Review

Andrea Peña-Malavera, Cecilia Bruno, Elmer Fernandez and Monica Balzarini*

Comparison of algorithms to infer genetic population structure from unlinked molecular markers

Abstract: Identifying population genetic structure (PGS) is crucial for breeding and conservation. Several clustering algorithms are available to identify the underlying PGS to be used with genetic data of maize genotypes. In this work, six methods to identify PGS from unlinked molecular marker data were compared using simulated and experimental data consisting of multilocus-biallelic genotypes. Datasets were delineated under different biological scenarios characterized by three levels of genetic divergence among populations (low, medium, and high F_{ST}) and two numbers of sub-populations (K=3 and K=5). The relative performance of hierarchical and non-hierarchical clustering, as well as model-based clustering (STRUCTURE) and clustering from neural networks (SOM-RP-Q). We use the clustering error rate of genotypes into discrete sub-populations as comparison criterion. In scenarios with great level of divergence among genotype groups all methods performed well. With moderate level of genetic divergence (F_{ST}=0.2), the algorithms SOM-RP-Q and STRUCTURE performed better than hierarchical and non-hierarchical clustering. In all simulated scenarios with low genetic divergence and in the experimental SNP maize panel (largely unlinked), SOM-RP-Q achieved the lowest clustering error rate. The SOM algorithm used here is more effective than other evaluated methods for sparse unlinked genetic data.

Keywords: cluster analysis; multilocus-biallelic genotypes; plant breeding; self-organizing maps.

DOI 10.1515/sagmb-2013-0006

Introduction

Genetic variability analysis oriented to identify population genetic structure (PGS) in plant collections is a crucial step in the formation of core collections for genetic resources conservation, as well as for plant breeding association studies (Wang et al., 2005; Shriner et al., 2007; Odong et al., 2011). The increasing availability of data from multiple molecular markers allows an exhaustive exploration of genetic diversity in plant species (Bernardo and Yu, 2007). Individual multidimensional molecular marker profiles are analyzed under a knowledge discovery framework, an overall process of finding and interpreting patterns from data (Ultsch, 2005). Hence, cluster analysis (Hartigan, 1975; Gordon, 1999) has proved to be a powerful tool to investigate "natural" groups of genotypes.

A large number of algorithms have been developed to classify individual genotypes into sub-populations using only genetic data. In plant genetics, unsupervised classification of genotypes into discrete or fuzzy

***Corresponding author: Monica Balzarini,** Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba and CONICET (National Council of Scientific and Technological Research), cc 509, 5000 Córdoba, Argentina, e-mail: mbalzari@gmail.com
Andrea Peña Malavera and Cecilia Bruno: Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba and CONICET (National Council of Scientific and Technological Research), cc 509, 5000 Córdoba, Argentina
Elmer Fernandez: Facultad de Ingeniería, Universidad Católica de Córdoba and CONICET, Camino Alta Gracia Km 10, Cordoba, Argentina

nature groups is commonly applied. Here we focus on unsupervised clustering algorithms that do not use any information about the population to which individuals are expected to belong. All of these methods attempt to identify cluster of similar genotypes. Most of these methods are associated with other algorithms to estimate the underlying number of clusters. However, for simplicity, we assume that the number of clusters is known for all studied scenarios.

Some clustering algorithms treat marker information as independent (unlinked models), but others consider linkage disequilibrium (LD) (i.e. correlation between markers). In the context of dense genotyping (or sequencing), the latter has proven to be more powerful (Lawson and Falush, 2012). However, when the marker dataset is largely unlinked the adjustment for correlation between markers does not provide an improvement. Therefore, by assuming LD, several different clustering applications could be applied to describe the underlying PGS.

Several algorithms of different nature have been used to group genotypes by means of genetic data (Lee et al., 2009; Odong et al., 2011; Lawson and Falush, 2012). Hierarchical clustering is a commonly used method that can be directly applied on molecular data (Odong et al., 2011), or after eigenanalysis (Principal Component Analysis) of such data. The latter is to avoid discarding correlated markers (Patterson et al., 2006). Hierarchical and non-hierarchical clustering is applied in a two-stage strategy, since a distance matrix is first constructed and then clustering is performed. These distance-based methods can use different metrics of multivariate similarity between pairs of genotypes. The similarity between individuals may depend not only on their own genetic make-up, but also on that of the rest of the sample (which can be considered by means of allele frequencies) (Weir, 1996). In the first case, the average genetic distance between two genotypes is a simple interpretation of the relatedness (McVean, 2009). The widely used squared Euclidean distance between marker profiles reflects the amount of allele not shared by state between two genotypes (Bruno and Balzarini, 2010; Odong et al., 2011).

The model-based method implemented in STRUCTURE (Pritchard et al., 2000) depends on Bayesian clustering and is usually chosen to identify PGS in germplasm collections with high level of relatedness due to its fuzzy nature. The software STRUCTURE produces a clear visualization of the resulting clustering by means of a bar plot; each individual in the data set is represented in the bar plot by a single vertical line, which is partitioned into K colored segments representing the probability to belong to one specific cluster. To run STRUCTURE can be computationally intensive and time-consuming in scenarios with high numbers of genotypes (Odong et al., 2011). Another alternative for grouping molecular marker profiles is Self-Organizing Maps (SOM) (Kohonen, 1997). It is an unsupervised neural network algorithm able to find relationships between high dimensional data, grouping and mapping them topologically. The central idea supporting the SOM algorithm is that similar objects in the input space will map close to each other into the self-organizing map. In the SOM array, PGS is recognized as a group of nodes (cluster). The identification in SOM is mainly achieved through visualization methods (Ultsch, 2005). An additional algorithm known as SOM-RP-Q (Fernández and Balzarini, 2007) can be used to enhance and facilitate visualization and interpretability of SOM results. In plant breeding, SOM applications to explore relationship between genotypes are not as frequent as in other fields (Toronen et al., 1999).

