Acta Scientiarum

http://www.uem.br/acta
ISSN printed: 1679-9275
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v37i4.19746

Hierarchical genetic clusters for phenotypic analysis

Luiza Barbosa da Matta*, Lívia Gracielle Oliveira Tomé, Caio Césio Salgado, Cosme Damião Cruz and Letícia de Faria Silva

*Departamento de Biologia Geral, Universidade Federal de Viçosa, Av. Peter Henry Rolfs, s/n, 36570-000, Campus Universitário, Viçosa, Minas Gerais, Brazil. *Author for correspondence. E-mail: luiza.matta@ufv.br*

ABSTRACT. Methods to obtain phenotypic information were evaluated to help breeders choosing the best methodology for analysis of genetic diversity in backcross populations. Phenotypes were simulated for 13 characteristics generated in 10 populations with 100 individuals each. Genotypic information was generated from 100 loci of which 20 were taken at random to determine the characteristics expressing two alleles. Dissimilarity measures were calculated, and genetic diversity was analyzed through hierarchical clustering and graphic projection of the distances. A backcross was performed from the two most divergent populations. A set of characteristics with variable heritability was taken into account. The environmental effect was simulated assuming $x \sim N(0, \sigma^2)$. For hierarchical clusters, the following methods were used: Gower Method, average linkage within the cluster, average linkage among clusters, the furthest neighbor method, the nearest neighbor method, Ward's method, and the median method. The environmental effect and heritability of the analyzed variables had an influence on the pattern of hierarchical clustering populations according to the backcrossed generations. The nearest neighbor method was the most efficient in reconstructing the system of backcrossing, and it presented the highest cophenetic correlation. The efficiency of the nearest neighbor method was the highest when the analysis involved characteristics of high heritability.

Keywords: hierarchical clustering, genetic diversity, backcrossing.

Agrupamentos genéticos hierárquicos para análises fenotípicas

RESUMO. Visando auxiliar o melhorista na escolha da melhor metodologia para análises de diversidade genética em populações de retrocruzamento avaliaram-se os métodos baseados em informações fenotípicas. Os fenótipos foram simulados para 13 características, geradas em 10 populações com 100 indivíduos cada. Geraram-se informações genotípicas de 100 locos, dos quais 20 foram tomados ao acaso para determinar as características, manifestando dois alelos. Medidas de dissimilaridade foram calculadas, e analisou-se a diversidade genética por meio agrupamento hierárquico e projeção gráfica das distâncias. A partir das duas populações mais divergentes fez-se o retrocruzamento. Considerou-se um conjunto de características com herdabilidade variável. Simulou-se o efeito ambiental admitindo distribuição normal, com média zero e variância σ^2. Para as análises de agrupamentos hierárquicos utilizaram-se os métodos: Método de Gower, Ligação média dentro de grupo, Ligação média entre grupo, Vizinho mais distante, Vizinho mais próximo, Método de Ward, Método da mediana. O efeito ambiental e a herdabilidade das variáveis analisadas influenciaram o padrão de agrupamento de populações sob retrocruzamento. Em características de herdabilidade elevada, o método do Vizinho mais próximo foi o mais eficiente em reconstituir o retrocruzamento, além de ter apresentado a maior correlação cofenética, sendo considerada a melhor metodologia a priori e a posteriori.

Palavras-chave: métodos de agrupamento hierárquico, diversidade genética, retrocruzamento.

Introduction

The study of genetic diversity has been common in many fields of biology, and its importance has been emphasized in evolutionary and adaptive studies with indispensable information on differentiation, structure and interactions among populations and samples. Equally important, studies of diversity have been highlighted in breeding to identify suitable parents to cross to obtain hybrids with larger heterotic effects, which provide greater segregation in recombination and allow the emergence of transgressive hybrids. Moreover, information on diversity should be used in conservation of variability and management of germplasm banks (HARTL; CLARK, 1997).

Overall, the studies on genetic diversity consist of analyzing a set of accessions with regard to a series of information with genetic or phenotypic nature.

These accessions may represent individuals, populations, lines, hybrids or any genetic material of interest to the researcher. The accessions are evaluated in experiments or through direct observations in relation to phenotypic characteristics of different nature involving binary, multicategoric or continuous data. Studies from genetic information involving molecular markers that are dominant or codominant with two or more alleles are performed as well (CRUZ et al., 2011).

After collecting the data, it is necessary to properly analyze and interpret the results to make better use of genetic resources and to guide the best strategy of use and conservation of the available genetic variability. There are several methods to analyze and interpret the results, but cluster analysis for optimizing hierarchical or graphical dispersion is the most widely used. Generally, these methods are complementary and start from a common base, but the researcher faces many options. Choosing the wrong method could compromise its interpretation resulting in the inappropriate use of available resources (CRUZ et al., 2011).

