Heredity (2009), 1–12
© 2009 Macmillan Publishers Limited All rights reserved 0018-067X/09 $32.00
www.nature.com/hdy

REVIEW

Genetic markers in the playground of multivariate analysis

T Jombart, D Pontier and A-B Dufour
Université de Lyon, F-69000, Lyon, Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France

Multivariate analyses such as principal component analysis were among the first statistical methods employed to extract information from genetic markers. From their early applications to current innovations, these approaches have proven to be efficient for the analysis of the genetic variability in various contexts such as human genetics, conservation and adaptation studies. However, because multivariate analysis is a wide and diversified area of statistics, choosing a method appropriate to both the data and to the question being asked can be difficult. Moreover, some particularities of genetic markers need to be taken into account when using multivariate methods. As a consequence, multivariate analyses are often used as black boxes, which results in frequent mistakes in the literature. In this review, we provide a critical analysis of the application of multivariate methods to genetic markers, using a general framework that unifies all these methods for the sake of clarity. First, we focus on some common mistakes in these applications and ways to avoid these pitfalls. We then detail the most critical particularities of allele frequencies that demand adaptations of multivariate methods, and we propose solutions to the subsequent problems. Finally, we tackle several questions of interest in which multivariate analysis has a great role to play, such as the study of the typological coherence of different genetic markers, or the investigation of spatial genetic patterns.

Heredity advance online publication, 21 January 2009; doi:10.1038/hdy.2008.130

Keywords: multivariate analysis; ordination; genetic markers; principal component analysis; statistics; methods

Introduction

Statistical methods have long become an essential component of the toolbox of population geneticists (Fisher, 1952). Developments in statistical theories and the continual increases in cheap computing power provide numerous tools for genetic marker analysis, allowing geneticists to address new and challenging questions. *Multivariate analyses* (also called *ordinations in reduced space*) such as principal component analysis (Pearson, 1901) have been shown to be efficient in extracting information from genetic markers (Cavalli-Sforza, 1966; Johnson *et al.*, 1969; Smouse *et al.*, 1982) because of their ability to summarize multivariate genetic information into a few synthetic variables. From these early applications to current innovative developments (Patterson *et al.*, 2006; Pavoine and Bailly, 2007; Jombart *et al.*, 2008), these methods have proven to be useful in various fields, such as human genetics (Menozzi *et al.*, 1978; Bertranpetit and Cavalli-Sforza, 1991; Cavalli-Sforza *et al.*, 1993), conservation (Moazami-Goudarzi *et al.*, 1997; Escudero *et al.*, 2003; Laloë *et al.*, 2007), phylogeography (Hanotte *et al.*, 2002; Matsuoka *et al.*, 2002; Ciofi *et al.*, 2006), landscape genetics (Angers *et al.*, 1999; McRae *et al.*, 2005) and the identification of adaptations (Johnson *et al.*, 1969; Mulley *et al.*, 1979; Barker *et al.*, 1986).

Multivariate analysis has several advantages over other classical approaches used in population genetics, like the Bayesian clustering implemented in the software STRUCTURE (Pritchard *et al.*, 2000; Falush *et al.*, 2003). First, multivariate methods are exploratory, that is, they do not require strong assumptions about an underlying genetic model, such as the Hardy–Weinberg equilibrium or the absence of linkage disequilibrium. Although clustering approaches suppose that genotypes are structured in discrete populations, ordinations in reduced space simply aim at summarizing the genetic variability, and can therefore reveal any kind of genetic structuring including clines (for example, Jombart *et al.*, 2008). Multivariate methods are not computer-intensive, and can be applied to huge datasets (such as *'hundreds of thousands of markers and thousands of samples'* in Patterson *et al.* (2006)), for which Bayesian clustering would be impractical. Moreover, multivariate analysis can address complex questions such as identification of adaptation, by linking genetic variability to environmental data (Barker *et al.*, 1986; Angers *et al.*, 1999), whereas the impossibility of formulating an explicit model of adaptation would make Bayesian clustering methods inapplicable, in most cases. Lastly, multivariate methods have been developed and used extensively for more than a century in various fields, such as psychometry and ecology (Pearson, 1901). Currently, multivariate analysis represents a whole, rich and diversified area of statistics offering a wide choice of methods, each with its own

Correspondence: Dr T Jombart, UMR CNRS 5558–LBBE, 'Biométrie et Biologie Évolutive', UCB Lyon 1—Bât. Grégor Mendel, 43 bd du 11 novembre 1918, 69622 VILLEURBANNE cedex, France.
E-mail: jombart@biomserv.univ-lyon1.fr

Received 25 July 2008; revised 24 November 2008; accepted 8 December 2008

properties (Takeuchi *et al*., 1982; Jambu, 1991; Legendre and Legendre, 1998).

The unfortunate consequence of this diversity of methods is that multivariate analyses are often used as black boxes when applied to genetic markers, leading to frequent mistakes that sometimes question the results of an entire study. In fact, it can be difficult to know which method can be efficiently applied to extract information from genetic markers, what precautions need to be taken and how the results should be interpreted. Moreover, there is no doubt that multivariate analysis has been under-utilized and has much more to offer to the study of the genetic variability. The purpose of this paper is to critically review the use of ordination in reduced space to infer biological structures from genetic markers.

First, we attempt to clarify the rationale for these methods and provide an overview of their current application to genetic markers. Frequent mistakes regarding the utilization of these methods are then detailed, and guidelines are provided to avoid these pitfalls. The following section focuses on some particularities of genetic markers that should be taken into account to improve their multivariate analysis. The rest of this review covers the use of multivariate analyses to tackle specific questions of interest, such as the coherence of the information of different genetic markers, linkage of genetic markers to other types of data, and the study of spatial genetic patterns. We conclude by examining some promising perspectives offered by these approaches to answer challenging questions in various fields, such as conservation, spatial genetics and molecular ecology.

Multivariate analysis of genetic markers

Rationale of multivariate analysis

Throughout this paper, the terms 'ordination in reduced space' and 'multivariate analysis' are used interchangeably. However, the first term is certainly more accurate than the second because ordinations in reduced space represent a particular class of multivariate methods, another being, for instance, hierarchical clustering. The purpose of these methods is to summarize a strongly multivariate dataset into a small set of uncorrelated *synthetic variables*. In other words, ordinations in reduced space aim to provide a simplified, yet meaningful, picture of complex information that is impossible to perceive. This task implies a necessary loss of information, and the crucial point in all these methods is to define a criterion that is optimized by the synthetic variables seeked. For instance, in principal component analysis (Pearson, 1901; Takeuchi *et al*., 1982, pp 185–224), synthetic variables best preserve the variance among observations, whereas the χ^2 distances are preserved in the correspondence analysis (Greenacre, 1966). Below, we introduce general concepts required to describe multivariate analyses with accuracy.