Because there is a large and diverse set of clustering techniques available, it is necessary to compare the performance of methods belonging to different families to support applications (Lee et al., 2009, Odong et al., 2011). The relative performance of different methods may depend on specific features of the underlying PGS which can be characterized by the number of subpopulations (K) and the genetic similarity among them. Evanno et al. (2005) carried out a simulation study to evaluate the ability of STRUCTURE to recognize PGS under several biological scenarios derived from different migration patterns. Odong et al. (2011) evaluated traditional hierarchical clustering algorithms with molecular marker data under different levels of genetic divergence in a plant germplasm collection using the results of STRUCTURE as a gold standard. Lee et al. (2009) compared PCA, STRUCTURE and non-hierarchical clustering techniques in a simulation study.

Given that methods belonging to different families of clustering techniques differ in the way they handle the clustered objects and have different outputs, the measures of agreements between the

groupings could vary (Milligan and Cooper, 1985). Under a fixed K, the level of agreement between the true grouping of genotypes and the cluster structure obtained by a particular technique can be considered an indicator of the method performance. The scope of this paper was to evaluate the SOM-RP-Q algorithm and other widely used methods in terms of their performance in clustering genotypes from sparse unlinked genetic data.

Materials and methods

Simulated data

Molecular marker data were simulated using QMSim (Sargolzaei and Schenkel, 2009) and scenarios involving numbers of genotypes and molecular markers that mimic maize plant breeding data were designed. A genome with 300 multilocus-biallelic markers, with a random selection design and random mating was used. Six biological scenarios were artificially created, corresponding to three levels of genetic divergence among populations (low, medium, and high F_{ST}) and two numbers of populations (K=3 and K=5). The number of genotypes per scenario was 180. Such population size is common in plant breeding studies. Simulated data were created from an historical population (200 individuals) and the mating system was based on the union of gametes randomly sampled for many generations. The average coancestry was low. Different levels of population genetic divergence were produced by variation of the number of generations from a founder population. Simulated data were coded as 0 and 1, for each marker, and each scenario was replicated several times. The average F_{ST} statistic (Wright, 1951) from the analysis of molecular variance (AMOVA) (Excoffier et al., 2009) was used to quantify the degree of genetic differentiation among populations in each scenario (Table 1).

Experimental data

A public set containing n=334 maize inbred lines genotyped by p=210 SNPs loci was used to identify PGS (Hansey et al., 2011). In a previous work on the same material Hansey et al. (2011) identified eight clusters or sub-populations that were verified from biological knowledge of maize genetic resources. Such classification was used as a gold standard in the current work. Therefore, all of the analyses on the empirical data set were done assuming the existence of eight well known clusters.

A LD heat map plot was performed to evaluate LD (Figure 1). A low level of correlation among the 210 SNPs was found.

Table 1 Number of populations and genetic divergence characterizing the population genetic structure in six simulated scenarios of multilocus-biallelic genotypes.

Scenario	Number of populations (K)	Genetic divergence	
		F_{ST} statistic	Level
I	3	0.07	Low
II	5	0.06	Low
III	3	0.23	Medium
IV	5	0.17	Medium
V	3	0.38	High
VI	5	0.38	High

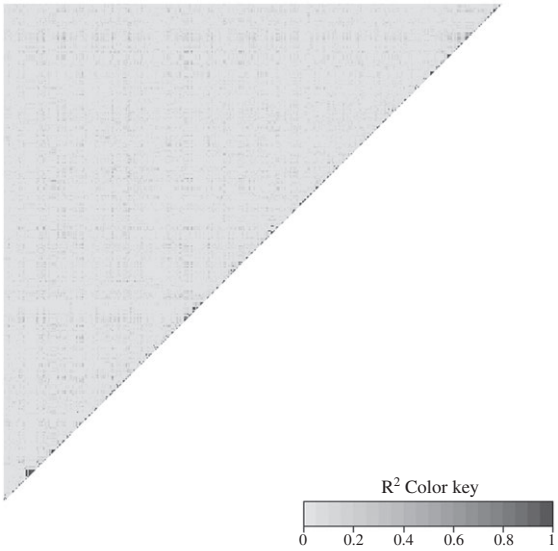

Figure 1 LDheatmap for experimental data. Each element of the upper triangular matrix is a LD measure (R^2) between SNPs. Low R^2 indicates uncorrelated markers.

Clustering algorithms

Distance-based clustering

Hierarchical and non-hierarchical cluster methods were implemented. Two hierarchical algorithms: Average Linkage or UPGMA (Sokal and Michener, 1958) and Ward (1963), both on the molecular markers data were used. The Ward algorithm was also applied to the principal components of the molecular data previously recognized as significant according to Tracy-Widom (TW) distribution (Tracy and Widom, 1994). The principal component analysis together with a modern statistic (TW theory) provides a fast and effective way to answer if individuals can be regarded as randomly chosen from a homogeneous population or if the data implies the existence of underlying PGS. Patterson et al. (2006) modeled the eigenvalues using the TW distribution, which describes the largest eigenvalue of the matrix ZZ^T, where Z is a matrix of molecular marker data. Therefore, TW can be used for assessing significant principal components. The combination of the eigenanalysis as a first step of the Ward algorithm was denoted as PCA+Ward.

