Genome analysis

Identifying molecular markers associated with classification of genotypes by External Logistic Biplots

J. R. Demey¹,*, J. L. Vicente-Villardón², M. P. Galindo-Villardón² and A. Y. Zambrano³

¹Centro de Biotecnología, Instituto de Estudios Avanzados (IDEA), Caracas, Venezuela, ²Departamento de Estadística, Universidad de Salamanca, Salamanca, España and ³Instituto Nacional de Investigaciones Agrícolas (INIA-CENIAP), Maracay, Venezuela

Received on April 26, 2008; revised on September 30, 2008; accepted on October 22, 2008

Associate Editor: Dmitrij Frishman

ABSTRACT

For characterization of genetic diversity in genotypes several molecular techniques, usually resulting in a binary data matrix, have been used. Despite the fact that in Cluster Analysis (CA) and Principal Coordinates Analysis (PCoA) the interpretation of the variables responsible for grouping is not straightforward, these methods are commonly used to classify genotypes using DNA molecular markers. In this article, we present a novel algorithm that uses a combination of PCoA, CA and Logistic Regression (LR), as a better way to interpret the variables (alleles or bands) associated to the classification of genotypes. The combination of three standard techniques with some new ideas about the geometry of the procedures, allows constructing an External Logistic Biplot (ELB) that helps in the interpretation of the variables responsible for the classification or ordination. An application of the method to study the genetic diversity of four populations from Africa, Asia and Europe, using the HapMap data is included.

Availability: The Matlab code for implementing the methods may be obtained from the web site: http://biplot.usal.es.

Contact: jhony.demey@gmail.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

To characterize and evaluate the genetic diversity, various molecular techniques have been employed, including Restriction Fragment Length Polymorphisms (RFLP), Random Amplified Polymorphic DNAs (RAPD), Amplified Fragment Length Polymorphisms (AFLP), Sequence Tagged Sites (STS), Simple Sequence Repeats (SSR) or microsatellites, Single Nucleotide Polymorphisms (SNPs) (Avise, 2004).

Molecular information is analyzed according to the type of marker and organism. Generally, a binary data matrix is obtained from amplified fragments coded as ‘presence’ (1’s) or ‘absence’ (0’s). Genetic relationships among genotypes are investigated using different techniques of classification/ordination such as unweighted pair group method with arithmetic mean (UPGMA) or Neighbor-joining clustering algorithm (Saitou and Nei, 1987; Sneath and Sokal, 1973) and Principal Coordinates Analysis (PCoA) (Gower, 1966). Despite its extensive use, these techniques of classification/ordination do not permit to study genotype–allele and allele–allele relations appropriately, i.e. it is not possible to determine which alleles or bands are responsible for (or associated to) the classification of genotypes.

An approach that facilitates the genetic interpretation, compared to the classic techniques of classification/ordination, is provided by the Biplot methods (Chapman et al., 2002; Gabriel, 1971; Sharov et al., 2005), that is, a simultaneous graphical representation of the rows (individuals) and the columns (variables) of a given data matrix. The main uses are exploratory, although it has also been used as a graphical representation for more formal models (Gabriel, 1998). The biplot can be fitted by performing alternating regressions and interpolations (Gabriel and Zmir, 1979; Gower and Hand, 1996; Jongman et al., 1995). However, when data are binary, like those obtained in the analysis of molecular information, Classical Linear Biplots and Principal Components Analysis (PCA) are not suitable because the response along the dimensions is linear. This is the same reason why linear regression is not appropriate for binary or categorical data.

Several strategies can be used in order to fit biplots from binary data matrices: Multiple Correspondence Analysis (MCA) can be considered as a particular form of biplot for a binary matrix, where the prediction regions are based on distances from the individual points to the category points (Gower and Hand, 1996); when modeling a two-way table using a bilinear model, the parameter estimates are obtained by an iterative process of alternating generalized row and column regressions (Falguerolles, 1998; Gabriel, 1998; van Eeuwijk, 1995a, b; Vicente-Villardón et al., 2006). Nevertheless MCA depends on the chi-squared distance, which does not reflect the structure of our data. In this case, PCoA provides a more flexible alternative, because we can use different similarity/dissimilarity measures to extract the genetic relationships among genotypes.

Vicente-Villardón et al. (2006) described the geometry of a linear biplot for binary data in which the response along the dimensions is logistic (Logistic Biplots, LB). In the LB, each individual is represented as a point and each variable as a direction through the origin. The projection of an individual point onto a character direction predicts the probability of presence of that character. The method is related to LR in the same way that biplot analysis is related to linear regression.