As formalized by the *duality diagram* framework (Escoufier, 1987; Dray and Dufour, 2007), most multivariate analyses are particular cases of a general algorithm, and can be described using a small set of concepts. The terminology we employ encompasses the most common terms, which can be found in reference textbooks (for examples, Takeuchi *et al*., 1982; Jambu, 1991; Legendre and Legendre, 1998; Lebart *et al*., 2004). Central to the analysis of a dataset of n objects and p descriptors is the question of whether we seek a description of the relationships among the objects or among the descriptors. When analysing genetic markers, the main interest is in finding relationships among objects (genotypes or populations) using p alleles. In this case, data are seen as a cloud of n points embedded inside a p-dimensional space, where each dimension is defined by an allele. Inside this space, *inertia* measures the dispersion of n points with respect to a given distance: this measurement of variability is used as a criterion that is optimized by the analysis. The directions inside this space reflecting the highest 'variability' (that is, with maximum inertia) among objects are the *principal axes*, also referred to as the *factors* of the analysis. By extension, a plane formed by two principal axes is often called a *factorial plane*. Each principal axis is defined by p coordinates inside the p-dimensional space, representing the *loadings* of the alleles. The principal axes are orthonormal (that is, perpendicular and with length one), and can therefore be used as a new basis to represent the n objects. The set of coordinates of the objects in this new basis are the *principal components*, but the terms *scores* (of objects) and *synthetic variables* are also commonly used. Each principal component is associated with an *eigenvalue* that quantifies the amount of inertia contained in the component. Eigenvalues can also be expressed as proportions of the total inertia of the analysis to indicate what fraction of the entire genetic variability is represented by the corresponding principal components. The plot of the eigenvalues sorted in decreasing order (the screeplot) is the basic tool used to choose which principal components to interpret: it describes how the total inertia is distributed across the principal axes. The basic idea is that a boundary between true structure and random noise would be indicated by a sharp decay between two successive eigenvalues. However, this is a simplistic view, and such a boundary rarely exists in practice: the *screeplot* merely provides insight about which component likely contains interesting structures, and which does not. Hence, the screeplot and the proportions of inertia associated with the principal components are two complementary tools, respectively indicating the genetic structures to be retained and their magnitude. The last criterion for interpreting principal components is that of the biological meaning, and is sometimes more useful than statistical criteria. In some cases, the first principal components (associated to large inertia) may indicate a trivial structuring, and provide little biological insight. Conversely, principal components associated to smaller eigenvalues might contain biologically relevant information; the interpretation of such components should not be discarded on the basis of a small inertia.

If multivariate analyses are unified by a single algorithm, the core difficulty is in choosing the method that best matches the nature of the data and the questions asked. Because of the variety of questions and data, numerous ordinations in reduced space are used to analyse genetic markers.

Applications to genetic markers

Multivariate analyses are natural tools to extract biological structures from genetic markers, as these data typically contain large numbers of genotypes or

Table 1 Multivariate analyses applied to genetic markers

Method	*Criterion*	*Application*	*Data*
Principal component analysis (PCA)	Variance (same as squared Euclidean distances)	Cavalli-Sforza (1966)	Allozymes
Principal coordinates analysis (PCoA)	Any Euclidean distance	Sanchez-Mazas and Langaney (1988)	Allozymes
Non-metric dimensional scaling (NMDS)	Ordering of objects	Lessa (1990)	Roger's and Nei's distances
Correspondence analysis (CA)	χ^2 distance	She *et al.* (1987)	Allozymes
Discriminant analysis (DA)	Variance between groups/total variance	Smouse *et al.* (1982)	Allozymes
Constant-row total multiple correspondence analysis (CRT-MCA)	Correlation ratio	Guinand (1996)	Allozymes
Factor analysis (FA)	'Common effect' in allele frequencies	Taylor and Mitton (1974)	Allozymes
Canonical correspondence analysis (CCA)	χ^2 distances in predicted data	Angers *et al.* (1999)	Microsatellites
Redundancy analysis (RDA)	Variance of predicted data	Kölliker *et al.* (2008)	AFLP and SSR
Canonical correlation analysis (CCorA)	Squared correlation between pairs of scores	Johnson and Schaffer (1973)	Allozymes
Co-inertia analysis (COA)	Squared covariance between pairs of scores	Jarraud *et al.* (2002)	AFLP
Multiple co-inertia analysis (MCOA)	Squared covariance between a set of scores	Laloë *et al.* (2007)	Microsatellites
Spatial principal component analysis (sPCA)	Product of variance and spatial autocorrelation	Jombart *et al.* (2008)	Microsatellites

Abbreviations: AFLP, amplified fragment length polymorphism; SSR, single sequence repeats.
Each method is indicated by its most frequent name and abbreviation. The 'criterion' is the quantity optimized by the principal components of the method. The 'application' column gives the reference of an early and representative publication using the method to analyse genetic markers.

populations described by hundreds of alleles (in terms of absolute or relative frequencies). A summary of the application of these methods to genetic markers is provided in Table 1.

Ordinations in reduced space are primarily used to find a few principal components that reflect as much of the genetic variability as possible. PCA was first employed to infer population structuring (Cavalli-Sforza, 1966) and spatial genetic structuring (Menozzi *et al.*, 1978; Bertranpetit and Cavalli-Sforza, 1991; Cavalli-Sforza *et al.*, 1993) in humans. PCA was also used early to infer adaptations from allozyme frequencies, by testing the correlations between principal components of genetic data and principal components of a PCA of environmental variables (Johnson *et al.*, 1969). In disease studies, regression onto the principal components of the PCA has been recently proposed to correct for population stratification (Price *et al.*, 2006). Another method commonly used to infer genetic structuring among genotypes or populations is principal coordinates analysis (PCoA, Gower, 1966; Sanchez-Mazas and Langaney, 1988; Warnes, 2003). Although PCA preserves the canonical Euclidean distance among the studied entities, PCoA can be employed to summarize any Euclidean genetic distance between genotypes or populations, but does not provide a representation of the alleles. This offers the advantage of using measures of genetic variability that are directly related to a population genetics model; for instance, PCoA has been used to summarize matrices of pairwise F_{ST} (Zhivotovsky *et al.*, 2003) and of Roger's distance (Baker and Moeed, 1987). Non-metric dimensional scaling (NMDS, Cox and Cox, 2001) has also been employed to analyse matrices of genetic distances (Baker and Moeed, 1987; Lessa, 1990). However, NMDS differs from PCoA in that it attempts to preserve the ordering of objects based on their genetic distance rather than their genetic distance *per se*; in this respect, NMDS can be thought of as a non-linear form of PCoA (Lessa, 1990). It is noteworthy that unlike other multivariate analyses, the NMDS solution is not analytical: an iterative algorithm aims at finding a good solution, but does not guarantee that this solution is the best. As an alternative to PCA of allele frequencies and PCoA (or NMDS) of genetic distances, correspondence analysis (CA, Greenacre, 1966) can be used to analyse a table of allele counts per population (She *et al.*, 1987; Li *et al.*, 2002). The last multivariate analysis commonly applied to genetic markers is discriminant analysis (DA, Lachenbruch and Goldstein, 1979). DA is not a fully exploratory approach, in that groups of genotypes must be known in advance. However, it can be used to achieve the best discrimination between groups inside a reduced space, to test for genetic differentiation, and for assignment purposes (Smouse *et al.*, 1982; Beharav and Nevo, 2003).

Other methods have remained somewhat unnoticed, such as constant-row multiple correspondence analysis (CRT-MCA, Guinand, 1996; Guinand *et al.*, 1996), factor analysis (FA, Taylor and Mitton, 1974; Mulley *et al.*, 1979) and distance-based redundancy analysis (db-RDA, Legendre and Anderson, 1999; Geffen *et al.*, 2004). The reason for this may be historical, or could arise from problems associated with using these approaches. For instance, CRT-MCA aims at finding synthetic variables with maximum F_{ST}, but only proposes an approximate solution. Denoting f as a set of frequencies of an allele for q populations, $\bar{f}$ as the mean frequency computed across populations and $\mathrm{var}(f)$ as the variance between populations of f, F_{ST} is defined as $\mathrm{var}(f)/\bar{f}(1-\bar{f})$, where $\bar{f}(1-\bar{f})$ is the theoretical variance of f (Weir, 1996, p 166). Unfortunately, the quantity optimized by CRT-MCA is $\mathrm{var}(f)/s_f^2$, where s_f^2 is the empirical variance of f ($s_f^2 = \frac{1}{q}\sum_{i=1}^{q}(f_i - \bar{f})^2$). While for arbitrarily large samples s_f^2 converges towards $\bar{f}(1-\bar{f})$, these quantities differ in practice, and the principal components yielded by CRT-MCA do not optimize F_{ST}. A possible cause for the minimal use of FA is that it was introduced to correlate patterns in allele frequencies with environmental variables (Taylor and Mitton, 1974), which is not the purpose of this method. In fact, FA estimates a model in which allele frequencies are expressed as a sum of two components: a part common to every allele and a residual part representing allele-specific effects (Seal, 1966, pp 153–180). Lastly, it is not clear why db-RDA has not been applied more often to genetic markers, but this could simply be due to its recent application (Geffen *et al*., 2004).