When the basic question of the experiment is to compare the efficiency of different clustering methods, the use of a simulation involving a known genetic pattern becomes indispensable. A type of controlled genetic design widely applied in breeding is the use of backcross generations, which are often used to increase the probability of obtaining higher progenies that provide more success in the selection process (LORENCETTI et al., 2006). Other applications are comprised of transferring target alleles (RANGEL et al., 2010). However, an important feature of backcross generations is the prior knowledge of similarity between individuals of populations of different generations, including recurring and non-recurring parents, using probabilistic principles related to gametic union resulting from meiosis to know in advance the hierarchical structure of the assessed cluster.

The use of predictive methods can be used to study the genetic diversity, especially those based on morphological, physiological and molecular differences where the diversity is expressed by coefficients of (dis)similarity and genetic distance measures. Once a dissimilarity matrix is obtained, a suitable clustering method is applied.

Hierarchical methods are routinely used in studies on genetic diversity in situations when there is no concern about the optimal number of groups because the major interest is in the "tree" and the obtained ramifications. In general, the boundaries can be determined by a visual inspection of the dendrogram that evaluates points with a higher level of change taking the number of individuals for a given group as delimiters. Although in some cases, a high level change is a criterion difficult to see, thereby preventing interpretation of the groups (FALCONER; MACKAY, 1996).

Thus, the objective of this study was to evaluate the methods of analysis of genetic diversity based on phenotypic information applied to a hierarchical system previously known. Moreover, the potential of dissimilarity measures for the assessment of genetic diversity in breeding as well as the techniques of hierarchical cluster analysis were studied. This study aimed to link all of these aspects to a theoretical and practical reference to assist researchers in their decision making regarding the best cluster method used in breeding.

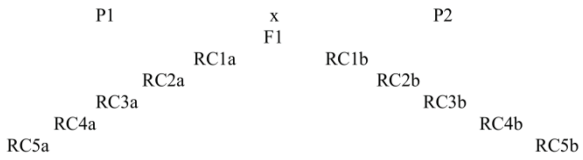
Material and methods

Population simulation

Genotypic data were originally simulated for 10 populations in Hardy-Weinberg equilibrium with 100 individuals each. Information on 100 loci expressing two codominant alleles was generated. This preliminary data set was used for the calculation of a genotypic dissimilarity measure and then to study the genetic diversity analysis by hierarchical clustering, optimization and graphic projection of the distances. The most divergent population pair was taken for generating a structured and hierarchical system designated as recurring and non-recurring parents.

The relations of family relationships and hierarchical structuring were established considering genetically divergent parent populations, F1 hybrids and five generations of backcrossing in relation to each of the parents, thereby establishing parameters of effectiveness of the assessed methodologies.

For comparison of measurements and the dissimilarity pattern clustering method, a structured system of population in which the relationship or hierarchy levels are previously known was established. Thus, from the pair of diverging parents (P1 and P2), 11 populations were generated according to the following model:



Simulation of phenotypic values of quantitative traits

A set of 13 characteristics with heritability values of 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60,

0.65, 0.70, 0.75 and 0.80 with a known mean and coefficient of variation was considered. These characteristics were established by the action of alleles of 20 loci randomly selected from 100 loci previously genotyped with a differential additive effect with weights established from a binomial distribution and average degree of null dominance.

From previous values of heritability, the coefficient of variation and means were established, and estimates of genetic and environmental variance were calculated. By principles of quantitative genetics, genetic variances expected in backcross generations can be predicted by genotype frequencies. Additive and dominant effects were considered null in this study. The additive/dominant model was used to generate the genotypic value of each individual, and a simulated environmental effect was added assuming normal distribution with a mean of zero, and the variance (σ^2) provided the phenotypic value submitted to analyses.

The simulations and data analyses were conducted in genes program.

Sequential, agglomerative, hierarchical and non-overlapping clustering methods.

Simple linkage method or nearest neighbor method

In this method, the dendrogram is established by the genotypes with the highest similarity, and the distance between an individual (k) and a group formed by individuals i and j is calculated as follows: $d_{(ij)k} = \min \{d_{ik}; d_{jk}\}$; where is $d_{(ij)k}$ is the smallest element of the set of the distances of the pairs of the individuals ((i and k) and (j and k)). The connections between individuals and groups or among groups are made by the distance between the groups defined as that between the most similar individuals in these groups.

The distance between two groups is calculated as follows:

$$d_{(ij)(kl)} = \min\{d_{ik}; d_{il}; d_{jk}; d_{jl}\}$$

The distance between the two groups formed by individuals, (i and j) and (k and l), is the smallest element of the set of which the elements are the distances among the pairs of individuals ((i and k), (i and l), (j and k) and (j and l)).

Complete linkage method or furthest neighbor method

The complete linkage method is a complete antithesis to the simple linkage method. In the complete linkage method, the similarity between the two groups is given by the individuals of each group that resemble the least. This method usually leads to compact and discrete groups with relatively small similarity values.

