Rev. FCA UNCUYO. ISSN 0370-4661. Tomo 43. N° 1. Año 2011. 261-275.

Review

Making genetic biodiversity measurable: a review of statistical multivariate methods to study variability at gene level

Cuantificando diversidad genética: una revisión de métodos estadísticos multivariados para estudiar variabilidad a nivel de genes

Mónica Balzarini [1]
Ingrid Teich [2]
Cecilia Bruno [1]
Andrea Peña [1]

INDEX **Pág.**

Originales: Recepción: 29/04/2011 - Aceptación: 31/05/2011

1 Cátedra de Estadística y Biometría. mbalzari@agro.unc.edu.ar
2 Centro de Relevamiento y Evaluación de Recursos Agrícolas y Naturales.
Facultad de Ciencias Agropecuarias. Universidad Nacional de Córdoba. Av. Valparaíso s/n. Ciudad Universitaria. C. C. 509. (5000) Córdoba. Argentina.

ABSTRACT

Measures of agro-ecosystems genetic variability are essential to sustain scientific-based actions and policies tending to protect the ecosystem services they provide. To build the genetic variability datum it is necessary to deal with a large number and different types of variables. Molecular marker data is highly dimensional by nature, and frequently additional types of information are obtained, as morphological and physiological traits. This way, genetic variability studies are usually associated with the measurement of several traits on each entity. Multivariate methods are aimed at finding proximities between entities characterized by multiple traits by summarizing information in few synthetic variables.

In this work we discuss and illustrate several multivariate methods used for different purposes to build the datum of genetic variability. We include methods applied in studies for exploring the spatial structure of genetic variability and the association of genetic data to other sources of information. Multivariate techniques allow the pursuit of the genetic variability datum, as a unifying notion that merges concepts of type, abundance and distribution of variability at gene level.

Keywords
ordination • clustering • multivariate association • spatial variability

RESUMEN

Obtener estimaciones confiables de la diversidad genética en los agroecosistemas es esencial para tomar decisiones basadas en el conocimiento científico que permitan proteger los servicios ecosistémicos que éstos brindan. Para construir el dato de variabilidad genética es necesario trabajar con gran cantidad de variables de distinta naturaleza. Los marcadores moleculares proveen datos multidimensionales que generalmente son complementados con otros tipos de información, por ejemplo datos morfológicos o fisiológicos. Así, los estudios sobre variabilidad genética están frecuentemente asociados a la medición de muchos caracteres en una misma entidad biológica. De especial interés son los métodos multivariados diseñados para analizar similitudes entre entidades caracterizadas por múltiples variables que permiten resumir la información en pocas variables sintéticas informativas de la variabilidad total.

En este trabajo se discuten e ilustran distintos métodos multivariados utilizados en la construcción del dato de variabilidad genética. Se incluyen métodos aplicados a la exploración de la estructura espacial de la variabilidad genética y métodos para estudiar la asociación de los datos genéticos con otras fuentes de información. Las técnicas multivariadas en esta revisión permiten abordar el problema de construir al dato de variabilidad genética como un concepto donde convergen mediciones sobre tipo, abundancia y distribución de la variabilidad a nivel de genes.

Palabras clave
ordenación • agrupamiento • asociación multivariada • variabilidad espacial

INTRODUCTION

Biodiversity in agro-ecosystems can be characterized at different levels of organization which are affected by different factors, ranging from landscape heterogeneity to land use and management. Knowledge of the relationship that exists between management practices and biodiversity is necessary to recommend management practices that minimize the loss of stability in the production. In consequence, sustainable agriculture requires information on biodiversity, comprising the appearance, structure and function of all levels of biological organization, including ecosystems, species and genes.

This has been acknowledged by the World Conservation Union (IUCN), which recommends conservation of biological diversity at the three levels (McNeely *et al.*, 1990) and by the recently held Convention on Biological Diversity (2010) which has implemented goals with respect to the genetic level. There are many empirical examples showing how the amount and distribution of genetic variation affects not only species and ecosystems, but also genes.

The genetic variability datum is essential to determine the real magnitude of spatial and temporal changes of biodiversity in agroecosystems and to sustain scientific-based actions and policies tending to protect the ecosystem services that they provide. Diversity is a concept that should reflect the number and abundance of different types in a collection of objects (populations, individuals, loci).

Genetic diversity is a part of biodiversity that contributes to various levels and therefore, can also be measured at several scales including individuals, populations,species and even ecosystems. Therefore, to measure gene diversity, in addition to identifying the object of study, it is vital to identify and characterize the variables measured (e.g. genotypes, haplotypes, alleles) and to establish the temporal and/or spatial dimension of the study. Usually, the genetic variability datum, as a unifying notion that merges concepts of type, abundance and distribution of the evaluated variables in the set of biological entities of interest.

The information necesary to build the genetic variability datum is statistical since many entities should be evaluated. Each entity can be characterized by different types of data, such asmolecular, morphological and physiological,with several variables of each type measured on every entity.