Although multivariate analyses can be efficiently used to extract information from genetic markers, choosing a method appropriate to the data and the question being asked is sometimes difficult. As a matter of fact, a number of mistakes occur quite frequently in such applications. In the following, we point out the major pitfalls, as well as strategies to avoid them.

Misuses, misinterpretations and specific issues

Ensuring reproducibility

A first concern in data analysis is to ensure reproducibility, or at least to provide all the elements required to evaluate the relevance of the results. Unfortunately, the literature regularly provides examples of studies in which it is almost impossible to know which analyses were actually performed.

The first problem lies in the absence of an accurate description of the method used: reference articles are rarely cited, and abbreviations sometimes do not match the name of the method. For instance, 'PCA' is used to refer to principal coordinates analysis (PCoA) in Pariset *et al*. (2003). Such confusion adds to the ambiguities that already exist between some methods, such as those between PCoA and NMDS. PCoA is also sometimes called 'metric dimensional scaling' (MDS), whereas NMDS is indifferently abbreviated MDS or NMDS (Legendre and Legendre, 1998). This is all the more confusing since PCoA is routinely used to initialize the algorithm of NMDS (Baker and Moeed, 1987). Papers demonstrating an ambiguity between PCoA and NMDS are not uncommon (for example, Preziosi and Fairbairn, 1992; Zhivotovsky *et al*., 2003).

Although required, providing a correct reference to a method is usually not sufficient. Some methods exist in different variants, according to the initial transformations of the data. This is particularly true for PCA: although centring (subtracting the mean allele frequency from all observations) is always achieved, scaling of the alleles (dividing each observation by allelewise values) is optional and can be performed in several ways. Scaling can drastically change the results of a PCA, but is rarely disclosed (for example, Mitton, 1978; MacHugh *et al*., 1998; Grivet *et al*., 2008). In PCoA and NMDS, the genetic distance employed should always be specified, and in the case of NMDS, how the algorithm was initialized should be indicated. An example of such an application can be found in Baker and Moeed (1987), who used an NMDS initialized by a PCoA of Roger's distances of allozyme data to explore the genetic variation among populations of common minas (*Acridotheres tristis*). Lack of accuracy in the description of the method always complicates interpretation of the results, and sometimes brings their validity into question. For instance, some papers show principal components of a PCA that were clearly not centred (their range of variation did not include zero), which indicates an error in the computations of the analysis and invalidates the results (for instance, MacHugh *et al*., 1997, 1998; Pariset *et al*., 2003). Moreover, it is difficult to ascertain precisely where the problem came from, as the software used for the computation was not mentioned in these publications.

Making graphics

Another classical problem lies in the graphical display of results. As mentioned previously, the screeplot is the basic tool used to assess which principal components should be interpreted, but it is most often omitted in publications. The amount of inertia associated with each principal component is often indicated, but this information is complementary to the screeplot and cannot be used as a substitute. For instance, in their study of the genetic differentiation among different yak (*Poephagus grunniens*) populations, Xuebin *et al*. (2005) presented a scatterplot of PCA displaying 80% of the whole variability, but this scatterplot was merely uninformative in terms of genetic differentiation. Conversely, two principal components of PCA containing less than 10% of total inertia provided insights about the phylogeny of different maize subspecies in Matsuoka *et al*. (2002). When used alone, the amount of inertia can therefore be a misleading criterion for choosing the principal components to interpret (see 'Interpreting genetic structures').

Another widespread custom is the use of 3-dimensional scatterplots (van Pijlen *et al*., 1995; Xuebin *et al*., 2005). Although these representations add a fancy touch to multivariate analyses, they also have the unfortunate effect of sacrificing the mathematical properties of an analysis, and thus its interpretability. By definition, principal axes and the associated principal components provide the best possible planar representation of the data. If three principal components are retained, their representation requires two factorial planes, with one axis being redundant. Scatterplots in three dimensions are ultimately always viewed on a screen or on a sheet of paper, and are thus re-projections of three principal components in two dimensions. The obtained representations are necessarily worse than the true representation of principal components because they no longer have maximum inertia nor orthogonality. Hence, 3-dimensional visualization should be restricted to interactive data analysis (where it can be useful), and is better avoided in publications.

Apart from these pitfalls common to every multivariate analyses, some specific issues also arise when certain methods are applied to genetic markers.

Some specific issues

A first particular issue concerns the use of CA. This method is appropriate for the analysis of a contingency table, that is, a matrix of positive integers (Greenacre, 1966), and is thus appropriate for the analysis of a table of allele counts. A good example of such an application is provided by She *et al*. (1987), who used a CA of allozyme data to investigate the genetic differentiation between populations of teleost fishes. Interestingly, this study also showed that 'correspondences' highlighted by CA can reflect linkage disequilibrium existing between alleles. In some cases, CA has been used for allele (relative) frequencies (Li *et al.*, 2002), which has been proven to significantly alter the results of the analysis (Perrière and Thioulouse, 2002). In such a case, it seems much more appropriate to use PCA or PCoA. However, even when CA is correctly used, one should be aware that scarce descriptors are given a stronger weight by the χ^2 distance, which is optimized by the analysis (Legendre and Legendre, 1998, p 285). The typical consequence is that a population possessing a rare allele would appear as an outlier in CA components. Simple simulations show that such an artifactual pattern arises even when studying groups of genotypes are randomly chosen from the same population (results not shown). A way to avoid this problem is to remove rare alleles from the data prior to the CA, although this solution requires some investigations regarding which frequency should be considered as 'rare' from the point of view of the CA.

A second specific issue occurs in DA. This method finds principal components maximizing the variance between populations while keeping the variance inside populations constant, assuring optimal discrimination between the populations (Krzanowski and Marriott, 1995, pp 1–56). However, this method involves computation of the Mahalanobis metric (Beharav and Nevo, 2003), which is the inverse of the matrix of covariances between alleles. For this inverse to exist, the covariance matrix must be of full rank, that is, of rank p if there are p alleles (Harville, 1997, p 80). This is never the case for allele frequencies: each marker spans a space of at most one dimension less than the number of its alleles because any frequency is entirely defined by all the others. That is, if there are k markers, the rank of the covariance matrix is at most $\min(p-k, n)$. Thus, the discriminant analysis can only be performed on a matrix of allele frequencies after removing a given number of alleles, and assuring that there are more objects (genotypes or populations) than alleles. In fact, the number of objects n should be consequently larger than the number of alleles p: (Williams and Titus, 1988) reported that n should be at least three times larger than p for DA to yield reliable results. Multicollinearity can also exist among alleles (that is, when alleles are correlated), especially when linkage disequilibrium occurs. In these cases, the Mahalanobis metric is said to be ill-conditioned, resulting in numerical instability. As a result, principal axes and principal components of DA cannot be computed with accuracy, and small changes in allele frequencies induce large changes in the results (Seber, 1977, pp 319–322). As a consequence, the alleles used in DA should be carefully selected before performing the analysis. An empirical approach consists of retaining only the most frequent allele of each locus (Sagnard *et al.*, 2002), but this does not ensure that the subset of alleles obtained is optimal with regard to discrimination. In fact, it does not even ensure that the multicollinearity problem is solved, as linkage disequilibrium can still exist between the most frequent alleles. One can preferentially use statistical approaches that are especially devoted to the selection of variables in DA (Lachenbruch and Goldstein, 1979), where such approaches proved useful for selecting a subset of best discriminating alleles (Fahima *et al.*, 1999; Beharav and Nevo, 2003). However, investigations should be carried out to assess whether a particular variable selection procedure is preferable to the others in the case of allele frequencies.

