Distances and their visualization in studies of spatial-temporal genetic variation using single nucleotide polymorphisms (SNPs)

Arthur Georges[1*], Luis Mijangos[1], Hardip Patel[2], Mark Aitkens[3] and Bernd Gruber[1]

[1] Institute for Applied Ecology, University of Canberra, ACT 2601, Australia

[2] National Centre for Indigenous Genomics, Australian National University, ACT 2601, Australia

[3] RPS Australia Asia Pacific, Unit 2A, 45 Fitzroy Street, Carrington NSW 2294, Australia

*Corresponding Author: georges@aerg.canberra.edu.au

Running Title: SNPs and genetic distances

Abstract

1. Distance measures are widely used for examining genetic structure in datasets that comprise many individuals scored for a very large number of attributes. Genotype datasets composed of single nucleotide polymorphisms (SNPs) typically contain bi-allelic scores for tens of thousands if not hundreds of thousands of loci.
2. We examine the application of distance measures to SNP data (both genotypes and sequence tag presence-absence) and use real datasets and simulated data to illustrate pitfalls in the application of genetic distances and their visualization.
3. Missing values arise from ascertainment biases in the SNP discovery process (null alleles in the case of SNP genotyping; true missing data in the case of sequence tag presence-absence data). Missing values can cause displacement of affected individuals from their natural groupings and artificial inflation of confidence envelopes, leading to potential misinterpretation. Failure of a distance measure to conform to metric and Euclidean properties is important but only likely to create unacceptable outcomes in extreme cases. Lack of randomness in the selection of individuals and lack of independence of both individuals and loci (e.g. polymorphic haploblocks), can lead to substantial and otherwise inexplicable distortions of the visual representations and again, potential misinterpretation.
4. Euclidean Distance is the metric of choice in many distance studies. However, other measures may be preferable because of underlying models of divergence, population demographic history, linkage disequilibrium, because it is desirable to down-weight joint absences, or because of other characteristics specific to the data or analyses. Distance measures for SNP genotype data depend on the arbitrary choice of reference and alternate alleles (e.g. Bray-Curtis distance) should be avoided. Careful consideration should be given to which state is scored zero when applying binary distance measures to fragment presence-absence data (e.g. Jaccard distance). Filtering on missing values then imputing those that remain avoids distortion in visual representations. Presence of closely related individuals or polymorphic haploblocks in the genomes of target species with limited genomic information occasionally emerge as challenges that need to be managed.

Keywords: Principal Components Analysis, Principal Coordinates Analysis, Genetic distance, Metric distance, Euclidean distance

Introduction

Population genetics is the study of the interplay of genetic drift, gene flow, recombination and selection, and to a lesser extent mutation, as they come to influence the contemporary genetic composition of populations. Finite populations can be expected to vary in genetic composition in space and through time under the influence of genetic drift alone. Rate of genetic drift strongly depends on overall population size and the level of spatial sub-structuring within populations. Populations are not static, and the history of population size fluctuations has a bearing on contemporary patterns of divergence evident in distance analyses, particularly if there has been a recent reduction in size or a sustained bottleneck. Divergence in allelic profiles can be reinforced and accelerated (or in some cases, impeded) by local selection (Edwards &

Cavalli-Sforza, 1967). The effect of these evolutionary processes on genetic distance are moderated by demographic events such as gene flow between populations of the same species, and hybridization and introgression between closely related species. Mutation sustains allelic variation in the context of allelic loss through drift and/or selection, but its effects are likely to be small compared to other influences in the context of population genetics (Edwards & Cavalli-Sforza, 1967).

Many measures of genetic similarity and dissimilarity have been developed to quantify inter-individual and inter-population variation. The concept of genetic distance (Sanghvi, 1953) is now a fundamental tool in genetics (Nei & Kumar, 2000). Genetic distances fall into several broad classes. Some are obtained by direct measurement, such as immunological distance (Faith, 1985) or DNA-DNA hybridization (de Ley et al., 1970; Hirayama et al., 1996; Kirsch et al., 1990). However, most genetic distances are calculated from character states (genotypes) arranged as a matrix of individuals (as entities) by attributes (genetic loci). Genetic distances can be further classified by whether they will be used to infer patterns of ancestry and descent among species (phylogenetics), the structure and relationships among populations of a species at various scales of divergence with or without gene flow (population genetics), or relationships among individuals (e.g. kinship).

Even if the focus is on elucidating contemporary divergence among populations, genetic distances between populations and the genetic distances between the individuals that comprise those populations are interdependent. In this paper, we deal with distances defined for both individuals and populations, in the broader context of divergence among populations.

The array of available measures of distance and similarity/dissimilarity (Deza & Deza, 2009) is daunting. In this article, we specifically address analyses of single nucleotide polymorphisms (SNPs), markers that are commonly used in studies of spatial or temporal variation among individuals and populations of a species or closely related species. SNPs have particular characteristics that influence the choice of a distance measure. SNPs can have more than two alleles (i.e. are multiallelic) but in practice, sites with more than two allelic states are filtered out during the selection of markers as part of pipelines to eliminate non-homologous sequence tags. Fortunately, multiallelic SNP sites are rarely observed. As a result of this filtering, SNP markers are bi-allelic, typically scored as the frequency of the alternate allele – 0 for homozygous reference allele, 1 for heterozygotes, and 2 for homozygous alternate allele (in diploid organisms). Such biallelic markers have characteristics that substantially limit options for a distance measure. For example, the values of 0 and 2 carry equal weight because the choice of reference allele and alternate allele is arbitrary. Any distance measure that gives differential weight to joint zeros (e.g. Bray & Curtis, 1957 Distance) will yield values that depend on the arbitrary choice of which allele is reference and which is alternate at each locus. Such distance measures can be eliminated from options available for SNP genotype data. Standardization or normalization across attributes (loci) is not required because the attributes (loci) are all measured as genotypes on the same scale (0, 1 or 2); this is not the case under some circumstances in multiallelic systems. That said, the biallelic nature of SNP markers permits the easy calculation of the maximum distance possible, which permits scaling distances to fall conveniently in the range of [0,1] while maintaining comparability across studies.

Finally, issues that arise in a multiallelic context do not apply in a biallelic context. For example, Rogers D (Rogers, 1972), and therefore Standard Euclidean Distance, can yield undesirable results when two populations are both polymorphic at a site, but share no alleles (Nei & Kumar, 2000:246). This situation does not arise in the biallelic case, and so Standard Euclidean Distance (or Rogers D) is often the distance of choice in SNP studies.

A second form of informative genetic data comprises presence or absence of the sequence tag typically denoted as 1 for presence and 0 for absence. Absence of a sequence tag in this case is not because of missing data *per se* (absence of evidence), but instead is the true absence of sequence tag in a sample (evidence of absence). Technical artefacts such as low DNA quality or quantity, or failure of experimental protocol resulting in shallow sequencing depth, may occasionally lead to the missing data owing to the absence of evidence. Typically, this is overcome by quality control processes, and data filtering is routinely performed to distinguish between the true absence and missing data for sequence tags. Therefore, absence of a sequence tag is assumed to arise as a null allele, which signifies a mutation at one or both of the restriction enzyme recognition sites. Provided technical reproducibility is achieved in the generation of data, sequence tag presence-absences are valid genetic markers in their own right. Since sequence tag presence-absence data are binary data, many binary distance measures (Deza & Deza, 2009) have been co-opted for genetic studies of presence-absence SNP data (e.g. Jaccard Distance).

In this paper, we consider genetic distances commonly used in analyses of SNP data (both genotypic and tag presence-absence) as they apply to individuals and populations. Although genetic distances are used in a wide range of contexts (Jansen & van Hintum, 2007; Libiger et al., 2009; Yin, 2020), we focus (not exclusively) on the application of distances in studies of spatial and temporal genetic structure among populations where allelic profiles are governed principally by recent or contemporary processes of drift, selection and gene flow. We do not consider distance measures used to reveal deeper historical patterns of ancestry and descent among lineages on independent evolutionary trajectories where the pattern of mutational events dominates (species-level phylogenetics). Nor do we consider distance measures used to quantify distances between pairs of SNPs as opposed to pairs of individuals or populations (Müller et al., 2005).

We further examine approaches for visualizing distances in multivariable space, in particular, Principal Components Analysis (PCA) (Hotelling, 1933; Jolliffe, 2002; Pearson, 1901) and Principal Coordinates Analysis (PCoA) (Gower, 1966) collectively referred to as ordination. These techniques have found application in many diverse fields such as ecology, economics, psychology, meteorology, oceanography, human health and genetics as a descriptive and an exploratory tool to generate hypotheses for further examination (Jolliffe & Cadima, 2016) rather than a formal statistical analysis (but see Patterson et al., 2006).

With the relatively recent advent of large SNP matrices, many applications of ordination examine spatial and temporal population structure in SNP datasets. We pay particular attention to underlying assumptions of the application of distance analyses and visualization of spatio-temporal structure using ordination, which are poorly appreciated.

These assumptions include the properties of various distance measures and the importance of metric and/or Euclidean properties for visualization, the impact of missing values and the importance of randomness in sampling and independence of both those individuals' genotypes and the loci selected for screening. We briefly review the most commonly used distance measures as they apply to SNP datasets and their usefulness in analysing population structure.

In what follows, the term "SNP data" includes both SNP genotype data and SNP sequence tag presence-absence data. We use the concept of distance loosely to encompass the notions of measures of dissimilarity through to metric distances and rigid Euclidean distances (Gower & Legendre, 1986). Where the distinction is necessary, a distance is referred to as a non-metric distance, a metric distance, or a Euclidean distance. In describing a SNP matrix, we refer to individuals, samples or specimens as entities, the SNP loci that are scored for each entity as attributes, and the scores themselves as states.