In the UPGMA the clustering process starts from a multivariate distance matrix containing all pairwise distances between genotypes. Then, genotypes are clustered using a grouping criterion based on average unweighted distances. In the Ward method the process is similar, but the grouping criterion is based on weighted (by covariances) average distances. To display hierarchical clustering, dendrograms are used, i.e., tree-based representations that track the history of union between genotypes and clusters (Johnson and Wichern, 1998).

Our paper also included the non-hierarchical K-Means method (MacQueen, 1967). The non-hierarchical clustering algorithm starts with an initial (random) partition of the genotypes in K groups and continues by reallocating each genotype in one of the clusters such that the distance between the genotype and the centroid of the cluster to which it was assigned is lower than the distance to any other centroid. Then it brings together individuals in K clusters so that the difference between clusters is maximized and the differences within each cluster are minimized (objective function). Results are visualized by plotting the objective function values for different K.

Model-based clustering

The fuzzy clustering method implemented in the software STRUCTURE (Pritchard et al., 2000) was also evaluated. This method based on Bayesian Markov Chain theory is used to assign genotypes to a *K* specific structure. Linkage equilibrium between markers and Hardy-Weinberg equilibrium are assumed. Genotypes in the collection are assigned probabilistically to these groups, or jointly by two or more populations if their genotype indicates a mixture of molecular patterns. STRUCTURE (Pritchard et al., 2000) produces a visualization of the resulting PGS by means of a bar plot; each individual in the data set is represented by a single vertical line, which is partitioned into *K* colored segments that represent that individual's estimated membership fraction in each of the *K* inferred clusters. One of the good properties of fuzzy clustering is that posterior probabilities indicate the uncertainty of the cluster assignment.

Heuristic clustering

A Self-Organizing Map (SOM) (Kohonen, 1997) is an artificial neural network capable of converting high dimensional data into a two-dimensional map in which data points that are found close together on the map are more similar than those that are farther away. A SOM consists of two layers of artificial neurons, the input layer and the output layer (Paini et al., 2010). In the SOM, the input layer is essentially the raw data and comprises *p* neurons (one neuron for each molecular marker in our case), with each neuron connected to all *n* individual genotypes. The output layer is the two-dimensional map comprising *n* artificial number of neurons (nodes), laid out in a grid. Each of the genotypes occupies a particular point in the *p* dimensional space. The SOM projects its nodes onto this space via neuron weight vectors; each SOM neuron occupies a point in the same multidimensional space as the genotypes, thereby interacting with the genotypes (Worner and Gevrey, 2006). When the analysis is initiated, each raw data point is assessed and the neuron that is closest to this data point in this multidimensional space is deemed to be the winning node or best matching node. The neuron weight vector of the winning node is adjusted so that it moves closer to the data point. Because all neurons are connected together, the process of the neuron moving exerts force that drags other neurons in the SOM. When the analysis is complete each data point will have a winning node, which is its closest neuron. In this way, genotypes that have very similar molecular marker assemblages will be located close together in the multidimensional space and will have the same winning node. In SOM, clusters are identified through specific visualization algorithms, such as the RP-Q method (Fernández and Balzarini, 2007). In this method a relative position (RP) is attached to each node in the SOM structure. The RP is a new node adaptive attribute that moves in a two-dimensional virtual space imitating the movement of the neurons. The final RP of the net nodes is shown in a scatter plot. Each node is represented by a circle whose diameter is proportional to the "activation frequency" (frequency of winning) during the learning phase. To assist in the identification of similarities, the nodes are linked. The length of the segment joining two nodes provides information about distances between them. The nodes will be linked to each other to form a cluster if the distance between them is less than or equal to a specified threshold value. The Q-statistic is used to obtain the classification. The Q-statistic handles isolated nodes for which codebook vectors are often a mixture of patterns from other clusters instead of representing a true pattern in the input space (transition nodes) (Fernández and Balzarini, 2007).

Algorithm specifications

To implement hierarchical and non-hierarchical clustering we used the squared Euclidean distance between genotypes. Hierarchical and non-hierarchical clustering were performed using Info-Gen (Balzarini and Di Rienzo, 2004) assuming a known number of clusters (K=3 and K=5 for simulated data, and K=8 for experimental data). STRUCTURE was run under the assumption of an admixture model with an independent allele

frequency model using the number of clusters (K) as prior information with a burn-in time and with a replication number of 50,000 for each run. The L-SOM software was used for training network in the SOM algorithm and RelPos visualization tool of Matlab® to implement the SOM-RP-Q method (Fernández and Balzarini, 2007). An array of 10×10 nodes (net sizes) was used to start the learning process.