Interpreting genetic structures

A major concern in multivariate analysis of genetic markers lies in interpreting the results. This issue can be illustrated by examining one case of misinterpretation, raising the question of which result of a multivariate analysis could be interpreted as genetic structuring.

In the controversy regarding the relevance of definiting human races based on genetic information (Lewontin, 1978; Mitton, 1978; Powell and Taylor, 1978), Mitton (1978) argued that genetic differentiation between 'human races' was important because they clearly appeared as distinct groups on the factorial map of a PCA. This misinterpretation of the results is related to the common mistake of not displaying the screeplot of the analysis along with the values of inertia associated with principal components. Ordinations in reduced space do not summarize the essential part of the genetic variability: they attempt to show as much genetic variability as possible in a few axes, which is different. Mitton (1978) showed that 'racial' groups were well separated on the factorial plane and that two principal components were sufficient to assign each population to a given group. However, this did not contradict the well-acknowledged fact that the genetic variability within 'races' is much larger than between 'races' (Edwards, 2003), as suggested by the author. For example, it would be possible to perfectly discriminate two populations using only one allele, but this allele may represent only 1% of the variability of a dataset containing 100 alleles. This point was discussed by (Edwards, 2003), who emphasized the fundamental difference between being able to assign genotypes to taxonomic groups, and observing larger genetic variability between these taxonomic groups.

We can ask what criterion the principal components of an analysis should meet to be considered as true genetic structuring. The relative amount of inertia cannot be used as a single criterion, because it depends directly on the number of alleles considered. As stressed previously, the screeplot can be used to assess which principal components likely contain interesting structures. Recently, Patterson *et al.* (2006) tested the significance of the eigenvalues from a PCA of genetic markers to infer population stratification. Another testing approach to select interpretable principal components of a PCA has been proposed by (Dray, 2008), and could also be used to identify significant genetic structures. It is noteworthy that both approaches are reserved to PCA (Patterson *et al.*, 2006; Dray, 2008), and it would be valuable to extend these tests to other multivariate methods. Another way of assessing relevant genetic structures emerging from an ordination method is to quantify the amount of genetic differentiation contained in the principal components. The main difficulty is then identifying clusters of genotypes from the principal components retained. This can be achieved using a given clustering algorithm (Legendre and Legendre, 1998, pp 303–381), such as the unweighted arithmetic average clustering (UPGMA, Rohlf, 1963). It is then possible to measure the amount of genetic differentiation between the obtained clusters of genotypes using classical approaches like F_{ST}. Note that the obtained statistics can only be used to quantify genetic differentiation, but not to test it, because the principal components are by definition, optimized with regard to some measurement of genetic differentiation. To conclude this point, the identification of interpretable structures is a major question in multivariate analysis, and is of particular interest when seeking genetic structures from molecular markers.

As we have seen, the application of multivariate analysis to genetic markers can be improved by avoiding a number of pitfalls. However, further improvements can be gained by adapting multivariate methods to several particularities of genetic markers.

Respecting the very nature of data

Scaling in PCA

In many cases, genetic markers are analysed as allele frequencies, which are subjected to a PCoA or a PCA. PCoA is usually well suited to genetic markers because several genetic distances can be used to summarize the genetic variability. In this case, it is necessary to use a Euclidean distance like Roger's (Weir, 1996, p 197), so that genetic relationships among entities can be fully represented in a plane and to choose a distance whose underlying model best matches the data (see for instance Weir, 1996, pp 190–198). In the case of PCA, attention must be devoted to the transformations of data: if centring of allele frequencies is almost mandatory, the scaling of allele frequencies can be discussed. The general reason for scaling is to compensate for trivial differences that occur in the variance of the descriptors, for instance, when descriptors are expressed in different units. A reason for not scaling allele frequencies is that doing so is not necessary (scales of variation are inherently the same for every allele), and could mask differences in the genetic variability contained by informative and non-informative markers, ultimately hiding structures in the data. Nonetheless, one good argument for scaling allele frequencies would be to compensate for differences in variance among alleles due to their underlying binomial nature: the theoretical variance associated with the jth allele frequency, f_j ($j = 1,\dots p$ where p is the total number of alleles), is proportional to $f_j(1-f_j)$. The result is that the variance of an allele frequency is expected to be 'naturally' larger for frequencies close to 0.5, and smaller for frequencies close to 0 or 1. The PCA seeking linear combinations of alleles with maximum variance, alleles with frequencies closer to 0.5 would be favoured by the analysis, although not necessarily reflecting a genetic structure. One way to correct this is to divide f_j by $\sqrt{f_j(1-f_j)}$, as has been previously proposed (Cavalli-Sforza *et al.*, 1994, pp 41–42). Mulley *et al.* (1979) used a related standardization of allele frequencies, which does not amount to unit theoretical variance, but accounts for the number of genotypes used to compute frequencies in each population. Interestingly enough, the variance between populations of the allele frequency standardized by $\sqrt{f_j(1-f_j)}$ is exactly the classical F_{ST} (Weir, 1996, p 166). Therefore, the between-class PCA (Dolédec and Chessel, 1987), which maximizes the variance between populations, would yield principal components with maximum F_{ST} if performed on allele frequencies centred to a mean of zero and scaled by $\sqrt{f_j(1-f_j)}$. Even though between-class PCA has only recently been applied to genetic markers by Parisod and Christin (2008, presented as 'inter-class PCA') and Jombart (2008), this method seems promising for investigating genetic differentiation between groups of genotypes.

Compositional data

The principal particularity of allele frequencies may be that they are sets of compositional data, that is, data with a constant sum for each population and locus. This feature induces non-independence between allele frequencies inside each locus (a frequency can always be deduced from all the others), and has several consequences on ordinations in reduced space. Developments in the multivariate analysis of compositional data were led by the work of Aitchison (Aitchison, 1983, 1999, 2003; Aitchison and Greenacre, 2002), but remained mostly ignored in genetics, apart from a few exceptions (Romano *et al.*, 2003; Reyment, 2005). As stressed before, allele frequencies at a given locus are not independent, as one can be entirely deduced from the others. Populations described by p_j alleles at the jth locus are not embedded inside a p_j-dimensional space, but are instead inside a space whose maximum dimensionality is (p_j−1), known as a simplex space (Aitchison, 2003, pp 24–28). A variety of problems can occur when directly computing an ordination in reduced space in the simplex space (or in a set of simplex spaces in the case of several loci), like the impossibility of identifying structures that are intrinsically non-linear and the numerical instability of principal components. The solutions proposed to account for these problems rely on transforming frequencies (mostly using logarithms) and performing a classical analysis like PCA of the obtained data. Reyment (2005) showed that the results of PCA could be strikingly improved by such practices, even when considering a simple log transformation of the data. Henceforth, these approaches should be considered when analysing allele frequencies.

Diversity inside the diversity

A portion of the literature in conservation biology stresses the idea that different genetic markers can provide different information about the genetic diversity of a set of populations (Moazami-Goudarzi and Laloë, 2002). In fact, genetic markers are usually taken as a whole to seek a global, common typology of individuals or populations, without trying to assess if such a common typology exists. There are, however, good reasons for this typology not to occur, the first being that selection can affect different loci in different ways. If this is obvious for selected markers like allozymes, it can also be true for supposedly neutral markers that are physically linked to selected regions of the genome. Interestingly, the first studies linking the genetic variability in allozymes to environmental features analysed each locus separately by PCA (Johnson *et al.*, 1969; Johnson and Schaffer, 1973).