The construction of the dendrogram is similar to the simple linkage method as follows: the most similar individuals are established at each stage. However, the distance between an individual (k) and a group consisting of individuals i and j is calculated as follows:

$$d_{(ij)k} = \max (d_{ik}; d_{jk})$$

where $d_{(ij)k}$ is the highest element in the set of individual pairs ((i and k) and (j and k)).

The distance between two sets is calculated as follows:

$$d_{(ij)(kl)} = \max\{d_{ik}; d_{il}; d_{jk}; d_{jl}\}$$

The distance between two groups formed by the individuals ((i and j) and (k and l)) is determined by the highest element in the set of which the elements are the distances among the pairs of the individuals in the clusters ((i and k), (i and l), (j and k) and (j and l)).

The average linkage between cluster method or the unweighted pair-group method using arithmetic averages (UPGMA)

This method is a clustering technique that uses arithmetic means (unweighted) of the dissimilarity measures, thus avoiding characterization of the dissimilarity by extreme values (minimum and maximum) between the considered genotypes.

The construction of the dendrogram is determined by genotype with the greatest similarity. However, the distance between an individual (k) and a group consisting of individuals i and j is calculated as follows:

$$d_{(ij)k} = \text{mean } (d_{ik}; d_{jk}) = \frac{d_{ik} + d_{jk}}{2}$$

where

$d_{(ij)k}$ is the average of the set of the distances of the pairs of individuals ((i and k) and (j and k)).

The distance between the two clusters is calculated as follows:

$$d_{(ij)(klm)} = \text{mean } (d_{ik}; d_{il}; \text{dim } d_{jk}; d_{jl}; d_{jm}) = \frac{d_{ik} + d_{il} + d_{im} + d_{jk} + d_{jl} + d_{jm}}{6}$$

The distance between two groups formed by individuals ((i and j) and (k, l and m)) is determined

by the mean of the elements of the set of which the elements are the distances among pairs of individuals of groups ((i and k), (i and e), (i and m), (j and k), (j and l) and (j and m)).

An overall expression for the unweighted mean among groups can be represented as follows:

$$d_{(ij)k} = \frac{n_i}{n_i + n_j} d_{ik} + \frac{n_j}{n_i + n_j} d_{jk}$$

where

$d_{(ij)k}$ is defined as the distance between the group (ij) with inner size n_i and n_j, respectively, and group k. In this equation, the i, j and k indices are characterized as individuals or groups, and this interpretation should be the same for subsequent methods.

Thus, to calculate the distance, d(12)3, in which the group formed by accessions 1 (ni = 1) and 2 (nj = 1) are considered, the following equation is used:

$$d_{(12)3} = \frac{1}{1+1} d_{13} + \frac{1}{1+1} d_{23} = \frac{d_{13} + d_{23}}{2}$$

Weighted pair-group method using arithmetic averages (WPGMA)

In this method, a weighted average of the new coefficients of distance is achieved at each step, which is recalculated for producing the "new" distance matrix. The equation used to calculate the weighted average is as follows:

$$d_{(ij)k} = \frac{d_{ik} + d_{jk}}{2}$$

Thus, to calculate the distance, d(12)3, the following equation is used:

$$d_{(12)3} = \frac{d_{13} + d_{23}}{2}$$

The following equation is used to calculate the distance, d(12.3)4, in which the group formed by accessions 1 and 2 joined with 3:

$$d_{(12.3)4} = \frac{d_{(12)4} + d_{34}}{2} = \frac{\frac{d_{14} + d_{24}}{2} + d_{34}}{2} = \frac{d_{14} + d_{24} + 2d_{34}}{2}$$

Median method or weighted pair-group centroid method (WPGMC)

The weighted centroid method, also called the median method, was proposed by Gower (1967) to avoid unequal contribution of centroids of different clusters in the formation of a new centroid (group) candidate. Khattree and Naik (2000) reported that if a cluster is considered small in terms of numbers of individuals, its contribution to the formation of a group with a new centroid will not be different from the mean (centroid) of the group with the largest number of individuals.

The median can be achieved by using the following equation:

$$d_{(i,j)k} = \frac{d_{ik} + d_{jk}}{2} - \frac{1}{4} d_{ij}$$

Ward's minimum variance method

For the initial formation of the group in Ward's minimum variance method, individuals provide the lowest sum of squares of deviations. It is assumed that any stage can be quantified by the ratio between the sum of the squares of the deviations within the cluster in formation and the total sum of the squares of the deviations. The sum of the squares of the deviations within is calculated considering only the accessions within the cluster in formation, and the sum of squares of total deviations is calculated considering all the individuals available for cluster analysis.

Clustering is made from the sum of squares of deviations between accessions or, alternatively, from the square of the Euclidean distance because the following relation is found:

$$SQD_{ii'} = \frac{1}{2} d^2_{ii'}$$

where:

$$SQD_{ii'} = \sum_{j=1}^{v} SQD_{j(ii')}$$

where:

$SQD_{j(ii')}$ is the sum of squares of deviation for the j^{th} variable considering accessions i and i'; and

$$d^2_{ii'} = \sum_{j=1}^{v} (X_{ij} - X_{i'j})^2$$

where:

$d^2_{ii'}$, is the square of the Euclidian distance between i and i' genotypes;

V is the number of the assessed traits; and

X_{ij} is the value of j trait for i genotype.

