Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Comparison of multivariate statistical algorithms to cluster tomato heirloom accessions

L.S.A. Gonçalves[1], R. Rodrigues[1], A.T. Amaral Júnior[1], M. Karasawa[2] and C.P. Sudré[1]

[1]Laboratório de Melhoramento Genético Vegetal,
Centro de Ciências e Tecnologias Agropecuárias,
Universidade Estadual do Norte Fluminense Darcy Ribeiro,
Campos dos Goytacazes, RJ, Brasil
[2]Empresa Pernambucana de Pesquisa Agropecuária,
Instituto Agronômico de Pernambuco, Belém de São Francisco, PE, Brasil

Corresponding author: L.S.A. Gonçalves
E-mail: lsagrural@yahoo.com.br

Genet. Mol. Res. 7 (4): 1289-1297 (2008)
Received September 11, 2008
Accepted October 15, 2008
Published November 18, 2008

ABSTRACT. Use of multivariate statistical algorithms is considered an important strategy to quantify genetic similarity. Local varieties and traditional (heirloom) seeds of genotypes are key sources of genetic variation. The Universidade Estadual do Norte Fluminense (UENF), Rio de Janeiro, Brazil, has a tomato gene bank with accessions that have been maintained for more than 40 years. We compared various algorithms to estimate genetic distances and quantify the genetic divergence of 40 tomato accessions of this collection, based on separate and joint analyses of discrete and continuous variables. Differences in continuous variables and discrete and joint analyses were calculated based on the Mahalanobis, Cole Rodgers and Gower distances. Although opinions differ regarding the validity of joint analysis of discrete and continuous data, we found that analyzing a larger number of variables together is viable and can help in the discrimination of accessions; the information that is generated is relevant and promising for both, the accessions conservation and the use of genetic resources in breeding programs.

Key words: *Solanum lycopersicum*; Gower dissimilarity;
Morphoagronomic descriptors; Molecular markers

INTRODUCTION

One of the main concerns of plant breeders is to quantify the degree of dissimilarity in genetic resources (Reif et al., 2005; San-San-Yi et al., 2008), since knowledge concerning genetic distances is necessary for optimum organization of gene banks and for identifying parental combinations that produce progenies with maximum genetic variability, thereby increasing the chances of obtaining superior individuals (Mohammadi and Prasanna, 2003; Crossa and Franco, 2004).

Use of multivariate statistical algorithms is considered an important strategy to quantify genetic similarity. Multivariate techniques permit standardization of multiple types of information of a set of characteristics (Podani and Schmera, 2006). The most widely used algorithms are principal component and canonical variable analysis, as well as clustering methods (Mohammadi and Prasanna, 2003; Sudré et al., 2007).

The principle of clustering methods is to join genotypes into groups, so that there is uniformity within and heterogeneity among groups. These methods depend on previous estimates of dissimilarity measures derived from discrete and continuous (or categorical) variables. These categorical variables can be defined as binary, nominal or ordinal (Núñez et al., 2003; Crossa and Franco, 2004; Podani and Schmera, 2006).

The most widely used dissimilarity distances for continuous variables are the Euclidian and Mahalanobis' generalized distances. The latter has the advantage of considering the residual variances and co-variances (Bedrick et al., 2000; Abbey et al., 2005). The arithmetic complement of the Jaccard index has found widespread use for binary data (Mohammadi and Prasanna, 2003; Buso et al., 2008), while the Cole-Rodgers distance (Cole-Rodgers et al., 1997) has been recommended for multistate variables (Knezovic et al., 2005; Cruz and Carneiro, 2006).

Distance measures that analyze different types of variables simultaneously have not been frequently used to quantify genetic dissimilarity (Vieira et al., 2007), possibly due to skepticism by researchers and a lack of free software for these analyses. In 1971, Gower proposed simultaneous analysis of continuous and categorical variables, using a 0 to 1 scale, regardless of the number of variables, as a basis for data standardization (Crossa and Franco, 2004); this facilitates construction of a dendrogram (Mason et al., 2005). The Gower algorithm (1971) provides a semi-defined positive matrix and is available both as a part of SAS (Mumm and Dudley, 1995) as well as in a free-access software, designated as "R" ("Project for Statistical Computing"). Satisfactory results have been obtained by calculating Gower's distance for the grouping of some crops, including *Brassica napus* (Rodríguez et al., 2005).