Comparison criterion

To simultaneously compare the performance of methods to infer the underlying grouping of genotypes, a clustering error rate (CER) was used. CER is defined as follows:

$$\text{CER}=\frac{\sum_{i}^{K} Er_i}{N}$$

where N is the number of genotypes, K is the number of groups, Er_i is the cluster error rate in the i-th group, with i=1, …, K. It is based on the difference between the true number of individuals belonging to the i-th group (N_i) and the individuals correctly classified into that group (C_i). The classification error rate for the i-th group was estimated as $Er_i=(N_i-C_i)/N_i$. CER is the average of the classification error rate through the K groups. For STRUCTURE, genotype classification was based on the highest cluster proportion for that individual. For SOM-RP-Q, classification of nodes into clusters (and consequently classification of genotypes) was based on the number of individuals of each underlying group at a given node. If the highest proportion of individuals belonged to group i, then the node was assumed to be a component of cluster i. The proportion of individuals of group i in the node classified as part of cluster i (named posterior probability) was used as indicator of the certainty in cluster assignment. The average, across nodes, of its complement was used as a measure of uncertainty of the cluster assignment by using SOM-RP-Q.

Results

A multidimensional scaling was performed on genomic data to display the level of genetic divergence of each simulated scenario. Figure 2 shows the scatter plots of the first two axes resulting from the multidimensional scaling of data from six scenarios with different levels of F_{ST} (low, medium, and high, from top to bottom, respectively) and numbers of clusters (K=3 and K=5).

In Figure 3 real genotypes are ordered in the plane formed by the first two axes resulting from the multidimensional scaling of maize experimental data. This figure provides information regarding genetic distance between genotypes and the group to which these genotypes were classified by Hansey et al. (2011). The F_{ST} statistic calculated from the data was 0.02, i.e., closer to the simulated values that indicate low divergence.

Analysis of simulated data

In scenarios where genetic divergence was low (F_{ST}<0.10), the SOM-RP-Q method performed best, with CER below 0.25, whereas the clustering error of the hierarchical methods reached 0.73 (Table 2). The estimate of posterior probability in the SOM-RP-Q algorithm indicated that, even under low genetic differentiation, the uncertainty of the cluster assignment was lower than 0.10.

For the PCA+Ward method the average number of significant components decreased with increasing F_{ST} statistic, i.e., with greater divergences between populations. In scenarios I and II, with low genetic difference, the TW statistic identified more than 20 significant components. Clustering errors were similar to those observed for other hierarchical methods used without previous eigenanalysis. For more complex structures

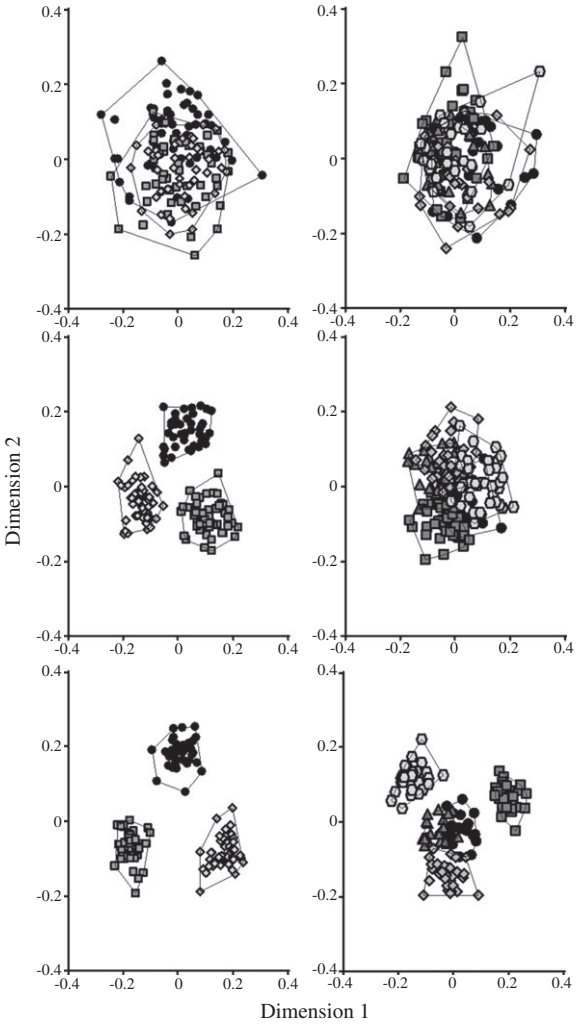

Figure 2 Scatter plots of the first two axes resulting from principal coordinates analysis of molecular data. On the left scenarios involving three populations, while at the right stage scenarios with five populations. From top to botton low (F_{ST}=0.06–0.07), medium (F_{ST}=0.23–0.17), and high (F_{ST}=0.38).

(K=5), the CER increased in all methods (Figure 4). With strong population structure (scenarios V and VI) all methods performed well, having low clustering error rates (CER<0.02) (Table 2). Hierarchical methods were competitive only in scenarios with high levels of divergence (V and VI), both when applied directly on genomic data and when used on the significant synthetic variables.

The trend in the CER was dependent on the level of genetic divergence between populations (Figure 4). The SOM-RP-Q method (black solid line) showed the lowest clustering error (Figure 4). For F_{ST} values close to 0.3, the error reached negligible values. CER trends for K-Means and STRUCTURE were similar for both values of K. However, the posterior probability given by STRUCTURE indicates uncertainty of the cluster was only neglected with high F_{ST}, but these reached average values of 0.40 under the low F_{ST} scenario. On the simulated data sets considered here, STRCUTURE and SOM had similar performances under the same biological scenarios.