Methods

For the purposes of illustration, a dataset was constructed from a SNP matrix generated for the freshwater turtles in the genus *Emydura*, a recent radiation of Chelidae in Australasia. The dataset includes populations that vary in level of divergence to encompass variation within species and variation between closely related species. Populations (i.e. sampling localities) with evidence of admixture between species were removed. Monomorphic loci were removed, and the data was filtered on call rate (<95%), repeatability (< 99.5%) and read depth ($5x <$ read depth $< 50x$). Where there was more than one SNP per sequence tag, only one was retained at random. The resultant dataset had 18,196 SNP loci scored for 381 individuals from 7 populations (sampling localities) – *Emydura victoriae* [Ord River, NT, n=15], *E. tanybaraga* [Holroyd River, Qld, n=10], *E. subglobosa worrelli* [Daly River, NT, n=25], *E. subglobosa subglobosa* [Fly River, PNG, n=55], *E. macquarii macquarii* [Murray Darling Basin north, NSW/Qld, n=152], *E. macquarii krefftii* [Fitzroy River, Qld, n=39] and *E. macquarii emmotti* [Cooper Creek, Qld, n=85]. The missing data rate was 1.7%, subsequently imputed by nearest neighbour (Beretta & Santaniello, 2016) to yield a fully populated data matrix.

The above manipulations were performed in R package dartR Version 2.0.3 (Gruber et al., 2018; Mijangos et al., 2022). Principal Components Analysis was undertaken using the glPCA function of the R adegenet package (as implemented in dartR, Gruber et al., 2018; Jombart, 2008; Jombart & Ahmed, 2011). Principal Coordinates Analysis was undertaken using the pcoa function in R package ape (Paradis & Schliep, 2019) implemented in dartR Version 2 (Mijangos et al., 2022).

To exemplify the effect of missing values on SNP visualisation using PCA (Figure 6), we performed computer simulations in dartR Version 2.0.3 (Mijangos et al., 2022). Ten simulated populations reproducing over 200 non-overlapping generations were placed in a linear series with low dispersal between adjacent populations (one disperser every ten generations). Each population had 100 individuals, of which 50 individuals were sampled at random. Genotypes were generated for 1000 neutral loci on one chromosome.

The data for the Australian Blue Mountains skink *Eulamprus leuraensis* (Figure 7) were generated for 372 individuals collected from 17 swamps isolated to varying degrees in the Blue Mountains region of New South Wales. A total of 13,496 loci were scored which reduced to 7,935 after filtering out secondary SNPs on the same sequence tag, filtering on reproducibility (threshold 0.99) and call rate (threshold 0.95), and removal of monomorphic loci.

The algorithms discussed in this paper form the basis of the distance analysis implemented in R software package dartR (Gruber et al., 2018; Mijangos et al., 2022).

The Concept of Distance

Standard Euclidean distance is a common-sense notion derived as an abstraction of physical distance. It is possible to calculate the Standard Euclidean distance between two points from their coordinates in a two-dimensional space defined by orthogonal Cartesian axes (Figure 1). The distance between two points in space is calculated by applying Pythagoras' rule to their projection onto the Cartesian axes (Figure 1a).

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

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

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

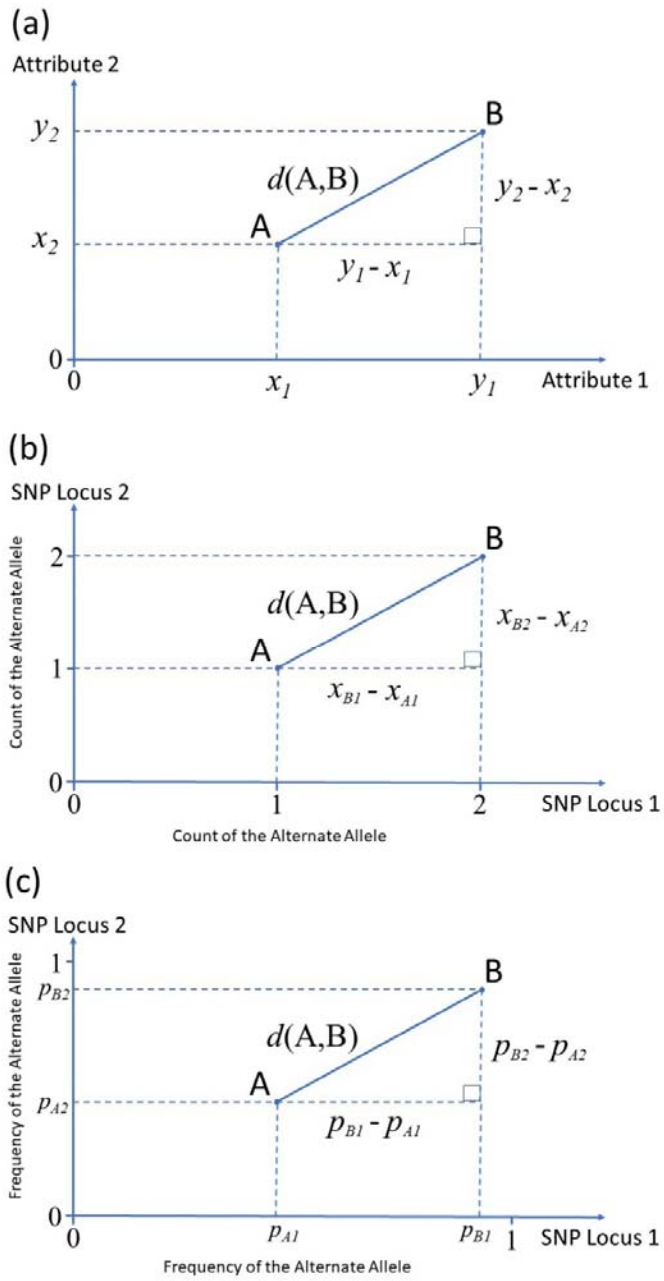

Figure 1. Distance between two points A and B represented in two-dimensional space (a) can be calculated from their Cartesian coordinates using Pythagoras' rule – the square of the hypotenuse of a right-angled triangle is equal to the sum of the squares of the two adjacent sides. Each axis can be considered to represent a locus (b), with the value taken by an individual (A or B) on that axis called as x_{Ai} and $x_{Bi} = 0$, 1 or 2 for SNP genotype. As such, each individual is represented by a point in a multidimensional space defined by the $i = 1$ to L loci. Populations A and B can be similarly depicted, as in (c), with the relative frequency of the alternate allele in the population (p_{Ai} and p_{Bi}) as the value taken on each SNP axis. Representation for sequence tag presence-absence data can be similarly defined.

Applying Standard Euclidean distance to SNP data is straightforward. Where points A and B represent two individuals (Figure 1b), the horizontal axis represents SNP Locus 1 and the values x_{A1} and x_{B1} represent the scores for that locus (0 or 1 or 2 for SNPs; 0 or 1 for tag presence-absence) for individuals A and B respectively; the vertical axis represents SNP Locus 2 with the values x_{A2} and x_{B2} similarly defined.

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

Standard Euclidean Distance can be similarly defined for populations,

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

where p_{Ai} and p_{Bi} are the relative frequencies of the alternate allele at locus i in populations A and B respectively (Figure 1c).

The equations for *d(A,B)* and *D(A,B)* can be scaled to fall in the range [0,1] on noting that its maximum is achieved for SNP data when all x_{Ai} are 0 and all x_{Bi} are 2; for tag presence-absence data the maximum is achieved when all x_{Ai} are 0 and all x_{Bi} are 1. Note also that the equations for *d(A,B)* and *D(A,B)* are symmetric with respect to choice of which allelic state is assigned to reference (0 for homozygous reference) and which is assigned to alternate (2 for homozygous alternate). Interchanging the reference and alternate alleles (that is applying transformation $x' = 2 - x$ for SNP genotype data or $x' = 1 - x$ for presence-absence data) has no impact on the value of the distance between A and B.

Visualization of SNP data

Principal Components Analysis (Hotelling, 1933; Jolliffe, 2002; Pearson, 1901) takes a SNP data matrix (genotypes or presence-absence data), represents the entities (individuals or samples or specimens) in a space defined by the *L* loci, centres and realigns that space by linear transformation (rotation) to yield new orthogonal axes ordered on the contribution of variance (represented by their eigenvalues) in their direction (defined by their eigenvector). This process maintains the relative positions of the entities. Because the resultant axes are ordered on the amount of information they contain, the first few axes, preferably two or three, tend to contain information on any structure in the data (signal) and later axes tend to contain only noise (Gauch, 1982). This is a powerful visual technique for examining structure in the SNP dataset. An example of a PCA is presented in Figure 2. Important variations include the combination of PCA with discriminant analysis (DAPC, Jombart et al., 2010) and adjustment for confounding factors (AC-PCA, Lin et al., 2016).