To tackle the question of the typological coherence of genetic markers, the locus must be considered as the unit of analysis. In this perspective, if there are K markers, *K* analyses should be performed and compared. A class of multivariate analyses, called the *K*-table methods (Dray *et al*., 2007), is devoted to this particular task. Such methods were introduced in genetics by Laloë *et al*. (2007), who used multiple co-inertia analysis (Chessel and Hanafi, 1996) to compare the typological information provided by different microsatellites. This study showed that microsatellites could provide different pictures of the genetic diversity among populations: whereas some microsatellites reveal the entire genetic structure, some perceive only particular aspects of the genetic diversity and others are simply not informative in terms of genetic differentiation patterns. The typological value of a marker can be used to quantify the extent to which this marker contributes to displaying a particular genetic structure (Laloë *et al*., 2007). The application of *K*-table approaches to genetic markers was further developed by (Pavoine and Bailly, 2007), who introduced other *K*-table methods coupled with a multivariate analysis of biodiversity (Pavoine *et al*., 2004). Their results confirmed the fact that summing the information coming from different genetic markers, as is usually performed for ordinations in reduced space, does not always provide the most accurate picture of biodiversity. Note that if *K*-table methods can suggest that loci experience different selective pressures, they cannot be used as a direct test for these differences. In fact, *K*-table approaches are first and foremost designed to identify common typologies, and not discrepancies, among a set of markers.

If *K*-table methods are more complex tools than single-table analyses, their use in genetics should be considered with attention. Note that the linkage of multilocus genetic information to environmental features like in Johnson *et al*. (1969) still raises challenging questions in terms of data analysis: How can we describe the genetic-environment relationships at several loci? What are the different patterns of adaptation among loci?

Linking genetic markers to other data

One of the greatest applications of ordinations in reduced space is in the linkage of genetic markers to other types of data (Johnson *et al*., 1969; Taylor and Mitton, 1974; Mulley *et al*., 1979; Barker *et al*., 1986; Jarraud *et al*., 2002). This is typically the case in the study of genotype-environment relationships, where multivariate methods can be used to investigate correlations between genetic data and environmental features (Johnson *et al*., 1969; Mulley *et al*., 1979). Another application of such an approach is to relate genetic information to phenotypic data (Jarraud *et al*., 2002). Note that when patterns of selection are being investigated, the genetic diversity should be inferred from non-neutral rather than neutral markers. Various methods are available for coupling two different kinds of information, some of which have been introduced into population genetics. These can be divided into two categories, depending on whether they treat both types of information symmetrically or not. Approaches like DA and between-class PCA are also methods for coupling genetic markers with a different information (some partitions of individuals). However, because their aim is very different from the methods presented below (their purpose is to investigate the genetic differentiation between groups of genotypes), DA and between-class PCA are not presented in this section.

Asymmetric methods

The first type of method is formed by constrained ordinations, which are devoted to investigating the variability in one dataset that can be explained by another dataset. This is achieved by a multivariate regression of a 'response' dataset onto an 'explanatory' dataset (Ter Braak, 1986). These methods are thus asymmetric, in that the variability in one dataset is explained by another. There are two main techniques in this context: redundancy analysis (RDA, Rao, 1964), which is a constrained version of PCA, and canonical correspondence analysis (CCA, Ter Braak, 1986), which is based on CA. RDA and CCA therefore inherit their properties from PCA and CA: RDA can be used for allele frequencies, whereas CCA is more appropriate to analysis of tables of allele counts. Both RDA (Kölliker *et al*., 2008) and CCA (Angers *et al*., 1999) have proven useful in population genetics, mostly to investigate the portion of the genetic variability that can be explained by a set of environmental variables. For instance, in Angers *et al*. (1999), the CCA revealed that the genetic diversity among a set of brook charr populations (*Salvelinus frontalis*) was mainly driven by the structure of the hydrographic network and by a few environmental variables. Another interest of this study is that analyses were applied to two different levels, to study the effects of hydrographic and environmental features on the genetic diversity inside, and between populations.

Like discriminant analysis, RDA and CCA involve computation of the Mahalanobis metric which is, in this case, the matrix of covariances between explanatory variables (Legendre and Legendre, 1998). These analyses therefore require that the number of explanatory variables (for instance, environmental variables) be fairly lower than the number of studied objects (genotypes or populations) to be computable. Following the previously cited study of Williams and Titus (1988) concerning DA, we can recommend that the number of objects should be at least three times larger than the number of explanatory variables. RDA and CCA also demand that the explanatory variables are reasonably uncorrelated to achieve numerical stability and interpretability of the results. As a rule of thumb, we suggest avoidance of correlations greater than 0.7, so that no more than one half of the variability of any predictor could be explained by another predictor (that is, $R^2 < 0.5 \Leftrightarrow |r| < \sqrt{0.5} \simeq 0.7$). Note that genetic markers could also be used as explanatory variables, for example, with an 'explained' dataset of phenotypic traits. In such cases, the dimension of the genetic information should be reduced, either by applying a standard variable selection procedure (for example, forward selection) to the allele frequencies, or by reducing the genetic data to a few principal components using PCA or PCoA.

When the above conditions are respected, constrained ordinations can be efficiently used to explain one kind of variability by another. However, when the purpose of a study is to investigate common patterns of variability in two datasets, or when RDA and CCA cannot be used for technical reasons, an alternative can be found in certain symmetric approaches.

Symmetric methods

Symmetric methods allow one to study the structures common to two datasets by treating the two types of information similarly. They differ from constrained ordinations in the same way that linear regression differs from correlation. Symmetric approaches include canonical correlation analysis (CCorA, Hotelling, 1936; Takeuchi *et al*., 1982, pp 225–280) and co-inertia analysis (COA, Dolédec and Chessel, 1994; Dray *et al*., 2003a). CCorA was introduced by Johnson and Schaffer (1973) to describe and test the correlations between allele frequencies in allozymes and a set of environmental features. The principle of this technique is to find two sets of orthogonal axes (one for each dataset), such that the obtained pairs of principal components have a maximum squared correlation (Takeuchi *et al*., 1982, pp 225–229). It is worth noting that Johnson and Schaffer (1973) were following another pioneering work (Johnson *et al*., 1969) in which the same authors used correlations between principal components of two PCAs (one of allele frequencies, one of environmental variables) to test genetic-environment relationships. A series of subsequent papers provided remarkable illustrations of the insights that CCorA can bring to the study of adaptation (Schaffer and Johnson, 1974; McKechnie *et al*., 1975; Mulley *et al*., 1979). A nice example is provided by Mulley *et al*. (1979), which used the CCorA to investigate patterns of adaptation in populations of *Drosophila buzzatii*. The authors have shown that the allelic variation observed at some allozyme loci was significantly correlated to climate descriptors, which strongly suggested the existence of local adaptations in these populations. A recurrent problem in these studies is that gene-flow can act as a confounding effect when assessing genetic-environment correlations. Schaffer and Johnson (1974) addressed this issue by regressing allele frequencies onto spatial coordinates prior to the analysis, and hence removing linear spatial trends from the data. Note that more efficient methods of removing spatial patterns have since been developed, some of which are described in the next section.

A typical problem in CCorA is that, like RDA and CCA, it requires to compute the Mahalanobis metric of both datasets: it cannot be used when there are more descriptors than studied objects and it requires descriptors to be uncorrelated to yield interpretable results. In some of these cases, a CCorA can still be performed after selecting a small subset of uncorrelated variables (for example, Mulley *et al*., 1979). A common criticism of CCorA is that pairs of principal components with maximum squared correlation could have a very small variance, and therefore have in general no real biological meaning (Taylor and Mitton, 1974). Taylor and Mitton (1974) suggested that a symmetric analysis should yield pairs of principal components reflecting both a fair amount of variance and be correlated with each other, that is, reflecting common parts of the variability in the two datasets. This is the definition of a method developed later in ecology; the co-inertia analysis (Dolédec and Chessel, 1994; Dray *et al*., 2003a).