In figure 1 (p. 264), two examples of common data-sets in studies designed to explore genetic variability are shown. In the first case (left), variables represent molecular markers; one or zero indicate if the molecular marker, or biomarker used to explore the genome, is present (1) or absent (0). If the biological entity in which the variable has been measured is a population, this information can be expressed by relative frequencies as it is shown in figure 1 (right), where 7 populations were explored with 7 markers genotyping several individuals within populations.

Ind	M1	M2	M3	M4	M5	M6	M7
1	1	0	1	1	0	1	0
2	1	0	1	1	0	1	1
3	1	1	0	1	0	1	1
4	0	0	1	1	0	1	1
5	1	0	1	1	1	0	1
6	1	0	1	0	0	0	1
7	1	0	1	1	0	0	0

Pop	M1	M2	M3	M4	M5	M6	M7
1	0,14	0,00	0,16	0,20	0,32	0,12	0,06
2	0,20	0,00	0,07	0,13	0,41	0,09	0,10
3	0,13	0,60	0,00	0,00	0,00	0,07	0,20
4	0,00	0,13	0,13	0,13	0,15	0,36	0,10
5	0,12	0,19	0,50	0,08	0,01	0,00	0,00
6	0,14	0,30	0,06	0,00	0,00	0,50	0,00
7	0,13	0,20	0,27	0,18	0,30	0,10	0,20

Figure 1. An example of typical molecular multivariate data where columns correspond to molecular markers and rows correspond to biological entities. Data corresponds to absence (0) or presence (1) of marker per individuals (left) and to marker relative frequencies per population (right).

In addition to this, other covariates could be linked to the data, like spatial and temporal coordinates which should be also incorporated in the biodiversity analysis. As consequence of the multidimensional nature of information; statistical multivariate algorithms have gained acceptance among biologists and agronomists. Several software allow the use of multivariate algorithms to handle several variables simultaneously. Moreover, free statistical software specifically designed to perform these techniques with genetic data have been made available. The examples in this work were processed with Info-Gen (Balzarini and Di Rienzo, 2004), software freely available at www.info-gen.com.ar. However, all the illustrated analysis can be also obtained using the R software (http://cran.r-project.org/). The objective of this review is to identify several multivariate techniques currently used to quantify genetic variability.

Multivariate exploratory analyses

Multivariate ordination and classification methods are aimed at finding proximities between entities characterized by multiple variables by summarizing information in few synthetic variables (Johnson and Wichern, 1998). Each synthetic variable is created by combining the original variables measured on each object, usually by means of a linear combination in which each one is differentially weighted. It is usually expected that most of the information in the set of original variables can be captured by a small number of these new synthetic variables expressed as linear combinations of the original ones. The efficiency of these methods is dependenton the ability of users to interpret the synthetic variables and whether a small number of those synthetic variables captures enough information or an important portion of the total variability.

Multivariate methods have been shown to be efficient in extracting information from gene level (Cavalli-Sforza, 1966; Johnson *et al*., 1969; Smouse *et al*., 1982) because of their ability to summarize multivariate information derived from the use of genetic markers. From these early applications to current innovative developments (Patterson *et al*., 2006; Pavoine and Bailly, 2007; Jombart *et al*., 2008, Jombart *et al*., 2010), these methods have proven to be useful in various type of biodiversity studies, including conservation (Moazami-Goudarzi *et al*., 1997; Escudero *et al*., 2003; Laloë *et al*., 2007), 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). In essence, dimension reduction techniques extract successive components from a multivariate similarity/dissimilarity matrix containing information about all pair wise individual profile comparisons. The synthetic variables are used as new axes of low-dimensional spaces to graphically represent the variability among entities. Main differences between observations are visualized on the axes beginning by axis 1.

Ordination methods

Principal components analysis (PCA) was first introduced to the study of genetic data almost thirty years ago by Cavalli-Sforza (1966). PCA is applicable to quantitative, or at least ordinal, type of data.It finds an orthogonal basis for the data in such a way that the first axis is along the direction of greatest variation of the multidimensional data and subsequent axes maximize explained variance, given that they are orthogonal to previous axes. Therefore, with PCA we reduce a set of correlated variables to a small number of linear combinations of these variables (principal components or synthetic variables). Such components give scatter plots of observations with optimal properties to study the underlying variability and correlation. If **X** denotes the $n \times p$ data matrix, with nentities and p variables, the PCA operates on the unique factorization of matrix $\frac{1}{n}$ **X'X** (or transformation) giving a set of eigenvectors containing the weight coefficients to build the components. The singular value decomposition is the algorithm that all software use to build the principal components. These weights depend on the relative importance of variables to separate objects, *i.e.*, to explain variability among entities. The data may or may not be standardized leading to different PCA types.