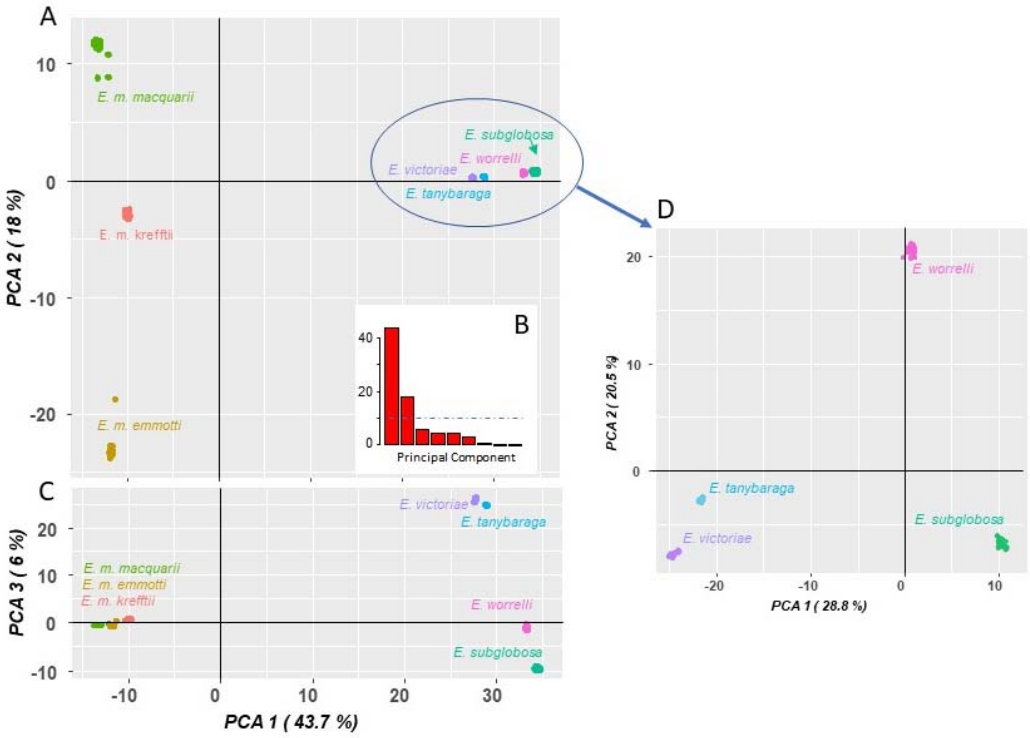

Figure 2. Principal Components Analysis (PCA) as applied to a SNP genotype dataset for the Australasian species of *Emydura* (Chelonia: Chelidae). Plot of individuals in a two-dimensional space (**a**) defined by Principal Component 1 (horizontal axis) and Principal Component 2 (vertical axis). Together they explain 61.7% of the total variance among individuals. A scree plot (**b**) shows the contributions to variance by the first nine principal components (those that exceed the Kaiser-Guttman threshold, Guttman, 1954) of which only two each explain more than 10% of variation. Proximity of *Emydura subglobosa/worrelli* to *E. victoriae/tanybaraga* in 2D obscures their distinction, evident when Principal Component 3 is considered (**c**). The variation explained by the first two principal components can be largely set aside by considering structure within the cluster *Emydura subglobosa/worrelli*/*victoriae/tanybaraga* only (**d**), effectively allowing consideration of deeper dimensions.

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 ordination would fail (the first two axes would each represent only a small and random percentage of the total variance). When there is structure, the absolute value of the percentage of variation explained by a principal component cannot be compared across studies as a measure of the strength of the result; the percentage variance explained by a principal component needs to be taken in the context of the average percent variation explained by all components (Guttman, 1954).

Note also that a PCA plot can be misleading if one chooses, for convenience, only two or three dimensions in which to visualize the solution. For example, separation in a 2-D plot can be accepted as real, but proximity cannot because further separation can occur in deeper

dimensions each coupled with a substantial explained variance (Figure 2c). A widely used approach to determining the number of dimensions for the final solution is to examine a scree plot (Cattell, 1966), where the eigenvalues (proportional to variance explained) associated with each of the new ordered dimensions are plotted (Figure 2b). It is usual to apply the Kaiser-Guttman criterion (Guttman, 1954) whereby only those dimensions with more than the average eigenvalue are considered, or to apply a related but less conservative approach to take into account sampling variability (Jolliffe, 1972). Even so, this may result in many informative dimensions. One must decide how much information to discard (e.g. keeping only those components that explain more than 10% of total variation) or adopt, as a threshold, a visually-evident sudden drop in the percentage variation explained, commonly referred to as an elbow. More formal techniques (Cangelosi & Goriely, 2007; Jackson, 1993; Peres-Neto et al., 2005) include the broken-stick approach of Macarthur (1957), which provides a good combination of simplicity of calculation and evaluation of suitable dimensionality (Jackson, 1993). The broken-stick model retains components that explain more variance than would be expected by randomly dividing the variance into equal parts. Another related approach is to observe that the eigenvalues of lower "noise" dimensions are likely to decline geometrically, a trend that can be linearized by a log transformation. Informative dimensions are those that exceed this linear trend line (more strictly, exert disproportionate leverage on the regression). A more recent approach assesses the statistical significance of the variation explained by each Principal Component (Patterson et al., 2006). Under specified assumptions, the sampling distribution of the ordered eigenvalues, under the null hypothesis of no structure in the data, follows Tracy-Widom distribution (Tracy & Widom, 1993). Thus, it is possible to assign a probability to an observed eigenvalue and retain for consideration only those principal components that are statistically significant (Patterson et al., 2006). The technique is sufficiently robust to violations of its underlying assumptions (e.g. normality) to be applicable to large genetic biallelic data arrays.

Displaying the results in a two-dimensional plot is straightforward. Various software packages can display the results in three dimensions and allow rotation of the three axes to provide the best perspective (e.g. dartR, Gruber et al., 2018; Mijangos et al., 2022). Higher dimensions can be visualized by plotting the set of largest components in pairwise fashion. Alternatively, if there are strong groupings in the PCA plot in two dimensions, individuals in each of these groupings can be isolated and analysed by PCA separately (Georges & Adams, 1992) (Figure 2d).

Generalization of the concept of Distance

Standard 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 we have:

$$d(A,B) = 0 \text{ if and only if } A = B$$

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

$$d(A,B) \leq d(A,C) + d(B,C)$$

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 path between them. From these properties we can conclude that metric distances must be non-negative.

$$d(A,B) \geq 0$$

In essence, metric distances are well-behaved distances. Standard Euclidean Distance, as with all Euclidean distances, is a metric distance.

Graphically, the metric properties make complete sense for a distance (Figure 3). Given three points defined by the distances between them, the position of each of them is uniquely defined (Figure 3a). This is necessary (though not sufficient) if we are to draw an analogy between our distances and a representation in a linear physical space.

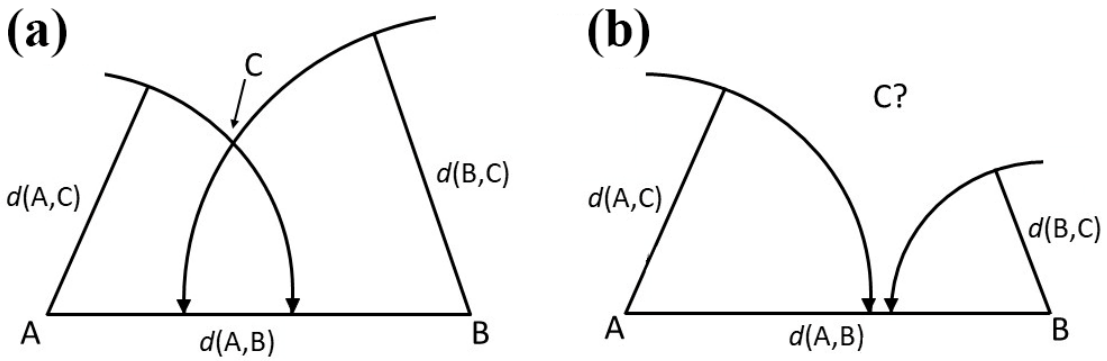

Figure 3. Visual representation of the triangle inequality as used to define a metric distance. **(a)** – if three distances between points A, B and C satisfy the metric property $d(A,B) \leq d(A,C) + d(B,C)$, then the position of each of A, B and C in space is well defined. **(b)** – if not, and $d(A,B) > d(A,C) + d(B,C)$, then point C is undefined.

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. Private alleles are those uniquely possessed by one population when compared to other populations. The number of private alleles possessed by population A compared to population B will typically be different from the number of private alleles possessed by population B compared with A. Thus the resultant distances will not satisfy the second metric criterion of symmetry. Other genetic distances in use 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 for the same reason, 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 (refer Legendre & Legendre, 2012-- tables 7.2 and 7.3).

Generalization of PCA

Principal Co-ordinates Analysis (PCoA, Gower, 1966) is an alternative visualization technique that represents a distance matrix in a Euclidean space defined by an ordered set of orthogonal axes, as does PCA. Again, the axes are ordered on 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 (Gauch, 1982). Important variations include adjustment for confounding factors (AC-PCoA, Wang et al., 2022) and application of iterative procedures to best match measured distances against distances in the visual representation for a specified number of dimensions (Belbin, 1991; Kruskal, 1964; metric, non-metric and hybrid MDS, Shepard, 1962).

The primary difference between PCA and PCoA is that PCA works with the covariance (or correlation) matrix derived from the original data whereas PCoA works with a distance matrix. Because the mathematics of PCA moves forward from the covariance (or correlation) matrix, the insight attributed to John Gower (1966) was to substitute any distance matrix at this point in the analysis, following a simple transformation. This yields an ordered representation of those distances, metric or otherwise, in multivariable space akin to PCA, greatly expanding the range of application of ordination. In PCA, the *N* individuals are represented in a space of *L* dimensions defined by the loci whereas in PCoA, the individuals are represented in an *N-1* space with coordinate axes based solely on their pairwise distances. Thus the PCoA is not implicitly connected to any raw SNP genotype or tag presence-absence data.

Choosing the number of dimensions to display in visualizing a PCoA is similar to PCA. Missing values are less disruptive for PCoA than classical PCA because they are accommodated in pairwise fashion rather than globally, but they nevertheless require consideration. Missing values result in variation in the precision of estimates of allele frequencies across loci and can break the Euclidean properties of a sample distance matrix even when the chosen metric is Euclidean in theory. Other considerations arise in PCoA from using distance matrices that are non-Euclidean. If a Euclidean distance matrix is used in PCoA then, in the absence of missing values, the distances in the input matrix are represented faithfully in the full ordinated space, that is, without distortion (Figure 4c). The same cannot be said of metric distances in general, as the metric properties ensure that the individuals can be represented in an ordinated space, but not in a rigid linear space (Figure 4a, after Gower, 1982) – some curvature of the space may be required to satisfy the triangle inequality (Figure 4b), which will potentially compromise ability to faithfully represent the distances between entities in a space defined by Cartesian coordinates. When applying PCoA to non-metric distances, the distortion in the representation can be severe.