COA has been imported into genetics to relate the genetic variability of several bacterial strains to the expression of toxin genes (Jarraud *et al*., 2002). It is worth noting that COA is closely related to Procustean analysis (Dray *et al*., 2003b), which has been proposed for the analysis of genetic markers coupled to other kinds of information (Cavalli-Sforza *et al*., 1994, p 41), although we were unable to find any applications of this technique to genetic markers. COA finds two sets of principal axes (one for each dataset), such that the pairs of principal components have a maximum squared covariance (that is, co-inertia). This criterion is particularly interesting as it amounts to maximizing the product of the variances of each principal component and their squared correlations (because $\mathrm{cov}^2(a,b) = \mathrm{var}(a)\mathrm{var}(b)\mathrm{cor}^2(a,b)$). Interestingly, the COA does not require inversion of a covariance matrix; consequently, it does not require the number of descriptors to be lower than the number of objects and it is not hampered by correlations among the descriptors. Moreover, COA relies on a modification of two separate analyses, each of which can be, for instance, a PCA, a PCoA, or a CA. For example, Jarraud *et al*. (2002) employed the co-inertia between a PCoA of a genetic distance matrix derived from AFLP markers and a PCA of distributions of toxin genes in several strains of *Staphylococcus aureus* to assess the evolution of virulence factors with respect to the genetic background of the strains. The COA appears to be a good alternative to RDA, CCA and CCorA when these methods cannot be applied for the reasons described above. In other cases, the COA may still be favoured whenever the squared covariance criterion is more satisfying than criteria used by other analyses, that is, when one is interested in identifying common patterns of variation between two different sources of information.

Spatial multivariate analysis

Many population genetics studies in which multivariate analyses were used involve georeferenced data. When processes related to gene-flow are being investigated—which may be the most common case—spatial genetic patterns are researched in neutral markers (Menozzi *et al*., 1978; Cavalli-Sforza *et al*., 1993). In contrast, when non-neutral markers are used to infer patterns of adaptation, spatial structures induced by gene flow can act as a confounding effect that would have to be removed (Schaffer and Johnson, 1974). As noted by Mulley *et al*. (1979), the drawback of this strategy is that *'if environmental factors with selective effects are strongly correlated to geographic location, adjustment for location may remove a major fraction of the selective effects'*. In such a case, it would be worthwhile to compare the selective effects detected with and without removing the effects of spatial patterns. Spatial information can be used in multivariate analysis of genetic markers, to investigate the part of the genetic variability that is or is not spatially structured.

Unfortunately, the methods commonly used to investigate spatial genetic patterns almost never take spatial information into account explicitly, that is, they do not incorporate spatial information as a component of the criterion optimized by the analysis (Jombart *et al*., 2008). This contrasts with other methodological frameworks such as analysis of molecular variance (Excoffier *et al*., 1992) or Bayesian clustering (Pritchard *et al*., 2000), in which spatially explicit methods are used (respectively, Dupanloup *et al*., 2002; François *et al*., 2006). However, spatial ordinations exist and are widely used in other domains, the closest to genetics being ecology. It is, therefore, not surprising that spatial ordinations were first proposed to analyse genetic markers in vegetation sciences (Escudero *et al*., 2003) and landscape genetics (Grivet *et al*., 2008).

Recently, Grivet *et al*. (2008) used the canonical trend surface analysis (Wartenberg, 1985) to detect spatial patterns using microsatellite markers. This approach relies on performing a CCorA to identify correlations between genetic and spatial data. Grivet *et al*. (2008) used polynomials of spatial coordinates as spatial predictors, although this approach was criticized in ecology (Borcard and Legendre, 2002; Dray *et al*., 2006), mainly because the obtained variables are generally correlated and can only model broad-scale patterns. Other spatial predictors, Moran's eigenvectors, are now used in ecology (Dray *et al*., 2006; Griffith and Peres-Neto, 2006). Contrary to polynomials of spatial coordinates, these spatial predictors are uncorrelated, and can model spatial patterns on a wide range of scales. To reveal spatial genetic patterns, Moran's eigenvectors can be used as explanatory variables in a CCA or an RDA of genetic markers. In studies in which spatial structures need to be removed to infer adaptations, Moran's eigenvectors could also be used as covariables in partial RDA or partial CCA (Legendre and Legendre, 1998, pp 769–779).

To our knowledge, the only spatial ordination developed within the genetic framework is the spatial principal component analysis (sPCA, Jombart *et al*., 2008). This method relies on a modification of PCA such that not only the variance of the principal components, but also their spatial autocorrelation, is optimized. Jombart *et al*. (2008) identified various kinds of spatial structuring that can arise in genetic data, and showed that sPCA can be efficiently used to reveal these patterns. In particular, a comparison between PCA and sPCA demonstrated that sPCA should be preferred to PCA whenever spatial genetic patterns are researched. Note that a similar approach was developed in the vegetation sciences by Dray *et al*. (2008), who proposed a spatial version of CA. Although sPCA is devoted to investigating spatial genetic patterns in allele frequencies, the approach of Dray *et al*. (2008) could be used to study spatial genetic patterns in allele counts.

Perspectives and conclusion

We reviewed how a multivariate analysis can be used to extract biological information from genetic markers. The large diversity of existing multivariate methods allow to investigate a wide variety of genetic structures, which depend on the nature of data as well as on the question being asked. One important observation emerging from this review is that application of multivariate methods to genetic markers could sometimes benefit from more rigorous practices. Methods should always be referred to clearly and with a distinction between the method itself and its implementation. An accurate description of an ordination in reduced space would include all data transformations, such as centring and scaling in PCA, the chosen distance in PCoA and NMDS, the selection of alleles in DA or algorithm initialization in NMDS. To facilitate reproducibility, free and script-based software should be favoured over other software. In this context, R software (R Development Core Team, 2008) is clearly an appealing choice: in addition to allowing exact reproducibility, it provides an interface between a large number of implemented multivariate methods (Chessel *et al*., 2004; Dray *et al*., 2007) and genetic marker data (Jombart, 2008), in addition to supporting the usual population genetics tools (Warnes, 2003; Goudet, 2005). From a more theoretical point of view, it seems important to further investigate the relationships between multivariate methods and genetic models. A step in this direction has been made by Patterson *et al*. (2006), who applied recent developments in statistics (Soshnikov and Fyodorov, 2005) to infer the number of populations in a set of genotypes and define a threshold for genetic structuring to be detectable by PCA.

More generally, several multivariate analyses developed in other disciplines can be adapted to search biological structures within genetic markers. This is clearly the case in spatial genetics, where constrained ordinations based on Moran's eigenvectors (Dray *et al*., 2006) could be used to investigate or correct for spatial genetic structures. It is also true for *K*-table methods, which were only recently introduced into population genetics (Laloë *et al*., 2007; Pavoine and Bailly, 2007), and open appealing perspectives for the study of the genetic diversity. These methods can also be used to investigate common patterns of variation inferred from genetic markers and other sources of information, like biological traits and environmental features. As noted by Patterson *et al*. (2006), multivariate analysis can analyse larger datasets than other usual approaches such as Bayesian clustering, and thus represents a relevant approach to extracting information from huge datasets produced by the detailed mapping of genetic variation for a large number of genotypes. This is the case, for instance, with the '1000 Genomes' project (http://www.1000genomes.org/), which aims at sequencing one thousand human genotypes to provide high-resolution information that is directly valuable for disease studies. Promisingly, a wide range of questions are raised by or through genetic markers, some of which can currently be solved by existing methods. Some of these questions will undoubtedly require specific developments in which multivariate models will have to closely match the genetic concerns, which makes the multivariate analysis of genetic markers a whole area of research in biometry.

