Acta Scientiarum

http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v42i1.42483

Comparison of projection of distance techniques for genetic diversity studies

Isabela de Castro Sant'Anna[1*], Gabi Nunes Silva[2], Vinícius Quintão Carneiro[3], Daiana Salles Pontes[1], Moysés Nascimento[1] and Cosme Damião Cruz[1,3]

[1]Laboratório de Bioinformática, Departamento de Estatística, Universidade Federal Viçosa, Avenida P.H. Rolfs, s/n, 36570-900, Viçosa, Minas Gerais, Brazil. [2]Departamento de Matemática e Estatística, Universidade Federal de Rondônia, Ji- Paraná, Rondonia, Brazil. [3]Laboratório de Bioinformática, Departamento de Biologia Geral, Universidade Federal Viçosa, Viçosa, Minas Gerais, Brazil. *Author for correspondence. E-mail: isabelacsantanna@gmail.com

ABSTRACT. The objective of this study was to compare different graphical dispersion analysis techniques in two- or three-dimensional planes. In this study, the data from different published works were used in order to determine the best methodology for analyzing the genetic diversity of different species. In this study, efficiency is measured by the amount of original distance absorbed by the projection of distances technique, which in the case of major components is equal to the amount of total variation originally available and retained by the principal components used for dispersion purposes. The projection of dissimilarity measurement technique, principal component analysis (PCA), and principal coordinate analysis (PCoA) were used. Considering the analysis by means of three orthogonal axes, the graphical dispersion efficiency was 82.22 for PCA, 87.22 for PCoA, and 85.25 for the projection of distances technique. For the 2D analysis, considering the two main axes, the mean dispersion efficiency was 69.90 for the PCA, 75.06 for the projection technique, and 78.16 for PCoA. Considering the studies carried out with experimental data of six different species, it is concluded that the principal coordinate analysis is superior.

Keywords: adaptive methods; dissimilarity; multivariate; crops; statistical.

Received on April 20, 2018.
Accepted on July 2, 2018.

Introduction

Studies regarding the discrimination of plant and animal populations are of great importance for the development of breeding programs and biodiversity conservation. Analysis of genetic diversity, through phenotypic characteristics, has guided the selection of suitable parents at the early stages of breeding programs, leading to the optimization of selective gains. In addition, analysis of genetic diversity has allowed the quantification of variability and has facilitated the management of germplasm banks, saving time and resources (Cruz, Ferreira, & Pessoni, 2011). Currently, there are several methodologies available for the quantification and evaluation of diversity in population studies, either from phenotypic information or genotypic data. However, due to the wide variety of information and given the particularities of each biological material, the choice and the application of the most appropriate methodology are important to obtain reliable results (Mohammadi & Prasana, 2003).

Generally, individual information is lost with clustering, and only information referring to the group means remains. However, when working with many individuals, the numbers obtained for similarity/dissimilarity estimates are relatively large, making it difficult to recognize homogeneous groups (Cruz et al., 2011). The choice falls on individual clusterings based on dispersions in relation to Cartesian axes obtained by the principal components, principal coordinate and projection of distance techniques.

Representation in plane, in a smaller number of dimensions, proves to be helpful for the identification of related genotypic groups (Cruz et al., 2011). Several studies used principal coordinates to demonstrate the divergence between genotypes using molecular data (Aitken et al., 2018; El-Esawi et al., 2018; Guzmán, Segura, Aradhya, & Potter, 2018; Taglioti et al., 2018;Wu et al., 2018; Ferreira et al., 2016; Zaher et al., 2011; Van Inghelandt, Melchinger, Lebreton, & Stich, 2010). Other studies were based on principal component analyses (Muller et al., 2018; Valcárcel, Peiró, Pérez-de-Castro, & Díez, 2018; Osawaru, Ogwu, & Dania-Ogbe, 2013; Santi et al., 2012; Jombart, Devillard, & Balloux, 2010). Additionally, studies have been conducted that used the projection of distance technique in a two or three-dimensional plane (Nicky, Carvalho, Assis, & Carvalho, 2008; Sant'Anna & Cruz, 2015).

With regard to plants, in addition to enabling the study of the genetic diversity of a group of accessions, principal component analysis has the advantage of allowing the evaluation of the importance of each characteristic studied based on the total variation available among genotypes. The interest in this evaluation lies in the possibility of discarding traits that contribute little to the discrimination of the evaluated genotypes, reducing the labor, time and costs necessary for agricultural experimentation.

The principal component analysis (PCA) is based only on the individual information of each accession, without the need for replicated data, and it has been frequently used in many multivariate procedures whose purposes are not restricted to the study of genetic diversity. However, a criticism of this technique is related to its association in terms of graphical dispersion by the Euclidean distance. Thus, according to Cruz et al. (2011), variance explained by the first two principal components corresponds to the complement of the statistical stress. This stress is measured by the ratio of the sum of the square of the Euclidean distance estimated from the principal component scores and the square of the Euclidean distance estimated from the original variables.