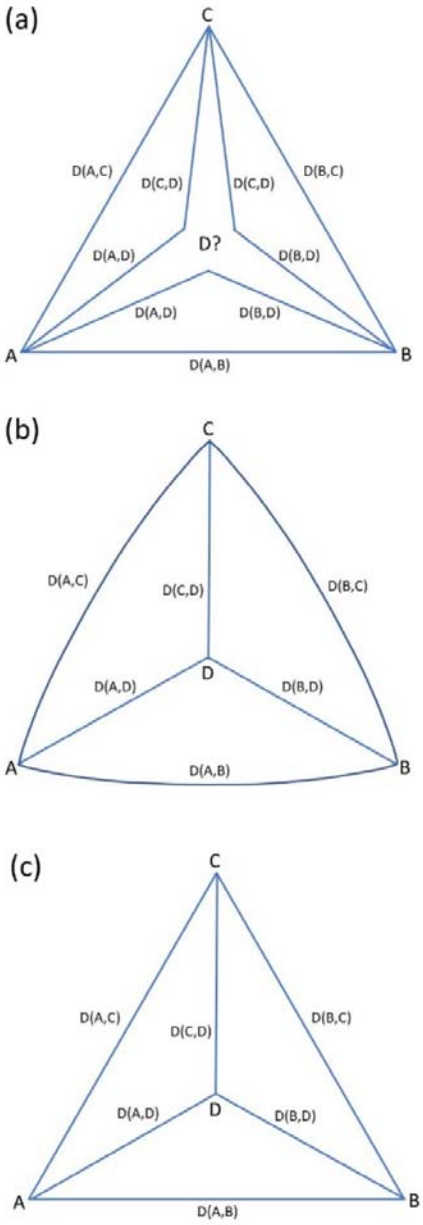

Figure 4. Metricity is not sufficient to represent distances in a rigid space defined by Cartesian coordinates; the distances must be Euclidean. **(a) –** distances between four individuals that satisfy the metric properties can nevertheless not be represented in a linear space defined by Cartesian coordinates because the position of individual D is not defined by the interindividual distances (after Gower, 1982). **(b)** – this distortion can be resolved by allowing non-linear links (geodesics) to represent distances between individuals as, in the case of four points shown, the edges of an irregular Reuleaux tetrahedron in three dimensions. **(c) –** in contrast, Euclidean distances between four individuals can always be represented by linear segments without distortion, as edges of an irregular conventional tetrahedron in three dimensions.

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; some eigenvalues will be negative (representing imaginary eigenvectors) (Gower & Legendre, 1986). However, the level of distortion is likely to be of concern only if the absolute magnitude of the largest negative eigenvalue is greater than that of any of the dimensions chosen for the reduced representation (Cailliez & Pages, 1976). A few small negative eigenvalues do not detract much from the visual representation if only a few of the highest dimensions are retained in the final solution (Sibson, 1979). Thus, departure from theory needs to be addressed in practice only if it causes serious issues. Note also that a distance measure does not need to be metric or Euclidean in theory for a sample distance matrix itself to be metric or Euclidean.

Negative eigenvalues complicate interpretation of the variance contributions. 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 (Legendre & Legendre, 2012:506).

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

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 the square root (Legendre & Legendre, 2012:501), Cailliez transformation (Cailliez, 1983; Gower & Legendre, 1986) and the Lingoes transformation (Gower & Legendre, 1986; Lingoes, 1971).

Square root	$\mathrm{D}'(\mathrm{A},\mathrm{B}) = \sqrt{\mathrm{D}(\mathrm{A},\mathrm{B})}$	
Cailliez	$\mathrm{D}'(\mathrm{A},\mathrm{B}) = \mathrm{D}(\mathrm{A},\mathrm{B}) + \mathrm{c}$	for all $A \neq B$
Lingoes	$\mathrm{D}'(\mathrm{A},\mathrm{B}) = \sqrt{\mathrm{D}(\mathrm{A},\mathrm{B}) + \mathrm{c}}$	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.

Finally, the outcome of a PCoA with an input matrix comprised of Standard Euclidean Distances is identical to the outcome of a PCA (Cox & Cox, 2001:43-44). In this context, the interchangeability of the two, PCA and PCoA leads to considerable confusion on the distinction between the two analyses.

Genetic Distances for Individuals

Binary Data

Tag presence-absence data involves scoring SNP loci as "called" [1] or "not called" [0]. 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, or if it is amplified from a different start site, is no longer considered homologous during SNP pre-processing, and the SNP is called as missing for that individual. The individual is scored as 0 for that locus.

Taking individuals two at a time, we can count up the different cases,

$$N_{00} = \sum_{i=1}^{L} \begin{matrix} 1, \text{where } x_{Ai} = x_{Bi} = 0 \\ 0, \text{where } x_{Ai} \neq x_{Bi} \end{matrix}$$

.....(1)

where x_{Ai} and x_{Bi} are the tag presence-absence scores for individuals A and B respectively. N_{00} is the sum of loci scored 0 (absent) for both individuals;

$$N_{10} = \sum_{i=1}^{L} \begin{matrix} 1, \text{where } x_{Ai} = 1 \\ 0, \text{where } x_{Bi} = 0 \end{matrix}$$

.....(2)

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

$$N_{01} = \sum_{i=1}^{L} \begin{matrix} 1, \text{where } x_{Ai} = 0 \\ 0, \text{where } x_{Bi} = 1 \end{matrix}$$

.....(3)

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

$$N_{11} = \sum_{i=1}^{L} \begin{matrix} 1, \text{where } x_{Ai} = x_{Bi} = 1 \\ 0, \text{where } x_{Ai} \neq x_{Bi} \end{matrix}$$

.....(4)

that is, sum of 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

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

.....(5)

Based on the above intermediates [(1) – (5)], there are several ways to calculate a binary distance between two individuals (Choi et al., 2010), some of which are shown in Table 1. The two most commonly used distance measures are Standard Euclidean distance and Simple Matching Distance (Sokal & Michener, 1958). Both are based on the number of mismatches between two individuals ($N_{01} + N_{10}$), expressed as a proportion of the total number of loci considered (L), and used when there is symmetry (equivalence) in the information carried by 0 (absence) and 1 (presence). Simple Matching Distance is simply the Standard Euclidean distance squared (and hence is non-metric).

The Jaccard Distance is a variation on the Simple Matching Distance that down-weights joint absences and so is no longer symmetric with respect to 0 and 1 scores. Down-weighting absences of the sequence tags 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 tag presence-absence data, you might consider recoding the data ($x' = 1 - x$) so that 1 represents presence of a mutation at one or both of the restriction enzyme sites (i.e. absence of the amplified tag) and 0 represents absence of such a disruptive mutation (i.e. success in amplifying the sequence tag). Having made this simple adjustment, the Jaccard Distance will down-weight joint absence of a disruptive mutation. The Jaccard Distance is a metric distance (Levandowsky & Winter, 1971).

The Sørensen 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 sequence tag presence-absence data. Special attention may be required to manage missing values when applying the Sørensen and Jaccard Distances, because adjustment of the denominator in their formulae (Table 1) can lead to a potential systematic bias (Orlóci, 1978:62). The Sørensen Distance is not a metric distance (Orlóci, 1978:61).

Table 1. Some genetic distances commonly applied to binary SNP data, that is, to sequence tag presence-absence data. Variables are described in the text. Formulae are presented to illustrate the adjustments made to the denominator for Jaccard and Sørensen distances rather than their conventional algebraic form.

Name	Formula	Alternate Names	Reference
Standard Euclidean Distance	$d_E = \sqrt{\frac{N_{01} + N_{10}}{L}}$		
Simple Matching Distance	$d_{SM} = \frac{N_{01} + N_{10}}{L}$	Scaled Hamming Distance	(Sokal & Michener, 1958)
Jaccard Distance	$d_J = \frac{(N_{01} + N_{10})}{L - N_{00}}$	Marczewski-Steinhaus D (Marczewski & Steinhaus, 1959); Ružička D; Soergel D	(Jaccard, 1912)
Sørensen Distance	$d_{BC} = \frac{(N_{01} + N_{10})}{L - N_{00} + N_{11}}$	Dice D (Dice, 1945)	Sørensen D (Sørensen, 1948)

SNP Genotype Data

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

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

This scoring scheme is convenient computationally because the value adopted represents the count of the alternate allele.

Standard Euclidean Distance as applied to SNP genotype data is defined in the usual way with loci as the axes in a coordinate space, and the value on each axis is 0, 1 or 2 as defined above. The scaling factor of ½ arises because the maximum squared distance between two individuals at a locus is (2-0)2 = 4.

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

where x_{Ai} and x_{Bi} 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_{Ai} and x_{Bi} are non-missing.

The Simple Mismatch Distance uses the counts of shared alleles between two individuals i and j at a locus is given by

$$c_{i,j} = 0, \text{where no alleles are shared } [0,2]|[2,0]$$

$$= 1, \text{where one allele is shared } [0,1]|[1,0]|[2,1]|[1,2]$$
$$= 2, \text{where both alleles are shared } [0,0], [1,1], [2,2]$$

and is calculated as

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

where L is the number of loci non-missing for both individuals i and j. It is non-metric and similar to the Allele Sharing Distance (Gao & Starmer, 2007), differing from it by a factor of 2.

Czekanowski Distance (Czekanowski, 1913) is calculated by summing the scores across the axes

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

where x_{Ai} and x_{Bi} are the counts of the alternate allele at locus i for individual A and B respectively; L is again the number of loci for which both x_{Ai} and x_{Bi} are non-missing. Referred to also as the Manhattan D or the City Block D, Czekanowski Distance is a metric distance.

SNP genotype data can be converted to binary data by counting the shared alleles between two individuals i and j at a locus

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

whereby distance measures devised for binary data can be applied. These distances are in effect considering only fixed allelic differences between the two individuals.

Genetic Distances for Populations

Standard Euclidean Distance as applied to allele frequencies is defined in the usual way with loci as the axes in a coordinate space, and the value for the population on each axis is set to the frequency of the alternate allele for the respective locus (Table 2). An appropriate scaling factor is applied to bring the value of the distance into the range [0,1], which renders it identical to Roger's D (Rogers, 1972). Standard Euclidean distance is a model-free metric distance, in that its formulation makes no assumptions regarding evolutionary processes generating the genetic distances.

Nei's standard genetic distance (Nei, 1972) is favoured by some over Standard Euclidean Distance 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 (Nei, 1972). Nei's D is non-metric.

Reynolds genetic distance (Reynolds et al., 1983) 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 mutation. Thus, it may be more appropriate than Nei's distance for spatial population genetics (divergence on relatively short timescales) and in particular, representation of genetic similarity in trees (phenograms) or networks where branch lengths need to be interpretable. A better approximation (Reynolds et al., 1983) of the linear relationship with time is given by

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

Reynold's *D* is non-metric.