Among grouping methods, hierarchical clustering has been used most frequently, particularly the single linkage (SL), unweighted pair group method using arithmetic averages (UPGMA) and Ward (Mohammadi and Prasanna, 2003; Podani and Schmera, 2006) methods. The reliability of clustering methods depends on the magnitude of the cophenetic correlation, which is the association between the genetic distance matrix and the matrix based on genotype grouping (Sokal and Rohlf, 1962).

Local varieties and traditional (heirloom) seeds of genotypes are important sources of genetic variation; both the conservation and the characterization of these accessions are crucial, as they ensure the identification and preservation of useful genes for plant breeding. In Brazil, the tomato is a vegetable of great economic importance (Agrianual, 2007), and there is a great concern in maintaining the variability of tomato gene banks, which may contain specific genes for pest and disease resistance associated with good yield and organoleptic fruit quality. The Universidade Estadual do Norte Fluminense has maintained a tomato germplasm collection that gather accessions with more than 40 years, constituting an important source of genetic resources for crop improvement. Part of these accessions has been characterized using morphological, agronomic and molecular descriptors. We compared various algorithms to estimate genetic distances for the quantification of genetic divergence between these accessions, using both separate and simultaneous analysis of continuous and discrete variables.

MATERIAL AND METHODS

Forty *Solanum lycopersicum* accessions from a gene bank of the Universidade Estadual do Norte Fluminense (UENF) were characterized and evaluated based on continuous (morphological and agronomic traits) and discrete (morphological and molecular markers of the random amplified polymorphic DNA (RAPD) type) variables. Morphological and agronomic characteristics were measured in field trials made in Campos dos Goytacazes, Rio de Janeiro, Brazil, in a randomized block design with three replications and 16 plants per plot. Twenty-five descriptors proposed by the Bioversity International were used (Table 1).

Table 1. Descriptors used for characterization and evaluation of 40 tomato accessions from the gene bank of the UENF, Brazil.

Traits	Description
A. Vegetative descriptors	
Plant growth type	1-4 scale (1 = Dwarf, 2 = Determinate, 3 = Semi-determinate, 4 = Indeterminate)
Foliage density	1-3 scale (1 = Sparse, 2 = Intermediate, 3 = Dense)
Leaf type	1-3 scale (1 = Dwarf, 2 = Potato leaf type, 3 = Standard)
Number of days to flowering	From sowing until 50% of plants have at least one open flower in a plot
B. Inflorescence and fruit	
Corolla color	1-4 scale (1 = White, 2 = Yellow, 3 = Orange)
Exterior color of immature fruit	1-5 scale (1 = Greenish-white, 2 = Light green, 3 = Green, 4 = Dark green, 5 = Very dark green)
Exterior color of mature fruit	1-5 scale (1 = Green, 2 = Yellow, 3 = Orange, 4 = Pink, 5 = Red)
Flesh color of pericarp (interior)	1-5 scale (1 = Green, 2 = Yellow, 3 = Orange, 4 = Pink, 5 = Red)
Fruit size	1-5 scale (1 = Very small, 2 = Small, 3 = Intermediate, 4 = Large, 5 = Very large)
Predominant fruit shape	1-8 scale (1 = Oblate, 2 = Slightly flattened, 3 = Rounded, 4 = High rounded, 5 = Heart-shaped, 6 = Long oblong, 7 = Pear-shaped, 8 = Plum-shaped)
Presence of green shoulder on the fruit	0 (Absent), 1 (Present)
Intensity of green shoulder	1-3 scale (1 = Slight, 2 = Intermediate, 3 = Strong)
Radial cracking	1-4 scale (1 = Corky lines, 2 = Slight, 3 = Intermediate, 4 = Severe)
Concentric cracking	1-4 scale (1 = Corky lines, 2 = Slight, 3 = Intermediate, 4 = Severe)
Number of locules per fruit	Counted on at least 10 fruits: 0 (Absent), 1 (Present)
Presence of open locules	Counting every fruit harvested from each plant
Total fruit number per plant	Total fruit number per plant and plant number ratio
Mean fruit number per plant	Assessed in 10 fruits considering all plants
Total fruit weight (g)	Total fruit weight and plant number ratio
Mean fruit weight (g)	Recorded from stem tip to flower tip at maturity, rounded to one decimal place
Fruit length (mm)	Recorded at the largest diameter of cross-sectioned fruits at maturity, rounded to one decimal place
Fruit width (mm)	
Number of days to maturity	From sowing until 50% of plants have at least one mature fruit
Number of flowers per inflorescence	Mean of 10 plants
Soluble solids	Measures in Brix units of two composite raw juice samples of at least five fruits per juice sample

UENF = Universidade Estadual do Norte Fluminense.