*To whom correspondence should be addressed.
In this article, we use a combination of PCoA, Cluster Analysis (CA) and External Logistic Biplots (ELB) on the principal coordinates, as a better way to identify the alleles or bands that are responsible for the classification of genotypes. The proposal is based on the fact that the column regression in the alternating procedure for binary data is simply a LR that can be fitted to the configuration obtained from PCoA. Although the whole alternating procedure could have been used, PCoA is simpler, more accessible to applied researchers and, in the authors’ experience, the results are similar. On the other hand, the alternating procedures, as described in the literature, although share the same geometry; need some adaptations for binary data matrices.

We have taken an exploratory point of view as opposed to the modeling approach in papers by van Eeuwijk (1995a, b), Gabriel (1998) or Falguerolles (1998). The main aim is to analyze a data matrix (individuals by variables) rather than to model a two-way (contingency) table using a bilinear model. Our proposal is closely related to MCA and some psychometric latent variable procedures such as item response theory or latent traits.

Measures of the quality of the representation of individuals, groups of individuals and variables (alleles or bands) are also defined. It is shown that this approach facilitates the genetic interpretation as compared to traditional clustering methods.

Some theory is developed for the proposal and a simulation study shows the performance of the methodology. An application of the method to study the genetic diversity of four populations from Africa, Asia and Europe, using the HapMap data is included.

### 2 METHODS

Let \( X \) be the matrix of binary data obtained from amplified fragments that were scored as present or absent (1 or 0), in which the rows correspond to \( n \) individuals or entries (genotypes) and the columns to \( p \) binary characters (alleles or bands). Let \( S = (s_{ij}) \) be a matrix containing the similarities among genotypes, obtained from the binary data matrix \( X \), and let \( \Delta = (\delta_{ij}) \) be the corresponding dissimilarity/distance matrix, taking for example \( \delta_{ij} = 1 - s_{ij} \).

The algorithm starts with a PCoA, as a technique of ordination of the individuals (genotypes). PCoA is concerned with the problem of constructing a configuration of \( n \) points in an Euclidean space in such a way that the distance between any two points of the configuration approximates, as closely as possible, the dissimilarity \( \delta_{ij} \) between genotypes represented by these points. The objective is then to find a configuration \( Y \) in a lower dimensional Euclidean space \( \mathbb{R}^k \) whose inter-point distance matrix \( D \) is as close as possible to \( \Delta \). When the observed dissimilarity/distance measured is ‘Euclidean’, it is possible to find an exact configuration in \( n - 1 \) dimensions. A lower dimensional approximation can be obtained projecting onto the first \( k \) principal coordinates (usually \( k = 2 \)). The theoretical considerations and demonstrations of the method can be found in Mardia et al. (1979).

In PCoA, it is known that the proportion of the total variance explained by \( k \) dimensions (overall goodness of fit or overall quality of representation) can be considered as an average of the \( n \) points in the graphical representation. However, a good overall fit does not imply that all the individuals have the same quality of representation and then that the interpretation of the positions of all the points in the diagram is equally reliable—this has not received sufficient attention in published research articles and major statistical packages. We consider that an individual is well represented when most of its information (measured through the variability) is accounted for in the reduced dimension. As the representation is centered at the origin, the variability of each individual is measured by its squared distance to the center, so that the quality of representation can be measured by the ratio between the squared distance in the reduced dimension and the squared distance in the complete space, that is:

\[
CR^2_i = \frac{\sum_{j=1}^{p} s_{ij}^2}{\sum_{j=1}^{p} s_{ij}^2} \times 100\%
\]  

(1)

where \( s_{ij} \) denotes the principal coordinates of individual \( i \) in the \( j \)-th dimension. Geometrically, it is the squared cosine of the angle between the vector in the complete space and its projection onto the representation space.

For groups of individuals, the quality of representation (on the PCoA ordination diagram) is calculated as in (1) using its centroid-
\( \bar{y}_g \) the average of the coordinates in \( j \)-th dimension for the group \( g \).

Unlike PCA in its Biplot version, where the new axes can be interpreted in terms of the original variables, in PCoA, the axes have no direct meaning. Therefore it is not possible to interpret the relationship between genotype–allele/band and allele/band–allele/band. It can be shown that PCA configurations are also obtained applying PCoA to the matrix of Euclidean distances. A classical biplot is obtained by fitting linear regressions to that configuration as described in Vicente-Villardon et al. (2006). Hence, an immediate heuristic generalization in this context is to use LR's and its graphical representation on the PCoA, called ELB, rather than linear regressions.