There is some confusion arising from the translation of Bray-Curtis Distance (Bray & Curtis, 1957) from an ecological perspective to a genetics perspective. The Bray-Curtis Distance applies to abundances and down-weights joint absences in a manner analogous to the Jaccard Distance. When applied to SNP data, the Bray-Curtis Distance uses the abundance of the reference allele which renders it asymmetric with respect to the arbitrary choice of which is reference and which is alternate allele. Bray-Curtis Distance defined in this way should thus not be used on SNP data. Some authors have suggested that the Bray-Curtis equation be applied separately to the counts of each allele, averaged for the locus, for computing genetic distance populations (Allele Frequency Difference or AFD, Berner, 2009; Bray-Curtis/Allele Frequency Difference or BCAFD, Sherwin, 2022). However, when this formulation is applied to either individuals or populations, it becomes algebraically equivalent to the Czekanowski Distance (=Manhattan D). It no longer has the properties of Bray-Curtis Distance (being neither non-metric nor asymmetric). To avoid confusion, members of this class of distances should not be referred to as Bray-Curtis, AFD or BCAFD, but rather as Czekanowski Distance or the more familiar Manhattan D, given that naming of a distance measure should not be context dependent. The same issue has arisen in ecology (Yoshioka, 2008). Czekanowski Distance applied to SNP data and corrected for maximum value dependency is referred to as ^{A}A Distance (Sherwin, 2022).

Table 2. Some genetic distances commonly applied to populations. 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. q_{Ai} and q_{Bi} are similarly defined for the reference alleles. L is the number of called loci.

Name	Formula	Reference
Standard Euclidean Distance	***Binary P/A*** $D_E(A,B) = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(p_{Bi} - p_{Ai})^2}$ ***SNP genotypes*** $D_E(A,B) = \frac{1}{2}\sqrt{\frac{1}{L}\sum_{i=1}^{L}(p_{Bi} - p_{Ai})^2}$	
Czekanowscki Distance (Manhattan Block)	$D_{cz}(A,B) = \frac{1}{2L}\sum_{i=1}^{L}(\lvert p_{Bi} - p_{Ai}\rvert)$	(Berner, 2009, as AFD; Czekanowski, 1913)
Nei Standard Genetic Distance	$D_{Nei}(A,B) = -ln\left(\frac{\sum_{i=1}^{L}([p_{Ai}p_{Bi} + q_{Ai}q_{Bi}]}{\sqrt{\sum_{i=1}^{L}[(p_{Ai}^2 + q_{Ai}^2]}\sqrt{\sum_{i=1}^{L}[(p_{Bi}^2 + q_{Bi}^2]}}\right)$	(Nei, 1972)
Reynolds Distance	$D_{Reynolds}(A,B) = \sqrt{\frac{\sum_{i=1}^{L}[(p_{Ai} - p_{Bi})^2 + (q_{Ai} - q_{Bi})^2]}{2\sum_{i=1}^{L}(1 - p_{Ai}p_{Bi} - q_{Ai}q_{Bi})}}$	(Reynolds et al., 1983)
Chord Distance	$D_{Chord}(A,B) = \sqrt{1 - \frac{1}{L}\sum_{i=1}^{L}\sqrt{p_{Ai}p_{Bi}} + \sqrt{q_{Ai}q_{Bi}}}$	(Edwards & Cavalli-Sforza, 1964)
Wright's paired F_{ST}	Paired F_{ST} requires estimation with approaches such as that used by R package *hierfstat* [genet.dist(gl, method="WC84")]	(Wright, 1951)

Chord Distance (Edwards & Cavalli-Sforza, 1964) (Table 2) assumes divergence between populations is via drift alone, and so again may be more appropriate than Nei's *D* for spatial population genetics. Chord Distance is based on Geodesic or Angular Distance (Bhattacharyya, 1946; Edwards, 1971; Edwards & Cavalli-Sforza, 1967),

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

where α is the Angular Distance. For a geometric interpretation using biallelic data, see Nei & Kumar (2000:267). Chord Distance approximates Angular Distance by replacing the arc distance with the length of the corresponding straight-line segment (Edwards & Cavalli-Sforza, 1964) (Table 2). It is a metric distance that can be transformed to be approximately Euclidean (Edwards, 1971). As with Reynold's *D*, Chord Distance is proportional to shallow divergence time (for relatively small values of D) under specified assumptions (Edwards, 1971). It may be preferred over the model-free Standard Euclidean Distance where an underlying genetic model of divergence with time is preferred.

Wright's *F-statistics* (Wright, 1951) describe the distribution of genetic diversity within and between populations (Holsinger & Weir, 2009). *F-statistics* are defined in terms of the proportion of heterozygotes observed (H_{obs}) and the proportion of heterozygotes expected under Hardy-Weinberg equilibrium (H_{exp}), as follows:

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

A deficit in the observed proportion of heterozygotes compared with that expected under Hardy-Weinberg equilibrium will yield a Wright's *F* that deviates from zero. If, at a single locus, observed and expected heterozygosity are in agreement, then *F=0*. If, at the extreme, no heterozygotes are observed, then *F=1*. If there is an excess of heterozygotes, F will be less than zero. When *F* is averaged across a large number of independent loci for a population, *F* will typically fall between zero and one. If two populations that differ in allelic profiles are combined (pooled), then Hardy-Weinberg equilibrium is not a sensible null expectation. Divergence of the two allele frequency profiles will manifest as non-random mating and a departure from Hardy-Weinberg equilibrium (the "Wahlund Effect", Wahlund, 1928). Wright's *F* applied to the pooled populations is thus informative when examining population subdivision, because the resultant deficit in heterozygotes is an indication of genetic structure. *F* applied in this way incorporates two components -- one arising from departure from Hardy-Weinberg expectation within each population and the second arising from departure from Hardy-Weinberg expectation because of structure between the two populations or sub-populations. For this reason, Wright's pairwise F_{ST} is most commonly used as a measure of genetic distance between two populations or sub-populations (see Nei, 1977; Nei & Kumar, 2000:236). F_{ST} *is* a measure of the reduction in heterozygosity attributable to differences in allelic frequency profiles between populations or sub-populations, having partitioned out the contributions from departure from Hardy-Weinberg expectation within populations or sub-populations. Although not a genetic distance in the strict sense, F_{ST} can be viewed a non-metric distance that varies in value between 0 and 1. Several methods to estimate F_{ST} have been developed and some are complex (Excoffier, 2001; Weir & Cockerham, 1984), but there are

many software packages available to estimate F_{ST} for SNP genotype data (e.g. R package hierfstat, Goudet, 2005).

Finally, SNP genotype data at the population level can be converted to binary data by counting the shared alleles between two populations *A* and *B* at a locus

$$c_{A,B} = 0, \text{no alleles are shared by the two populations at that locus}$$
$$= 1, \text{one or both alleles are shared at that locus}$$

whereby distance measures devised for binary data can be applied. These distances are, in effect, considering only fixed allelic differences between the two populations.

Missing values

Missing data are problematic for distance analyses. Techniques like classical PCA that access the raw data matrix cannot accommodate missing data. PCoA, which accesses a distance matrix, is affected in ways that are not immediately transparent – potentially breaking metric or Euclidean properties, and varying the precision of estimates of population allele frequencies across loci. The trade-off is one of balancing data loss with bringing distortion of the visual representation within acceptable levels.

SNP datasets typically have substantial numbers of missing values. With SNP genotypes, missing data can arise because the read depth is insufficient to detect SNP-containing sequence tags consistently, or because mutations at one or both of the restriction enzyme recognition sites in some individuals result in allelic drop-out (null alleles). Because those that arise from mutation are inherited, they are subject to genetic drift (and potentially local selection) differentially within each sampled population, and so are not randomly distributed across the entire dataset. Indeed, in some populations the mutation(s) leading to missing values may become fixed. Filtering on call rate with a threshold applied globally will potentially admit high frequencies of missing data at particular loci at the level of single populations, which has consequences as outlined below.

In SNP sequence tag presence-absence data, the null alleles are themselves the data, scored as presence or absence of the amplified tag. In this context, missing values arise because the read depth is insufficient to be definitive about the absence of a particular sequence tag.

Because classical PCA will not accept missing values, when a locus is not scored for a particular individual, either the data for the entire locus must be deleted or the data for the entire individual must be deleted. This is clearly very wasteful of information, and unacceptable loss when working with SNP datasets. The data loss can be controlled 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 may still need to be managed. Numerous ways for handling missing data in PCA have been suggested (Dray & Josse, 2015), 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).

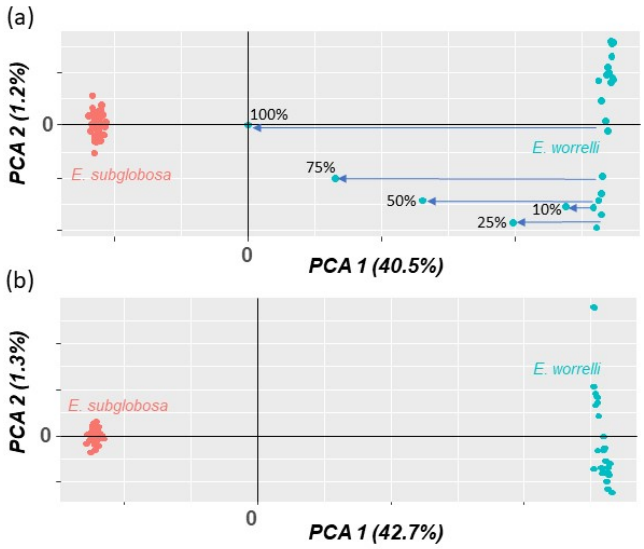

Figure 5. Principal Components Analysis (PCA) applied to two populations where five individuals in one population (*Emydura worrelli*) have had percentages of SNP loci ranging from 10% to 100% set to missing. **(a)** – the impact of this on PCA where missing data are filled with the average global allele frequencies (mean-imputed missing data) is clear, and subject to misinterpretation as hybridization or various levels of introgression. **(b)** – the issue can be resolved by local imputation, in this case by nearest-neighbour imputation.