For molecular characterization, 300 mg of the leaves was collected from 35-day-old tomato plants grown in a greenhouse. DNA was extracted based on the protocol of Doyle and Doyle (1987). The process of DNA amplification followed Williams et al. (1990). Electrophoretic process gels were stained with ethidium bromide and photographed under UV light using Eagle Eye II - Stratagene equipment. The following primers were selected: OPPA 03, OPAA 04, OPAA 18, OPAB 05, OPAB 07, OPAB 09, OPAB 14, OPAC 06, OPAH 01, OPC 08, OPC 09, OPC 11, OPC 15, OPE 06, OPE 07, OPE 18, OPG 16, IPO 12, OPK 16, NPO 06, NPO 08, OPO 10, OPT 16, OPW 06, OPW 13, and OPV 12, of Operon Technologies.

For the continuous variables, an analysis of variance was performed at a probability of 1%, and Mahalanobis' generalized distances were calculated to construct the genetic distance matrix. The discrete variables were analyzed based on the distance proposed by Cole-Rodgers (Cole-Rodgers et al., 1997), disregarding the combination 0-0 and considering agreement or disagreement when values were >1. In this situation, this distance is equal to the arithmetic complement of the Jaccard index.

The estimate of the genetic distance matrix for the joint analysis of variables was obtained based on the Gower algorithm (1971), given by:

$$Sij = \frac{\sum_{k=1}^{p} Wijk.Sijk}{\sum_{K=1}^{p} Wijk} \qquad \text{(Equation 1)}$$

where K is the number of variables (k = 1, 2,…, p); i and j any two individuals; W_{ijk} is a weight attributed to comparison ijk, assigning 1 if comparisons are valid and 0 if comparisons are invalid (if the value of the variable is absent in one or both individuals); S_{ijk} is the contribution of variable K to the similarity between individuals i and j, with values between 0 and 1. If the value of variable K is the same for both individuals i and j, for a nominal variable, then S_{ijk} = 1, otherwise S_{ijk} = 0; for a continuous variable

$$S_{ijk} = 1 - \frac{|X_{ik} - X_{jk}|}{R_K} \qquad \text{(Equation 2)}$$

where X_{ik} and X_{jk} are values of variable K for the individuals i and j, respectively, and R_K is the range (minimum subtracted from the maximum value) of variable K in the sample. The division by R_K eliminates the differences between variable scales and produces a value in the range [0, 1] with equal weights.

The accessions were clustered by the SL, UPGMA and Ward methods. The groups were established using a horizontal cut-off value of the distance means for each group formed. The groups were validated by the cophenetic correlation coefficient (Sokal and Rohlf, 1962) and the *t*-test (Steel and Torrie, 1980). The distance matrices were compared with 1000 permutations using the Mantel's correlation test (Mantel, 1967).

The data were analyzed with the program R (http://www.r-project.org), using the package "clusters" and the procedure proposed by Daisy (Maechler, 2007).

RESULTS AND DISCUSSION

Except for soluble solids, the other continuous variables were significant at 1% probability by the F-test, which means that these genotypes have significant variability. Variability in multistate traits was also found in these accessions, since they differed for all traits except for plant growth habit and corolla color; all of them had an indeterminate growth habit and yellow corollas.

The binary traits consisted of RAPD markers; 131 markers were observed with 26 primers. Of these, 89 were polymorphic; therefore, each primer generated a mean of 3.42 polymorphic bands (range 1-6). UPGMA hierarchical clustering was more reliable than Ward and SL in all analysis situations, whether the variables were considered to be separately or jointly, based on cophenetic correlations (Table 2). Sokal and Rohlf (1962) indicated that a cophenetic correlation above 0.80 indicates a good adjustment of the original distance matrices; Mohammadi and Prasanna (2003) and Podani and Schmera (2006) came to the same conclusion.