To search for the variables associated to the ordination obtained in PCoA, we can look for the directions in the ordination diagram that better predict the probability of presence of each allele.

More formally, define \( \pi_{ij} = E(s_{ij}) \) as the expected probability that the allele \( j \) be present at genotype for a genotype with coordinates \( y_{ij} (i = 1, \ldots, n; j = 1, \ldots, k) \) on the ordination diagram, then

\[
\pi_{ij} = \frac{b_0 + \sum_{k=1}^{l} b_{kj}s_{ij}}{1 + e^{-b_0 + \sum_{k=1}^{l} b_{kj}s_{ij}}}
\]

where \( b_{kj} \) are the LR coefficients that correspond to the \( j \)-th variable (alleles or bands) in the \( x \)-th dimension. The model is a generalized linear model having the logit as a link function.

\[
\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = b_0 + \sum_{k=1}^{l} b_{kj}y_{ik} = b_0 + y_{ik}'b_j
\]

where \( y_{ij} = (y_{i1}, \ldots, y_{ik})' \) and \( b_j = (b_{j1}, \ldots, b_{jk})' \). \( y \)' s and \( b \)' s define a biplot in logit scale. This is called External Logistic Biplot because the coordinates of the genotypes are calculated in an external procedure (PCoA). Given that the \( y \)' s are known from PCoA, obtaining the \( b \)' s is equivalent to performing a LR using the \( j \)-th column of \( X \) as a response variable and the columns of \( y \) as regressors.

The regression equation predicts the probability that an allele will be present in that genotype. Geometrically, the \( y \)'s can be represented as points in the reduced dimension space and the \( b \)'s are the vectors showing the directions that best predict the probability of presence of each allele \( \pi_{ij} \).

For a complete explanation of the geometrical properties of the ELB (see Vicente-Villardon et al., 2006).

The prediction of the probabilities is made in the same way as in a linear Biplot, i.e. the projection of a genotype point on the direction of an allele vector predicts the probability of presence of that allele in the genotype. To facilitate the interpretation of the graph, fixed prediction probabilities points are situated on each allele vector. To simplify the graph, in our application, a vector joining the points for 0.5 and 0.75 are placed; this shows the cut point for prediction of presence and the direction of increasing probabilities. The length of the vector can be interpreted as an inverse measure of the discriminatory power of the alleles or bands, in the sense that shorter vectors correspond to alleles that better differentiate individuals. Two alleles pointing in the same direction are highly correlated, two alleles pointing in the same direction are highly correlated.
pointing in opposite directions are negatively correlated, and two alleles forming an angle close to 90° are almost uncorrelated.

For each allele, the ordination diagram can be divided into two separate regions predicting presence or absence, the two regions are separated by the line that is perpendicular to the allele vector in the Biplot and cuts the vector at the point predicting 0.5. The alleles associated to the configuration are those that predict the presences adequately.

In a practical situation not all the alleles are associated to the ordination. Due to the high number of alleles usually studied, it is convenient to situate on the graph only those that are related to the configuration, i.e. those that have an adequate goodness of fit after adjusting the LR.

A goodness-of-fit (or quality of the representation) criterion, to select the alleles, is the ‘percentage of correct classifications’ calculated as the percentage of coincidences between the binary data matrix and the expected binary matrix obtained from the LR models. When the percentage of correct classification is added for all the alleles, the overall goodness-of-fit of the logistic Biplot is obtained. Additionally, the pseudo $R^2$-Squared performed according to Nagelkerke/Cragg & Uhler’s (Long, 1997) for the regressions of categorical outcome variables are used as measures of the ‘quality of the representation’ and this is interpreted in the manner commonly used in correspondence analysis (Tenenhaus and Young, 1985).

Additionally, a Bonferroni correction can be used as criterion of selection of alleles with higher discriminatory power. With this method, only those alleles that have a given significance level ($P \leq 0.05$) with the number of alleles) will be included in the biplot.

For large data sets $P$-values are highly affected by the sample size and the number of alleles. In these cases, it is better to use the pseudo $R^2$ with a highly restrictive value, for example $R^2 > 0.9$, because pseudo $R^2$ is less sensitive to the sample size.