The PCA with standardized data operates on the correlation matrix and is useful when variables are not commensurable. The PCA without standardization is implemented on the covariance matrix of the data; consequently it allows analyzing variances beside correlations. For data of allele frequencies or other type of proportions, it is common to center the relative frequencies by the mean of each column data. The PCA can be applied either on the correlation or covariance matrix of variables, or over the correlation or covariance matrix of observations. The former provides an ordering of observations, while the second provides an ordering of thevariables. Both ordinations can be visualized simultaneously bymeans of biplots (Gabriel, 1971) where objects and variables (markers) are representedin a common space. For the population data showed in figure 1 (right, p. 264), the biplot for the centered data was built (figure 2, left, p.266).

PCA preserves the canonical Euclidean distance among the studied entities. In contrast, Principal Coordinates Analysis (PCO) cansummarize any distance between entities. The algorithm used to build the Principal coordinates is also the singular value decomposition, but operating on a user defined distance matrix.

The technique is a type (metric) of multidimensional scaling. It is useful to determine similarities among entities or traits and to depict similarities/distances on reduced spaces where the inter-distances in the full space are reproduced. PCO offers the opportunity of using different distance metrics including measures of genetic variability that are directly related to a population genetic model, such as

the Fst statistics,which measures genetic differentiation between populations, and several genetics distances between individuals as the Roger´s distance (Baker and Moeed, 1987).For instance, PCO has been used to summarize matrices of pairwise Fst between populations and of Rogers' distance between multiallele molecular genotypes. However, PCO does not provide a representation (in the same space) of the variables used in the characterization (figure 2, right).

Minimum spanning trees (MST) (Gower and Ross, 1969) is another tool commonly used to improve the visualization in ordination analyses, mainly PCO configurations. The algorithm produces a collection of straight lines that link each point on the graph in such a way that there are no closed loops and that each point is connected to every other point either directly or indirectly. Segments are connected in such a way that the sum of the lengths of the individual segments is minimized. Computed on the full dimension of data but showed on the reduced space, the MST provides information on the quality of the projection on the low dimensional space, showing relationships that may not be seen by inspection on the reduced space.If many branches and segments cross each other, it suggests distortion problems in the projection which could bias regular interpretations (figure 2, right).

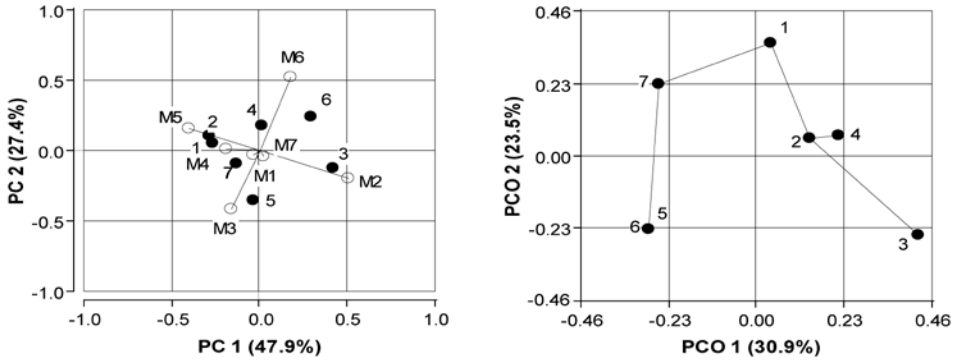

Figure 2. Biplot by PCA of population centered data showing allele markers (white dots) and entities (black dots) in the same space. PC1 and PC2 are synthetic variables (Principal Components) that explain 47.9% and 27.4% of total variability, respectively (left). In the right panel PCO is used to visualize variability between entities characterized by binary marker information. The Roger´s distance was used for measuring profile similarity and a MST was over-imposed on the PCO ordination.

Clustering methods

Several clustering algorithm shave proved to be a powerful tool to investigate natural clusters of genetic data (Peña *et al.*, 2010). These algorithms range from hierarchical clustering (Eisen *et al*., 1998; Levenstien *et al*., 2003), non hierarchical clustering such as k-means and clustering by simulated annealing (Lukasin *et al*., 2001) to neural net

work algorithms such as Self-Organizing Maps(SOM; Töronen, 1999; Fernandez and Balzarini, 2007). All clustering algorithms tend to join objects of interest into clusters such that the elements in a cluster are more alike than elements in different clusters. Clustering techniques do nota *priori* assume any grouping and demand the selection of a distance/similarity metric between objects and between clusters. To cluster *n* samples, a *n*×*n* distance matrix between samples is used, mean while to cluster *p* markers the input distances are arranged in a *p*×*p* matrix. With agglomerative hierarchical clustering procedures, the objects are grouped in a pairwise mode as a function of their similarities; the process begins by joining two objects with most similarity and continues joining the other objects or clusters until all elements belong to the same group. The distance between two clusters can be defined either directly or by an equation for updating a distance matrix when two clusters are joined. The mathematical expression of the distance between two clusters defines the clustering method. Most of the applications with genetic data use the following methods: Unweigthed Pair-Group Arithmetic Average (UPGMA, Sokal *et al.*, 1958), Ward's minimum variance (Ward, 1963), and nearest neighbor (Wong *et al*., 1983).