In some cases, the investigator has reasons not to adopt the Euclidean distance as the dissimilarity measure; consequently, he or she will have the same reasons not to adopt the technique of principal components as the best graphical representation of the distance in the evaluated accessions. A common situation in genetics and breeding, in which the Euclidean distance has not been recommended, refers to studies by molecular markers with dominant expression in which a pattern of agreement of band absence should not be counted as similarity between accessions. Moreover, in this case, measures such as the arithmetic complement of the Jaccard index similarity index and Nei and Li distances have been preferred (Jaccard, 1908; Nei & Li, 1979).

The use of projection of distances and principal coordinate (PCoA) techniques has been recommended to graphically represent any dissimilarity measure that properly represents the standard and the particularity of the studied data. However, there is no information on the effectiveness of one technique over the other, and these techniques are often arbitrarily chosen by the researcher. For this reason, it is important to carefully choose the methodology that has greater adaptability in order to ensure that the result is not an artifact of the technique. Thus, the objective of this study was to compare different techniques of graphical dispersion analysis in two and three-dimensional planes to determine the best method in order to assist in further genetic diversity analyses based on the analyses of different crops.

Material and methods

The present study was carried out in the Bioinformatics Laboratory of the Institute of Biotechnology Applied to Agriculture (BIOAGRO) of the Federal University of Viçosa, through the software Genes (Cruz, 2013) and R (R Core Team, 2018). Thus, phenotypic data from previously published studies were used. Analyses were carried out for six plant species: Brachiara, coffee, cassava, Commom bean, eucalyptus, and soybeans (Table 1).

Table 1. Description of the researches carried out by different authors, showing the studied crop, number of considered variables and number of genotypes in each work.

Crop	Variables	Number of genotypes	Author
Brachiaria spp.	18	6	(Assis, Euclydes, Cruz, & Do Valle, 2003)
Coffee	14	40	unpublished data
Eucalyptus saligna	18	5	unpublished data
Eucalyptus saligna	14	3	unpublished data
Common Bean	15	25	unpublished data
Common Bean	10	23	unpublished data
Common Bean	15	3	unpublished data
Cassava	8	6	(Gomes, Carvalho, Jesus, & Custódio, 2007)
Soybean	6	20	unpublished data
Soybean	6	20	unpublished data
Soybean	6	20	unpublished data
Soybean	10	11	unpublished data
Soybean	6	5	unpublished data
Soybean	6	20	unpublished data
Soybean	10	11	unpublished data
Soybean	6	20	unpublished data

*In the Brachiaria study, 301 accessions of *Brachiaria* species were evaluated: 150 *B. brizantha*, 46 *B. decumbens,* 36 *B. humidicola*, 31 *B. jubata*, 28 *B. ruziziensis*, and 10 *B. dictyoneura.*

Species were used only to provide datasets with different numbers of accessions and traits, which were evaluated using different experimental accuracies. The hypothesis considered that the superiority of a particular technique used in the 17 experimental sets could be indicative of its potentiality and of the possibility of a generalized recommendation of its use. These works were chosen for being fairly complete, for having been carried out over one or more evaluation years, and for being crops that are widely used in breeding programs. Traits used in this study are described (Suplementar Material).

For the genetic diversity analyses carried out by the principal coordinate analysis and the projection of distances on the plane, the dissimilarity matrix was used, the elements of which were given by the Euclidean distance mean. For the principal component analysis, the means of the accessions available in the referred studies were used.

Principal components

The principal component analysis consists of converting an original set of variables to another set of equivalent dimensions but with important properties that are of great interest in some improvement studies. Each principal component is a linear combination of the original variables. Moreover, they are independent and are estimated in order to retain, in an estimate order, the maximum information in terms of the total variance contained in the original data.

Projection of distances

In this procedure, dissimilarity measures are converted into scores of two or three variables, which, when represented in dispersion graphics, will reflect distances originally obtained from the v-dimensional space (v = number of traits used for obtaining distances) in the two- or three-dimensional space.

To create a graphical representation of similarity measures, the coordinate of the most divergent measures is calculated. Then, the coordinate of those that showed, in descending order, the greatest diversity with the genotypes previously considered is calculated, as described by (Cruz et al., 2011), where i and j are the most divergent genotypes. It is arbitrarily considered that the coordinate of i is equal to (0.0) and the coordinate of j is equal to (d_{ij}, 0). A third genotype k presents the coordinates (X_k, Y_k), and X_k are Y_k mathematically estimated. The other coordinates are established by statistical processes with allocation error minimization.

Principal Coordinates

Analyses carried out by principal coordinate techniques used the matrix of genetic distances to produce a coordinate plot in which accessions or populations are represented by points in the Cartesian plane (Gower & Hand, 1966). Similar to the cluster analysis, this analysis allows clear visualization of the genetic diversity level of accessions. In the present study, the Euclidean distance mean was chosen for the dispersions carried out in the 2D and 3D planes. The D'Center module was used to transform the distance matrix in scalar products in order to allow the calculation of the eigenvalues and eigenvectors used in the analysis of the principal coordinates (Gower & Hand, 1966). During this transformation, all elements of the triangular matrix (d_{ij}), except for the principal diagonal, are replaced by -1/2 $d^2{}_{ij}$. Afterwards, the mean for each line and the mean of each column are subtracted from the value of each element, and finally, the matrix mean is summed. Thus, the focusing matrix was obtained. This matrix, through the EIGEN module, was used to calculate the eigenvalues and eigenvectors, which allow the representation of the principal coordinates in two and three dimensions. The consistency of the results was evaluated by the percentage of total variance explained by the first vectors of the principal coordinates.