Frequently the analysis also obtains groupings; Cluster Analysis (CA) can be applied using the initial distance matrix $A$ or the fitted Euclidean distance matrix $D$ obtained from the PCAo. The partition obtained is represented on the PCoA ordination using the convex hulls of the points belonging to each cluster. It could be argued that using the principal coordinates $Y$ or the fitted distance $D$ for additional analysis can result in a loss of information. This can also be thought of as a way to separate the signal from the noise; the loss of information that entails the use of principal coordinates it is compensated by the noise level that is reduced, additionally we guarantee the orthogonality of the regressors. Chae and Warde (2006) show that the retrieval abilities of the logistic Biplot is obtained.

In summary, the general algorithm for ELB works as follows: (i) make a PCAo of the binary data matrix, using the most adequate similarity coefficient for the data; (ii) calculate standard LR using the principal coordinates as independent variables and each allele or band as dependent; (iii) plot the Biplot filtering the variables using the Bonferroni correction; and (iv) draw the groups using the principal coordinates and the most adequate clustering algorithm.

### 3 SIMULATION STUDY

#### 3.1 Method

In order to show the behavior of ELB in identifying molecular markers associated with the classification of genotypes simulated data will be used. Simulations are powerful tools for evaluating the performance of a method because we know the a priori structure of groups and the variables responsible for the classification. Three basic scenarios typical of real data sets (worst case scenarios) were investigated. In order to facilitate the visualization of the Biplot representation, a moderate number of genotypes ($n = 50$) was included.

Matrices of binary data with known group structure were generated and scored as present or absent (1 or 0); the rows correspond to 50 individuals or entries (genotypes) and the columns to 2 binary characters (alleles or bands). Each group was characterized by a set of alleles in such a way that the characteristic alleles were present in that group and absent in the rest. External and internal noises were added to the matrices. The external noise consisted of adding a set of supplementary alleles. These supplementary alleles were generated using a uniform distribution (0,1), the alleles were considered present if the simulated value of the samples $x_i \geq 0.5$; and absent otherwise. An internal noise was added to the total set of alleles. It consisted in modifying, at random, 5%, 10% and 20% of the values assigned to each individual by allele, i.e. the values of presence in the matrices were replaced by absence or vice versa in the indicated percentage. The internal noise is just the percentage of genotyping errors. Although the 20% error can be considered too large for advanced genotyping technology, that percentage would be useful for another application like the detection call in Affymetrix microarrays. Using a high percentage is also useful to assess the reliability of the proposal even in extreme situations.

DICE similarity coefficient and the UPGMA clustering algorithm were used.

Twenty-seven scenarios were simulated in total. Each scenario was repeated 1000 times.

The performance of the proposed method was evaluated using the following criteria: the variance accounted by the first two dimensions and its comparison with the standard for this type of experiments, the projection of a group on the direction of an allele vector that theoretically is present in the group and therefore predicts the probability of presence of allele in the group, the sensitivity of the method measured through the quality of the representation of the alleles (QRIAlleles), the quality of the representation of the alleles with higher discriminatory power. With this method, only those alleles with a given significance level ($P \leq 0.05$) with the number of alleles) will be included in the biplot.

...
are not affected even with small variance absorption. Hesa and Gabriel (2001) demonstrated that graphical displays of multivariate data often clearly exhibit features of the expectations even though the data themselves are poorly fitted by the displays. Thus, it often occurs that ordinations and biplots that poorly fit the sample data still reveal salient characteristics such as clusters of similar individuals and patterns of correlation.

As an example, Figure 2 shows the Biplot representation of the relations among the individuals and the alleles that determine the group structure in some of the simulations. It can be observed that the alleles that have been marked as those of importance in the analysis of variance for the groups that they define. The supplementary alleles are projected on the Biplot in irregular form and with higher lengths indicating a low discriminatory power (see practical interpretation rules in the Supplementary Material). After Bonferroni corrections (5%) almost all supplementary alleles were eliminated.

Only in the case of the Figure 2c, Group 2 does not show the alleles associated with its formation, and this is due to the fact that, generally, g − 1 or less axes are needed to correctly retain the structure of groups. In this case, it is probable that we would need to include a third axis to be able to observe more clearly the variables that are associated with the structure of this group. All the alleles that defined groups had high quality of representation and percentage of coincidences between the binary data matrix and the expected binary matrix obtained from the LR models.