The total sum of the squares of deviations is calculated as follows:

$$SQDTotal = \frac{1}{g}\sum_{i<}^{g}\sum_{i'}^{g} d_{ii'}^2$$

where:

g is the number of accessions to be clustered.

In this cluster analysis, the pair of accessions that provides the lowest sum of squares of deviations is identified in matrix D (where the elements are the squares of the Euclidean distances; $d_{ii'}^2$) or matrix S (where the elements are the sums of squares of deviations; $SQD_{ii'}$). With these accessions clustered, a new dissimilarity matrix of lower dimension is recalculated as follows:

$SQD_{(ijk)} = \frac{1}{k} d_{(ijk)}^2$ (k is the number of accessions in the group, which is 3 in this example).

$$d_{(ijk)}^2 = d_{(ij)}^2 + d_{(ij)k}^2 = d_{ij}^2 + d_{ik}^2 + d_{jk}^2$$

$SQD_{(ijkm)} = \frac{1}{k} d_{(ijkm)}^2$ (k is the number of accessions in the cluster, which is 4 in this example).

$$d_{(ijkm)}^2 = d_{ij}^2 + d_{ik}^2 + d_{jk}^2 + d_{im}^2 + d_{jm}^2 + d_{km}^2$$

In this procedure, cluster analysis is carried out by providing the bg^{-1} steps of the clustering to form the dendrogram. This method is similar to the nearest neighbor method for using distances between accessions and to directly apply the formulas.

For each clustering method, maximum fusion levels of 50 and 75% were considered.

Results and discussion

Differentiation among populations by features of different heritabilities

The analysis of variance was performed for the characteristics of low (h^2 = 0.20 to 0.50) and high (h^2 = 0.55 to 0.80) heritabilities to check for significant differences among means of populations representing the parent generations, F1 generations and backcrosses to infer the relevance of these traits for the study of genetic diversity (Tables 1 and 2). All assessed traits showed significant effects when submitted to analysis of variance indicating the existence of genetic variability among populations, which can be significantly differentiated considering each of the variables used either of high or low heritability.

In this analysis, the variation within and among populations provided by the variation among individuals was considered. Thus, the analysis was performed so that the coefficients of variation had small magnitudes and that the overall means were similar and approximately 100, as was firstly assumed.

Table 1. Result of analysis of variance of seven characteristics of low heritability (0.20–0.50) obtained by simulation and evaluated in thirteen populations.

FV	DF	Y_{H20}	Y_{H25}	Y_{H30}	Y_{H35}	Y_{H40}	Y_{H45}	Y_{H50}
Treatment	12	.5409*	7.461**	.5188**	.4656**	2.874**	.7459**	2.116**
Residue	637	.3033	.2426	.2107	.1726	.1490	.0952	.0781
Mean		100.34	100.50	100.46	100.38	100.33	100.60	100.66
CV(%)		0.55	0.49	0.45	0.41	0.38	0.30	0.27

**and *significant at 1 and 5% levels, respectively, according to an F-test.

Table 2. Result of analysis of variance of six characteristics of high heritability (0.55–0.80) obtained by simulation and evaluated in thirteen populations.

FV	DF	Y_{H55}	Y_{H60}	Y_{H65}	Y_{H70}	Y_{H75}	Y_{H80}
Treatment	12	.8261**	1.058**	.3925**	.7643**	.7101**	1.996**
Residue	637	.1398	.1201	.1265	.1059	.0696	.0550
Mean		100.28	100.30	100.25	100.38	100.18	100.58
CV(%)		0.37	0.34	0.35	0.32	0.26	0.23

**and *significant at 1% and 5% levels, respectively, according to an F-test.

Performance of the individual analyses of variance for each phenotypic trait was followed by multivariate analyses to identify the pattern of clusters through hierarchical techniques that are widely used in studies of genetic diversity. For each method employed, the resulting patterns of similarity were represented in dendrograms shown in Figure 1.

Pattern dissimilarity by hierarchical clustering techniques

The objective of this study was to identify if the clustering methods would be able to reveal the hierarchical structure of the generations of backcrossing by keeping the similarity expected according to probabilistic principles associated with chromosome segregation during meiosis. Thus, the F1 intermediate positioning in relation to their parents was expected to show great genetic diversity in this study. The RC1, RC2, RC3, RC4 and RC5 backcrossed generations should present similarity in relation to the recurring parent equal to 75, 87.5, 93.75, 96.87 and 98.43%, respectively. It was expected that this pattern can be revealed by both characteristics of high and low heritability when submitted to different hierarchical clustering methods. To prove this hypothesis, the analyses were made from two sets of characteristics involving those with low heritability (below 50%) and high heritability (above 50%) (Figure 1).

