis is in the direction of maximal variation in the data; the second new axis, orthogonal to the first, lies in the direction of maximal remaining variation; the third new axis, orthogonal to the first two, lies in the direction of maximal remaining variation, and so on. The new, N-1 ordered axes are then adopted as the basis for our new reference system as shown in Figure 7(c).

Because the new orthogonal axes ordered on the contribution of variance in their direction, we can examine important patterns of variation in a relatively few dimensions, preferably 2 or 3. This too is a very powerful visual technique, extended to apply to any well-behaved distance matrix.

The analysis yields the following information.

Eigenvalues	eigenvalues give the component of variation in the direction of each dimension of the reduced space representation, which can be expressed as a percentage of the sum of the eigenvalues.
Scores	scores are the coefficients of the linear relationships that define the new PCA axes and are used to plot the entities in the ordinated space.

PCoA does not retain a link to the data used to generate the distance matrix. It is possible to generate a Pearson correlation between the original data axes (the loci) and the new PCoA axes. These are called Loadings and provide an indication of which loci are providing the variation contributing to that represented by each retained PCoA axis.

Note that were the data to have been drawn from a panmictic population (arguably the null proposition), each of the original variables would, on average, be expected to capture the same quantity of variance, and the PCoA would fail (the first two axes would each represent only a small percentage of the total variance). The visualization is informative because the different populations differ in genetic composition. There is structure to uncover.

All of the same considerations on choice of dimension for the final reduced space that applied to PCoA apply to PCA. PCA is also equally intolerant of missing data, which need to be managed.

The result of a PCoA with an input matrix comprised of Euclidean Distances[16:43-44] is identical to a PCA. In this context, the interchangeability of the two, PCA and PCoA, leads to considerable confusion on the distinction between the two analyses.

Impact of missing values

PCoA is not as sensitive to missing values as PCA because it works from a distance matrix, and distance matrices are typically complete. As such, missing values are important insofar as they influence the distance matrix used for PCoA.

First of all, in the presence of missing data, the distance matrix may no longer be metric or Euclidean even if the distance measure used for the computations is theoretically metric or Euclidean. In extreme cases, this may lead to negative eigenvalues and distortion of the representation of entities in the ordinated space.

The second issue is that the distances calculated between entities in pairwise fashion will vary in sample size depending on the frequency of missing values for the particular pair, and so the distances in the matrix will vary in precision.

Neither of these issues are likely to be serious if there is adequate filtering of the raw data based on call rate, say to ensure that the call rate for loci is > 95% and for individuals is > 80%. Distance measures that take into account the number of loci in calculating the distance between pairs of individuals should be chosen if possible.

In a sense, PCoA is more robust to missing values than is PCA. This is in part because the potential impact of missing values on PCA is global, whereas the impact of missing values on PCoA can be constrained to pairwise comparisons.

Implementation in dartR

Distances of relevance to population genetics can be calculated for SNP genotype data and SNP presence-absence data using the dartR functions gl.dist.ind() and gl.dist.pop().

When it receives a matrix or distance object (class dist), the script gl.pcoa() in package dartR uses the pcoa() function of {ape}[17] with default parameters to undertake a PCoA.

A sample script for undertaking a PCoA on presence-absence data might be:

```
gs <- gl.filter.callrate(gs, method="loc", threshold=0.95)
gs <- gl.filter.callrate(gs, method="ind", threshold=0.80)
D <- gl.dist.ind(gs, method="jaccard", scale=TRUE, flip=TRUE)
pc <- gl.pcoa(D)
```

Non-metric Distances (Dissimilarity Measures)

The advent of PCoA admits the use of almost any metric distance or dissimilarity measure for generating reduced dimension representations of multivariable data. A great number of measures of genetic dissimilarity have been devised over time[1], many of which apply well to SNP data. Each distance has different properties that may make it more suitable for a particular purpose than does another measure.

While the metric properties of a distance are clearly important, many measures used in genetics are non-metric. An example of a dissimilarity measure that fails to satisfy the symmetry condition is one defined on private alleles. The number of private alleles possessed by population X when compared to population Y will typically be different from the number of private alleles possessed by population Y in comparison with X. The resultant 'distances' will not satisfy the second metric criterion of symmetry.

Many genetic distances do not satisfy the triangle inequality. For example, percent fixed differences satisfy the first two conditions of a metric distance, but not the triangle inequality and so is non-metric. Nor is Nei's D a metric distance, but the common alternative of Rogers' D is metric. F_{ST} is non-metric. The Bray-Curtis dissimilarity measure is non-metric but is rank-order similar to the Jaccard distance, which is metric. And so on.

In the following sections, distance measures of relevance to SNP data, both genotypes and presence-absence data, are introduced.

Non-Metric vs Metric vs Euclidean Distances

Strictly speaking, Euclidean Distance is required for entities to be faithfully represented in a space defined by Cartesian coordinates. Only then are we guaranteed to have all of our entities (individuals or populations) accurately positioned in the space of $(N\text{-}1)$ dimensions. Euclidean distances are a special class of metric distance, because they allow unambiguous representation of our entities (=individuals, specimens or samples) in a space defined by the familiar Cartesian coordinates, *without any distortion*. The distances in the original Distance Matrix and the distances in the reduced space will agree.

Entities defined by their relationship as a Metric Distance can be uniquely and accurately represented in a non-linear space regardless of the metric distance measure chosen, but their representation in a space defined by Cartesian coordinates typically comes at the expense of some distortion. This is analogous to representing points on a restricted area on the surface of the earth in a 2D map – the challenge of cartographers. In most cases, an adequate representation is possible in Euclidean space. The level of distortion is acceptable.

Entities defined by their relationship as a non-metric dissimilarity measure are potentially problematic because they cannot necessarily be uniquely represented in a space at all. Attempting to represent these entities (individuals or populations) in a space defined by Cartesian coordinates can lead to considerable distortion, which may mislead interpretation.

To illustrate the above points, consider the case presented in Figure 8. Here we have distances between four points, X, Y, Z and V (Figure 8A). Taken three at a time, the distances between these points satisfy the metric conditions, yet it is still not possible to represent all four in a Cartesian space. In Figure 8A, point V is

not defined. So, although it is possible to represent any 4 points in a 3D space, with distances between them meeting the metric conditions, not all sets of distances between 4 points can be represented in 3D space even if the metric properties hold. They can be represented in 3D space if we admit a level of curvature (Figure 8B), but not in a flat space represented by Cartesian coordinates.