Although the group structure was known, the UPGMA method was used here to calculate the error rate of classification in order to check the sensitivity of the method when the structure is not known ‘a priori’. In the same manner as the other sensitivity criteria, the error rate of classification had a similar behavior with respect to internal noise. The good sensitivity of the method is clearly to check the sensitivity of the method when the structure is not known 'a priori'. In the same manner as the other sensitivity criteria, the error rate of classification had a similar behavior with respect to internal noise. The good sensitivity of the method is clearly demonstrated by the fact that in all scenarios the error rates are not >25% in the worst case and <9% when the genotyping error is <10%.

Table 1. Scenarios simulated

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Groups</th>
<th>Number of individuals/groups</th>
<th>Number of alleles(^a)</th>
<th>Supplementary alleles/ external noise</th>
<th>Total number of alleles</th>
<th>Internal noise (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1((a_1,b_1,c_1))</td>
<td>2</td>
<td>(20,30)</td>
<td>12</td>
<td>30</td>
<td>42</td>
<td>(a_1,b_1,c_1) : 10</td>
</tr>
<tr>
<td>S1((a_2,b_2,c_2))</td>
<td>2</td>
<td>(20,30)</td>
<td>12</td>
<td>48</td>
<td>60</td>
<td>(a_2,b_2,c_2) : 10</td>
</tr>
<tr>
<td>S1((a_3,b_3,c_3))</td>
<td>2</td>
<td>(20,30)</td>
<td>12</td>
<td>66</td>
<td>78</td>
<td>(a_3,b_3,c_3) : 10</td>
</tr>
<tr>
<td>S2((a_1,b_1,c_1))</td>
<td>3</td>
<td>(10,15,25)</td>
<td>20</td>
<td>50</td>
<td>70</td>
<td>(a_1,b_1,c_1) : 10</td>
</tr>
<tr>
<td>S2((a_2,b_2,c_2))</td>
<td>3</td>
<td>(10,15,25)</td>
<td>20</td>
<td>80</td>
<td>100</td>
<td>(a_2,b_2,c_2) : 10</td>
</tr>
<tr>
<td>S2((a_3,b_3,c_3))</td>
<td>3</td>
<td>(10,15,25)</td>
<td>20</td>
<td>110</td>
<td>130</td>
<td>(a_3,b_3,c_3) : 10</td>
</tr>
<tr>
<td>S3((a_1,b_1,c_1))</td>
<td>4</td>
<td>(6,10,14,20)</td>
<td>34</td>
<td>85</td>
<td>119</td>
<td>(a_1,b_1,c_1) : 10</td>
</tr>
<tr>
<td>S3((a_2,b_2,c_2))</td>
<td>4</td>
<td>(6,10,14,20)</td>
<td>34</td>
<td>136</td>
<td>170</td>
<td>(a_2,b_2,c_2) : 10</td>
</tr>
<tr>
<td>S3((a_3,b_3,c_3))</td>
<td>4</td>
<td>(6,10,14,20)</td>
<td>34</td>
<td>187</td>
<td>221</td>
<td>(a_3,b_3,c_3) : 10</td>
</tr>
</tbody>
</table>

\(^a\) Alleles that define group’s structure.