In the present study, efficiency is a parameter that can be used to evaluate the best technique. Thus, the efficiency of techniques is measured by the quantity of the original distance absorbed by the graphical distance, which, in the case of principal components, is equal to the quantity of total variation originally available and retained by the principal components used for graphical dispersion purposes.

Results and discussion

Comparison of techniques

In this study, three biometric techniques were used for graphical representation of genetic diversity, and the choice of technique depends on the type of the available data or the author's preference without adopting a scientific criterion. This study clarified the situations in which the available methodologies can be chosen so that experimental accuracy is not compromised, and the results are the diversity quantification and evaluation, considering the particularities of the biological material and the program's goals.

In the analyses carried out for 16 different crops with different numbers of variables and genotypes, the principal coordinate technique was more efficient in 93.75% and 70% of the analyses in two and three dimensions, respectively. Additionally, the principal coordinate technique was equally efficient in the analysis with two main axes in comparison with the other methods in 18.75% of cases. In contrast, in most analyses, the principal coordinate analysis was far superior to the other two methods (Table 2). The projection of distances technique was successfully used to represent genetic diversity by Sant'Anna and Cruz (2015), who found that in simulated populations, this technique placed the parents and the five backcross generations in the Cartesian plane, respecting the genetic similarities, which ranged from 75% to 98%, exactly as expected by the Mendelian principles. Nicky et al. (2008) also observed the projection of distances technique's efficiency when applying it together with other genetic diversity methods in green pepper and pepper populations.

The principal coordinate analysis has been widely used to represent genetic diversity in breeding programs in species such as maize (Van Inghelandt et al., 2010), sweet sorghum (Murray, Rooney, Hamblin, Mitchell, & Kresovich, 2009), rice (Choudhury et al., 2014; Roy et al., 2015), sweet potato (Su et al., 2017), and barley (Ferreira et al., 2016), among others. In all these works, associations among genotypes were revealed with the principal coordinate analysis.

Table 2. Description of the works carried out by different authors, showing the studied crop, the number of considered variables and the number of genotypes in each work, and the results obtained by projection of distances, principal components and principal coordinates in two and three orthogonal axes. The results are presented in distortion values.

			3D			2D		
Variables	Genotypes	Crop	PCoA	PD	PCA	PCoA	PD	PCA
6	20	Soybean	5.83	6.7	6	11.78	13.46	13.6
6	20	Soybean	2.76	0.03	3.85	16.13	18.71	16.7
6	20	Soybean	4.77	5.12	5.84	9.31	10.25	11.77
10	11	Soybean	7.46	14.48	9.18	17.62	19.84	24.17
6	5	Soybean	3.83	9.45	5.48	14.12	18.41	18.2
6	20	Soybean	3.77	4.61	4.71	10.17	11.04	11.96
10	11	Soybean	8.55	10	9.44	19.42	21.09	22.08
6	20	Soybean	7.46	8.5	9.32	18.05	8.5	24.72
10	23	Common bean	35.11	30.16	40.28	34.19	41.55	55.1
15	3	Common bean	33.05	31.04	43.23	34.71	40.49	59.86
15	25	Common bean	11.21	26.66	35.16	32.2	36.36	47.41
18	6	Brachiaria	30.62	10.07	12.35	18.28	20.78	27.54
14	40	Coffee	19.57	24.31	30.42	30.7	38.58	46.93
3	9	Eucalyptus	14.64	18.43	19.87	26.72	32.86	32.98
3	9	Eucalyptus	12.71	16.24	22.01	24.79	30.9	39.22
8	6	Cassava	4.87	36.25	30.61	31.31	36.25	45.43
		mean	12.89	15.75	17.98	21.84	24.94	31.10

Considering the analysis by means of three orthogonal axes, dispersion graphical efficiency had a mean of 82.22 for the principal component analysis, 87.22 for the principal coordinate analysis, and 85.25 for the projection of distances technique. In the case of 2D analyses, considering two main axes, dispersion efficiency was, on average, 69.9 for the principal components analysis, 78.16 for the principal coordinate analysis, and 75.06 for the projection of distances technique. The three techniques were successful when projected on three main axes. However, usually, the principal component analysis cannot be used due to the necessity of projecting distances, which are calculated considering the limitations caused by biological and experimental conditions (Cruz et al., 2011).

However, the main components technique has the advantage of evaluating the importance of each studied feature on the total available variation among the evaluated accesses, allowing fewer discriminating characters to be discarded, since the characters are already correlated with other variables redundant or by their invariance. For example, Valcárcel et al. (2018) studied the morphological characterization of 206 cucumber accessions using five plants per accession that were characterized with 17 qualitative and nine quantitative descriptors; eight of them referred to plant traits, and 18 related to the fruit. According to the PCAs, the traits related to the weight and the shape of the fruit are key to explain the variability found among the accessions included in each group. Besides that the morphological characterization allowed the selection of 67.2% of the collection, eliminating the most similar accessions.