How much weight you place on choosing a non-metric dissimilarity measure, a metric distance or a Euclidean distance depends on whether or not other considerations dominate. For example, you may place value on some underlying model of evolutionary divergence. Nei's D, for example, is roughly linear with time of divergence, assuming drift-mutation equilibrium. This may outweigh considerations of its non-metricity in practice.

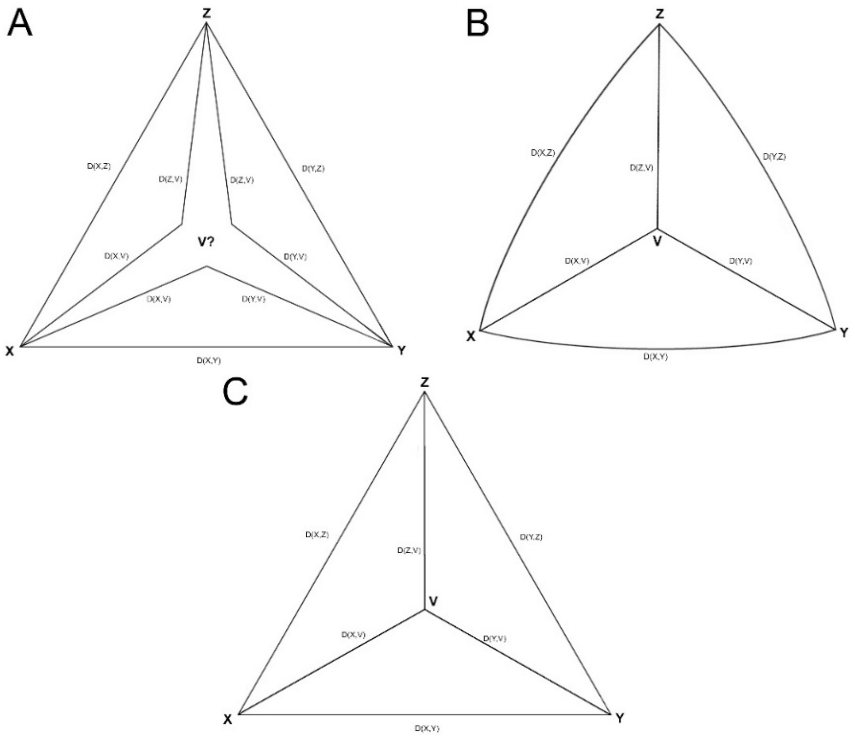


Figure 8. Metricity is not sufficient to represent distances in a rigid space defined by Cartesian coordinates. **A**, distances between four individuals satisfy the metric properties can nevertheless not be represented in three dimensions. **B**, this distortion can be resolved by allowing non-linear links to represent distances between individuals. **C**, Euclidean distances between four individuals can be represented without distortion. [after Gower[18]]

The distortion arising from using non-Euclidean distances manifests as displacement of the points in the visualization, so that the distances among them no longer fundamentally represent the input values, and as negative eigenvalues (imaginary eigenvectors)[19]. However, the level of distortion is only likely to be of concern if the absolute magnitude of the largest negative eigenvalue is less than that of any of the dimensions chosen for the reduced representation[20]. So in

practice, a few small negative eigenvalues do not detract much if only a few dimensions are retained in the final solution[21].

Because the distortion arising from using non-Euclidean distances also manifests as negative eigenvalues, interpretation of the variance contributions is challenging. In particular, one can no longer calculate the percentage variation explained by a PCoA axis by expressing the value of its eigenvalue over the total sum of the eigenvalues. A correction is necessary[22:506].

$$\%\ explained\ = \frac{e_i + k}{\sum_{i=1}^{N} e_i + (N-1)k}$$

where e_i is the eigenvalue for PCoA axis i, N is the number of entities, and k is the absolute magnitude of the largest negative eigenvalue.

If negative eigenvalues are considered problematic for the reduced space representation, a transformation can render them all positive and the distance matrix Euclidean. Common transformations put to this purpose are:

Square root[22:501] $\mathrm{D}(\mathrm{A},\mathrm{B}) = \sqrt{\mathrm{D}(\mathrm{A},\mathrm{B})}$

Cailliez[19,23] $\mathrm{D}(\mathrm{A},\mathrm{B}) = \mathrm{D}(\mathrm{A},\mathrm{B}) + \mathrm{c}$ for all $A \neq B$

Lingoes[19,24] $\mathrm{D}(\mathrm{A},\mathrm{B}) = \sqrt{\mathrm{D}(\mathrm{A},\mathrm{B}) + \mathrm{c}}$ for all $A \neq B$

The value of c is chosen to be the smallest value required to convert the most extreme negative eigenvalue to zero.

A final point to note is that it is a departure from theory requires addressing in practice only if it causes serious issues. In particular, distances do not need to be measured using a metric that is metric Euclidean in theory for the distance matrix to be metric or Euclidean. It is important to look at the diagnostics to assess if there is a problem of sufficient magnitude to require remedial action.

Genetic Distances for Individuals

Let's apply these ideas now to formulate distances between individuals.

Binary Data

Where the data are in the form of presences [1] and absences [0], the data are said to be binary. Such is the case where we are scoring SNP loci as "called" or "not called". They are called because the two restriction enzymes find their mark (in DArTSeq or ddRAD), the corresponding sequence tags are amplified and sequenced, and the SNP is scored. The individual is thus scored as 1 for that locus.

If, however, there is a mutation at one or both of the restriction enzyme sites, then the restriction enzyme does not find its mark, the corresponding sequence tag in that individual is not amplified or sequenced, and the SNP is called as missing for that individual. The individual is scored as 0 for that locus.

We might have, for example,

	Loc01	Loc02	Loc03	Loc04	Loc05	Loc06	Loc07
IndA	0	0	1	1	0	1	0
IndB	0	1	0	1	0	1	0

We can count up the different cases,

$$N_{00} = \sum_{i=1}^{L} \begin{matrix} 1, where\ x_i = y_i = 0 \\ 0, where\ x_i \neq y_i \end{matrix}$$

that is, sum loci scored 0 (absent) for both individuals;

$$N_{10} = \sum_{i=1}^{L} \begin{matrix} 1, where\ x_i = 1 \\ 0, where\ y_i = 0 \end{matrix}$$

that is, sum loci scored 1 (present) for Individual A and 0 (absent) for Individual B;

$$N_{01} = \sum_{i=1}^{L} \begin{matrix} 1, where\ x_i = 0 \\ 0, where\ y_i = 1 \end{matrix}$$

that is, sum loci scored 0 (absent) for Individual A and 1 (present) for Individual B;

$$N_{11} = \sum_{i=1}^{L} \begin{matrix} 1, where\ x_i = y_i = 1 \\ 0, where\ x_i \neq y_i \end{matrix}$$

that is, sum loci scored 1 (present) for both individuals. These summations do not include loci for which data are missing (NA) for one or both individuals.