Table 2. Simulation results

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Average QR individuals</th>
<th>QRA Alleles/ groups</th>
<th>QRA Alleles/ supplementary</th>
<th>% CC Alleles/ groups</th>
<th>% CC Alleles/ supplementary</th>
<th>ERC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1a1</td>
<td>30.87</td>
<td>0.89</td>
<td>0.15</td>
<td>95.16</td>
<td>60.66</td>
<td>0.00</td>
</tr>
<tr>
<td>S1a2</td>
<td>26.55</td>
<td>0.89</td>
<td>0.13</td>
<td>95.10</td>
<td>59.65</td>
<td>0.01</td>
</tr>
<tr>
<td>S1a3</td>
<td>23.90</td>
<td>0.88</td>
<td>0.12</td>
<td>95.11</td>
<td>59.09</td>
<td>0.02</td>
</tr>
<tr>
<td>S1b1</td>
<td>27.34</td>
<td>0.76</td>
<td>0.15</td>
<td>90.11</td>
<td>60.69</td>
<td>0.13</td>
</tr>
<tr>
<td>S1b2</td>
<td>23.70</td>
<td>0.76</td>
<td>0.13</td>
<td>90.12</td>
<td>59.65</td>
<td>0.20</td>
</tr>
<tr>
<td>S1b3</td>
<td>21.41</td>
<td>0.76</td>
<td>0.12</td>
<td>90.03</td>
<td>59.20</td>
<td>0.27</td>
</tr>
<tr>
<td>S2c1</td>
<td>20.99</td>
<td>0.53</td>
<td>0.15</td>
<td>79.99</td>
<td>60.74</td>
<td>3.81</td>
</tr>
<tr>
<td>S2c2</td>
<td>18.57</td>
<td>0.52</td>
<td>0.13</td>
<td>79.63</td>
<td>59.87</td>
<td>4.71</td>
</tr>
<tr>
<td>S2c3</td>
<td>17.06</td>
<td>0.51</td>
<td>0.12</td>
<td>79.31</td>
<td>59.42</td>
<td>5.90</td>
</tr>
<tr>
<td>S2a1</td>
<td>29.16</td>
<td>0.88</td>
<td>0.08</td>
<td>95.03</td>
<td>57.20</td>
<td>0.25</td>
</tr>
<tr>
<td>S2a2</td>
<td>24.84</td>
<td>0.87</td>
<td>0.08</td>
<td>94.82</td>
<td>57.27</td>
<td>0.72</td>
</tr>
<tr>
<td>S2a3</td>
<td>22.06</td>
<td>0.86</td>
<td>0.08</td>
<td>94.59</td>
<td>57.34</td>
<td>1.68</td>
</tr>
<tr>
<td>S2b1</td>
<td>25.42</td>
<td>0.73</td>
<td>0.09</td>
<td>89.59</td>
<td>56.71</td>
<td>1.69</td>
</tr>
<tr>
<td>S2b2</td>
<td>21.85</td>
<td>0.72</td>
<td>0.09</td>
<td>89.21</td>
<td>57.76</td>
<td>3.60</td>
</tr>
<tr>
<td>S2b3</td>
<td>19.52</td>
<td>0.70</td>
<td>0.09</td>
<td>88.76</td>
<td>57.74</td>
<td>5.86</td>
</tr>
<tr>
<td>S2c1</td>
<td>19.22</td>
<td>0.45</td>
<td>0.11</td>
<td>78.08</td>
<td>59.00</td>
<td>14.26</td>
</tr>
<tr>
<td>S2c2</td>
<td>16.92</td>
<td>0.43</td>
<td>0.11</td>
<td>77.22</td>
<td>58.71</td>
<td>17.65</td>
</tr>
<tr>
<td>S2c3</td>
<td>15.44</td>
<td>0.41</td>
<td>0.10</td>
<td>76.58</td>
<td>58.47</td>
<td>20.04</td>
</tr>
<tr>
<td>S3a1</td>
<td>28.50</td>
<td>0.82</td>
<td>0.07</td>
<td>94.20</td>
<td>56.54</td>
<td>1.52</td>
</tr>
<tr>
<td>S3a2</td>
<td>23.90</td>
<td>0.81</td>
<td>0.07</td>
<td>94.10</td>
<td>56.65</td>
<td>2.55</td>
</tr>
<tr>
<td>S3a3</td>
<td>20.75</td>
<td>0.81</td>
<td>0.07</td>
<td>93.89</td>
<td>56.65</td>
<td>4.00</td>
</tr>
<tr>
<td>S3b1</td>
<td>24.73</td>
<td>0.67</td>
<td>0.07</td>
<td>88.88</td>
<td>56.78</td>
<td>4.86</td>
</tr>
<tr>
<td>S3b2</td>
<td>20.71</td>
<td>0.66</td>
<td>0.07</td>
<td>88.59</td>
<td>56.86</td>
<td>6.87</td>
</tr>
<tr>
<td>S3b3</td>
<td>18.07</td>
<td>0.65</td>
<td>0.07</td>
<td>88.18</td>
<td>56.89</td>
<td>8.78</td>
</tr>
<tr>
<td>S3c1</td>
<td>17.97</td>
<td>0.40</td>
<td>0.09</td>
<td>77.53</td>
<td>57.71</td>
<td>16.27</td>
</tr>
<tr>
<td>S3c2</td>
<td>15.44</td>
<td>0.38</td>
<td>0.09</td>
<td>76.56</td>
<td>57.71</td>
<td>19.94</td>
</tr>
<tr>
<td>S3c3</td>
<td>13.82</td>
<td>0.36</td>
<td>0.09</td>
<td>75.73</td>
<td>57.63</td>
<td>23.15</td>
</tr>
</tbody>
</table>

Fig. 1. Distribution of the accounted variance values for different scenarios: (a) all eigenvalues and (b, c, d) first two dimensions.
Finally, we can conclude that independently of the number of alleles or amplification fragments that are evaluated, the method recovers those of importance in the definition of the structure or natural grouping of the individuals in the first principal coordinates and that the CA almost always achieves nearly 100% accuracy for assigning individuals to their true groups when the genotyping error is small.