Analysis of experimental data

The methods SOM-RP-Q, Ward and STRUCTURE recognized the underlying structure (eight clusters) well, with clustering errors of 0.09, 0.21, and 0.23, respectively. By contrast, the methods UPGMA, K-Means and

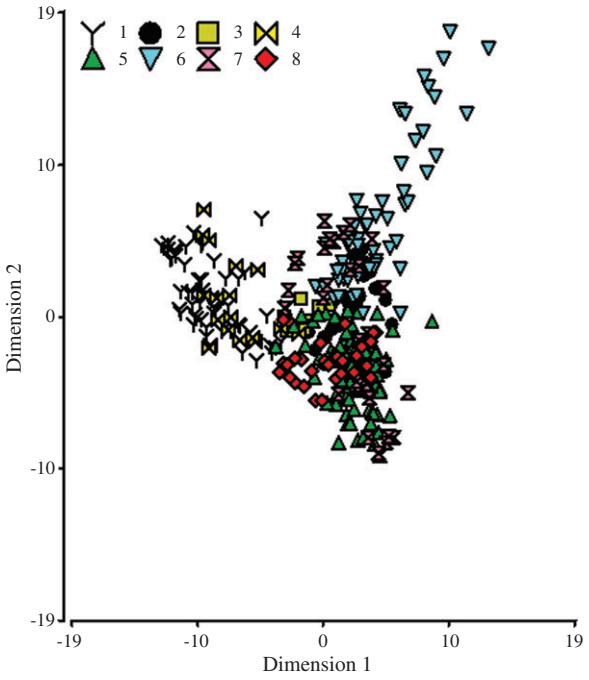

Figure 3 Scatterplot of the first two axes resulting from principal coordinates analysis of molecular data (SNPs) from eight groups of experimental maize data.

Table 2 Proportion of classification error of the evaluated algorithms to identify PGS from simulated molecular data.

Genetic divergence		Populations	Clustering method					
Level	F_{ST} statistic	*K*	UPGMA[a]	Ward[b]	PCA+Ward[c]	K-Means[d]	STRUCTURE[e]	RP-Q-SOM[f]
Low	0.07	3	0.61	0.56	0.57 (22)	0.52	0.48	0.17
Low	0.06	5	0.73	0.68	0.71 (20)	0.69	0.68	0.22
Medium	0.23	3	0.09	0.03	0.09 (8)	0.01	0.02	0.01
Medium	0.17	5	0.47	0.22	0.15 (9)	0.08	0.09	0.03
High	0.38	3	0	0	0 (7)	0	0	0
High	0.38	5	0.04	0.01	0 (6)	0	0	0

[a]Unweighted Pair Group Method using Arithmetic Average.
[b]Ward Method.
[c]Ward algorithm applied to the principal components (PC) which were significant according to Tracy-Widom (TW), the average number of PCs used is shown between brackets.
[d]Non hierarchical method K-Means.
[e]Bayesian method implemented in STRUCTURE.
[f]The relative position self organizing map method.

PCA+Ward, were not able to identify the underlying PGS. The dendrogram plot, which is shown in the top-right panel, is the resulting diagram from the hierarchical clustering analysis. The "x" axis in the dendrogram indicates distance between individual genotypes and clusters (Figure 5).

In the Figure 5 bottom panel, a bar plot from STRUCTURE is shown, where vertical lines represent individuals. On each line, the probability of belonging to each group is represented by a color code. The figure also shows the graphical output of SOM-RP-Q applied to the experimental maize data. In the SOM-RP-Q plot (top-left panel), small circles represent SOM nodes and are grouped into clusters (marked with ellipses) according their relative position. Node sizes are proportional to the winning frequency on the node during the training phase, and the numbers in each node represent their position in the SOM structure. The arbitrary RP space is shown on the "x" and "y" axis.