The environmental effect on the quantitative characteristics had a great impact on the clustering standard by hierarchical methods by making them less efficient in reconstituting the clustering hierarchy when the heritability was of low magnitude.

Analyzing the dendrogram obtained from the characteristics of low heritability (Figure 1A to G), which were strongly influenced by the environment, showed that no biological patterns of similarity between populations reconstructed the hierarchical structure of clusters related by backcrossing in reference cutoffs (50 and 75% of the level of maximum fusion of the clustering method). The theoretical expectation was the occurrence of the following groups representing the F1 generation: P2 was clustered with their descendants (RC5b, RC4b, RC3b, RC2b and RC1b); P1 was grouped with their descendants (RC5a, RC4a, RC3a, RC2a and RC1a); and the backcross generations (RC5, RC4 and RC3) were more similar to the recurring parents than to the RC1 and RC2 generations. The inefficiency of hierarchical clustering when there is high environmental effect can be found by the absence of branches visible at 50 and 75%, thereby separating backcross cluster 1, reciprocal backcrossing and F1 individuals.

In the clustering pattern analyses by means of low heritability characteristics, the nearest neighbor clustering method (Figure 1G) was the method that best reflected the actual clustering structure for backcrosses 1 and 2, for reclustering all individuals from the RCb backcross together with P2, for reclustering all individuals from the RCa backcross together with P1, and for maintaining F1 apart from both. There was no rigor in the order in which individuals of the RC3, RC4 and RC5 generations were presented in the dendrogram due to the high degree of similarity that they had with the recurring parent with values of 93.75%, 96.88% and 98.44%, respectively. The other methods (Figure 1A to F) did not correctly reproduce the clustering, and the RC1b generation was not clustered together with the other backcrossing generations with its respective recurring parent.

When dendrograms constructed from characteristics of high heritability were analyzed, the average linkage clustering method within the group and Ward's method only reconstructed two hierarchical groups, thereby discriminating accessions referring to representative populations of backcrosses 1 (relative to parent 1) and 2 (relative to parent 2). The F1 population was not identified as an individualized cluster but was grouped together with one of the backcrossed generations (Figures 1H and I). Giannotti et al. (2005) used Ward's method to perform the clustering of information related to cattle to analyze their heritability estimates. De Albuquerque et al. (2006) proposed a system for the study and interpretation of the stability of the methods of cluster analysis through various clustering algorithms of plant data. These authors used the simple linkage, complete linkage, means of the distances, centroid, median, Ward, and cophenetic correlation methods, and they verified a certain coincidence in these methods, except in the centroid method.

The hierarchical clustering methods, namely WPGMA and furthest neighbor method, were more efficient than those previously cited for clearly defining the three hierarchical groups with a cut made at 50%. However, this same separation was not maintained when a cut at 75% was established (Figure 1J and K).

The WPGMC, UPGMA and nearest neighbor clustering methods were the methods that best grouped generations derived from the breeding genetic system established by generations of backcrosses because a cut made at 75% still allowed differentiation of the three hierarchical groups (Figures 1L, M and N). By means of multivariate techniques, Dos Santos et al. (2011) evaluated the genetic divergence among soybean genotypes, and they found variability among the genotypes tested using the UPGMA method.

Therefore, the best clustering method when considering high heritability was the nearest neighbor method. This method effectively separated the three groups of accessions with cuts at 50% and 75%, which was also accomplished with the UPGMA and WPGMC methods. However, the nearest neighbor method has potential to be efficient at cuts made with values even higher than 90%.

Pattern of dissimilarity by graphic dispersion

Figures 2 and 3 show the projected measures of dissimilarity in the plan. In these figures, the graphic dispersion shows coefficient of distortion values of 1.493 and 2.054%, and they show stress values of 2.089 and 3.059% for analyses made from traits with low and high heritability, respectively. The cophenetic correlations, which express the relationship between the original distances and graphically distances for traits with low and high heritability, had values equal to 0.999 (significant at 1% of probability according to the Mantel test).