4 APPLICATION TO HAPMAP DATA

The practical merit of our methodology is additionally illustrated in a real data study of the genetic diversity of four populations from Africa, Asia and Europe, using the genotype data generated by the International HapMap Consortium (2003).

4.1 Genotype data

All our analyses are based on the HapMap phase 3 genotypes for chromosome 22 from samples of four populations: 171 Maasai (MKK); 82 Han Chinese in Beijing, China (CHB); 82 Japanese in Tokyo, Japan (JPT) and 162 with Western European ancestry (CEU), available for bulk downloads at http://www.hapmap.org/ (Supplementary Table 1). Previously to the analysis, the matrix was depurated eliminating the monomorphic SNPs and those with missing values. Data were scored like a binary matrix using only SNPs common to all populations obtaining a matrix of 497 individuals corresponding to 7333 genotypes. The alleles were labeled using the SNP name and the allele: rs1314-T and rs1314-G, respectively. (The complete list of variables is shown in Supplementary Table 2).

4.2 Method

Genetic relationships among the 497 individuals were investigated using PCoA, CA and ELB as described above. To compare their behavior in the proposed context, the following coefficients were used: Dice, Jaccard, Rogers and Tanimoto and Simple Matching (Sneath and Sokal, 1973). The correlation between the observed and fitted distances, based on genetic dissimilarity, was used to establish a criterion for the selection of the number of axes, measure the grouping analysis performance and to evaluate the stability of the constructed relationship. Although CA is not necessary because the group structure is known, the UPGMA cluster algorithm was used to check the group structure obtained from the principal axes retained. Measures of quality of representation of groups and variables (alleles) were obtained as described above.

4.3 Results

The distribution of the correlation values among observed and expected distance matrices for different dissimilarity coefficients and several combinations of principal coordinates retained showed that for the three principal coordinates, the simple matching coefficient is the one that better defines the DNA sequence variation patterns using high-density SNPs genotyping arrays. The correlation values among observed and expected distance matrices are interpreted in a similar form to Cophenetic correlation values.

Figure 3 shows 3D representation of the first three principal coordinates. The four populations are clearly separated using the first three principal coordinates. The first five dimensions account for 15.96%, 10.00%, 1.27%, 0.97% and 0.91% of the total variance. Plane 1–2 differentiates between MKK, CEU and the Asian (CHB and JPT, that are mixed in the graphical representation). In our analysis, when we introduce the third axis it is possible to separate the CHB from JPT. Further axes do not contribute with...
Classification of genotypes by ELB

Fig. 3. 3D representation of the first three principal coordinates showing the genetic relationships among individuals based on simple matching dissimilarity matrix, whole dataset.

discriminatory information, i.e. adding new axes does not introduce any information about the differences among populations.

The centroid of a group could be taken as the representative of the group, the more compact is the group the better is the summary. In the graphical representation, all the individuals belonging to the same group are close in the graph, showing that all share a common DNA pattern. The fact that the four populations are clearly separated using just the first three principal coordinates means that the main source of variation in the SNPs patterns is the differentiation among groups, i.e. the differences among populations are higher than the differences within populations. Supplementary Figure 1 shows more specifically the projection obtained from PCoA of the individuals on the planes 1–2 (1a), 1–3 (1b) and 4–5 (1c).

The projection of the alleles on the principal coordinates solution using ELB, on the planes 1–2 (2a) and 1–3 (2b) is showed in Supplementary Figures 2. The alleles are represented by lines pointing in the direction of increasing predictions of the probabilities. As mentioned previously, the beginning of the line is the point predicting 0.5 and the end, the point predicting 0.75. The length of each is related to capacity to predict presence or absence of the alleles in each individual (see practical interpretation rules in the Supplementary Material). After adjusting the ELB, the global goodness-of-fit as a percentage of correct classifications was 81.04%, i.e. if the biplot is used to predict each genotype for each individual, 81.04% of them will be correct; the percentage is high even when many of the alleles are not associated with any pattern in data. If we consider each allele separately, the percent of correct classifications of 75% of the alleles was higher than 69.42%.