The number of loci L is given by

$$L = N_{00} + N_{01} + N_{10} + N_{11}$$

There are several ways to calculate a binary dissimilarity between two individuals[25], some of which follow:

Euclidean Distance

$$d_E = \sqrt{N_{01} + N_{10}}$$

Properties: Metric. Range [0,$\sqrt{L}$). Symmetric in choice of coding for reference and alternate allele.

Scaled Euclidean Distance

The number of mismatches $N_{01} + N_{10}$ achieves its maximum of L when there is a mismatch at all loci, so the Euclidean Distance can be conveniently scaled to fall within the range of [0,1] as

$$d_E = \sqrt{\frac{N_{01} + N_{10}}{L}}$$

Properties: Metric. Range [0,1]. Symmetric in choice of coding for reference and alternate allele.

Simple Matching Distance[26]

$$d_{SM} = \frac{N_{01} + N_{10}}{L}$$

which is the sum of the mismatches across loci over the total number of non-missing loci.

Properties: Non-metric. Range [0,1]. Symmetric in choice of coding for reference and alternate allele. Accommodates missing data (NA) in part by expressing mismatches as a proportion of the total number of loci considered (i.e. scaled to fall between 0 and 1).

Notes: This simple matching distance is used when there is symmetry (equivalence) in the information carried by 0 (absence) and 1 (presence).

Jaccard Distance[27]

$$d_J = \frac{(N_{01} + N_{10})}{L - N_{00}}$$

which is the sum of the mismatches over the total number of non-missing loci for which at least one of the individuals scores a 1 (presence).

Properties: Metric distance[28]. Range [0,1]. Accommodates missing data (NA) in part by expressing mismatches as a proportion of the total number of loci considered (i.e. scaled to fall between 0 and 1). Not symmetric in choice of coding for reference and alternate allele. Requires consideration.

Notes: The Jaccard Distance down-weights the joint absences, which is arguably what you do not want for data comprised of counts of sequence tag absences arising from a positive event, that of a mutation at one (or both) of the restriction enzyme sites.

If you wish to use the Jaccard Distance on DArT or ddRAD presence/absence data, you might consider recoding the data so that 1 represents presence of a mutation at one or both of the restriction enzyme sites and 0 represents absence of such a disruptive mutation. This application of the Jaccard Distance will down-weight joint absence of a disruptive mutation leading to loss of the particular sequence tag.

Alternative names: Marczewski-Steinhaus Distance[30]; Ružička Distance; Soergel Distance

Bray-Curtis Distance[31]

$$d_{BC} = \frac{(N_{01} + N_{10})}{L - N_{00} + N_{11}}$$

Properties: Not a metric distance[29:61], so it is strictly a dissimilarity measure, but it is rank consistent with the Jaccard metric. Range [0,1]. Accommodates missing data (NA) in part by expressing mismatches as a proportion of the total number of loci considered. Not symmetric in choice of coding for reference and alternate allele. Requires consideration.

Notes: Bray-Curtis Distance adjusts the denominator to down-weight the joint absences (0,0) and up-weight joint presences (1,1). As with the Jaccard Distance, you might consider reversing the scores for absence (0) and presence (1) to 1 and 0 respectively when dealing with DArT or ddRAD data.

Alternative names: Sørensen Distance[32], Dice Distance[33]; also equivalent to the Nei & Li Distance[34] as applied to two individuals.

Impact of Missing Values

The package dartR uses the above algorithms with deletion of loci scored as missing for one or both of the individuals in a pair. This pairwise deletion is accommodated in distance measures that are corrected for the number of loci examined (L in the denominator). Adjustment of the denominator in the Jaccard and Bray-Curtis distances can lead to a potential systematic bias arising from missing values[29:62]. Unscaled Euclidean metric is severely affected by missing values, and should be used only for complete data.

Implementation in dartR

The above algorithms are implemented in dartR with the script gl.dist.ind() as it applies to binary presence-absence data (e.g. SilicoDArT).

SNP Data

Unlike binary data, SNP data take on three values at a locus

0, homozygous reference allele
1, heterozygous
2, homozygous alternate allele

This is a scoring scheme that is convenient because the value represents the frequency of alternate allele, and so can be considered to be measured at least at the ordinal level.

With SNPs scored in this way, a property of a distance measure, in addition to the desirable metric properties, is

> *Shared homozygous reference alleles (0) and shared homozygous alternate alleles (2) need to contribute equally to the distance measure.*

from which it follows that the distance measure needs to be invariant under choice of which allele is considered the reference allele and which is considered the alternate allele (DArT typically chooses the most frequent allele as the reference allele).

This additional condition eliminates from consideration a number of potential metric distances (e.g. Jaccard Distance) and non-metric dissimilarity measures (e.g. Bray-Curtis Distance) that are appropriate for binary presence/absence data[35].

Euclidean Distance

Euclidean distance is defined by:

$$d(A,B) = \sqrt{\sum_{i=1}^{L}(x_i - y_i)^2}$$

where x_i and y_i are the counts of the alternate allele at locus i for individual A and B respectively, and L is the number of loci for which both x_i and y_i are non-missing.

Properties: Metric distance. Range [0,$4L$). Symmetric in choice of coding for reference and alternate allele.

Scaled Euclidean Distance

The maximum squared distance between two individuals at a locus is $(2\text{-}0)^2 = 4$, so the sum of squared distances over L loci has a maximum of $4L$. The Euclidean Distance applied to SNPs can thus be rescaled to fall within the range of [0,1] as follows:

$$d_{Euclidean}(A,B) = \frac{1}{2}\sqrt{\sum_{i=1}^{L}\frac{(x_i - y_i)^2}{L}}$$

where x_i and y_i are the counts of the alternate allele at locus i for individual A and B respectively; L is the number of loci for which both x_i and y_i are non-missing.

Properties: Metric distance. Range [0,1]. Symmetric in choice of coding for reference and alternate allele.

Simple Mismatch Distance

The count of shared alleles between two individuals i and j at a locus is given by

$$\begin{aligned} c_{i,j} &= 0, where\ no\ alleles\ are\ shared\ [0,2]||[2,0] \\ &= 1, where\ one\ allele\ is\ shared\ [0,1]||[1,0]||[2,1]||[1,2] \\ &= 2, where\ both\ alleles\ are\ shared\ [0,0], [1,1], [2,2] \end{aligned}$$

The Simple Mismatch Distance is given by

$$d_{SM}(A,B) = 1 - \frac{1}{2L}\sum_{i=1}^{L} c_{ij}$$

where L is the number of loci non-missing for both individuals i and j.