Mean-imputation of missing data can lead to the individuals (or samples or specimens) with a high proportion of missing data being drawn out of their natural grouping and toward the origin, leading to potential misinterpretation (Yi & Latch, 2021). 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 when the missing values are replaced by the global average allele frequencies (Figure 5). Its location intermediate to the two clusters might be falsely interpreted as a case of admixture. Individuals with missing data corrected by mean-imputation will also distort confidence envelopes as applied to clusters with consequences for interpretation (Figure 6).

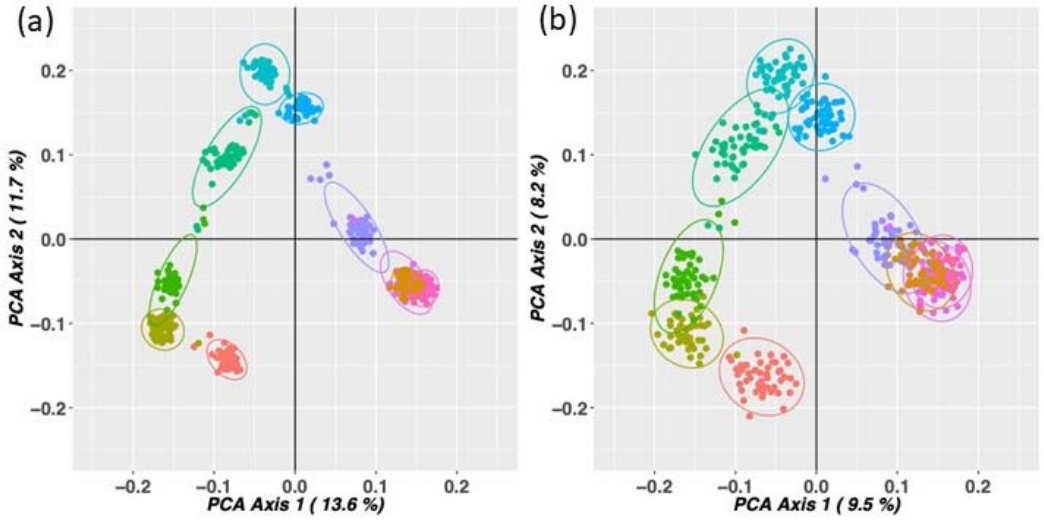

Figure 6. A PCA plot of simulated data showing aggregations and their associated 95% confidence limits for (a) data with no missing values and (b) data with 50% missing values. Note the inflationary effect missing values have on the spread of the points in each aggregation, manifested as inflation of the 95% confidence limits, an artifact. This is of particular relevance to studies of population assignment.

Better ways than global mean-imputation for handling missing values in SNP datasets include local filtering for call rate or local imputation. There are several options:

1. Apply a threshold for an acceptable call rate separately to each population, noting that a missing value rate of <10% causes only modest distortion (Figure 5a). Filtering on call rate by population with a threshold of 95% can be expected to constrain the distortion satisfactorily.

2. Replace missing values for an individual with the mean observed 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.

3. Replace missing values for an individual with the expected value for the population to which the individual belongs, based on the assumption of Hardy-Weinberg equilibrium. The individual will again be displaced toward the centroid of the population from which it was drawn, not the origin of the PCA.

4. Replace missing values for an individual by the value at the same locus from its nearest neighbour (the individual closest to the focal individual based on Standard Euclidean genetic distance) (Beretta & Santaniello, 2016). The focal individual will be drawn toward its nearest neighbour, typically an individual within the same population. This method has the advantage of filling missing values even where all individuals in a population are missing for a given locus.

Strictly, methods of local imputation should be applied to each 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 each aggregation that appears as distinct in a preliminary PCA.

If imputation is not desirable or heavy filtering on call rate considered too wasteful of data, an alternative approach is to apply pairwise deletion of loci rather than the global deletion dictated by classical PCA. This can be done by calculating a matrix of Standard Euclidean distances for individuals taken pairwise, removing loci with a missing value for one or both individuals. Principal Coordinates Analysis (PCoA) can then be applied to the distance matrix to deliver the ordination. This approach capitalizes on the observation that PCA and PCoA, using Standard Euclidean distance, yield the same visualizations (Cox & Cox, 2001:43-44). 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.

Linkage

Distance analyses usually assume that each locus contributes independent information to the overall distance value (i.e. they segregate independently). With the large genotype arrays typical of SNP datasets, some SNP loci are likely to be linked (Waples et al., 2022). In extreme cases, linkage disequilibrium can seriously distort the genetic structure, confounding interpretation. Large blocks of sequence with limited haplotype variation (presumably a result of limited recombination) have been observed in humans (Daly et al., 2001; Patil et al., 2001) and plants (Battlay et al., 2022). If many markers have been sequenced in such a haploblock, they will be tightly linked, and support for any population structure that they represent will be proportionately inflated. This will be evident in a PCA or PCoA as artifactual structure that will potentially defy explanation when working with organisms with little genomic information. An example of such artifactual structure in a PCA plot generated from linked SNP markers is provided by polymorphism in a large inversion in human chromosome 8. SNP markers associated with this inversion generate a coordinated signal which manifests as a three-group pattern, one corresponding to inverted homozygotes, one corresponding to heterozygotes and one corresponding to non-inverted homozygotes (Amos & Ma, 2012; Battlay et al., 2022). Such an explanation was invoked to explain the disaggregation of individuals of the Australian dragon *Pogona vitticeps* into two mirrored clusters that could not be explained by location of capture, sex effects or batch effects (Wild et al., 2022). The signature three-group pattern characteristic of a large polymorphic inversion was evident also in a SNP study of genetic variation across swamps occupied by the Australian Blue Mountains Water Skink *Eulamprus leuraensis* (Figure 7).

Linkage arising from SNPs residing on a shared, non-recombining block in the genome can be resolved by identifying loci that are strongly correlated with the Principal Component that separates out the artifactual groupings and removing those loci from the analysis (Wild et al., 2022).

Relatedness among individuals

Representing distance between populations on the basis of a sample is subject to random sampling error provided the individuals are sampled at random and their genotypes are independent. However, systematic errors can arise if some sampled individuals are more closely related than are individuals selected at random from their population; these individuals can be expected to separate out from the main body of individuals in a PCA or PCoA (Figure 8a). This can occur if for example, individuals are selected from a single school of fish that may be more closely related than individuals chosen at random. The effect can be pronounced if parents and their offspring (siblings) are among the sampled individuals (Figure 8a). These closely related individuals should be identified and all, but one removed from the analysis.

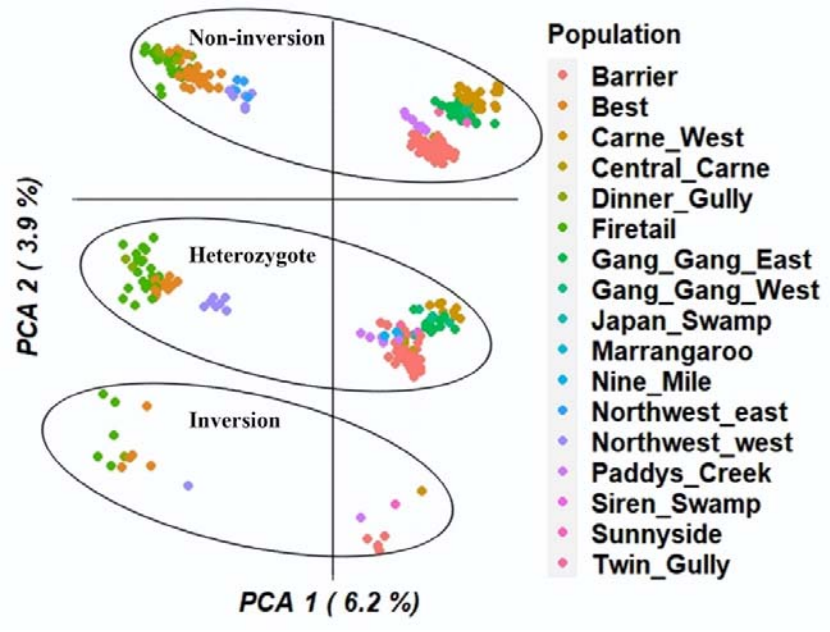

Figure 7. An ordination of SNP genotypes for the Australian Blue Mountains Water Skink *Eulamprus leuraensis* collected from 17 locations with varying levels of isolation. The ordination shows the three-group structure characteristic of a large autosomal inversion (Amos & Ma, 2012). For sake of illustration, we have assumed that the inversion is the least frequent of the two polymorphic states. This interpretation remains speculative until sufficient genomic information is available for this species to demonstrate the existence of the inversion and to associate the SNP loci strongly correlated with PCA 2 to that inversion.

When comparing two or more populations, relatedness among sampled individuals to a greater degree than among individuals in their population as a whole will potentially affect the estimate of genetic distance between that population and others. This can lead to misinterpretation by artificially displacing the affected population from what otherwise would

be a spatio-temporal trend. An important assumption of these analyses is that the sampled individuals are drawn at random from their populations and that their genotypes are independent. Datasets should be screened and filtered for close familial relationships in studies of spatio-temporal genetic variation at the level of populations. Care should also be exercised when interpreting clinal structure if the unit of dispersal comprises groups of related individuals (Fix, 1997).