Using the Bonferroni correction, 9319 alleles (63.55%) were selected (Supplementary Figure 3). As a consequence of the large sample size, most of those P-values are associated with low explanatory power and are probably not very useful for the discrimination among populations. The quality of the representation measured through the pseudo $R^2$, was <0.5 in 8219 (92.83%) of the selected alleles and >0.9 in 1.78% of the cases, i.e. from the 14666 alleles, 260 have the highest discriminatory power. This more restrictive selection criterion is less sensitive to the sample size and permitted selecting the most important alleles to determine the patterns of DNA sequence variation among the four populations. Figure 4 shows the 3D representation of the alleles with a $R^2 > 0.9$. The bidimensional projections into planes 1–2 (4a) and 1–3 (4b) are depicted in Supplementary Figure 4 and the list of selected alleles is shown in Table 3 of the same.

The quality of representation of groups calculated with the three first retained dimensions was >99% for all the populations. The patterns of DNA sequence variation among the four populations generated with all the SNPs (14666) and with the small subset of SNPs (260) are similar (Fig. 5). Supplementary Figure 5 shows the bidimensional projections into planes 1–2 (5a) and 1–3 (5b). Global goodness-of-fit as a percentage of correct classifications for the reduced matrix was 98.69%.

It has been demonstrated that the proposed methodology is useful when evaluating large datasets such as data from the HapMap project. It allows recovering the structure of the studied populations with small dataset. This helps to reduce the problem of multiple comparisons that arises from testing tens to hundreds of thousands of SNPs and haplotypes for disease associations. In such settings, it is often desirable to reduce the number of markers needed for structure identification and the identification of the functionally important SNPs. Additionally, the method helps to visualize the data structure in a reduced dimension.

The described procedure has been successfully applied to different types of binary data and contexts—with or without knowing a previous group structure—for example: detection call in Affymetrix microarrays (Vicente-Villardón et al., 2006) and genetic diversity studies in plants collections (Demey et al., 2006).
Selection procedures using univariate statistics to compare the populations with each allele are commonly used by most researchers. For example, Weir et al. (2005) using $F_{ST}$ as a measurement of genetic population structure also found high similarity between the HapMap Han Chinese from Beijing (HCB) and Japanese from Tokyo (JPT) for the majority of the chromosomes studied including the 22. However, it is not clear how this approach can be applied to the selection of informative markers when the parental information is unknown (Rosenberg et al., 2003).

Our approach uses the multivariate nature of the data offering some advantages over the classical methods: (i) by using the principal coordinates the main patterns of genetic variation among populations are summarized in just three combined variables; (ii) the graphical representations permit not only global exploration of the main patterns and the variables associated to the discrimination, but also the direction of the association and the selection of small subsets of SNPs that have a similar behavior in relation to the discrimination; (iii) it is possible to study the correlation structure among alleles; and (iv) it is possible to know the population structure without any prior knowledge about the parental information.

Paschou et al. (2007, 2008) used a multivariate approach (PCA) to infer population structure using data from the HapMap project; however, although they indicate that the algorithm can be used to identify a small set of structure informative markers, they do not use the Biplot properties of the Singular Value Decomposition to interpret the SNPs responsible for the discrimination. Additionally they use a linear technique for continuous data to a categorical data matrix in which the elements coded as $+1$, $0$ or $-1$ is questionable. When the data are already genotyped, and therefore it is categorical, PCA is not suitable. Our approach is a generalization of the PCA method and the Singular Value Decomposition (Biplot) that can handle the genotyped data in an appropriate way.

5 FINAL REMARKS

In summary, the proposed methodology using a combination of PcoA, CA and ELB represents an improvement over traditional methods for classifying genotypes using DNA molecular markers, since it produces groups, calculates a measure of the quality of the groups, identifies the alleles or bands that are responsible for the classification of genotypes (allowing the study of individual–individual, individual–variable and variable–variable relations more appropriately), and facilitates the genetic interpretation of the results.

The complementarity nature between PcoA, CA and LB yields a holistic comprehension of the data structure and facilitates the interpretations of the results. Consequently, a combined use of this set of techniques is highly recommended for a thorough description of data in studies of genetic diversity using DNA markers.

ACKNOWLEDGEMENTS

We are thankful to Prof John Gower by the critical reading and the suggestions made about this article.

Funding: Proyecto de Biotecnología BID-FONACIT II, Venezuela.

Conflict of Interest: none declared.

REFERENCES