On the other hand, when the projection of distances and the principal coordinates are considered, it is necessary to make a conscious choice since both techniques can be calculated using the same types of data, and researchers should base their choice on the most appropriate scientific criteria.

Graphical analysis in the 2D plane

Individual clustering based on dispersions, with respect to Cartesian axes obtained by the principal component analysis, principal coordinate analysis and projection of distances technique, facilitates the visualization of the genetic diversity of homogeneous groups without a great loss of individual information. Therefore, these techniques are chosen over others that also quantify genetic diversity (Cruz et al., 2011). Among the results, the study of the coffee crop, including 40 genotypes and 14 agronomic characteristics, showed greater discrepancy in the comparison between techniques. Therefore, these results were used to illustrate the differences between the techniques in 2D graphical analyses (Figures 1, 2 and 3). Analysis by means of two main axes of the 40 coffee genotypes, in which 14 characteristics were analyzed, presented a distortion of 38.58 for the projection of distances technique, while the principal components and coordinates presented 46.93 and 30.70, respectively. It was observed that individuals 17 and 35 as well as 12 and 13 are similar to each other and divergent in relation to the others, and they remained more distant from the other individuals.

This behavior is similar for the three techniques, with few differences. However, there is a nearby group consisting of individuals 6, 7, 11, and 19. Although this group is observed in the three techniques, it is not very different from the remaining individuals in the projection of distances technique, similar to the principal main component and coordinate techniques. Figures 1, 2, and 3 are quite similar when they are rotated. However, it is up to the author to decide which technique best represents the data, since the results obtained by the utilized technique, despite being quite similar, present discrepancies that may be quite relevant. Thus, the peculiarities of the material and the technique used in the study must be considered. For example, if it is necessary to preserve the most divergent group, the projection of distance technique can provide three points without any distortion (Cruz et al., 2011).

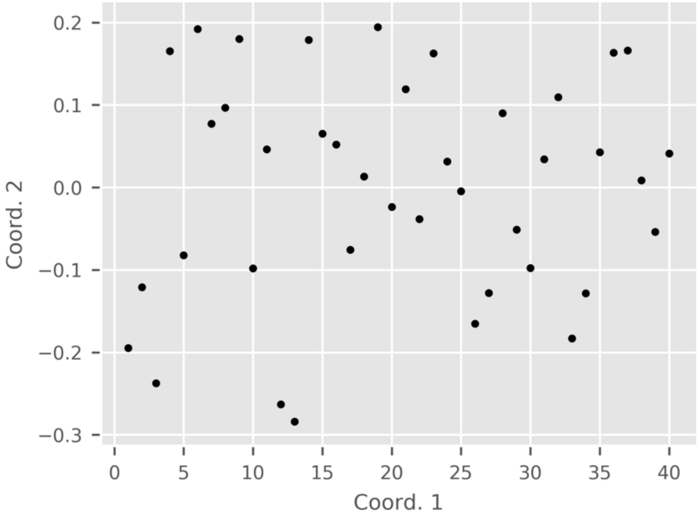

Figure 1. Graphical dispersion of the 40 coffee genotypes in relation to the first and second principal components, established by linear combination of 14 agronomic characteristics.

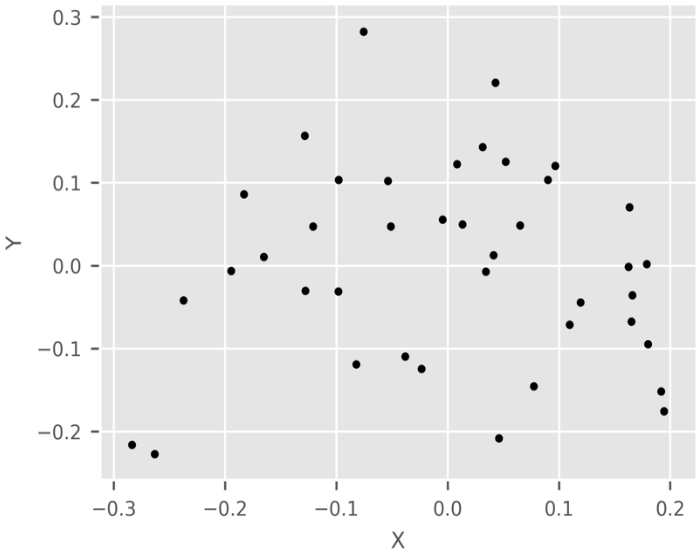

Figure 2. Graphical dispersion of the 40 coffee genotypes in relation to the Euclidean distance mean expressed in two-dimensional plane.