Acknowledgements

We are very grateful to Christian Biémont, F Stephen Dobson, Gilles Yoccoz and Sébastien Devillard for providing constructive comments on an earlier version of the article. We also address many thanks to F Stephen Dobson for improving the English of this paper.

References

Aitchison J (1983). Principal component analysis of compositional data. *Biometrika* **70**: 57–65.

Aitchison J (1999). Logratios and natural laws in compositional data analysis. *Math Geol* **31**: 563–589.

Aitchison J, Greenacre M (2002). Biplot of compositional data. *J R Stat Soc Ser C, Appl stat* **51**: 375–392.

Aitchison J (2003). *The Statistical Analysis of Compositional Data*. The Blackburn Press: Cladwell, New Jersey.

Angers B, Plante M, Bernatchez L (1999). Canonical correspondence analysis for estimating spatial and environmental effects on microsatellite gene diversity in brook charr (*Salvelinus fontinalis*). *Mol Ecol* **8**: 1043–1053.

Baker AJ, Moeed A (1987). Rapid genetic differentiation and founder effect in colonizing populations of common mynas (*Acridotheres tristis*). *Evolution* **41**: 525–538.

Barker JSF, East PD, Weir BS (1986). Temporal and microgeographic variation in allozyme frequencies in a natural population of *Drosophila buzzatii*. *Genetics* **112**: 577–611.

Beharav A, Nevo E (2003). Predictive validity of discriminant analysis for genetic data. *Genetica* **119**: 259–267.

Bertranpetit J, Cavalli-Sforza LL (1991). A genetic reconstruction of the history of the population of the Iberian Peninsula. *Ann Hum Genet* **55**: 51–67.

Borcard D, Legendre P (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol Modell* **153**: 51–68.

Cavalli-Sforza LL (1966). Population structure and human evolution. *Proc R Soc Lond Ser B* **164**: 362–379.

Cavalli-Sforza LL, Menozzi P, Piazza A (1993). Demic expansions and human evolution. *Science* **259**: 639–646.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994). *The History and Geography of Human Genes*. Princeton University Press: Princeton.

Chessel D, Hanafi M (1996). Analyses de la co-inertie de *K* nuages de points. *Revue de statistique appliquée XLIV* **2**: 35–60.

Chessel D, Dufour AB, Thioulouse J (2004). The ade4 package-I-one-table methods. *R News* **4**: 5–10.

Ciofi C, Wilson GA, Beheregaray LB, Marquez C, Gibbs JP, Tapia W *et al*. (2006). Phylogeographic history and gene flow among giant galápagos tortoises on southern Isabela Island. *Genetics* **172**: 1727–1744.

Cox RF, Cox MAA (2001). *Multidimensional Scaling*. Chapman & Hall/CRC: Bora Raton, Florida.

Dolédec S, Chessel D (1987). Rythmes saisonniers et composantes stationnelles en milieu aquatique. I. description d'un plan d'observation complet par projection de variables. *Acta Oecologica, Oecologia Generalis* **8**: 403–426.

Dolédec S, Chessel D (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshw Biol* **31**: 277–294.

Dray S, Legendre P, Peres-Neto P (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol Modell* **196**: 483–493.

Dray S, Dufour AB (2007). The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw* **22**: 1–20.

Dray S, Chessel D, Thioulouse J (2003a). Co-inertia analysis and the linking of ecological data tables. *Ecology* **84**: 3078–3089.

Dray S, Chessel D, Thioulouse J (2003b). Procrustean co-inertia analysis for the linking of multivariate datasets. *Ecoscience* **10**: 110–119.

Dray S, Dufour AB, Chessel D (2007). The ade4 package—II: Two-table and *K*-table methods. *R News* **7**: 47–54.

Dray S (2008). On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Comput stat data anal* **52**: 2228–2237.

Dray S, Saïd S, Debias F, Chessel D (2008). Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *J Veg Sci* **19**: 45–56.

Dupanloup I, Schneider S, Excoffier L (2002). A simulated annealing approach to define the genetic structure of populations. *Mol Ecol* **11**: 2571–2581.

Edwards AWF (2003). Human genetic diversity: Lewontin's fallacy. *BioEssays* **25**: 798–801.

Escoufier Y (1987). The duality diagramm: a means of better practical applications. In: Legendre P, Legendre L (eds). *Development in Numerical Ecology.* NATO advanced Institute, Serie G. Springer Verlag, Berlin. pp 139–156.

Escudero A, Iriondo JM, Torres ME (2003). Spatial analysis of genetic diversity as a tool for plant conservation. *Biol Conserv* **113**: 351–365.

Excoffier L, Smouse PE, Quattro JM (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**: 479–491.

Fahima T, Sun GL, Beharav A, Krugman T, Beiles A, Nevo E (1999). RAPD polymorphism of wild emmer wheat populations, *Triticum dicoccoides*, in Israel. *Theor Appl Genet* **98**: 434–447.

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Fisher RA (1952). Statistical methods in genetics. *Heredity* **6**: 1–12.

François O, Ancelet S, Guillot G (2006). Bayesian clustering using hidden markov random fields in spatial population genetics. *Genetics* **174**: 805–816.

Geffen E, Anderson MJ, Wayne RK (2004). Climate and habitat barriers to dispersal in the highly mobile grey wolf. *Mol Ecol* **13**: 2481–2490.

Goudet J (2005). HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Mol Ecol Notes* **5**: 184–186.

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

Greenacre M (1966). *Theory and Applications of Correspondence Analysis*. Academic Press: London.

Griffith DA, Peres-Neto P (2006). Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology* **87**: 2603–2613.

Grivet D, Sork VL, Westfall RD, Davis FW (2008). Conserving the evolutionary potential of California valley oak (*Quercus lobata* Née): a multivariate genetic approach to conservation planning. *Mol Ecol* **17**: 139–156.

Guinand B (1996). Use of a multivariate model using allele frequency distributions to analyse patterns of genetic differentiation among populations. *Biol J Linnean Soc* **58**: 173–195.

Guinand B, Bouvet Y, Brohon B (1996). Spatial aspects of genetic differentiation of the European chub in the Rhone River basin. *J Fish Biol* **49**: 714–726.

Hanotte O, Bradley DG, Ochieng JW, Verjee Y, Hill EW, Rege JEO (2002). African pastoralism: genetic imprints of origins and migrations. *Science* **296**: 336–339.

Harville DA (1997). *Matrix Algebra From a Statistician's Perspective.* Springer: New York.

Hotelling H (1936). Relations between two sets of variates. *Biometrika* **28**: 321–327.

Jambu M (1991). *Exploratory and Multivariate Data Analysis*. Academic Press Inc.: Orlando, Florida.

Jarraud S, Mougel C, Thioulouse J, Lina G, Meugnier H, Forey F *et al.* (2002). Relationships between *Staphylococcus aureus* genetic background, virulence factors, *agr* groups (alleles), and human disease. *Infect Immun* **70**: 631–641.

Johnson FM, Schaffer HE, Gillaspy JE, Rockwood ES (1969). Isozyme genotype-environment relationships in natural populations of the harvester ant, *Pogonomyrmex barbatus*, from Texas. *Biochem Genet* **3**: 429–450.

Johnson FM, Schaffer HE (1973). Isozyme variability in species of the genus drosophila. VII. Genotype-environment relationships in populations of *D. melanogaster* from the eastern United States. *Biochem Genet* **10**: 149–163.

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

Jombart T, Devillard S, Dufour AB, Pontier D (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101**: 92–103.

Krzanowski WJ, Marriott FHC (1995). *Multivariate Analysis. Part 2: Classification, Covariance Structures and Repeated Measurements*. Halsted Press, John Wiley & Sons: Edward Arnold, London.

Kölliker R, Bassin S, Schneider D, Widmer F, Fuhrer J (2008). Elevated ozone affects the genetic composition of *Plantago lanceolata* L. Populations. *Environ Pollut* **152**: 380–386.