Properties: Non-metric. Ranges [0,1]. Symmetric in choice of coding for reference and alternate allele.

Alternative names: Similar to the Allele Sharing Distance[36], differing from it by a factor of 2.

Absolute Mismatch Distance

The count of shared alleles between two individuals i and j at a locus is given by

$$c_{i,j} = 0, no\ alleles\ are\ shared\ [0,2]$$
$$= 1, one\ or\ both\ alleles\ are\ shared\ [0,0]|[0,1]|[1,2]|[2,2]$$

The Absolute Mismatch Distance is given by

$$d_{AM}(A,B) = 1 - \frac{1}{L}\sum_{i=1}^{L} c_{ij}$$

where L is the number of loci non-missing for both individuals i and j.

Properties: Non-metric. Ranges [0,1]. Symmetric in choice of coding for reference and alternate allele.

Czekanowski (Manhattan) Distance

Often referred to as the City Block Distance, this metric is calculated by summing the scores on each of the axes representing the loci.

$$d = \sum_{i=1}^{L} |x_i - y_i|$$

where x_i and y_i are the counts of the alternate allele at locus i for individual A and B respectively; L is the number of loci for which both x_i and y_i are non-missing.

Clearly, $|x_i - y_i| = 0$ when $x_i = y_i$ and achieves a maximum of 2 when $x_i = 0$ and $y_i = 2$ or vice versa. So we can scale this value to range between 0 and 1 by dividing by $2L$.

$$d_{Cz}(A,B) = \frac{1}{2L}\sum_{i=1}^{L} |x_i - y_i|$$

Properties: Metric. Ranges [0,1]. Symmetric in choice of coding for reference and alternate allele.

Impact of Missing Values

The package dartR uses the above algorithms with deletion of loci scored as missing for one or both of the individuals in a pair. This pairwise deletion is accommodated in all but one of distance measures by correcting for the number of loci examined (L in the denominator). The unscaled Euclidean metric is severely affected by missing values, and should be used only for complete data.

Implementation in dartR

The above algorithms are implemented in dartR with the script gl.dist.ind() as it applies to SNP genotypes (e.g. DArTseq).

Genomic Relationship

A pedigree provides knowledge of the genealogical relationships among individuals. Progeny receive a random half of each parents' genes and full-sibs are expected to share half their genes, on average. Genealogical relationship and genetic similarity are loosely connected, in that there can be considerable departure from the 0.5 expectation for shared alleles across independent loci in practice, because of sampling variation. Indeed, because genes are parcelled on to chromosomes, a parent of a species with a small chromosome number can expect to generate descendants with no direct inheritance of its genes after even relatively few generations. They are related to their descendants by descent genealogically, but can be unrelated genetically. It is for this reason that measures of genetic relatedness are considered more informative than knowledge of a pedigree alone.

This field has expanded rapidly because of its relevance to animal and plant breeding. The techniques are beyond the scope of this technical note, but the reader is directed to the work of VanRaden[37] and those who cite his work. A genetic relatedness matrix or G-matrix can be generated from a genlight object in dartR with gl.grm(), a wrapper for the A.mat function of package {rrBLUP}[38]. Alternatively, you may wish to look at other R packages, such as {snpReady}[39].

Genetic Distances for Populations

Binary Data

When dealing with populations and binary data, we consider the relative frequency p of presences (1) at each locus. If 10% of individuals are scored as a presence at a particular locus i for population A, then $p_{Ai} = 0.1$.

$$p_{Ai} = \frac{1}{N}\sum_{k=1}^{N} x_{ik}$$

where x_{ik} is the score (0 or 1) for individual k at locus i and N is the number of individuals in population A. The relative frequency of the absences is given by

$$q_{Ai} = 1 - p_{Ai}$$

The relative frequencies of presences at each locus is the fodder of binary distance as applied to populations (Figure 9).

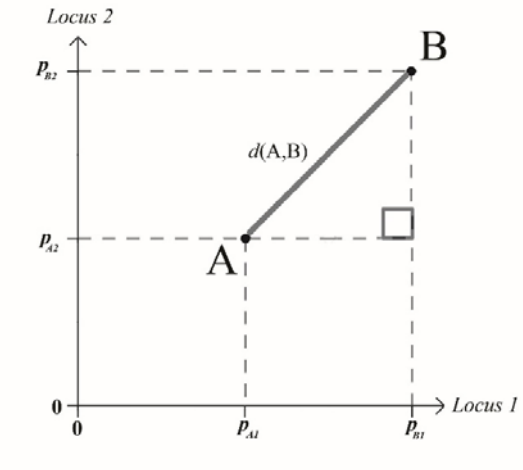


Figure 9. Two populations, A and B, plotted in a space defined by the proportions of Presences at Locus 1 and Locus 2. Euclidean distance between the two populations can be calculated from their Cartesian coordinates using Pythagoras' rule.

Euclidean Distance

The Euclidean Distance between population A and B in the space defined by Locus 1 and 2 (Figure 9) is given by

$$D(A,B) = \sqrt{(p_{B1} - p_{A1})^2 + (p_{B2} - p_{A2})^2}$$

which can be generalized for L loci as

$$D(A,B) = \sqrt{\sum_{i=1}^{L} (p_{Bi} - p_{Ai})^2}$$

and, noting that $0 < p < 1$, is scaled to the range $[0,1]$ as

$$D_{Euclidean}(A,B) = \sqrt{\frac{1}{L}\sum_{i=1}^{L} (p_{Bi} - p_{Ai})^2}$$

Properties: Metric. Range [0,1]. No underlying genetic model.

SNP Genotypes

Euclidean Distance

SNP data for populations comprise the proportion of alleles that are the alternate allele (Figure 10), so computationally the options are not that much different than for binary presence/absence data.

$$D_{Euclidean}(A,B) = \sqrt{\frac{1}{L}\sum_{i=1}^{L} (p_{Bi} - p_{Ai})^2}$$

where p_{Ai} is the proportion of the alternate allele for Locus i in population A, p_{Bi} is the proportion of the alternate allele for Locus i in population B and L is the number of called loci.

Noting that $0 < p < 2$, this can be scaled to the range $[0,1]$ as

$$D_{Euclidean}(A,B) = \frac{1}{2}\sqrt{\frac{1}{L}\sum_{i=1}^{L}(p_{Bi} - p_{Ai})^2}$$

Properties: Metric distance. Range [0,1]. No underlying genetic model.