Table 2. Cophenetic correlations* between the dissimilarity matrix and the three clustering methods used to estimate the genetic divergence in 40 tomato accessions of the UENF gene bank.

Distance	Clusters		
	UPGMA	SL	Ward
Mahalanobis	0.83	0.78	0.70
Cole-Rodgers	0.86	0.79	0.63
Gower	0.87	0.80	0.61

*All correlations were significant at P = 0.01 by the *t*-test. UENF = Universidade Estadual do Norte Fluminense; UPGMA = unweighted pair group method using arithmetic averages; SL = single linkage.

Four groups were clustered by the UPGMA method, based on continuous data (Figure 1). Accessions UENF 178, 201, 202, and 213 of the cherry group were allocated to group I; group II, the largest, was formed by 33 accessions with wide variability, especially for mean number of fruit (from 8 to 22 fruit per plant) and mean fruit weight (from 23 to 70); group III consisted of accession UENF 224 alone, and group IV comprised UENF 196 and 197. In a comparative analysis, 224 UENF did not join group IV, most likely due to the higher total number of fruit, total fruit weight and mean number of fruit compared to accessions UENF 196 and 197. Furthermore, the mean fruit weight and fruit width of the accessions of group IV were higher than for the other accessions. Nonetheless, UENF 197 and 196 were grouped together as they had the lowest values for mean fruit number (3 and 5). This shows that the genotypic superiority of this group in terms of fruit weight is not sufficient for commercial production, in view of their extremely low productivity.

Eleven groups were formed by the discrete (multistate and binary) data (Figure 2). The groups with the most genotypes were X and XI, both with 14 accessions; these genotypes generally had very similar fruit shapes. Slightly flattened and rounded types were predominant, with slight susceptibility to cracking and small or intermediate fruit size. Polymorphism for RAPD markers was greater in groups X and XI, with 33 and 42 polymorphic bands, respectively. This intrapopulational variability in these groups is therefore not negligible; these are groups that join accessions with a small genetic distance of 32% of the maximum multistate variability, e.g., between UENF 186 and 218 in group X, and between UENF 180 and 189 in group XI, with a genetic distance value of 0.09. The other groups contained one or two genotypes: groups I, II, III, V, VI, and VII with one accession, and groups IV, VIII, and IX with two accessions each. But even in groups that contained two accessions, there was a high magnitude of intragroup distance, with values of 0.15, 0.17, and 0.15 among genotypes allocated to groups IV, VIII, and IX, respectively.

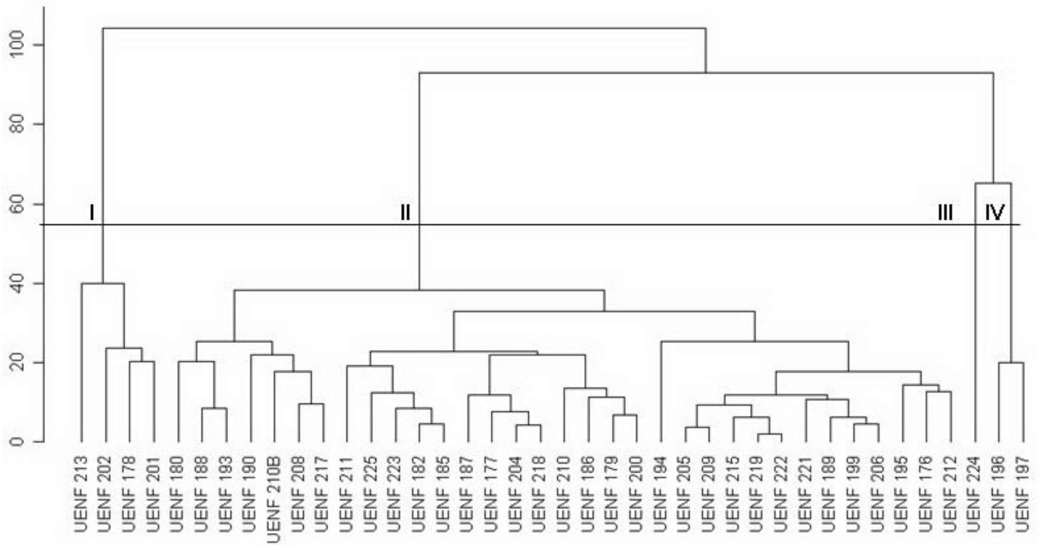

Figure 1. UPGMA dendrogram based on the analysis of 40 accessions using the Mahalanobis distance and continuous variables. UENF = Universidade Estadual do Norte Fluminense.