Lachenbruch PA, Goldstein M (1979). Discriminant analysis. *Biometrics* **35**: 69–85.

Laloë D, Jombart T, Dufour AB, Moazami-Goudarzi K (2007). Consensus genetic structuring and typological value of markers using multiple co-inertia analysis. *Genet Sel Evol* **39**: 545–567.

Lebart L, Morineau A, Piron M (2004). *Statistique Exploratoire Multidimensionnelle*. DUNOD: Paris.

Legendre P, Legendre L (1998). *Numerical Ecology*. Elsevier Science B.V.: Amsterdam.

Legendre P, Anderson DJ (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol Monogr* **69**: 1–24.

Lessa EP (1990). Multidimensional analysis of geographic genetic structure. *Syst Zool* **39**: 242–252.

Lewontin RC (1978). Single-locus and multiple-locus measures of genetic distance between groups. *Am Nat* **112**: 1138–1139.

Li MH, Zhao SH, Bian C, Wang HS, Wei H, Liu B *et al*. (2002). Genetic relationships among twelve chinese indigenous goat populations based on microsatellite analysis. *Genet Sel Evol* **34**: 729–744.

MacHugh DE, Shriver MD, Loftus RT, Cunningham P, Bradley DG (1997). Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics* **146**: 1071–1086.

MacHugh DE, Loftus RT, Cunningham P, Bradley DG (1998). Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers. *Anim Genet* **29**: 333–340.

Matsuoka Y, Vigouroux Y, Goodman MM, Jesus Sanchez G, Buckler E, Doebley J (2002). A single domestication for maize shown by multilocus microsatellite genotyping. *Proc Natl Acad Sci USA* **99**: 6080–6084.

McKechnie SW, Ehrlich PR, White RR (1975). Population genetics of euphydryas butterflies. I. genetic variation and the neutrality hypothesis. *Genetics* **81**: 571–594.

McRae BH, Beier P, Dewald LE, Huynh LY, Keim P (2005). Habitat barriers limit gene flow and illuminate historical events in a wide-ranging carnivore, the American puma. *Mol Ecol* **14**: 1965–1977.

Menozzi P, Piazza A, Cavalli-Sforza LL (1978). Synthetic maps of human gene frequencies in Europeans. *Science* **201**: 786–792.

Mitton JB (1978). Measurement of differentiation: reply to Lewontin, Powell and Taylor. *Am Nat* **112**: 1142–1144.

Moazami-Goudarzi K, Laloë D, Furet JP, Grosclaude F (1997). Analysis of genetic relationships between 10 cattle breeds with 17 microsatellites. *Anim Genet* **28**: 338–345.

Moazami-Goudarzi K, Laloë D (2002). Is a multivariate consensus representation of genetic relationships among populations always meaningful? *Genetics* **162**: 473–484.

Mulley JC, James JW, Barker JSF (1979). Allozyme genotype-environment relationships in natural populations of *Drosophila buzzatii*. *Biochem Genet* **17**: 105–126.

Pariset L, Savarese MC, Cappuccio I, Valentini A (2003). Use of microsatellites for genetic variation and inbreeding analysis in Sarda sheep flocks of central Italy. *J Anim Breed Genet* **120**: 425–432.

Parisod C, Christin PA (2008). Genome-wide association to fine-scale ecological heterogeneity within a continuous population of *Biscutella laevigata* (brassicaceae). *New Phytol* **178**: 436–447.

Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *PLoS genet* **2**: 2074–2093.

Pavoine S, Dufour AB, Chessel D (2004). From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *J Theor Biol* **228**: 523–537.

Pavoine S, Bailly X (2007). New analysis for consistency among markers in the study of genetic diversity: development and application to the description of bacterial diversity. *BMC Evolut Biol* **7**: 156.

Pearson K (1901). On lines and planes of closest fit to systems of points in space. *Philos Mag* **2**: 559–572.

Perrière G, Thioulouse J (2002). Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res* **30**: 4548–4555.

Powell JR, Taylor CE (1978). Are human races 'substantially' different genetically? *Am Nat* **112**: 1139–1142.

Preziosi RF, Fairbairn DJ (1992). Genetic population structure and levels of gene flow in the stream dwelling waterstrider *Aquarius* (= *Gerris*) *remigis* (Emiptera: Geridae). *Evolution* **46**: 430–444.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria ISBN 3-900051-07-0. http://www.R-project.org.

Rao CR (1964). The use and interpretation of principal component analysis in applied research. *Sankhya, A* **26**: 329–359.

Reyment RA (2005). The statistical analysis of multivariate serological frequency data. *Bull Math Biol* **67**: 1303–1313.

Rohlf FJ (1963). Classification of *Aedes* by numerical taxonomic methods (diptera:culicidae). *Ann Entomol Soc Am* **56**: 798–804.

Romano V, Calí F, Ragalmuto A, D'Anna RP, Flugy A, De Leo G *et al.* (2003). Autosomal microsatellite and mtDNA genetic analysis in Sicily (Italy). *Ann Hum Genet* **67**: 42–53.

Sagnard F, Barberot C, Fady B (2002). Structure of genetic diversity in *Abies* alba Mill. from southwestern Alps: multivariate analysis of adaptive and non-adaptative traits for conservation in France. *For Ecol Manage* **157**: 175–189.

Sanchez-Mazas A, Langaney A (1988). Common genetic pools between human populations. *Hum Genet* **78**: 161–166.

Schaffer HE, Johnson FM (1974). Isozyme allelic frequencies related to selection and gene-flow hypotheses. *Genetics* **77**: 163–168.

Seal HL (1966). *Multivariate Statistical Analysis for Biologists*. Methuen and co.: London.

Seber GAF (1977). *Linear Regression Analysis*. John Wiley & Sons: New York.

She JX, Autem M, Kotulas G, Pasteur N, Bonhomme F (1987). Multivariate analysis of genetic exchanges between *Solea aegyptiaca* and *Solea senegalensis* (Teleosts, Soleidae). *Biol J Linnean Soc* **32**: 357–371.

Smouse PE, Spielman RS, Park MH (1982). Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. *Am Nat* **119**: 445–463.

Soshnikov A, Fyodorov YV (2005). On the largest singular values of random matrices with independent Cauchy entries. *J Math Phys* **46**: 033302.

Takeuchi K, Yanai H, Mukherjee BN (1982). The foundations of multivariate analysis: a unified approach by means of projection onto linear subspaces. Wiley Eastern Limited: New-Delhi.

Taylor CE, Mitton JB (1974). Multivariate analysis of genetic variation. *Genetics* **76**: 575–585.

Ter Braak CJF (1986). Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67**: 1167–1179.

van Pijlen IA, Amos B, Burke T (1995). Patterns of genetic variability at individual minisatellite loci in minke whale *Balaenoptera acutorostrata* populations from three different oceans. *Mol Biol Evol* **12**: 459–472.

Warnes GR (2003). The genetics package. *R News* **3**: 9–13.

Wartenberg DE (1985). Canonical trend surface analysis: a method for describing geographic patterns. *Syst Zool* **34**: 259–279.

Weir BS (1996). *Genetic Data Analysis II*. Sinauer Associates: Sunderland, Massachussetts.

Williams BK, Titus K (1988). Assessment of sampling stability in ecological applications of discriminant analysis. *Ecology* **69**: 1275–1285.

Xuebin Q, Jianlin H, Chekarova I, Badamdorj D, Rege JEO, Hanotte O (2005). Genetic diversity and differentiation of Mongolian and Russian yak populations. *J Anim Breed Genet* **122**: 117–126.

Zhivotovsky LA, Rosenberg NA, Feldman MW (2003). Features of evolution and expansion of modern humans, inferred from genomwide microsatellite markers. *Am J Hum Genet* **72**: 1171–1186.