The history of this algorithm is summarized in a dendrogram, asit is shown in figure 3 (Anderberg, 1973). The dendrogram shows that population 1 and 2 are more similar than any other pair of populations because they are joined at the smallest Euclidean distance, and population 3 is the most distintic one. The resulting dendrogram is usually presented with measures suggesting the goodness of the clustering with respect to the object distances in the full space (cophenetic correlation coefficient) (Rohlf and Fisher, 1968). Info-Gen gives such coefficient which allows researchers to quantity the correlation between the classification of entities on the dendrogram and the tree classification of entities in the multi dimensional space. The cophenetic correlation ranges from 0 to 4.

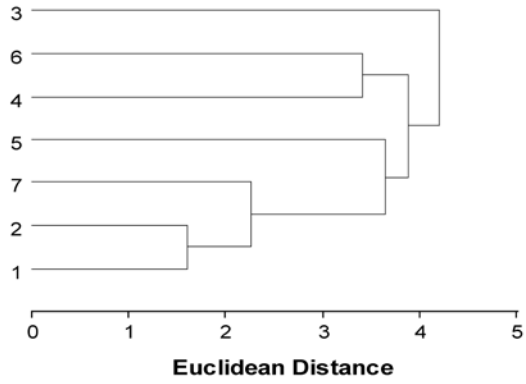

Figure 3. Dendrogram resulting by UPGMA algorithm with data of allele frequency of 7 populations extracted from figure 1, right (p. 264). Cophenetic correlation was 0.92.

Alternatively, objects can be grouped by means of machine learning techniques such as self organized maps (SOM) (Lippman, 1989) which couple different strategies to identify clusters. They providean opportunity to visualize groups when many traits are collected on each entity, such as thousand of markers. SOM is an unsupervised neural network algorithm able to find relationships between high dimensional data, grouping and mapping them topologically. It implements a linear projection from a space of *p*-dimensional entries to a low (1 or 2) dimensional space. The SOM algorithm creates a new representation of objects arranged in a net. Closer nodes are more similar than nodes topologically apart. Each node comprises a number of entities. A central challenge in analyzing genetic variability is to explore whether there is any evidence that the samples in the data are structured. For example we might be interested in determining if the individuals are from a homogeneous population or from a population containing subgroups that are genetically distinct. Understanding such structure may be important to key scientific issues, like uncovering the demographic history of the population under study. This knowledge is important for pest control programs. Additionally, the presence of undetected population structure in association mapping studies can lead to spurious associations and thus invalidate standard tests (Ewens and Spielman, 1995).

Patterson *et al.* (2006) showed recently that PCA can be successfully used to detect population structure, in particular in large datasets consisting of thousands of molecular markers. The idea is to run a PCA and after that, a cluster analysis using the significant PC as input variables. Recently,also Discriminant Analysis of Principal Components or Principal Coordinates (DAPC), a new methodological approach, has been introduced to study genetic structure (Jombart *et al.*, 2010). As well as Patterson´s proposal, DAPC relies on data transformation using ordination techniques as a prior step to Discriminant Analysis (DA), which ensures that variables submitted to DA are perfectly uncorrelated. Without implying a necessary loss of genetic information, this transformation allows DA to be applied to various types of genetic data.

Spatial analysis of genetic variability

The joint analysis of spatial and genetic data is rapidly becoming the norm in conservational biology and agronomy. More and more studies explicitly describe and quantify the spatial organization of genetic variation and try to relate it to underlying ecological processes and environments. The spatial detection and location of genetic discontinuities between biological entities (populations or individuals) is essential to provide information on how environmental features and management practices influence population genetic structure (Manel *et al*., 2003, Lao *et al*., 2008).

Synthetic variables obtained from multivariate techniques, like PCA, summarize in few uncorrelated quantitative variables the genetic variation expressed by many loci and can be used for the identification of spatial patterns in the genetic variability. Synthetic variables that summarize high percentages of the variability can be used as single variables to model the spatial autocorrelation of genetic variability allowing the inference of spatial parameters as the range and sill. The spatial interpolation of the major principal components derived from PCA leads to a “synthetic map” of genetic variability. Synthetic maps have been used to visualize pest dispersion in crops (Nansen *et al*. 2003) and for mapping genetic variability of livestock in Africa (Hanotte *et al*., 2002).