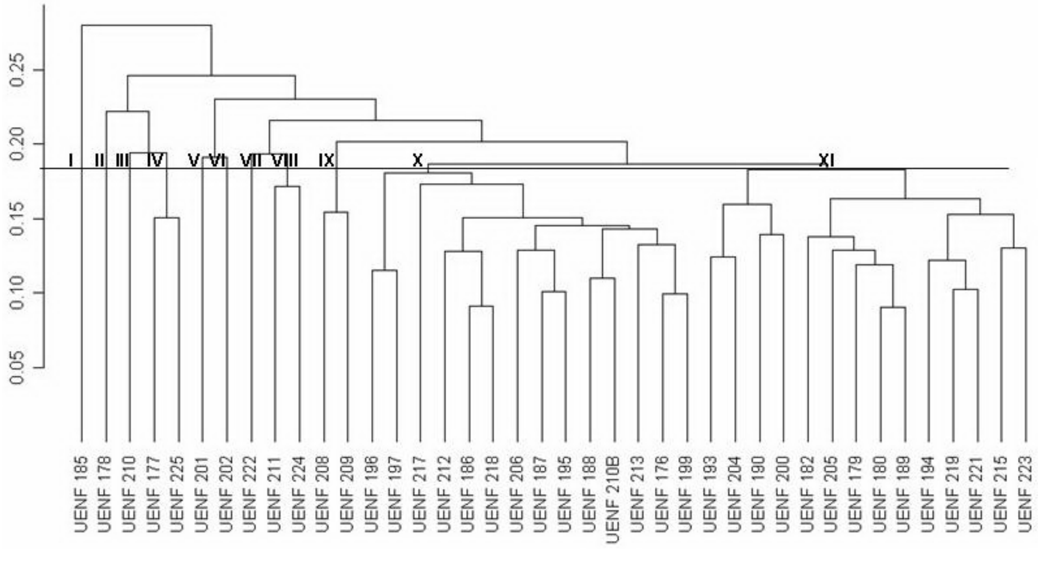

Figure 2. UPGMA dendrogram based on the analysis of 40 accessions using the Cole-Rodgers distance and the discrete (multistate and binary) variables. UENF = Universidade Estadual do Norte Fluminense.

Large differences in grouping can be seen in Figures 1 and 2. They not only had different numbers of groups, but the cherry group accessions, clustered together by the continuous variables, were grouped separately when evaluated by discrete variables. This was the case for the genotypes UENF 178, 201, 202, and 213, which were arranged in the groups II, V, VI and X, respectively (Figure 2). Group II, based on continuous variables, was separated into seven different groups by discrete variables. Genotype discrimination was, therefore, more detailed by discrete than by continuous variables.

Similarly to the discrete variable analysis, Gower's algorithm formed 11 groups (Figure 3). The relationships between Figures 2 and 3 can be summarized as: a) groups I, III, IV, IX, and XI by Gower's distance were similar, respectively, to I, VII, VIII, II and III, based on discrete variables; b) based on Gower's distance, groups V and VI contained 14 accessions each; of the 14 accessions of group V, 10 were grouped in group X, while 12 accessions of group VI composed group XI, by discrete variables, and c) accessions UENF 196 and 197, which formed group VIII by Gower's distance, were part of group X by discrete variables. In turn, accessions UENF 201 and 202, separated into different groups by discrete variables, composed a single group - group II - by Gower's algorithm. There was disagreement; accessions UENF 177 and 225, which were grouped together in group IV, as well as UENF 208 and 209 in group IX by discrete variables, were allocated into separate groups by Gower's algorithm.