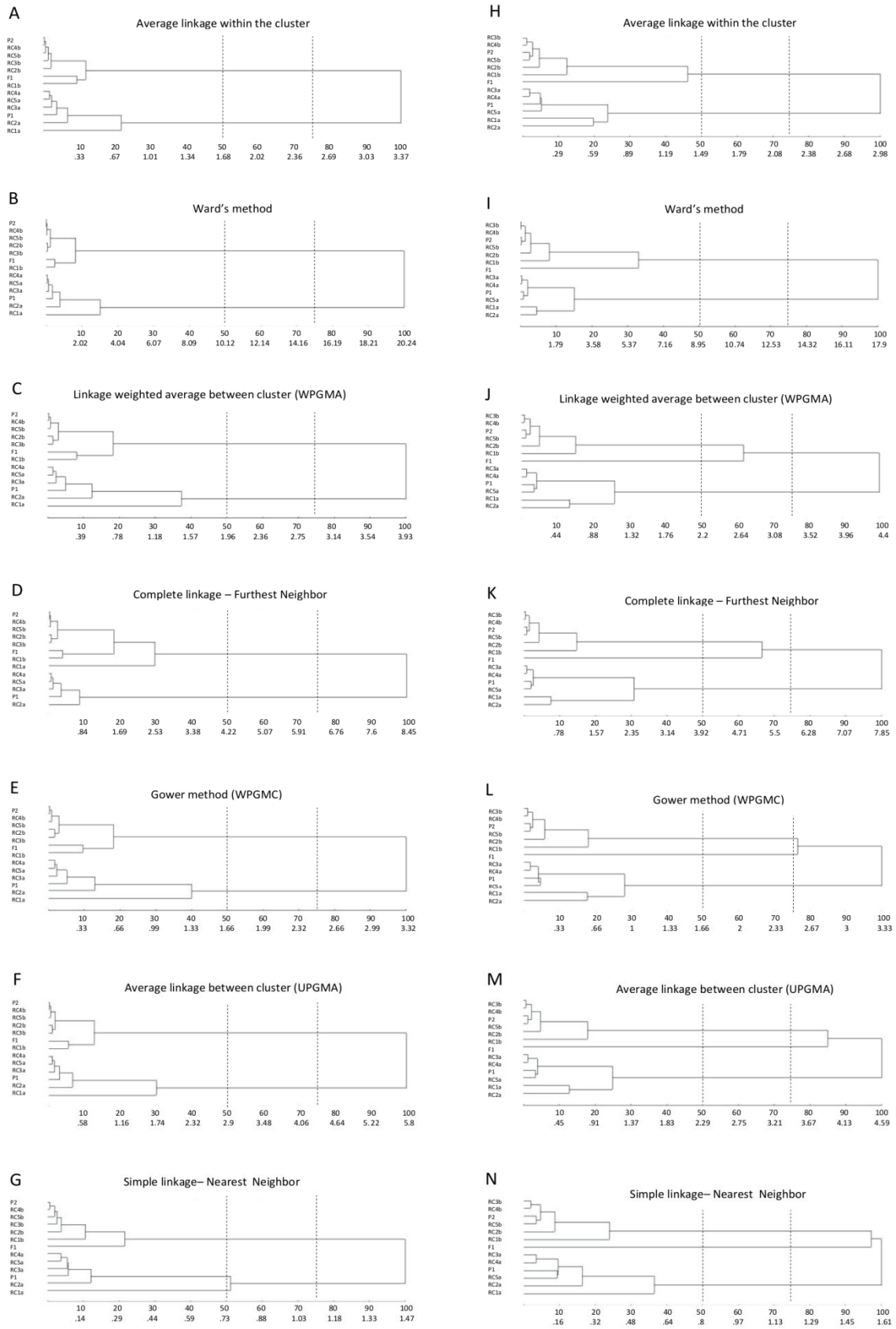


Figure 1. Dendrograms related to the seven hierarchical compared methodologies. The dendrograms on the left refer to the groups with low heritability (20 to 50%), and the dendrograms on the right show the groups of high heritability (55 to 80%).

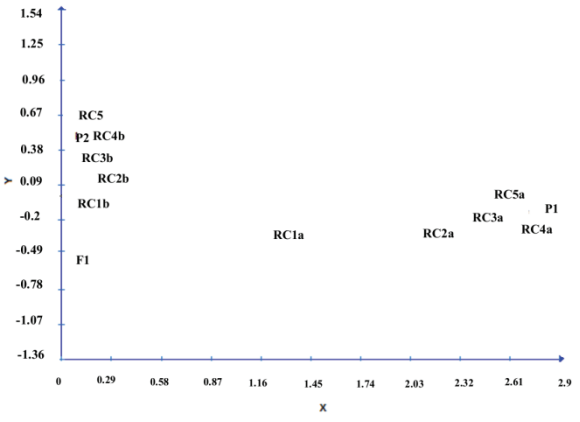


Figure 2. Graphical representation of the 2D projection of the distances in the group with low heritability (0.20 to 0.50) with their respective values of stress (2.089%), distortion (1.493%) and correlation (0.999).

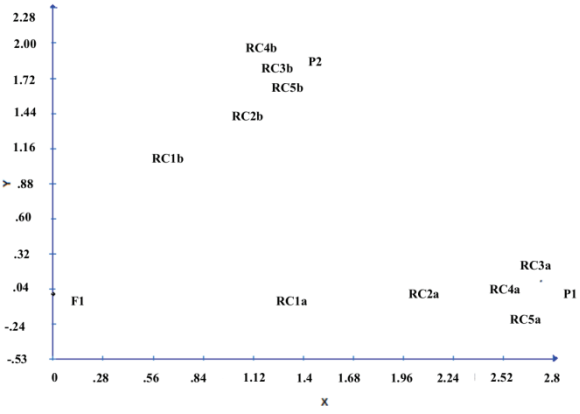


Figure 3. Graphical representation of the 2D projection of the distances in the group of high heritability (0.55-0.80) with their respective values of stress (3.059%), distortion (2.054%) and correlation (0.999).