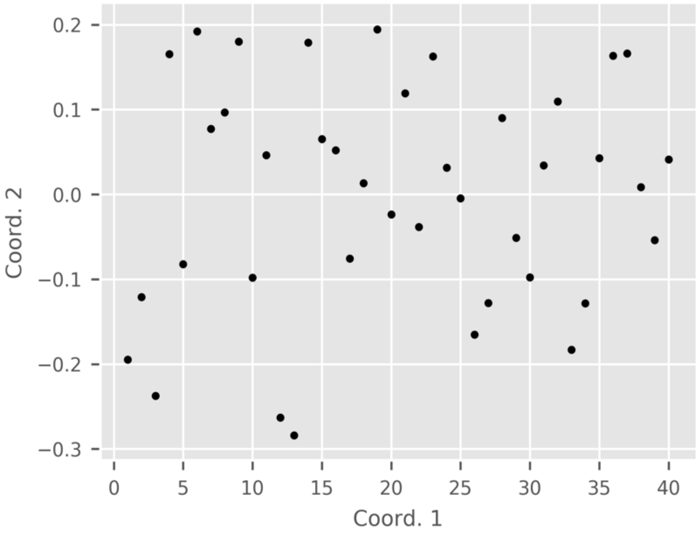

Figure 3. The two-dimensional PCoA plots show the low percentage of the total genetic diversity.

3D plane graphical analysis

In situations in which the researcher does not find the expected deviation for the data through projection techniques at a distortion level below 20% (which is recommended by the literature) (Cruz et al., 2011), it is recommended to use projection in three orthogonal axes to ratify the obtained conclusions. However, when stress or distortion is very high, above 20%, there might be no correlation between the results obtained by projections in two or three orthogonal axes, in relation to the divergence level found. Again, graphical analysis by means of three orthogonal axes of the 40 coffee genotypes in which 14 characteristics were analyzed presented distortion of only 19.57 principal coordinates, while a distortion of 30.62 and 24.31 for the principal components and projection of distance techniques was observed, respectively. Therefore, the relationships previously observed are not found in Figures 4, 5, and 6. However, when distortions for the 2D planes are below that recommended by the literature, it is possible to observe compatibility in the results. In the case of dispersions produced by the three techniques, there were discrepancies between the results, but all dispersions positioned individuals 12 and 13 near one another, which is similar to what happened to the 2D plane. This suggests the need to use other methods that evaluate genetic diversity so that the most suitable technique is chosen.

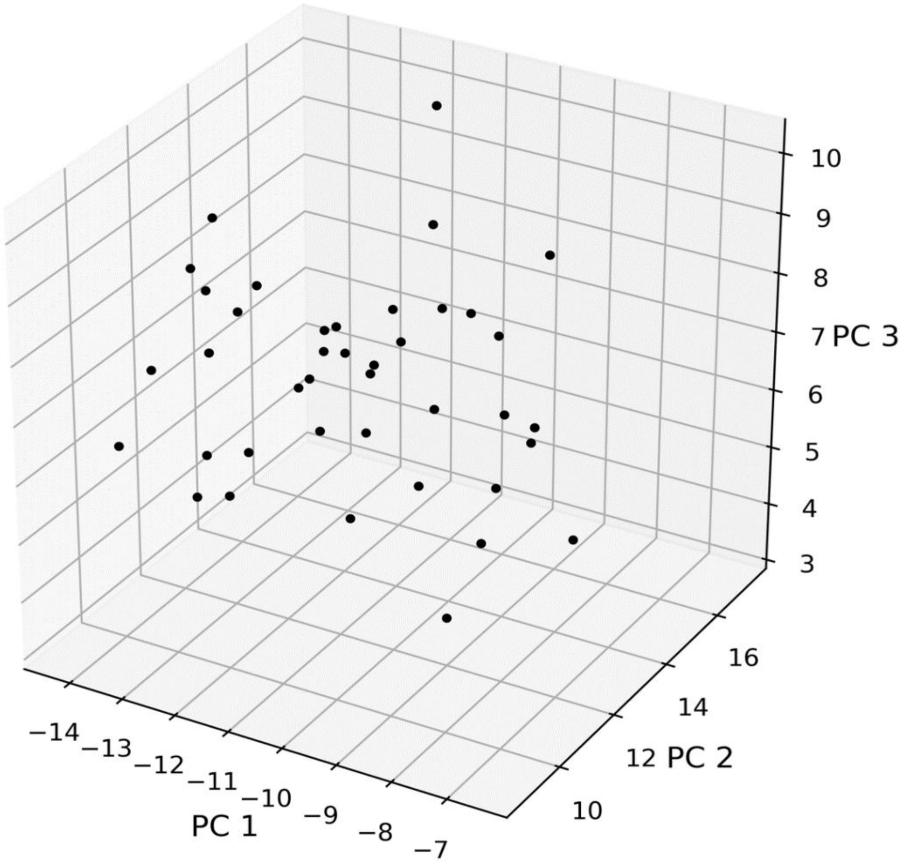

Figure 4. Graphical dispersion of the 40 coffee genotypes in relation to the first, second and third principal components, established by linear combination of 14 agronomic characteristics.