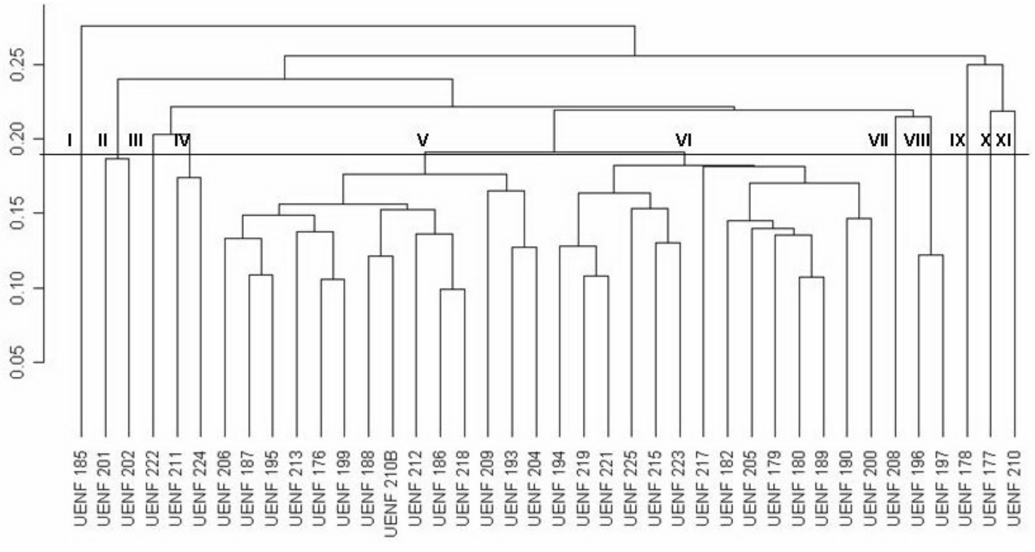

Figure 3. UPGMA dendrogram based on the analysis of 40 accessions using the Gower distance and the mixed (continuous and discrete) variables. UENF = Universidade Estadual do Norte Fluminense.

The greater similarity among groups based on discrete variables and Gower's distance, as opposed to the grouping by continuous variables, deserves attention because, in the former, the analyses of the discrete variables as well as the joint analysis of all variables were more efficient in the discrimination of dissimilar genotypes. However, this premise requires care in interpretation, since we only analyzed 10 continuous variables altogether, while the discrete and joint (continuous plus discrete), totaled 146 and 156 data points, respectively.

Thus, for the latter, there are 146 and 156 data points that may or may not be consistent among accessions, and in the distance calculation proposed by Cole-Rodgers as well as by Gower; what matters is the agreement or disagreement of discrete variables. This can generate improved evaluation of genetic dissimilarity based on discrete differences, as if it were a tandem procedure, and consequently favors a clearer separation of accessions. Nevertheless, the capacity of continuous variables for genetic discrimination is unquestionable, since we found high variability for these traits, based on the F-test.

The value of the Mantel correlation was only moderately high (r = 0.40) in the matrices of discrete and continuous variables, indicating that the Mahalanobis and Cole-Rodgers techniques sampled different genome regions. The situation was similar in the matrices of joint and continuous variables, based on the correlation of 0.49. On the other hand, the correlation between the joint matrix (Gower) and Cole-Rodgers was high (r = 0.88). This indicated that the matrices from discrete and joint data can be used indifferently. This may be true when there is considerable qualitative data and little quantitative data. Similarly, Vieira et al. (2007), in a study of 19 wheat genotypes using amplified fragment length polymorphism molecular and morphological quantitative markers, observed a moderate correlation between the morphological and molecular matrices, a high correlation between the matrices of the joint analysis (molecular and morphological markers) with the morphological matrix, and a moderate correlation between the matrices of the joint analysis with the morphological matrix. They concluded that, due to moderate correlation between the joint and the morphological data matrix, the genotype divergence data should be considered separately. However, Franco et al. (2001) suggested that genotypes are best discriminated by the simultaneous analysis of morphological and molecular data, determining *a priori* the minimum number of markers that lead to the same results as when combined with all markers.

Although opinions differ regarding the use of joint analysis of discrete and continuous data, we found that using a greater number of variables analyzed together is viable and can help separate accessions, provided that the information that is generated is relevant and useful both for the conservation of accessions and for the use of genetic resources in breeding programs.

ACKNOWLEDGMENTS

The authors thank FAPERJ (Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro), Brazil, for a scholarship granted to the first author.