Alternative names: Similar to Roger's D, differing from it by a constant factor.

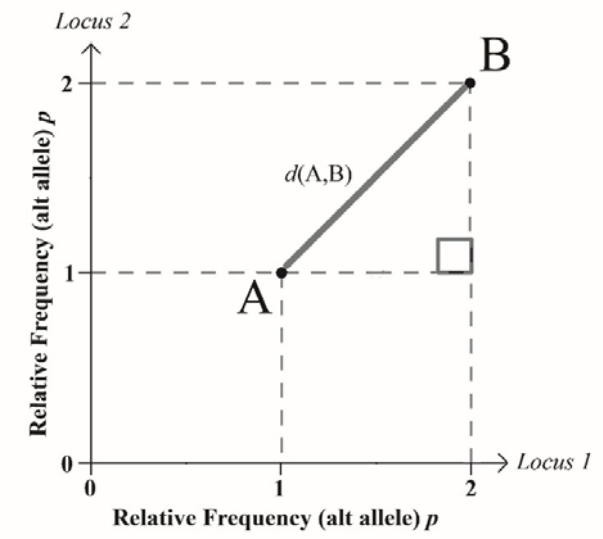


Figure 10. Two populations, A and B, plotted in a space defined by the relative frequency of the alternate allele at Locus 1 and Locus 2. Euclidean distance between the two populations can be calculated from their Cartesian coordinates using Pythagoras' rule

Nei's Standard Genetic Distance

Nei's standard genetic distance[40] is favoured by some because of its relationship to divergence time. When populations are in mutation-drift balance throughout the evolutionary process and all mutations result in new alleles in accordance with the infinite-allele model, Nei's *D* is expected to increase in proportion to the time after divergence between two populations.

$$D_{Nei}(A,B) = -ln\left(\frac{\sum_{i=1}^{L}([p_{Ai}p_{Bi} + q_{Ai}q_{Bi}]}{\sqrt{\sum_{i=1}^{L}[(p_{Ai}^2 + q_{Ai}^2]}\sqrt{\sum_{i=1}^{L}[(p_{Bi}^2 + q_{Bi}^2]}}\right)$$

Properties: Non-metric. Range [0,∞). The underlying genetic model incorporates both drift and mutation. Proportional to divergence time under specified assumptions[40].

Reynolds Genetic Distance

Reynolds genetic distance[41] is also approximately linearly related to divergence time in theory, but unlike Nei's Standard Genetic Distance, it is based solely on a drift model and does not incorporate mutations. As such, it may be more appropriate than Nei's distance for population genetics and in particular, representation of genetic similarity in trees or networks where branch lengths need to be interpretable.

$$D(A,B) = \sqrt{\frac{\sum_{i=1}^{L}[(p_{Ai}-p_{Bi})^2+(q_{Ai}-q_{Bi})^2]}{2\sum_{i=1}^{L}(1-p_{Ai}p_{Bi}-q_{Ai}q_{Bi})}}$$

A better approximation[41] of the linear relationship with time is given by

$$D_{Reynolds}(A,B) = -ln[1-D(A,B)]$$

which is used in dartR.

Properties: Non-metric. Range [0,1]. The underlying genetic model incorporates drift alone. Proportional to shallow divergence time under specified assumptions[41].

Chord Distance

Edward's Angular distance[42] assumes divergence between populations is via drift alone, and so again may be more appropriate than Nei's Distance for population genetics. It is calculated as

$$cos\,\alpha = \sqrt{p_{Ai}p_{Bi}} + \sqrt{q_{Ai}q_{Bi}}$$

where α is the Angular Distance. This can be approximated by the straight-line segment or Chord Distance[43] as follows:

$$D_{Chord}(A,B) = \sqrt{1-\frac{1}{L}\sum_{i=1}^{L}\sqrt{p_{Ai}p_{Bi}}+\sqrt{q_{Ai}q_{Bi}}}$$

Properties: Metric. Range [0,1]. It can be transformed to be approximately Euclidean[42]. Underlying genetic model incorporates drift alone. Proportional to shallow divergence time under specified assumptions[42].

Wright's F Statistics

Wrights F can be defined as

$$F = 1 - \frac{H_{obs}}{H_{exp}}$$

such that a deficit in the observed frequency of heterozygotes compared with that expected under Hardy-Weinberg equilibrium will yield a Wright's F less than one. If observed and expected heterozygosity are in agreement, then $F=0$. If, at the extreme, no heterozygotes are observed, then $F=1$. Wright's F can be interpreted as a measure of inbreeding.

When applied to multiple populations, for which Hardy-Weinberg equilibrium is not a sensible null expectation, Wright's F

is nevertheless informative, as any deficit in heterozygotes in the populations when pooled is an indication of structure among those populations.

In this context, Wright's F can be considered a [non-metric] distance when applied to populations in pairwise fashion.

We defer discussion of this distance to our treatment of spatial analysis and assessing structure across the landscape.

Impact of Missing Values

When dealing with population level data, random missing values are important only insofar as they reduce the sample size in calculations of the allele frequency distributions. This becomes important only when all individuals in a population are missing for a locus, in which case the information for that locus is discarded. Most analyses, including PCoA, that use a population-level distance matrix expect it to be complete.

Implementation in dartR

The above algorithms are implemented in dartR with the script gl.dist.pop() as it applies to SNP genotypes.

Distance of an Individual from a Population

Distance of an individual from a population is not as simple as it might seem. When variation among individuals within a population varies in the direction of particular axes (that is, the measurements represented by the axes are correlated), Euclidean Distance can be quite misleading. Consider the data presented in Figure 11. The two highlighted points A and B are equally distant from the population centroid, but clearly point A is an outlier whereas point B is quite within expectation for a point belonging to the population. So Euclidean Distance is not a good measure of the distance of an individual from a population.

The solution to this is to express the deviation of a point from the population centroid in units of standard deviation measured in the direction of the point from the centroid. You can see then how Point A of Figure 11 is much further away from the centroid than Point B.