However, PCA is not properly designed to investigate spatial patterns, using spatial information, such as latitude and longitude of sample locations, a posteriori of the multivariate analysis. The recently developed Spatial Principal Components Analysis (sPCA) relies on a modification of PCA such that not only the variance of the synthetic variables, but also their spatial autocorrelation, is optimized(Jombart *et al.*, 2008). It uses spatial information *a priori*, as input of the multivariate algorithm,and in this way the spatial patterns can be more clearly visualized. The new technique allows investigating spatial structures other than the most evident, by focusing on the part of the variability which is spatially structured and not in the total variability as PCA. With sPCA different kinds of spatial structure (global and local) that arise in genetic data can be efficiently revealed. In particular, comparison between PCA and sPCA demonstrated thats PCA should be preferred to PCA when spatial genetic patterns are researched. In figure 4, differences between PCA and sPCA as techniques to reveal spatial patterns is shown.

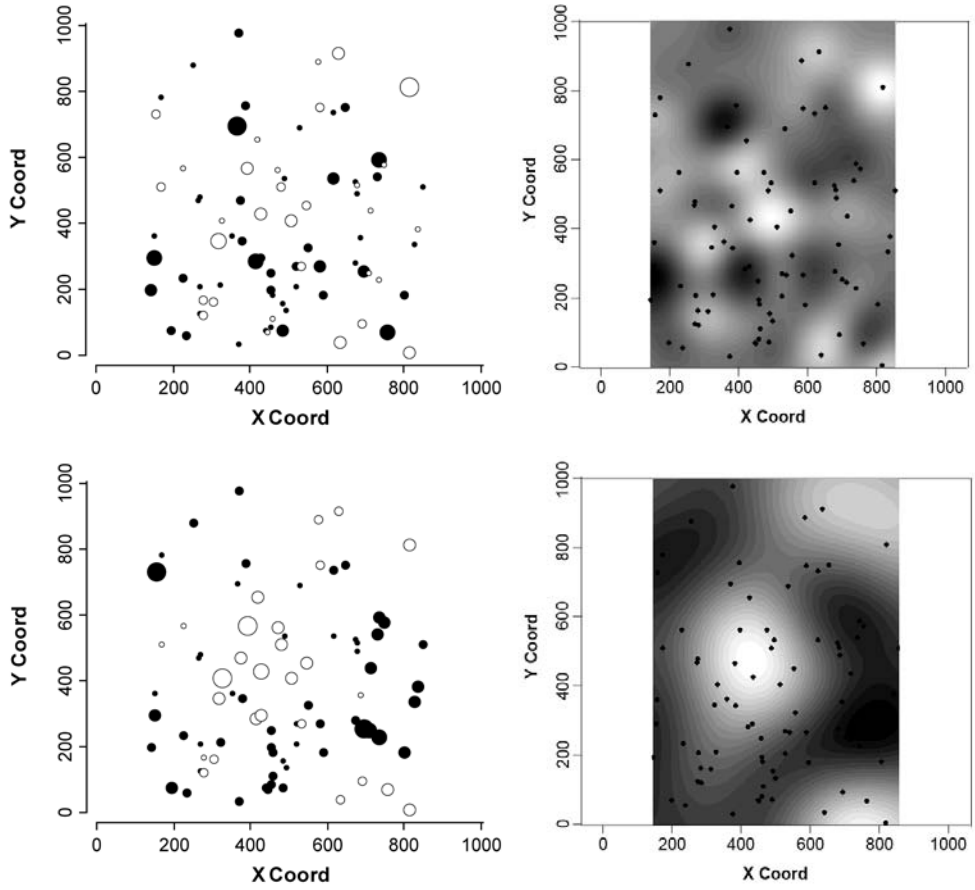

Figure 4. Scatter plots (left) and maps obtained by interpolation (right) of the first synthetic variable obtained by PCA (above) and sPCA (below). Positive and negative values are colored in white and black, respectively.

The values of the PC1 are plotted according to the latitude (Y Coord) and longitude (X Coord) of each sample point, the size of the circles shows the magnitude of the values and the color differentiates positive and negative values. Both, the scatter plot and the map built via the interpolation of the values, show spatial variability since in some areas individuals close in space have similar PC1. However, when sPCA is used to obtain the synthetic variable, the map reveals clearer spatial structure.

To study the spatial structure of genetic variability, other methods make use of multivariate genetic distances as input for geostatistical algorithms. Recently, a non-parametric variogram-based method for autocorrelation analysis between DNA samples that have been genotyped by means of multilocus-multiallele molecular markers has been proposed (Bruno *et al.*, 2008). This method addresses two important aspects of fine-scale spatial genetic analyses: the identification of a non-random distribution of genotypes in space, and the estimation of the magnitude of any non-random structure. The method uses a plot of the squared Euclidean genetic distances vs. spatial distances between pairs of DNA-samples as an empirical variogram (Figure 5). The underlying spatial trend in the plot is fitted by a non-parametric smoothing (LOESS, Local Regression). Finally, the predicted LOESS values are explained by segmented regressions (SR) to obtain classical spatial values such as the extent of autocorrelation. The fit by LOESS/ SR is simpler to obtain than parametric analyses since initial parameter values are not required during the trend estimation process.