REFERENCES

Abbey L, Joyce DC, Aked J and Smith B (2005). Evaluation of eight spring onion genotypes, sulphur nutrition and soil-type effects with an electronic nose. *J. Hort. Sci. Biotechnol.* 80: 375-381.

Agrianual (2007). Anuário da Agricultura Brasileira. FNP Consultoria & Comércio, São Paulo.

Bedrick EJ, Lapidus J and Powell JF (2000). Estimating the Mahalanobis distance from mixed continuous and discrete data. *Biometrics* 56: 394-401.

Buso GS, Paiva MR, Torres AC, Resende FV, et al. (2008). Genetic diversity studies of Brazilian garlic cultivars and quality control of garlic-clover production. *Genet. Mol. Res.* 7: 534-541.

Cole-Rodgers P, Smith DW and Bosland PW (1997). A novel statistical approach to analyze genetic resource evaluations using *Capsicum* as an example. *Crop Sci.* 37: 1000-1002.

Crossa J and Franco J (2004). Statistical methods for classifying genotypes. *Euphytica* 137: 19-37.

Cruz CD and Carneiro PCS (2006). Modelos Biométricos Aplicados ao Melhoramento Genético. 2nd edn. Universidade Federal de Viçosa,Viçosa.

Doyle JJ and Doyle JL (1987). Isolation of plant DNA from fresh tissue. *Focus* 12: 13-15.

Franco J, Crossa J, Ribout JM, Betran J, et al. (2001). A method for combining molecular markers and phenotypic attributes for classifying plant genotypes. *Theor. Appl. Genet.* 103: 944-952.

Gower JC (1971). General coefficient of similarity and some of its properties. *Biometrics* 27: 857-874.

Knezovic Z, Gunjaca J, Satovic Z and Kolak I (2005). Comparison of different methods for classification of gene bank accessions. *Agric. Conspec. Sci.* 70: 87-91.

Maechler M (2007). The cluster package. Available at [http://www.r-project.org]. Accessed January 8, 2008.

Mantel N (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27: 209-220.

Mason NWH, Mouillot D, Lee WG and Wilson JB (2005). Functional richness, functional evenness and functional divergence: the primary components of functional diversity. *Oikos* 111: 112-118.

Mohammadi SA and Prasanna BM (2003). Analysis of genetic diversity in crop plants - salient statistical tools and considerations. *Crop Sci.* 43: 1235-1248.

Mumm RH and Dudley JW (1995). A PC SAS computer program to generate a dissimilarity matrix for cluster analysis. *Crop Sci.* 35: 925-927.

Núñez M, Villarroya A and Oller JM (2003). Minimum distance probability discriminant analysis for mixed variables. *Biometrics* 59: 248-253.

Podani J and Schmera D (2006). On dendrogram-based measures of functional diversity. *Oikos* 115: 179-185.

Reif JC, Melchinger AE and Frisch M (2005). Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Sci.* 45: 1-7.

Rodríguez VM, Cartea ME, Padilla G, Velasco P, et al. (2005). The nabicol: A horticultural crop in northwestern Spain. *Euphytica* 142: 237-246.

San-San-Yi, Jatoi SA, Fujimura T, Yamanaka S, et al. (2008). Potential loss of unique genetic diversity in tomato landraces by genetic colonization of modern cultivars at a non-center of origin. *Plant Breed.* 127: 189-196.

Sokal RR and Rohlf FJ (1962). The comparison of dendrograms by objective methods. *Taxon* 11: 33-40.

Steel RGD and Torrie JH (1980). Principles and Procedures of Statistics: A Biometrical Approach. McGraw-Hill, New York.

Sudré CP, Leonardecz E, Rodrigues R, Amaral Júnior AT, et al. (2007). Genetic resources of vegetable crops: a survey in the Brazilian germplasm collections pictured through papers published in the journals of the Brazilian Society for Horticultural Science. *Hortic. Bras.* 25: 496-503.

Vieira EA, Carvalho FIF, Bertan I, Kopp MM, et al. (2007). Association between genetic distances in wheat (*Triticum aestivum* L.) as estimated by AFLP and morphological markers. *Genet. Mol. Biol.* 30: 392-399.

Williams JG, Kubelik AR, Livak KJ, Rafalski JA, et al. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* 18: 6531-6535.