The results were in agreement with those obtained by hierarchical clustering techniques. The results indicated great diversity between the parents (P1 and P2), and they showed that the F1 generation was intermediary as predicted by gametic union of their parents. The hierarchical structure of the backcrossed generations was viewed by the positioning of RCa populations near the recurrent parent P1 and by the positioning of RCb generations near the recurrent parent P2. The positioning of the early generations of RC1 and RC2 backcrosses contrasted the advanced generations of backcrosses (RC3, RC4 and RC5) because the latter generations were presented close to the recurrent parents due to the high similarity presented with them (all of them above 93%).

The dispersion process of the measures of dissimilarity in the plane is satisfactory when the coefficients expressing the degree of distortion and stress are lower than 20%. Estimates lower than those were obtained in this dataset. In this study, the values obtained by these statistics were adequate for the inferences made from the positioning graph due to the low distortion.

Cophenetic correlations

The cophenetic correlation coefficient has been widely used as a statistical criterion for selecting the hierarchical clustering method to be used, especially when there is no prior knowledge of the pattern of clustering or relevant information on the assessed genetic material.

The coefficients of cophenetic correlation obtained in this study for the analysis performed with both sets of characteristics with high and low heritability were above 0.8 with a range of variation from 0.83 to 0.89. Thus, in general, it was difficult to establish statistical superiority of one method over another taking this as a criterion of adequacy of the method. Despite the small variation of cophenetic correlations, the furthest neighbor clustering method was highlighted with the highest estimate of cophenetic correlation (0.8866) in the analyses performed with the set of characteristics with heritability between 0.20 and 0.50. For the analyses with the set of characteristics with heritabilities ranging from 0.55 to 0.80, the nearest neighbor method showed the highest estimate of cophenetic correlation (0.8786) (Table 3).

Table 3. Values of cophenetic correlation of the assessed hierarchical methods.

Clustering method	Low heritability traits	High heritability traits
Gower	0.8713**	0.8755**
Average linkage within the group	0.8717**	0.8510**
UPGMA	0.8726**	0.8771**
Furthest distance	0.8866**	0.8700**
Nearest neighbor	0.8686**	0.8786**
Ward	0.8703**	0.8354**
WPGMA	0.8718**	0.8679**

**significant at 1% level according to the T-test.

Rohlf and Fisher (1968) suggested that a cophenetic correlation above 0.8 shows the adequacy of the clustering method to represent the genetic diversity among the assessed populations. These authors studied the distribution of cophenetic correlation coefficients under the assumption that the individuals are randomly selected from a single multivariate normal distribution.

Rohlf and Fisher (1968) also found that, on average, cophenetic correlation coefficients tend to decrease as the number of individuals increases and that they are often independent of the number of variables. It cannot be inferred in this work if the amount of variables (phenotypic characteristics) may

influence the quality of the clustering because the heritability was the factor that most influenced the rebuilding of the hierarchy of backcrossing.

Rohlf (1982) warned that a cophenetic correlation near to 0.9 does not guarantee that the dendrogram is able to synthesize the phenetic relationship of individuals. As defined by Sneath and Sokal (1973), a phenetic relationship is the similarity among individuals based on a set of phenotypic characteristics.

The choice of a clustering method depends on the material and the objectives at matter because different clustering methods can lead to different results. No method is considered superior, but some methods are more suitable for certain situations than others (KAUFMAM; ROSSEEUW, 1990). In addition to the algorithm used and the material evaluated, the results of the clustering can be influenced by the selected coefficient of dissimilarity (JACKSON et al., 1989).

On some occasions, the breeder may have doubts about the methodology to be used to study genetic diversity because many of the methods are complementary or overlapping. Although similar, some methods are based on different principles and assumptions. Greater knowledge must be considered even when the analysis is performed with the aid of software. Due to lack of knowledge, in general, users choose analysis options that do not fit their datasets, which result in inconsistent answers. Thus, the scope of the studies, information, and methods has led to some difficulty in choosing methodologies, correctly applying the available methodologies, and conveniently interpreting the meaning of the results. Therefore, new theoretical and practical references that guide the use of computer applications and biometric resources to eliminate gaps and to interpret data are needed. Literature reviews, comparative studies and applied studies are relevant because they help clarify the main issues on the subject (CRUZ; CARNEIRO, 2003).

For data analysis in genetics and breeding, there is some difficulty in establishing the best methodology to be used. Often, the choice is made priori given the particularities of each method or posteriori by some statistical method as the cophenetic correlation coefficient. This difficulty is generated by not knowing the real hierarchy of the group at matter. Therefore, works aiming at finding the most appropriate methodology to reconstruct the clustering pattern of the data are of great importance (FARRIS, 1970).