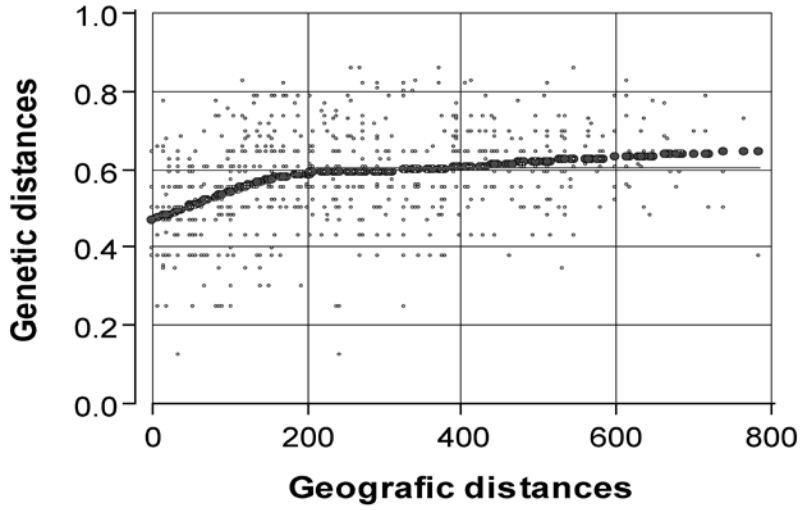

Figure 5. Multivariate genetic distances are plotted against spatial distances (in meters) between sampling points. The smoothed trend (dotted line) and the fitted model (solid line) suggest that spatial autocorrelation is present up to 200 m of distance between samples (range).

Association of genetic markers to other types of data

One of the greatest applications of ordinations in reduced space is the study of associations 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). The need to study organisms as a whole and find associations between data sets of different nature has increased dramatically with the need to observe the complete ecosystem and high throughput technologies that produce an incredible amount of data. In phenomics, for example, it is essential to analyze associations between genotype and phenotypes (Houle *et al.*, 2010).

In the study of genotype-environment relationships, multivariate methods can be used to investigate correlations between genetic data and environmental features (Johnson *et al*., 1969; Mulley *et al*., 1979). Statistical models of such complex interactions are difficult for both computational and biological interpretation aspects. On the contrary, multivariate methods can be used to filter the main signals of genetic data and to study the association between sets of genotypic, phenotypic and environmental data. They are more straightforward and can provide meaningful insight for later statistical modeling.

Generalised Procrustes Analysis (GPA), proposed by Gower in 1975, is an ordination technique used to determine relationships among observations from the simultaneous use of different data types. It allows handling ordinations of a same genotype under different types of descriptors and intends to establish agreement or consensus between them. GPA allows a deeper study of the relationships among relative ordinations of same entities under different types of descriptors to establish concordance between characterizations. In GPA, each data set is analyzed with an appropriate ordination metric according the nature of data. For example, for molecular markers encoded as binary data, PCO using binary similarities/distances metrics is recommended. In cases of morphological markers (continuous) a PCA may be used to represent relationships among entities. GPA then measures the agreement between ordinations, the PCO ordination for molecular markers and the PCA ordination for morphological markers. Bramardi *et al*. (2005) used GPA to determine the relationships amonggenotypes via the simultaneous use of agronomic traits and molecular markers data.

To illustrate the capacity of GPA to generate ordinations that conciliate alternative configurations of same entities, in figure 6 (p. 272) we show an ordination of 20 individuals by two different types of traits and the ordination produced by GPA. Data used in the analyses belong to three populations, which were first arranged according to a Principal Coordinate Analysis of molecular data (expressed as binary profiles of marker data) and then via PCA of quantitative morphological traits. In both ordinations, two groups appear but they are not the same. Only when both types of characterizations are simultaneously considered by GPA, the ordination of entities reflects the true underlying structure, allowing the differentiation of the three groups, which we knew were present in the data.