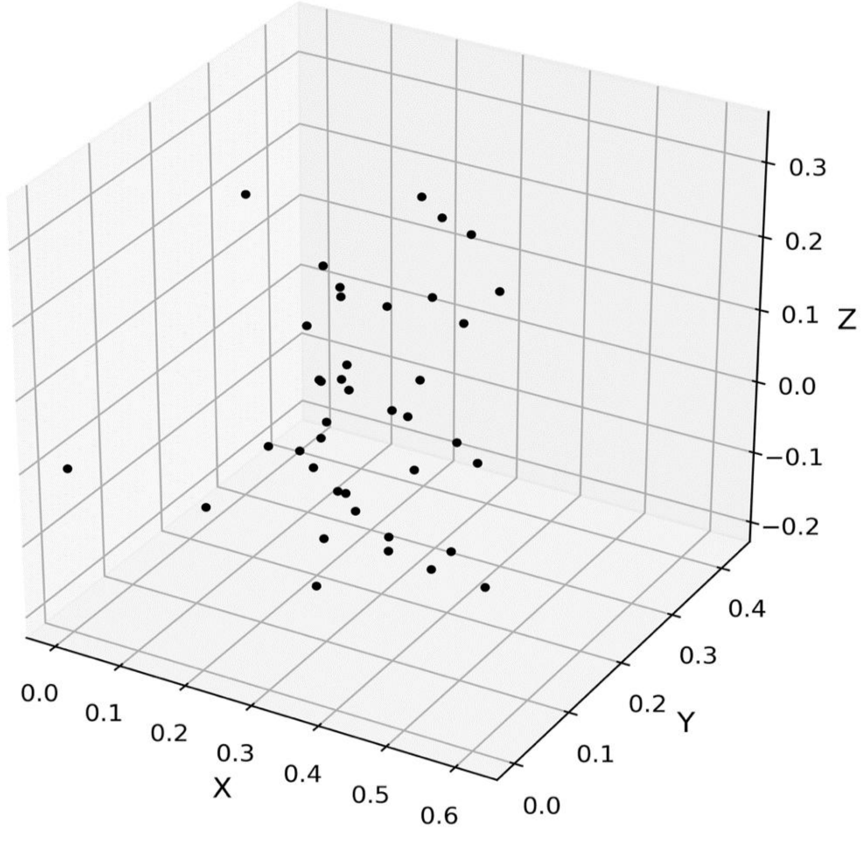

Figure 5. Graphical dispersion of the 40 coffee genotypes expressed by the Euclidean distance mean in three-dimensional plane.

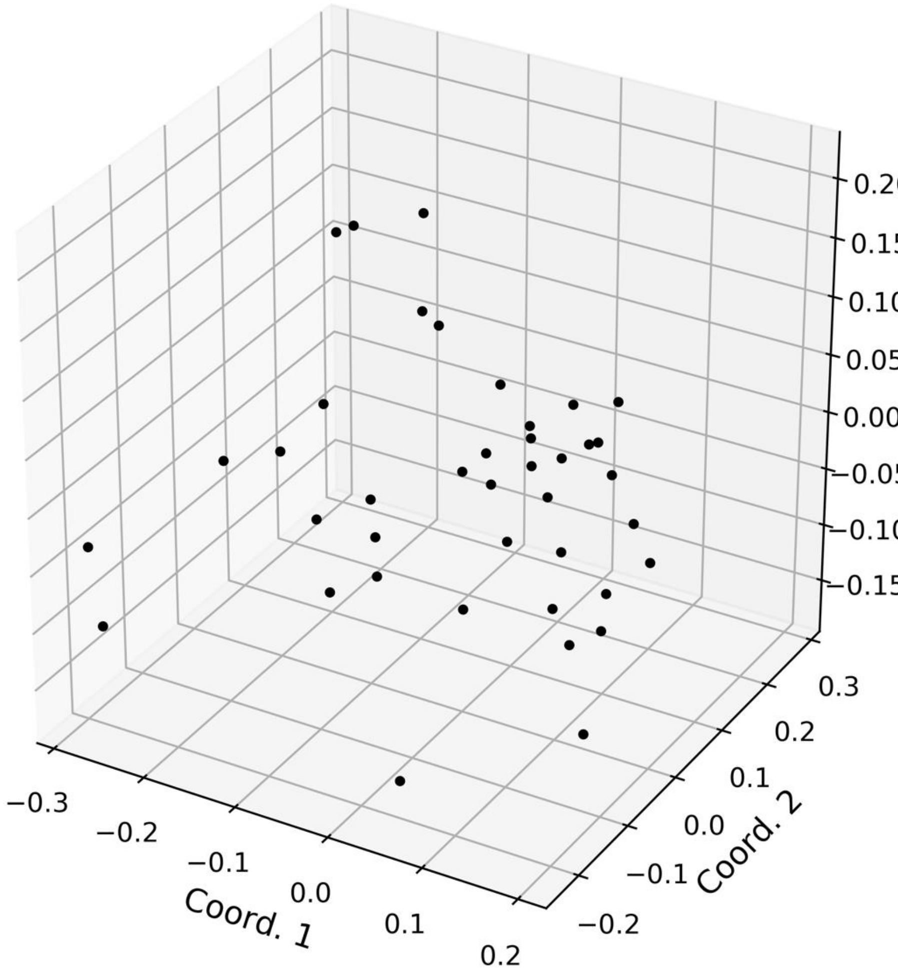

Figure 6. The three-dimensional PCoA plots show the low percentage of the total genetic diversity.