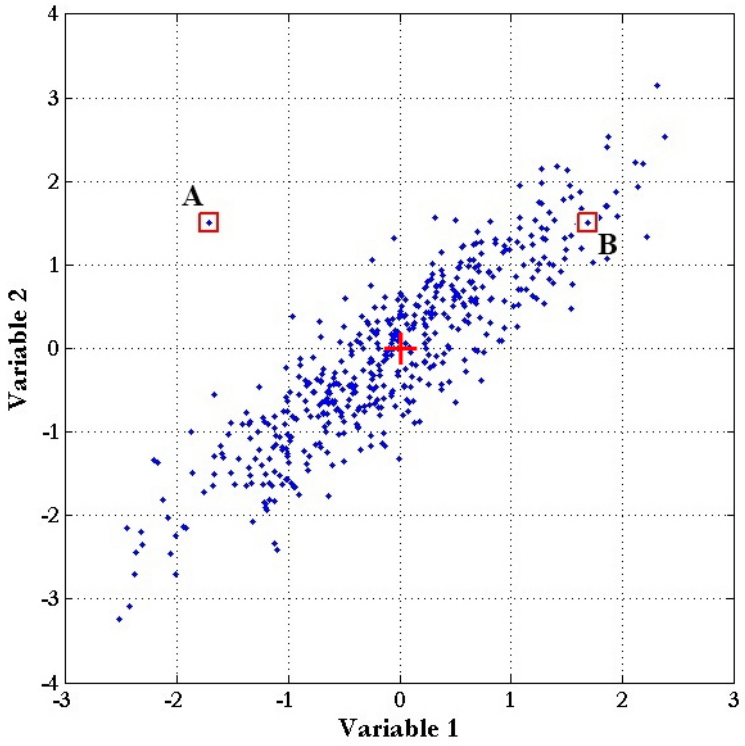


Figure 11. A bivariate plot of individuals belonging to a population in a space defined by two measurement variables. Individual B could be quite reasonably considered to belong to the population, whereas individual A is a clear outlier. Yet they are equally distant from the population centroid if distance is defined as Euclidean Distance.

The distance measure that does this is Mahalanobis Distance[44]. The formula for computing Mahalanobis Distance is complicated, but essentially is a generalization of the standardization of a univariate variable to a Z-score, to the multivariable case.

A way of visualizing this for those familiar with Principal Components Analysis is to first centre the data on the multivariate mean (centroid), ordinate the space to establish a series of uncorrelated axes, then standardize each those axes to a unit variance. After that transformation (converting the ellipsoid to a sphere), the Mahalanobis Distance is simply the Euclidean Distance.

The Mahalanobis distance is very useful for identifying outliers, and in population assignment for individuals of unknown provinence. It the data are multivariate normal, then the transformed data will follow a Chisquare distribution to good approximation, so that p values can be associated with each individual.

These techniques all work well for SNP data.

Genetic Distance for Loci

Just as individuals (entities) can be represented in a space defined by axes where each axis represents a locus (attributes), the loci can be represented in a space defined by axes where each axis represents an individual (Figure 12). The former analysis leads to what can be regarded as an R-mode distance matrix and the second analysis leads to what can be regarded as a Q-mode distance matrix. The distinction between the two is one of focus. In an

R-mode analysis, we are interested in the distance structure amongst the entities of interest – individuals or populations. In the Q-mode analysis, the focus is on the attributes – the loci scored for presence-absence or scored for SNP variants.

Application of Q-mode analysis of SNP sequence tag presence-absence using SNP variants is not well developed, apart from the possibility of visually identifying loci departing from linkage equilibrium, so we deal with it only briefly here.

To undertake a Q-mode analysis, one needs a measure of distance defined for loci on the basis of their state in the individuals.

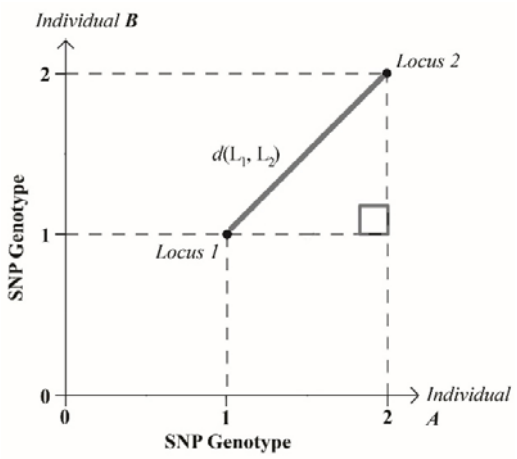


Figure 12. SNP Loci plotted in a space defined by axes representing the SNP scores over individuals.

Euclidean distance between two SNP loci based on their variation across individuals is given by

$$D_{Euclidean}(L_1, L_2) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_{L_1 i} - p_{L_2 i})^2}$$

where $\mathrm{p}_{\mathrm{L}_1\mathrm{i}}$ is the frequency of the alternate allele at locus L_1 for individual i; $\mathrm{p}_{\mathrm{L}_2\mathrm{i}}$ is similarly defined, and N is the number of individuals.

This distance is probably of little utility unless applied to individuals from a single population in Hardy-Weinberg equilibrium.

Genetic Distance for Sequence Tags

Sequence-level genetic distance as it applies in the context of SNP data refers to distance measures devised to capture variation among sequence tags. The most common distance measure used in this context is the Hamming Distance.

Hamming distance is defined as the number of base mismatches between the sequence tags for Locus 1 and Locus 2.

The DNA sequences being compared, the sequence tags, need to be aligned as they are because they all start at the first restriction enzyme site and should be of the same length. Sequence tags arising from double digestion are typically not of the same length, and DArT sequence tags for example range in length from 20 base pairs to 69 base pairs.

When confronted with sequence tags that differ in length, the package dartR truncates the longer sequence tag before calculating the Hamming Distance.

Clearly, a 5 bp difference between sequence tags of 20 bp length is not the same as 5 bp difference between sequence tags of 60 bp length. To overcome this in part, the Hamming Distance can be scaled to a proportion of the number of base pairs compared. Nevertheless, one might consider filtering sequence tags below some threshold length if Hamming Distance considerations are important.

Properties: Non-metric. Range [0, $Length$). If scaled, range [0,1]

Tree Distance

A final distance measure that is worth mentioning is tree distance. A tree distance, in addition to being a metric distance, satisfies the four-point condition[45]:

$$d(X,Y) + d(Z,V) \leq max[d(X,Z) + d(Y,V), d(Y,Z) + d(X,V)]$$

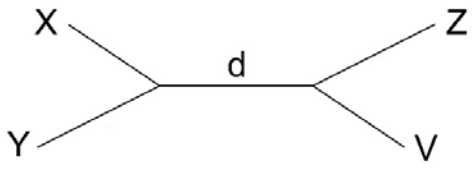


Figure 13. An unrooted tree showing the relationships among four individuals X, Y, V and Z. In order to satisfy the criteria of a tree distance, the internal node d must be defined.

Without labouring on the point, the four-point condition is a constraint that the internal node has a defined distance, d $\geq$ 0 (Figure 13). A distance matrix that satisfies the four-point condition can be represented, without distortion, as a unique bifurcating tree.