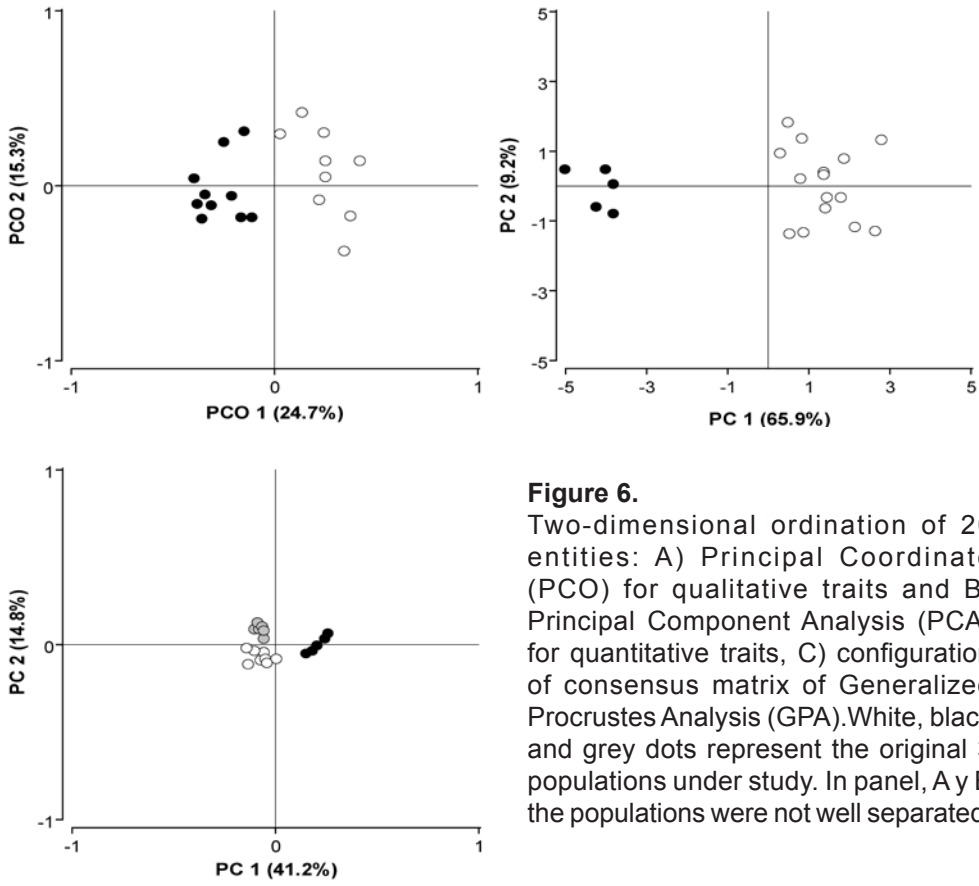

Figure 6.
Two-dimensional ordination of 20 entities: A) Principal Coordinate (PCO) for qualitative traits and B) Principal Component Analysis (PCA) for quantitative traits, C) configuration of consensus matrix of Generalized Procrustes Analysis (GPA).White, black and grey dots represent the original 3 populations under study. In panel, A y B the populations were not well separated.

CONCLUSIONS

Multivariate statistical analyses provide different tools to study genetic variability expressed among multidimensional entities. These techniques are specially designed to perform complex analyses in which different types of variables are involved and have some advantages to other statistical procedures. Firstly,exploratory multivariate techniques do not require distributional assumptions and do not require to work with independent variables; what is more, they benefit from strong correlation structures among variables. This is the situation in most biological studies concerning biodiversity at gene level, in which molecular marker information may not be independent. Secondly, dimension reduction techniques offer the possibility of summarizing the information of multiple traits in few synthetic variables. Such synthetic variables may be difficult to comprehend but the clarity that emerges in the study of variability patterns from these new variablesis invaluable. Synthetic variables that summarize the genetic variability allow the use of other techniques that are designed for univariate data to explore spatial patterns of biodiversity. In most of the cases, the simultaneous use of different

multivariate techniques together with other methodological approaches generates maximum knowledge. However, the exploratory analysis is not enough to statistically contrast scientific hypotheses. To provide trust worthy information it is important to apply these techniques properly and to build the conclusions supported by a complete understanding of the biological system. The adequacy of each technique to deal with genetic data depends not only on the objectives of the study but also in the nature of the data and on the existence of other covariates. The application of multivariate methods to explore genetic biodiversity is a developing field, with a wide range of biological concepts and analytical methods which offer the possibility of making the amount of genetic variability measurable within an ecosystem.

REFERENCES

Anderberg M. (1973) Cluster analysis for applications. New York. Academic Press.pp 359.

Angers B, Magnan P, Angers A, Desgroseiller L. (1999) Canonical correspondence analysis for estimating spatial and environmental effect on microsatellite gene diversity in brook charr (*Salvelinus fontinalis*). *Molecular Ecology*, 8:1043–1054.

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

Balzarini M, Di Rienzo J (2004) InfoGen. Statistical software for genetic data. Universidad Nacional de Córdoba, Argentina.

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

Bramardi S, Bernet G, Asíns M, Carbonell E. (2005) Simultaneous Agronomic and Molecular Characterization of Genotypes via the Generalized Procrustes Analysis: An Application to Cucumber. *Crop Sci* 45(4):1603-1609.

Bruno C, Macchiavelli R, Balzarini M. (2008) Non-parametric modelling of multivariate genetic distances in the analysis of spatial population structure at fine scale. *Theoretical and Applied Genetic*. 117(3): 435-447.

Cavalli-Sforza L. (1966) Population structure and human evolution. Roy. Soc.(Lond.) 164: 362-379. International journal of Biological Sciences 164(995):362–379.

Convention on Biological Diversity. (2010) Global Biodiversity Outlook 3. Montréal. (http://gbo3.cbd.int/ last accessed 19/05/2011).