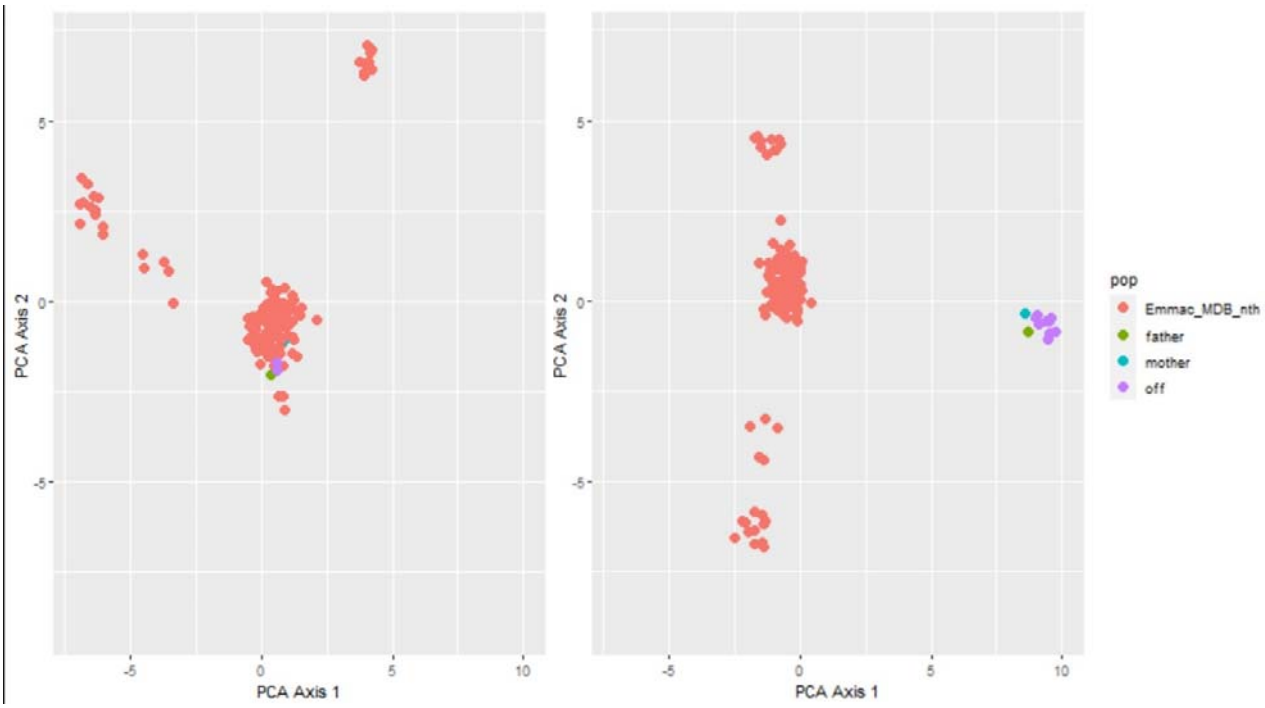

Figure 8. A series of sub-populations of *Emydura macquarii* from the northern basin of the Murray-Darling drainage (a) in which one subpopulation (central) has two individuals artificially added to be in a parent-offspring relationship versus eight individuals in a parent-offspring relationship (b). If such closely related individuals go undetected, and are retained, the spatio-temporal distance relationships between populations can be subject to misinterpretation.

Discussion

Species and populations usually do not constitute a single panmictic unit where individuals breed at random over their entire range. Population subdivisions typically exist, which may be hierarchically arranged if they reflect patterns of ancestry and descent, or not if they are each on independent random walks, perhaps periodically reset by episodic gene flow (Georges et al., 2018). Human genomic studies, where contributions of alleles to a phenotype are statistically measured using genome-wide association studies, require samples to be from one population without any substructure (Uffelmann et al., 2021). However, this is rarely the case given that self-reported ancestries and definition of social groups do not always capture the underlying genetics owing to admixture resulting in continuous population structure.

Genetic distance measures and their ordination using the continuous variation from PCA axes can be a powerful way to adjust genotypes and phenotypes based on ancestry to compute association statistic (Tian et al., 2008).

Isolation by geographic distance (Wright, 1943) in a homogenous and continuous landscape is perhaps the simplest instance of departure from panmixia in a widespread population, and is typically examined by comparing a genetic distance matrix with geographic distance between individuals, sub-populations or sampling localities. FST and Chord Distance are often chosen as the measures of genetic distance to examine any relationship with geographic distance (Séré et al., 2017). Such a relationship is often used as the null hypothesis against which to examine more complicated hypotheses to explain patterns of variation across the landscape (Manel et al., 2003; Spear et al., 2015). In particular, isolation by distance can be complicated by variation in resistance to gene flow in complex landscapes (McRae, 2006). Introduction of the concept of landscape resistance delivers a continuous set of circumstances between genetic structure arising from isolation by distance in otherwise freely breeding populations, and subdivision between populations of a species where gene flow is absent, rare or an episodic event. SNP studies of structure across the landscape typically use Standard Euclidean Distance (or Rogers D) for genotype data and Simple Matching Distance or Jaccard Distance for fragment presence-absence data. More novel applications of genetic distances include stratified sampling of a core collection of specimens from a larger pool of accessions to maximally capture genetic diversity (Rogers D: Jansen & van Hintum, 2007) and tracking the relationships of SARS-CoV-2 variants (Jaccard D: Yin, 2020).

While several distance measures are available for different types of genetic data (Libiger et al., 2009), the characteristics of data generated as SNP genotypes and restricting the scope to contemporary rather than historical processes governing variation among populations, dramatically reduce the options from which to select an appropriate distance. SNP markers are biallelic in practice, so distinct distance measures in a multiallelic context can become the same in a biallelic context. Additionally, symmetry considerations in the arbitrary choice of reference and alternate allele eliminate from use of distance measures that treat scores of homozygous reference (0) as fundamentally different from scores of homozygous alternate (2). Standard Euclidean Distance is typically the default choice of a genetic distance (equivalent to Roger's D). It is free of assumptions about the processes that generate the variation and, in the biallelic case, does not suffer the distortions that arise in multiallelic data when two populations are polymorphic at a site, but share no alleles (Nei & Kumar, 2000:246). Other distance measures may be selected because they consider underlying processes of mutation and genetic drift (e.g. Chord Distance), even though some lack some fundamental properties of a distance (e.g. Nei's D, Reynolds D). Reynold's Distance or Chord Distance for example may be preferred when time-dependence of the distances is important, such as when interpretable branch lengths on a phenogram are desirable.

Departures from the properties of a metric distance or Euclidean distance can present difficulties in faithfully representing those distances in an ordinated Cartesian space, but this can be overstated, as it is likely to be of practical consequence only in extreme cases such as when the sum of the negative eigenvalues is of a magnitude comparable to that of the

dimensions retained in the final solution (typically top 2 or 3). These departures can be diagnosed by examining eigenvalues, and the presence of negative eigenvalues can be redressed by simple transformations if required. Essentially, the distortion arising from only considering the top two or three dimensions from among a number of informative dimensions is likely to greatly exceed the influence of a few negative eigenvalues.

Binary data generated from the success or failure of sequence tag amplification in studies based on representational sequencing with restriction enzymes (e.g. ddRAD, DArTSeq) have true absences (0). For data such as these, Euclidean Distance (and its square, Simple Matching Distance) are complemented by other distance measures (e.g. Jaccard Distance) that down-weight joint absences should this be considered desirable. Consideration needs to be given on which state should be considered an absence if applying Jaccard Distance.

Arguably, missing values, and how they are managed, is the issue with the most significant consequences for calculating genetic distances and their visualisation. Classical PCA requires a complete input dataset, and filtering missing values requires removal of data from entire loci or individuals from the analysis. To overcome unacceptable loss of data, the missing values are typically infilled using the global average for the locus concerned. This can result in misleading displacement of individuals with high rates of missing data away from their natural groupings and toward the global centroid, with the risk of misinterpretation. It can also inflate confidence envelopes for aggregations of entities in a PCA, with serious consequences for analyses such as population assignment. We have reviewed the approaches to deal with missing values that avoid these distortions, the most effective of which appears to be replacement of a missing value with that of the nearest neighbour. Even replacement with a random value is preferable to replacement with the global average allele frequency (the most common approach) because random values simply add noise to the data which is driven to lower dimensions in any ordination.

Missing values can be managed on a pairwise basis, noting that PCA and PCoA with Euclidean Distance yield the same outcome. PCoA is less affected by missing values because they are eliminated from the computations taking the entities a pair at a time. However, PCoA is not immune to distortion from missing values, as their presence destroys the metric and Euclidean properties of the distance measure, and therefore the ability of PCoA to faithfully represent the distance matrix in a space defined by Cartesian coordinates. There are also issues with the variance of distances calculated locus by locus when the number of values differs among loci because of missing values.

In summary, we recommend selection of a distance measure for SNP genotype data that does not give differing outcomes depending on the arbitrary choice of reference and alternate alleles, and careful consideration of which state should be considered as zero when applying binary distance measures to fragment presence-absence data. Diagnostic examination of eigenvalues should be undertaken when a non-metric distance has been selected, or if the analysis is to include substantial missing values. Action should be taken if the sum of negative eigenvalues is substantial, to avoid distortion in the final visual representation. We strongly recommend filtering heavily on missing values, then imputing those that remain to create a full

matrix prior to undertaking a distance analysis. Failure to do so can substantially and artificially inflate confidence envelopes for populations or aggregations of populations and lead to other distortions that can lead to misinterpretation. Screening for closely related individuals (parent-offspring or sib relationships) is also important, and the impact of polymorphic haploblocks in the genomes of target species or populations occasionally emerges as a challenge for studies of species with limited genomic information.

Data Availability

Real datasets and the scripts used to generate simulated datasets used in this paper are available on Dryad (URL to be provided on acceptance).

Author Contributions

All authors contributed to the development of ideas presented in this manuscript. AG led the writing, LM and BG undertook the simulations, HP provided input on human studies, MA provided access to the skink data and provided associated context.

Acknowledgements

We would like to thank Bill Sherwin and Fred Allendorf for comments on an earlier draft of this manuscript, and for assisting with setting the scope and direction of the manuscript. We would like also to thank Springvale Coal Pty Ltd for allowing the publication of genetic data obtained for the Blue Mountains Water Skink Research and Management Program coordinated by RPS Australia East Pty Ltd for the Springvale Extension Project.

References

Amos, J., & Ma, C. I. (2012). Investigation of inversion polymorphisms in the human genome using Principal Components Analysis. *PLoS One*, *7*, e40224.

Battlay, P., Wilson, J., Bieker, V. C., Lee, C., Prapas, D., Petersen, B., Craig, S., van Boheemen, L., Scalone, R., de Silva, N. P., Sharma, A., Konstantinović, B., Nurkowski, K. A., Rieseberg, L., Connallon, T., Martin, M. D., & Hodgins, K. A. (2022). Large haploblocks underlie rapid adaptation in an invasive weed. *BioRxiv*, doi: https://doi.org/10.1101/2022.03.02.482376.