Distances that satisfy the four-point condition are referred to as additive.

Further Notes on Managing Missing Data

Missing data are problematic for distance analyses. Techniques like PCA that access the raw data matrix cannot accommodate missing data, and PCoA which accesses a distance matrix, is affected in ways that are not entirely transparent. Missing data can destroy the metricity of a distance matrix even though the distance measure upon which it is based is a metric distance. The same is true of a Euclidean distance matrix. There are three strategies

- Filter stringently on Call Rate
- Remove loci with missing values
- Impute missing values based on observed allele frequencies within populations.

The first strategy is to filter stringently on Call Rate (say with a threshold of 0.95 for loci; 0.80 on individuals) after applying whatever other filters you routinely apply. That may suffice to address any issues of practical consequence.

A further, more draconian step, is to remove loci with missing data for any individual so that your input matrix has no missing data. This is very wasteful of data but may be the solution that many algorithms apply in any case.

The third strategy is appropriate if the entities are grouped into populations. To avoid severe data loss and maintain the representation of the distance between your entities (individuals or populations), it might be sensible to impute the missing values that remain after stringent filtering on Call Rate.

The basic idea is to avoid distorting the outcome with the imputation, so one strategy is to consider each population separately, calculate allele frequencies, and draw randomly from them to substitute the missing value. Another way is to calculate the allele frequencies expected under Hardy-Weinberg Equilibrium, and draw from them randomly. A third way is to identify the individual that is genetically most similar to the individual with the missing value, and borrow its value.

Using one of these three methods, the accuracy of the representation is maintained, though the precision is slightly underestimated.

In summary, to make a dense matrix (no missing values) for analyses that are sensitive to such missing values, one might employ the following code.

```
gl <- gl.filter.callrate(gl, method="loc", threshold=0.95, verbose=3)
gl <- gl.filter.callrate(gl, method="ind", threshold=0.80, verbose=3)
gl <- gl.filter.allna(gl, by.pop=TRUE, verbose=3)
gl <- gl.impute(gl, method="frequencies", verbose=3)
```

Visualization

Heatmap

Genetic distances are typically presented as a distance matrix with N rows and N columns, where N is the number of individuals or populations being compared in pairwise fashion.

Visualizing genetic distances can be quite a challenge if the number of entities being compared (individuals or populations) is large. Poring over a large matrix of numbers is not the most enlightening exercise.

One way of looking for structure in a distance or dissimilarity matrix is to represent it as a heat map (Figure 14). Zero distances are shown in the lightest colour, the largest distances in bright red, with scaled colours for distances in between.

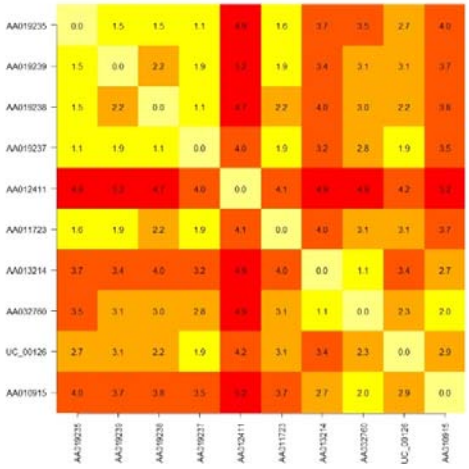


Figure 9. A heat map showing the distribution of distance values across a distance matrix defined, in this case, for individuals using unscaled Euclidean Distance. Light yellow represents zero distance (as along the diagonal, as $d(\mathrm{AA}) = 0$) through to dark red representing large distances. The matrix is symmetric around the diagonal (as $d(\mathrm{AB}) = d(\mathrm{BA})$).

Reordering the rows and columns of this distance matrix based on grouping like with like is perhaps more informative, when interpreted in the context of knowledge of the individuals or populations involved (Figure 10).

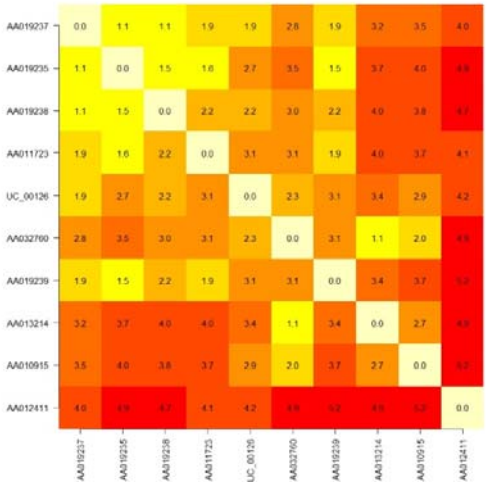


Figure 14. A heat map showing the distribution of distance values across a distance matrix defined, in this case, for individuals using unscaled Euclidean Distance. In this case, the distances are ordered.

Network Analysis

Another way of representing dissimilarities and distances is in a network diagram (Figure 15). Here, all the entities are regarded as nodes, linked together by edges. Various algorithms are applied to reduce the length of the links between entities that are similar in comparison to entities that are less similar. Needless to say, this will introduce some level of distortion between the graphical representation and the distances in the distance matrix. But the technique can serve to visually highlight clusters as an aid to communication.

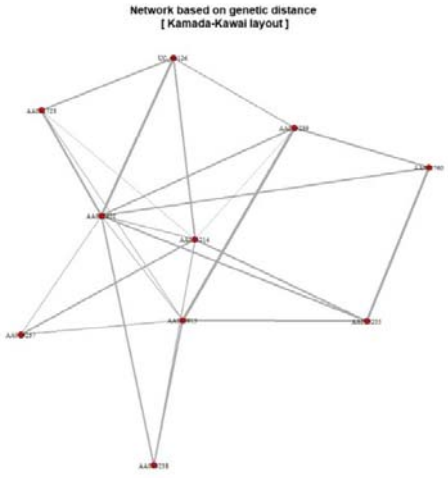


Figure 15. A network diagram where the lengths of the branches between entities reflects their similarity – the longer the branch, the greater the dissimilarity.

Networks have the advantage, together with heatmaps, of depicting asymmetric dissimilarity measures (bidirectional network plots). That option will not be covered here.

Trees

Genetic relationships can be summarized in the form of a tree, usually using the Neighbour-joining algorithm[46]. This algorithm uses the distance matrix to calculate the relative proximity of each of the entities to each other (closest neighbours) and then selects the pair of entities that are closest to form the first node of the tree. It then repeats the process for the node and remaining entities, and so on until it completes the tree. A good example is provided on Wikipedia (https://en.wikipedia.org/wiki/Neighbor_joining).