Sometimes the researcher has in mind the structure of the population through geographical dispersion, its deployment history or previous experimental data. Thus, the cluster analyses are useful for revealing the pattern of similarity and to test hypotheses regarding the causes of this grouping (CRUZ et al., 2011). Therefore, this study helps the breeder to take the best hierarchical clustering method for phenotypic data considering some quantitative parameters.

Conclusion

The hierarchical methodologies and graphic projections were efficient in clustering populations according to the expected clustering standard based on the genetic parents, hybrid generation and backcross derivatives from both parents.

Initially, the expected clustering of traits of low and high heritability was the same, but it was better established when characteristics of high heritability were used. High heritable traits revealed a more efficient hierarchical structure of populations demonstrating that the environmental effect and, therefore, the heritability have an important impact on groups established by the hierarchical methods.

The nearest neighbor method was the most efficient in reconstituting the reality of the backcross system, and it also presented the highest cophenetic correlation. The efficiency of the nearest neighbor method was higher when the analysis involved characteristics of high heritability.

Acknowledgements

The authors thank Fapemig (Foundation Support Research in the State of Minas Gerais) and CNPq (National Council for Scientific and Technological Development) for funding this research.

References

CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. **Biometria Aplicada ao Estudo da Diversidade Genética**. 1. ed. Viçosa: UFV, 2011.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. v. 2. Viçosa: UFV, 2003.

DE ALBUQUERQUE, M. A.; FERREIRA, R. L. C.; DA SILVA, J. A. A.; DE SOUZA, S. E.; STOSIC, B.; DE SOUZA, A. L. Estabilidade em análise de agrupamento: estudo de caso em Ciência florestal. **Revista Árvore**, v. 30, n. 2, p. 257-265, 2006.

DOS SANTOS, E. R.; BARROS, H. B.; FERRAZ, E. C.; CELLA, A. J. S.; CAPONE, A.; DOS SANTOS, A. F.; FIDELIS, R. R. Divergência entre genótipos de soja, cultivados em várzea irrigada. **Revista Ceres**, v. 58, n. 6, p. 755-764, 2011.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to quantitative genetics**. 4th ed. New York: Longman, 1996.

FARRIS, J. S. Methods for computing Wagner Trees. **Systematics Zoologies**, v. 19, n. 1, p. 83-92, 1970.

GIANNOTTI, J. D. G.; PACKER, I. U.; MERCADANTE, M. E. Z.; DE LIMA, C. G. Análise de agrupamento para implementação da meta-análise em estimativas de herdabilidade para características de crescimento em bovinos de corte. **Revista Brasileira de Zootecnia**, v. 34, n. 4, p. 1165-1172, 2005.

GOWER, J. C. A Comparision of Some Methods of Cluster Analysis. **International Biometric Society**, v. 23, n. 4, p. 623-637, 1967.

HARTL, D. L.; CLARK, A. G. **Principles of population genetics**. 3rd ed. Massachusetts: Sunderland Sinauer Associates, 1997.

JACKSON, A. A.; SOMERS, K. M.; HARVEY, H. H. Similarity coefficients: measures for cooccurrence and association or simply measures of occurrence? **American Naturalist**, v. 133, n. 1, p. 436-453, 1989.

KAUFMAN, L.; ROUSSEEUW, P. J. Finding groups in data. An introduction to cluster analysis. New York: John Wiley and Sons, 1990.

KHATTREE, R.; NAIK, D. N. **Applied multivariate statistics with SAS software**. 2nd ed. Cary: SAS Institute Inc., 2000.

LORENCETTI, C.; CARVALHO, F. I. F. D.; OLIVEIRA, A. C. D.; VALÉRIO, I. P.; HARTWIG, I.; MARCHIORO, V. S.; VIEIRA, E. A. Retrocruzamento como uma esratégia de identificar genótipos e desenvolver populações segregantes promissoras em aveia. **Ciência Rural**, v. 36, n. 4, p. 11-18, 2006.

RANGEL, P. H. N.; MOURA NETO, F. P.; FAGUNDES, P. R. R.; MAGALHÃES JUNIOR, A. M. D.; MORAIS,O. P. D.; SCHIMIDT, A. B.; MENDONÇA, J. A.; SANTIAGO, C. M.; RANGEL, P. N.; CUTRIM, V. D. A.; FERREIRA, M. E. Development of herbicide-tolerant irrigated rice cultivars. **Pesquisa Agropecuária Brasileira**, v. 45, n. 7, p. 701-708, 2010.

ROHLF, F. J. Consensus indices for comparing classifications. **Mathematical Bioscience**, v. 59, n.1, p. 131-144, 1982.

ROHLF, F. J.; FISHER, D. R. Test for hierarchical structure in random data sets. **Systematic Zoology**, v. 17, n. 4, p. 407-412, 1968.

SNEATH, P. H.; SOKAL, R. R. **Numerical taxonomy: the principles and practice of numerical classification**. San Francisco: W.H. Freeman, 1973.

Received on February 4, 2013.
Accepted on July 11, 2013.