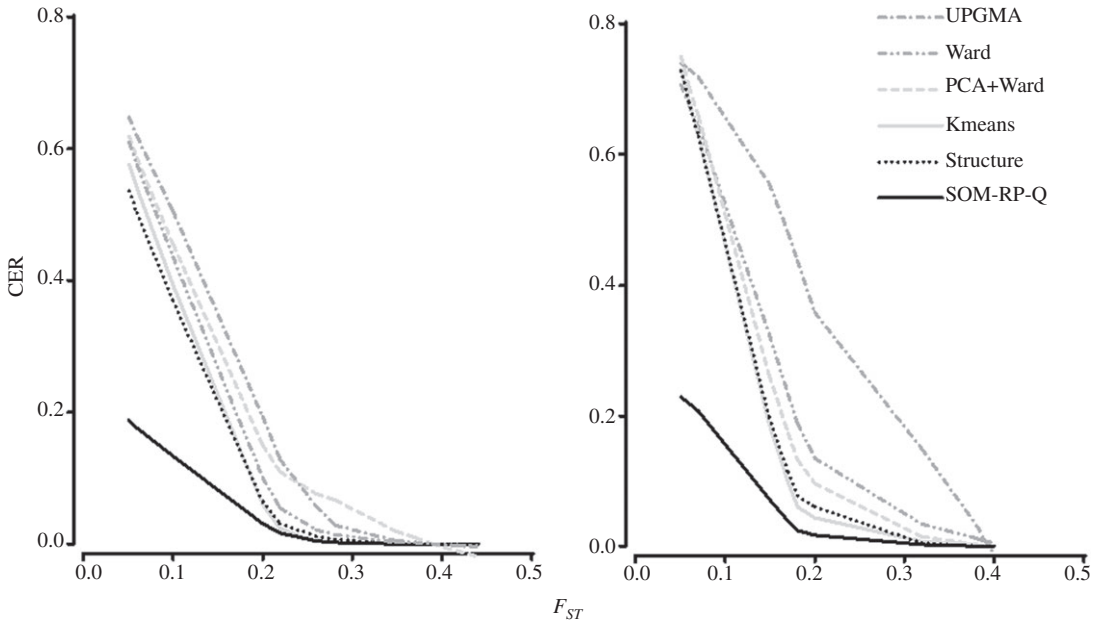

Figure 4 Clustering error rate with respect to the level of genetic divergence (F_{ST}) among populations for PGS with three (left) and five subpopulations (right). Six clustering procedures are compared.

Discussion

Identifying genotype clusters from marker profiles is a common task in genetics and molecular biology. Cluster analysis has proved to be a powerful tool to investigate "natural" groups of molecular marker data. In most cluster analyses, groups are not known a priori, and the interest is focused on finding them without the help of a response variable (unsupervised learning). Hierarchical (e.g. UPGMA and Ward's method) and non-hierarchical (e.g. K-Means) clustering algorithms provides a simple and powerful tool for determining the PGS of germplasm collections using molecular marker data (Odong et al., 2011). The analysis can be conducted directly from/onto the data sets and followed by other procedures to estimate the number of clusters in the underlying PGS (Gordon, 1999; Still and Bialek, 2004; Tibshirani et al., 2001).

Recent studies looking for PGS, show that molecular information can also be effectively handled by model-based clustering and principal component analysis. The former (probabilistically) assigns individuals in the sample to clusters, or jointly to two or more populations if their genotypes indicate that they are admixed. The widely used software STRUCTURE allows us to implement a model-based clustering that can be applied to most of the genetic markers, provided that they are not closely linked (Pritchard et al., 2000). The approach to the study of PGS from principal component analysis provides a formal test for the presence of structure in the genetic data (Patterson et al., 2006). The performance of PCA, as the other of distance-based clustering techniques, depends only on subgroup differentiation and does not require other biological assumptions.

In this work these generic clustering algorithms were compared, among them and with other clustering method from the artificial neural network family (SOM-RP-Q). The overall performance of the selected methods for cluster marker profiles were evaluated over several simulated scenarios and using a well-known experimental data set of maize inbred lines (Hansey et al., 2011). SOM-RP-Q produced a good estimation of the number of simulated clusters in the artificial data sets, being robust to the appearance of isolated nodes (Fernández and Balzarini, 2007). The identification of the nodes belonging to the clusters makes it possible to calculate uncertainty coefficients in a cluster assignment.

The results showed that UPGMA hierarchical clustering applied to molecular data was not efficient to detect population genetic structures in cases of low genetic divergence. At medium levels of population

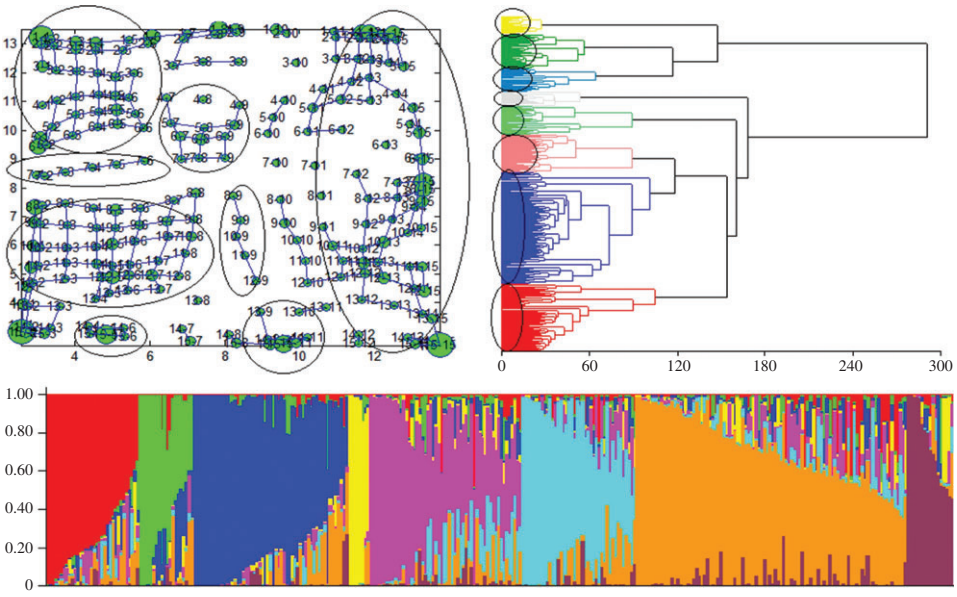

Figure 5 Graphical output of the SOM-RP-Q (top-left panel), hierarchical Ward (top-right panel) and the STRUCTURE (bottom panel) methods applied to the experimental maize data.

genetic divergence ($0.1<F_{ST}<0.25$), UPGMA improved its relative performance in the situations of simple population structure (e.g.: 3 subpopulations) but maintained high clustering error rates with more complex structures (e.g.: 5 subpopulations). Therefore, this cluster analysis cannot be the recommended as a choice for determining the genetic structure of a genotype set, except in situations in which high divergence among plant populations is expected. Odong et al. (2011) suggested that the poor performance of UPGMA method in recovering the original structure with low divergence is because it produces highly unbalanced clusters.