In the context of population genetics, these trees are phenograms that provide a visual summary of genetic similarity rather than phylograms that summarise the pattern of ancestry and descent. Nevertheless, when aggregations of individuals have been isolated and their allelic profiles have drifted apart over time, selecting among the distance measures based on a model incorporating drift (Reynolds Distance, Chord Distance) or drift and mutation (Nei Distance) is an option. The branch lengths on the tree will then reflect time since divergence, at least in theory.

Generating neighbour-joining trees has gained popularity over other methods of generating visual summaries of genetic similarity (e.g. UGPMA[47]) because it is not constrained by the assumption of equal rates of divergence (which under a drift model, translates to equal population sizes historically), it usually finds the tree of minimum overall length and reassuringly it will always recover the correct unique tree specified by a distance matrix that satisfies the four-point condition.

A detracting feature is that the neighbour-joining algorithm can produce negative branch lengths. This can be addressed by setting the negative branch length to zero, and then adding the difference to the adjacent branch length so that the total distance between an adjacent pair of terminal nodes remains unaffected[8<25].

You might wish to present your tree diagram together with a PCoA, in which case the same distance measure should be used for both analyses (usually Reynolds D).

References

1. Deza, M. M. & Deza, E. *Encyclopedia of Distances*. (Spinger-Verlag, 2009).
2. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).
3. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philisophical Mag.* **2**, 559–572 (1901).
4. Jolliffe, I. *Principal Component Analysis*. (Springer International Publishing, 2002).
5. Gower, J. C. Some distance properties of latent root and vector methods used in

multivariate analysis. *Biometrika* **53**, 325–338 (1966).

6. Gauch, H. G. J. Noise reduction by eigenvector ordinations. *Ecology* **63**, 1643–1649 (1982).

7. Cattell, R. B. The Scree Test For The Number Of Factors. *Multivariate Behav. Res.* **1**, 245–276 (1966).

8. Guttman, L. Some necessary conditions for common factor analysis. *Psychometrika* **19**, 149–161 (1954).

9. Georges, A. & Adams, M. A phylogeny for australian chelid turtles based on allozyme electrophoresis. *Aust. J. Zool.* **40**, (1992).

10. Dray, S. & Josse, J. Principal component analysis with missing values: a comparative survey of methods. *lant Ecol.* **216**, 657–667 (2015).

11. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).

12. Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).

13. Yi, X. & Latch, E. K. Nonrandom missing data can bias Principal Component Analysis inference of population genetic structure. *Mol. Ecol. Resour.* **22**, 602–611 (2021).

14. Gower, J. C. Statistical methods of comparing different multivariate analyses of the same data. in *Mathematics in the archaeological and historical sciences* (eds. Hodson, F. R., Kendall, D. G. & Tautu, P.) 138–149 (Edinburgh University Press, 1971).

15. Faith, D. P. A model of immunological distances in systematics. *J. Theor. Biol.* **114**, 511–526 (1985).

16. Cox, T. F. & Cox, M. A. A. *Multidimensional scaling*. (Chapman & Hall, 2001).

17. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

18. Gower, J. C. Euclidean distance geometry. *Mathermatical Sci.* **7**, 1–14 (1982).

19. Gower, J. C. & Legendre, P. Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.* **3**, 5–48 (1986).

20. Cailliez, F. & Pages., J. P. *Introduction `a l'analyse des donn´ees*. (Soci6t6 de Math6matiques Appliqu6es et de Sciences Humaines, 1976).

21. Sibson, R. Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. *J. R. Stat. Soc.* **41B**, 217–229 (1979).

22. Legendre, P. & Legendre, L. Numerical Ecology. *Dev. Environ. Model.* **24**, 1–990 (2012).

23. Cailliez, F. The analytical solution to the additive constant problem. *Psychometrika* **48**, 305–308 (1983).

24. Lingoes, J. C. Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika* **36**, 195–203 (1971).

25. Choi, S. S., Cha, S. H. & Tappert, C. C. A survey of binary similarity and distance measures. *Syst. Cybern. Informatics* **8**, 43–48 (2010).

26. Sokal, R. R. & Michener, C. D. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **28**, 409–1438 (1958).

27. Jaccard, P. The distribution of the flora in the alpine zone. *New Phytol.* **11**, 37–50 (1912).

28. Levandowsky, M. & Winter, D. Distance between sets [5]. *Nature* **234**, 34–35 (1971).

29. Orlóci, L. *Multivariate Analysis in Vegetation Research*. (Dr W Junk Publishers, 1978).

30. Marczewski, M. & Steinhaus, H. On the taxonomic distance of biotopes. *Zastos. matem* **4**, 195–203 (1959).

31. Bray, J. R. & Curtis, J. T. An ordination of upland forest communities of southern Wisconsin. *Ecol. Monogr.* **27**, 325–349 (1957).

32. Sørensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *K. Danske Vidensk. Selsk.* **5**, 1–34 (1948).

33. Dice, L. R. Measures of the amount of ecologic association between species. Ecology. *Ecology* **26**, 297–302 (1945).

34. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 5269–5273.

35. Kosman, E. & Leonard, K. J. Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Mol. Ecol.* **14**, 415–424 (2005).

36. Gao, X. & Starmer, J. Human population structure detection via multilocus genotype clustering. *BMC Genet.* **8**, 34 (2007).

37. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423 (2008).

38. Endelman, J. B. & Jannink, J.-L. Shrinkage estimation of the realized relationship matrix. *G3 Genes, Genomics, Genet.* **2**, 1405 (2012).

39. Granato, I. snpReady. https://www.rdocumentation.org/packages/snpReady/versions/0.9.6 (2018).

40. Nei, M. Genetic distance between populations. *Am. Nat.* **106**, 283–292 (1972).

41. Reynolds, J., Weir, B. S. & Cockerham, C. C. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* **105**, 767–779 (1983).

42. Edwards, A. Distances between populations on the basis of gene frequencies. *Biometrics* **27**, 873–881 (1971).

43. Edwards, A. W. F. & Cavalli-Sforza, L. L. Reconstruction of evolutionary trees. in *Phenetic and Phylogenetic Classification* 67–76 (Systematics Association, 1964).

44. Mahalanobis, P. C. On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India* **2**, 49–55 (1936).

45. Buneman, P. A note on the metric properties of trees. *J. Comb. Theory B* **17**, 48–50 (1974).

46. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

47. Michener, C. D. & Sokal, R. R. A quantitative approach to a problem of classification. *Evolution (N. Y).* **11**, 490–499 (1957).

48. Kuhner, M. K. & Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**, 459–468 (1994).