Belbin, L. (1991). Semi-strong hybrid scaling, a new ordination algorithm. *Journal of Vegetation Science*, *2*, 491–496.

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, *16 (Suppl.*(74), 198–208.

Berner, D. (2009). Allele frequency difference AFD–an intuitive alternative to Fst for quantifying genetic population differentiation. *Genes*, *10*, 308. https://doi.org/10.3390/genes10040308

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhya*, *7*, 401–406.

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

Cailliez, F. (1983). The analytical solution to the additive constant problem. *Psychometrika*,

48, 305–308.

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

Cangelosi, R., & Goriely, A. (2007). Component retention in principal component analysis with application to cDNA microarray data. *Biology Direct*, *2(2)*, https://doi.org/10.1186/1745-6150-2-2.

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

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

Cox, T. F., & Cox, M. A. A. (2001). *Multidimensional scaling* (2nd ed.). Chapman & Hall.

Czekanowski, J. (1913). *Zarys Metod Statystycznych w Zastosowaniu do Antropologii [Outline of statistical methods: as applied to anthropology]*. Prace Towarzystwa Naukowego Warszawskiego [Works of the Warsaw Scientific Society].

Daly, M., Rioux, J., Schaffner, S., Hudson, T., & Lander, E. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics*, *29*, 229–232.

de Ley, J., Cattair, H., & Reynaerts, A. (1970). The quantitative measurement of DNA hybridization from renaturation rates. *European Journal of Biochemistry*, *12*, 133–142.

Deza, M. M., & Deza, E. (2009). *Encyclopedia of Distances*. Spinger-Verlag.

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

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

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

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

Edwards, A. W. F., & Cavalli-Sforza, L. L. (1967). Analysis, phylogenetic procedures, models and estimation. *American Journal of Human Genetics*, *19*, 233–257.

Excoffier, L. (2001). Analysis of population subdivision. In D. J. Balding, M. Bishop, & C. Cannings (Eds.), *Handbook of Statistical Genetics* (pp. 271–307). John Wiley & Sons Ltd.

Faith, D. P. (1985). A model of immunological distances in systematics. *Journal of Theoretical Biology*, *114*, 511–526.

Fix, A. G. (1997). Gene frequency clines produced by kin-structured founder effects. *Human Biology*, *69*, 663–673.

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

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

Georges, A., & Adams, M. (1992). A phylogeny for australian chelid turtles based on allozyme electrophoresis. *Australian Journal of Zoology*, *40*(5). https://doi.org/10.1071/ZO9920453

Georges, A., Gruber, B., Pauly, G. B., White, D., Adams, M., Young, M. J., Kilian, A., Zhang, X., Shaffer, H. B., & Unmack, P. J. (2018). Genome-wide SNP markers breathe new life into phylogeography and species delimitation for the problematic short-necked

turtles (Chelidae: Emydura) of eastern Australia. *Mol Ecol*. https://doi.org/10.1111/mec.1492510.1111/mec.14925.

Goudet, J. (2005). Hierfstat, a package for R to compute and test variance components and F-statistics. *Molecular Ecology Notes*, *5*, 184–186.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, *53*, 325–338.

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

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

Gruber, B., Unmack, P. J., Berry, O. F., & Georges, A. (2018). dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources*, *18*(3), 691–699. https://doi.org/10.1111/1755-0998.12745

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

Hirayama, H., Tamaoka, J., & Horikoshi, K. (1996). Improved immobilization of DNA to microwell plates for DNA–DNA hybridization. *Nucleic Acids Research*, *24*, 4098–4099.

Holsinger, K., & Weir, B. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, *10*, 639–650. https://doi.org/10.1038/nrg2611

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417–441.

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

Jackson, D. A. (1993). Stopping rules in Principal Components Analysis: A comparison of heuristical and statistical approaches. *Ecology*, *74*, 2204–2214.

Jansen, J., & van Hintum, H. (2007). Genetic distance sampling: a novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. *Theoretical and Applied Genetics*, *114*, 421–428.

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I. Artificial data. *Applied Statistics*, *21*, 160–173.

Jolliffe, I. T. (2002). *Principal Component Analysis* (Second). Springer International Publishing.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, *374*, 20150202.

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

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

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(94), 1–15.

Kirsch, J. A., Springer, M. S., Krajewski, C., Archer, M., Aplin, K., & Dickerman, A. W. (1990). DNA/DNA hybridization studies of the carnivorous marsupials. I: The intergeneric relationships of bandicoots (Marsupialia: Perameloidea). *Journal of Molecular Evolution*, *30*, 434–438.

Kruskal, J. B. (1964). Non-metric multidimensional scaling: a numerical method. *Psychometrika*, *29*, 115–129.

Legendre, P., & Legendre, L. (2012). Numerical Ecology. *Developments in Environmental Modelling*, *24*, 1–990.

Levandowsky, M., & Winter, D. (1971). Distance between sets [5]. *Nature*, *234*(5323), 34–35. https://doi.org/10.1038/234034A0

Libiger, O., Nievergelt, C. M., & Schork, N. J. (2009). Comparison of genetic distance measures using human SNP genotype data. *Human Biology*, *81*, 389–406.

Lin, Z., Yang, C., Zhu, Y., Duchi, J., Fu, Y., Wang, Y., Jiang, B., & Zamanighomi, M. Xuming Xu, Mingfeng Li, Nenad Sestan, Hongyu Zhao hongyu.zhao@yale.edu, and W. H. W. (2016). Simultaneous dimension reduction and adjustment for confounding variation. *Proceedings of the National Academy of Sciences*, *113*, 14662–14667.

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

Macarthur, R. (1957). On the relative abundance of bird species. *Proceedings of the Natlional Academy of Sciences USA*, *43*, 293–295.

Manel, S., Schwartz, M. K., Luikart, G., & Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution*, *18*, 189–197.

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

McRae, B. H. (2006). Isolation by resistance. *Evolution*, *60*, 1551–1561.

Mijangos, J., Gruber, B., Berry, O., Pacioni, C., & Georges, A. (2022). dartR v2: an accessible genetic analysis platform for conservation, ecology, and agriculture. *Methods in Ecology and Evolution*, *3*, 2150–2158.

Müller, T., Selinski, S., & Ickstadt, K. (2005). Cluster analysis: a comparison of different similarity measures for SNP data. In *Technical Report, Universität Dortmund, Sonderforschungsbereich 475 - Komplexitätsreduktion in Multivariaten Datenstrukturen, Dortmund* (Vol. 14). http://hdl.handle.net/10419/22605

Nei, M. (1972). Genetic distance between populations. *American Naturalist*, *106*, 283–292.

Nei, M. (1977). F- statistics and analysis of gene diversity in subdivided populations. *Annals of Human Genetics*, *41*, 225–233.

Nei, M., & Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press.

Orlóci, L. (1978). *Multivariate Analysis in Vegetation Research* (2nd ed.). Dr W Junk Publishers.

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

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T. N., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R., Thomas, D. J., Trulson, M., Vyas, K. R., … Cox, D. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human Chromosome 21. *Science*, *294*, 1719–1723.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*, e190.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space.

Philisophical Magazine, *2*, 559–572.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How Many Principal Components? Stopping Rules for Determining the Number of Non-Trivial Axes Revisited. *Computational Statistics and Data Analysis*, *49*, 974–997.

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

Rogers, J. S. (1972). Measures of Genetic Similarity and Genetic Distance. *Studies in Genetics VII, University of Texas Publication, Austin, USA*, *7213*, 145–153.

Sanghvi, L. D. (1953). Comparison of genetical and morphological methods for a study of biological differences. *American Journal of Physical Anthropology*, *11*, 385–404.

Séré, M., Thévenon, S., Belem, A. M. G., & de Meeûs, T. (2017). Comparison of different genetic distances to test isolation by distance between populations. *Heredity*, *119*, 55–63.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, *27*, 125–139.

Sherwin, W. B. (2022). Bray-Curtis (AFD) differentiation in molecular ecology: Forecasting, an adjustment (AA), and comparative performance in selection detection. *Ecology and Evolution*, *12*, e9176. https://doi.org/10.1002/ece3.9176

Sibson, R. (1979). Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. *Journal of the Royal Statistical Society*, *41B*, 217–229.

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *28*, 409–1438.

Sørensen, T. (1948). 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. *Kongelige Danske Videnskabernes Selskab*, *5*, 1–34.

Spear, S. F., Cushman, S. A., & McRae, B. H. (2015). Resistance Surface Modeling in Landscape Genetics. In N. Balkenhol, S. A. Cushman, A. T. Storfer, & L. P. Waits (Eds.), *Landscape Genetics* (pp. 129–148). John Wiley & Sons Ltd.

Tian, C., Gregersen, P., & Seldin, M. (2008). Accounting for ancestry: population substructure and genome-wide association studies. *Human Molecular Genetics*, *17*(R2), R143-150.

Tracy, C., & Widom, H. (1993). Level-spacing distributions and the Airy kernel. *Physics Letters B*, *305*, 115–118.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vrie, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Review Methods: Primers*, *1*, 1–59.

Wahlund, S. (1928). Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas*, *11*, 65–106.

Wang, Y., Sun, F., Lin, W., & Zhang, S. (2022). AC-PCoA: Adjustment for confounding factors using principal coordinate analysis. *BMC Computational Biology*, *18*, e1010184.

Waples, R. S., Waples, R. K., & Ward, E. J. (2022). Pseudoreplication in genomic-scale data sets. *Molecular Ecology Resources*, *22*, 503–518.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population-structure. *Evolution*, *38*, 358–1370.

Wild, K. H., Roe, J. H., Schwanz, L., Georges, A., & Sarre, S. D. (2022). Evolutionary stability inferred for a free ranging lizard with sex-reversal. *Molecular Ecology*, *31*,

2281–2292.

Wright, S. (1943). Isolation by distance. *Genetics*, *8*, 114–138.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, *15*, 323–354.

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

Yin, C. (2020). Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics*, *112*, 3588–3596.

Yoshioka, P. M. (2008). Misidentification of the Bray-Curtis similarity index. *Marine Ecology Progress Series*, *368*, 309–310. https://doi.org/10.3354/meps07728