At higher levels of genetic divergence ($F_{ST}>0.2$), the performance of all procedures became similar and acceptable. The differences in classification error between Ward and UPGMA in real data confirmed the findings from simulation results that show better performance of Ward than UPGMA under lower genetic divergence. Jobson (1992) discusses the ability of Ward to keep outlying accessions within clusters. Therefore, Ward's algorithm was the preferred choice between the hierarchical clustering methods. However, our results indicate that SOM-RP-Q performed the best in classifying molecular marker profiles and discovering PGS. SOM-RP-Q recovered the underlying populations even for relatively low divergence among subpopulations ($F_{ST}<0.1$). SOM-RP-Q provided the means to map molecular markers profiles in a two-dimensional space. The inner net structure allows inference about the number of clusters emerging over the learned net. The performance of STRUCTURE was close to that of SOM-RP-Q, but the identification of the underlying PGS by means of SOM-RP-Q was faster. The model-based algorithm in the software STRUCTURE uses Markov Chain method for parameter estimation, which is computationally time-consuming with respect to neural networks algorithms (Roux et al., 2007; Lee et al., 2009).

Wang et al. (2002) concluded that by using neural networks as an intermediate step to analyze genome wide gene expression data, the gene expression patterns can be more easily revealed than by using hierarchical clustering and a non-hierarchical clustering. Evanno et al. (2005) stated that STRUCTURE was able to detect the subpopulations in simulated data sets according to an island migration model. Our results agree with findings of Evanno et al. (2005), but only when the subpopulation divergence is not very low. Zhao et al. (2007) reported that the model-based clustering method could not adequately define the structural complexity of the populations.

Lawson and Falush showed that model-based algorithms could outperform generic clustering approaches under LD among genetic markers and high relatedness between individuals. They discussed the problem of high statistical correlation between physically close sites and the subsequent loss

of information to calculate similarities between individuals. Model-based algorithm FineSTRUCTURE (Lawson et al., 2012), is recommended to be used for clustering with LD among markers. This method was not used here because, given the low amount of markers (mostly unlinked) (200–300 SNPs), the data sets could be largely unlinked. Lawson and Falush (2012) worked in the context of full sequenced data, simulating data sets that contain 500,000 SNPs for a small number of individuals (n=140). The methods included in our paper assume that each marker is independent of most of the other markers in the panel (>95%). Such unlinked models give equal weight to all markers in the study. By using QMSim, we simulated data in which statistical correlation between markers was low, and in the experimental maize data used as illustration, the levels of ancestry and the LD were also low. When high ancestry is expected, the similarity measures that include the linkage concept may provide an extra gain during the clustering process. The present study also showed that non-hierarchical cluster K-Means performed comparatively well with respect to SOM-RP-Q under the different scenarios. Similar results were reported by Lee et al. (2009), but the use of non-hierarchical cluster K-Means with the experimental data set was not successful to recovery PGS in the data set.

Based on the results of this work, we recommend using the algortihms SOM-RP-Q and the one implemented in the software STRUCTURE to infer PGS from molecular markers rather than other of the evaluated clustering procedures, especially with genotypes with low genetic divergence. SOM is an efficient method for analyzing systems ruled by complex non-linear relationships and provides an alternative to traditional statistical methods for classifying complex data (Nikolic et al., 2009; Worner and Gevrey, 2006). SOM is widely used for knowledge discovery, pattern recognition, and clustering and visualization of large multidimensional data sets. Worner and Gevrey (2006) found that SOM was able to reduce very high dimensional data into patterns that could be usefully interpreted. SOM-RP-Q analysis can not only perform significant data reduction but also could provide information about typical genotype marker profiles representing each cluster (Fernández and Balzarini, 2007). While other clustering techniques are more conventional methods used for the analysis of molecular marker profiles in plant breeding, SOM-RP-Q is an unsupervised learning algorithm able to uncover PGS that preserve topology of clusters into a two-dimensional form, which allows its visualization. This qualifies SOM to be a good tool for molecular marker profile clustering. The SOM algorithm used here resulted more effective than other evaluated methods for sparse unlinked genetic data.

Acknowledgments: The authors would like to acknowledge two anonymous reviewers for their critical review of the manuscript, the scientific contributions of Natalia de Leon, German Muttoni, Margot Tablada, Miguel Di Renzo, Ingrid Teich and Jorgelina Brasca and the databased depuration made by Marcos Perrachione. This work was supported by the National Council of Scientific and Technological Research (CONICET), Argentina.

References

Balzarini, M. and J. Di Rienzo (2004): Info-Gen 2010, Universidad Nacional de Cordoba, Córdoba.

Bernardo, R. and J. Yu (2007): "Prospects for genome-wide selection for quantitative traits in maize," Crop Sci., 47, 1082–1090.

Bruno, C. and M. Balzarini (2010): "Distancias genéticas entre perfiles moleculares obtenidos desde marcadores multilocus multialélicos," Revista de la Facultad de Ciencias Agrarias UNCuyo, 41, 11.

Excoffier, L., T. Hofer and M. Foll (2009): "Detecting loci under selection in a hierarchically structured population," Heredity, 103, 285–298.