Therefore, it is known that when there is no information regarding the genetic relationship between most genotypes, it is not possible to determine which clustering method is more accurate. Thus, when comparing results from different methods, it is possible to avoid erroneous inferences Cruz, Salgado & Bhering,(2014). The authors points out that the use of more than one clustering method, due to differences in the hierarchical classification, optimization and groups ranking, is essential since it allows the complementation of results, preventing erroneous inferences.

Additionally, the 2D projection of distances technique has the property of representing three distances without any distortion, being a technique of faithful reproduction of the most divergent genotypes. In the case of the 3D projection technique, it can represent six distances without distortion. Because of this, this technique is reliable for the selection of relatives. In a different approach, the principal coordinates analysis is a method to explore and to visualize similarities or dissimilarities in the data in which individual or group differences can be used to show outliers, represent patterns in the data, or separate the accessions into subspecies or populations. On the other hand, the principal component analysis offers an efficient criterion of selection, and once you have found patterns in the data, it can be used to reduce the number of dimensions without much loss of information (Shlens, 2014).

This study presents the limitation of not having examined molecular data. However, further studies will be carried out to indicate the best method to be used for any type of data, regardless of phenotypic or genotypic data. This is important because as verified by numerous studies, both morphological and molecular markers are informative tools to powerfully assess the genetic diversity in a breeding program. For example, Wang et al. (2013) assessed the genetic diversity of 142 sweet sorghum parent lines used in the hybrid breeding program of Heilongjiang Academy of Agricultural Sciences (Harbin, China) based on agronomical traits and simple sequence repeat (SSR) markers and concluded that both tools should be considered simultaneously for the diversity analysis in hybrid breeding programs. Studies have shown that it is better to analyze genetic diversity both with morphological and molecular traits; however, few studies have assessed genetic diversity using morphological traits and molecular markers simultaneously. In the present work, the results were very clear, and thus, it is concluded that the principal coordinates technique is superior for this type of analysis, as shown in Table 1.

Conclusion

In this study three biometric techniques were used for graphical representation of genetic diversity considering experimental data of six different species, it is concluded that the principal coordinate analysis is superior. In spite of that, the behavior is similar for the three techniques in many cases, with few differences. However, it is up to the author to decide which technique best represents the data, since the results obtained by the utilized technique, despite being quite similar, present discrepancies that may be quite relevant. Thus, the peculiarities of the material and the technique used in the study must be considered.

Acknowledgements

The authors thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Fundação de Pesquisa do Estado de Minas Gerais (FAPEMIG), and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support and scholarships.

References

Aitken, K., Li, J., Piperidis, G., Qing, C., Yuanhong, F., & Jackson, P. (2018). Worldwide genetic diversity of the wild species *Saccharum spontaneum* and level of diversity captured within sugarcane breeding programs. *Crop Science*, *58*(1), 218-229. DOI: 10.2135/cropsci2017.06.0339

Assis, G. M. L. D., Euclydes R. F., Cruz, C. D., & Do Valle C. B. (2003). Discriminação de espécies de *Brachiaria* baseada em diferentes grupos de caracteres morfológicos. *Revista Brasileira de Zootecnia*, *32*(3), 576-584.

Choudhury, D. R., Singh, N., Singh, A. K., Kumar, S., Srinivasan, K., Tyagi, R. K., ... Singh, R. (2014). Analysis of genetic diversity and population structure of rice germplasm from North-eastern region of India and development of a core germplasm set. *PLoS ONE*, *9*(11), e113094. DOI: 10.1371/journal.pone.0113094

Cruz, C. D., Salgado, C. C., & Bhering, L. L. (2014). Biometrics applied to molecular analysis in genetic diversity. *Biotechnology and Plant Breeding: Applications and Approaches for Developing Improved Cultivars*, 47-81. DOI: 10.1016/B978-0-12-418672-9.00003-9

Cruz, C. D. (2013). Genes: a software package for analysis in experimental statistics and quantitative genetics. *Acta Scientiarum. Agronomy*, *35*(3), 271-276. DOI: 10.4025/actasciagron.v35i3.21251

Cruz, C. D., Ferreira, F. M., & Pessoni, L. A. (2011). *Biometria aplicada ao estudo da diversidade genética.* Visconde do Rio Branco, MG: Suprema.

El-Esawi, M. A., Witczak, J., Abomohra, A. E. F., Ali, H. M., Elshikh, M. S., & Ahmad, M. (2018). Analysis of the genetic diversity and population structure of Austrian and Belgian wheat germplasm within a regional context based on DArT markers. *Genes*, *9*(1), 47. DOI: 10.3390/genes9010047

Ferreira, J. R., Pereira, J. F., Turchetto, C., Minella, E., Consoli, L., & Delatorre, C. A. (2016). Assessment of genetic diversity in Brazilian barley using SSR markers. *Genetics and Molecular Biology*, *39*(1), 86-96. DOI: 10.1590/1678-4685-GMB-2015-0148

Gomes, C. N., Carvalho, S. P., Jesus, M. A., & Custódio T. N. (2007). Caracterização morfoagronômica e coeficientes de trilha de caracteres componentes da produção em mandioca. *Pesquisa Agropecuária Brasileira, 42*(8),1121-1130.