Eisen M, Spellman P, Brown P, Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns. *PNAS* 95(25):14863-14868.

Escudero D, Cardenoso V, Bonafonte A. (2003) Experimental evaluation of the relevance of prosodic features in Spanish using machine learning techniques, *Eurospeech*, 2309-2312.

Ewens W, Spielman R. (1995) The transmission/disequilibrium test: history, subdivision, and admixture. *The American Journalof Human Genetics* 57:455–464.

Fernandez E, Balzarini M. (2007) Improving cluster visualization in Self-Organizing Maps: Application in Gene Expression Data Analysis. *Computers in Biology and Medicine*, Elsevier Science B.V., Amsterdam. 37(3):1677-1689

Gabriel K. (1971) The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58:453-467.

Gower J. (1975) Generalizad procrustes analysis. *Psychometrika* 40:33-51.

Gower J, Ross G. (1969) Minimum Spanning Trees and Single Linkage Cluster Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 18(1):54-64.

Hanotte O, Bradley D, Ochieng J, Verjee Y, Hill E, Rege J. (2002) African pastoralism: genetic

imprints of origins and migrations. *Science* 296:336–339.

Houle D, Govindaraju D, Omholt S. (2010) Phenomics: the next challenge. *Nature Reviews Genetics* 11:855-66.

Jarraud S, Mougel C, Thioulouse J, Lina G, Meugnier H, Forey F, Nesme X, Etienne J, Vandenesch F. (2002) Relationships between *Staphylococcus aureus* genetic background, virulence factors, *agr* type (alleles) and human disease type. *Infection and Immunity* 70:631-641.

Johnson F, Schaffer H, Gillaspy J, Rockwood E. (1969) Isozyme genotype relationships in natural populations of the harvester ant, *Pogonomyrmex barbatus,* from Texas. *Biochem*. *Genetics*3: 429-460.

Johnson R, Wichern D. (1998) Applied multivariate statistical analysis 4th Ed. *Prentice Hall*, Upper Saddle River, NJ.

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.

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

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

Lao O, Lu T, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden A, Gieger C, Wichmann H, Rüther A, Schreiber S, Becker C, Nürnberg P, Nelson M, Krawczak M, Kayser M. (2008). Correlation between Genetic and Geographic Structure in Europe. *Current Biology* 18(16): 1241-1248.

Levenstien M, Yang Y, Ott J. (2003) Statistical significance for hierarchical clustering in genetic association and microarray expression studies. *BMC Bioinformatics* 4: 62.

Lippman R. (1989) Review of Neural Networks for Speech Recognition.*Neural Computation1*(1). MIT 1-38.

Lukashin A, Fuchs R. (2001) Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters, *Bioinformatics* 17 (5):405–414.

Manel S, Schwartz M, Luikart G, Taberlet P. (2003) Landscape genetics: combining landscape ecology and population genetics. *TRENDS in Ecology and Evolution* 18(4):189-197.

McNeely J, Miller K, Reid W, Mittermeier R,Werner T. (1990) Conserving the World's Biological Diversit. IUCN, WRI, CI, WWF-US, The World Bank.

McRae K, Cree G, Seidenberg M, McNorgan C. (2005) Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments & Computers 37*: 547-559.

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

Mulley J, James C, Barker J. (1979) Allozyme genotype - environment relationships in natural populations of Drosophila buzzatii. *Biochem. Genetics* 17: 105-126.

Nansen C, Campbellc J, Phillips T, Mullen M. (2003) The Impact of Spatial Structure on the Accuracy of Contour Maps of Small Data Sets. *Journal of Economic Entomology* 96(6):1617-1625.

Patterson, N., Price AL and Reich D. (2006) Population Structure and Eigenanalysis. *PLoS Genetics*: 2(12): e190.

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 Evolutionary Biology* 7: 156.

Peña A, Bruno C, Teich I, Fernández E, Balzarini M. (2010) Análisis de conglomerados en la identificación de estructura genética a partir de datos de marcadores moleculares. *Revista TUMBAGA: Ciencia en Construcción* 5:225-237.

Rohlf F, Fisher D. (1968) Test for hierarchical structure in random data sets. *Systematic Zoology* 171: 407-412.

Smouse P, Spielman R, Park M. (1982) Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. *The American Journal of Human Genetics* 119:445.463.

Sokal R, Michener C. (1958) A Statistical Methods for Evaluating Systematic Relationships. *University of Kansas Science Bulletin* 38:1409-1438.

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

Toronen P, Kolehmainen M, Wong G, Castren E, (1999)Analysis of geneexpression data using self-organizing maps, FEBS Letters. 451142–146.

Ward J. (1963) Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58: 236-244.

Wong Y, Ludden T, Bell R. (1983) Effect of erythromycin on carbamazepine kinetics. *Clinical Pharmacology& Therapeutics*33:461-464.