Evanno, G., S. Regnaut and J. Goudet (2005): "Detecting the number of clusters of individuals using the software structure: a simulation study," Mol. Ecol., 14, 2611–2620.

Fernández, E. A. and M. Balzarini (2007): "Improving cluster visualization in self-organizing maps: Application in gene expression data analysis," Comput. Biol. Med., 37, 1677–1689.

Gordon, A. (1999): Clustering, 2nd edition, Chapman & Hall/HRC Press: London.

Hansey, C. N., J. M. Johnson, R. S. Sekhon, S. M. Kaeppler and Nd Leon (2011): "Genetic diversity of a maize association population with restricted phenology," Crop Sci., 51, 704–715.

Hartigan, J. A. (1975): Cluster algorithms, Wiley: New York.

Jobson, J. D. (1992): Applied multivariate data analysis: categorical and multivariate methods, Springer-Verlag, New York.
Johnson, R. A. and D. W. Wichern (1998): Applied multivariate statistical analysis, 3rd edition. Prentice Hall, New Jersey.
Kohonen, T. (1997): Self-organizing maps, 2nd edition, Springer: Berlin.
Lawson, D. J. and D. Falush (2012): "Population identification using genetic data," Annu. Rev. Genomics. Hum. Genet., 13, 337–361.
Lawson, D. J., G. Hellenthal, S. Myers and D. Falush (2012): "Inference of population structure using dense haplotype data," PLoS Genet., 8, e100245.
Lee, C., A. Abdool and C.-H. Huang (2009): "PCA-based population structure inference with generic clustering algorithms," BMC Bioinformatics, 10, S73.
MacQueen, J. (1967): "Some methods for classification and analysis of multivariate observations," *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability* 1, p. 17.
McVean, G. (2009): "A genealogical interpretation of principal components analysis," PLoS Genet., 5, e1000686.
Milligan, G. and M. Cooper (1985): "An examination of procedures for determining the number of clusters in a data set," Psychometrika, 50, 159–179.
Nikolic, N., Y. S. Park, M. Sancristobal, S. Lek and C. Chevalet (2009): "What do artificial neural networks tell us about the genetic structure of populations? The example of European pig populations," Genet. Res. (Camb), 91, 121–132.
Odong, T., J. van Heerwaarden, J. Jansen, T. van Hintum and F. van Eeuwijk (2011): "Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data?," TAG Theor. Appl. Genet., 123, 195–205.
Paini, D. R., S. P. Worner, D. C. Cook, P. J. De Barro and M. B. Thomas (2010). "Using a self-organizing map to predict invasive species: sensitivity to data errors and a comparison with expert opinion," J. Appl. Ecol., 47, 290–298.
Patterson, N., A. Price and D. Reich (2006): "Population structure and eigenanalysis," PLoS Genet., 2, e190.
Pritchard, J., M. Stephens and P. Donnelly (2000): "Inference of population structure using multilocus genotype data," Genetics, 155, 945–959.
Roux, O., M. Gevrey, L. Arvanitakis, C. Gers, D. Bordat and L. Legal (2007): "ISSR-PCR: tool for discrimination and genetic structure analysis of Plutella xylostella populations native to different geographical areas," Mol. Phylogenet. Evol., 43, 240–250.
Sargolzaei, M. and F. Schenkel (2009): "QMSim: a large-scale genome simulator for livestock," Bioinformatics, 25, 680–681.
Shriner, D., L. Vaughan, M. Padilla and H. Tiwari (2007): "Problems with genome-wide association studies," Science, 316, 1840–1842.
Sokal, R. and C. Michener (1958): "A statistical methods for evaluating systematic relationships," University of Kansas Science Bulletin, 38, 29.
Still, S. and W. Bialek (2004): "How many clusters? An information theoretic perspective" Neural Comput. 16, 2483–2506.
Tibshirani, R., G. Walther and T. Hastie (2001): "Estimating the number of clusters in a data set via the gap statistic," J. R. Statist. Soc. B., 63, 411–423.
Toronen, P., M. Kolehmainen, G. Wong and E. Castren (1999): "Analysis of gene expression data using self-organizing maps," FEBS Lett., 451, 142–146.
Tracy, C. A. and H. Widom (1994): "Level-spacing distributions and the Airy kernel," Comm. Math. Phys., 159, 151–174.
Ultsch, A. (2005). Clustering with SOM: U*C. WSOM 2005, Paris.
Wang, J., J. Delabie, H. Aasheim, E. Smeland and O. Myklebost (2002): "Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study," BMC Bioinformatics, 3, 36.
Wang, W., B. Barratt, D. Clayton and J. Todd (2005): "Genome-wide association studies: theoretical and practical concerns," Nat. Rev. Genetics, 6, 109–118.
Ward, J. (1963): "Hierarchical grouping to optimize an objective function," J. Am. Stat. Assoc., 58, 236–244.
Weir, B. S. (1996): Genetic data analysis II: methods for discrete population genetic data, Sinauer Assoc., Inc.: Sunderland, MA, USA.
Worner, S. P. and M. Gevrey (2006): "Modelling global insect pest species assemblages to determine risk of invasion," J. Appl. Ecol., e43, 858–867.
Wright, S. (1951): "The genetical structure of populations," Ann. Eugen., 15, 31.
Zhao, K., M. J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram and M. Nordborg (2007): "An arabidopsis example of association mapping in structured samples," PLoS Genet, 3, e4.