Gower, J. C., & Hand, D. J. (1996). *Biplots* (Monographs on Statistics and Applied Probability, 54). London, UK: Chapman and Hall.

Guzmán, F. A., Segura, S., Aradhya, M., & Potter, D. (2018). Evaluation of the genetic structure present in natural populations of four subspecies of black cherry (*Prunus serotina* Ehrh.) from North America using SSR markers. *Scientia Horticulturae*, *232*, 206-215. DOI: 10.1016/j.scienta.2018.01.013

Jaccard, P. (1908). Nouvelles researches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles, 44*), 223-270.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1). DOI: 10.1186/1471-2156-11-94

Mohammadi, S. A., & Prasanna, B. M. (2003). Analysis of genetic diversity in crop plants-salient statistical tools and considerations. *Crop Science*, *43*(4), 1235-1248. DOI: 10.2135/cropsci2003.1235.

Müller, T., Schierscher-Viret, B., Fossati, D., Brabant, C., Schori, A., Keller, B., & Krattinger, S. G. (2018). Unlocking the diversity of genebanks: whole-genome marker analysis of Swiss bread wheat and spelt. *Theoretical and Applied Genetics*, *131*(2), 407-416. DOI: 10.1007/s00122-017-3010-5

Murray, S. C., Rooney, W. L., Hamblin, M. T., Mitchell, S. E., & Kresovich, S. (2009). Sweet sorghum genetic diversity and association mapping for brix and height. *The Plant Genome*, *2*(1), 48-62. DOI: 10.3835/plantgenome2008.10.0011

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, *76*(10), 5269-5273. DOI: 10.1073/pnas.76.10.5269

Nick, C., Carvalho, M., Assis, L. H. B., & Carvalho, S. P. (2008). Genetic dissimilarity in cassava clones determined by multivariate techniques. *Crop Breeding and Applied Biotechnology*, *8*(2), 104-110.

Osawaru, M. E., Ogwu, M. C., & Dania-Ogbe, F. M. (2013). Morphological assessment of the genetic variability among 53 accessions of West African Okra [*Abelmoschus caillei* (A. Chev.) Stevels] from South Western Nigeria. *Nigerian Journal of Basic and Applied Sciences*, *21*(3), 227-238. DOI: 10.4314/njbas.v21i3.8

Roy, S., Banerjee, A., Mawkhlieng, B., Misra, A. K., Pattanayak, A., Harish, G. D., ... Bansal, K. C. (2015). Correction: Genetic diversity and population structure in aromatic and quality rice (*Oryza sativa* L.) landraces from North-Eastern India. *PLoS ONE*, *10*(10), e0141405. DOI: 10.1371/journal.pone.0141405

Sant'Anna, I. C., & Cruz, C. D. (2015). *Redes neurais artificiais na discriminação de populações.* Riga, LE: SIA Omni Scriptum Publishing.

Santi, A. L., Amado, T. J. C., Cherubin, M. R., Martin, T. N., Pires, J. L., Della Flora, L. P., & Basso, C. J. (2012). Análise de componentes principais de atributos químicos e físicos do solo limitantes à produtividade de grãos. *Pesquisa Agropecuária Brasileira*, *47*(9), 1346-1357. DOI

Shlens, J. (2014). *A tutorial on principal component analysis - Version 3.02* (arXiv preprint arXiv:1404.1100). Mountain View, CA: Google Research.

Su, W., Wang, L., Lei, J., Chai, S., Liu, Y., Yang, Y., ... Jiao, C. (2017). Genome-wide assessment of population structure and genetic diversity and development of a core germplasm set for sweet potato based on specific length amplified fragment (SLAF) sequencing. *PLoS ONE*, *12*(2), e0172066. DOI: 10.1371/journal.pone.0172066

Tagliotti, M. E., Deperi, S. I., Bedogni, M. C., Zhang, R., Carpintero, N. C. M., Coombs, J., ... Huarte, M. A. (2018). Use of easy measurable phenotypic traits as a complementary approach to evaluate the population structure and diversity in a high heterozygous panel of tetraploid clones and cultivars. *BMC Genetics*, *19*(1), 1-12. DOI: 10.1186/s12863-017-0556-9

R Core Team (2018). *R*: *A language and environment for statistical computing*. Vienna, AU: R Foundation for Statistical Computing.

Valcárcel, J. V., Peiró, R. M., Pérez-de-Castro, A., & Díez, M. J. (2018). Morphological characterization of the cucumber (*Cucumis sativus* L.) collection of the COMAV's Genebank. *Genetic Resources and Crop Evolution, 65*(4), 1293-1306. DOI: 10.1007/s10722-018-0614-9

Van Inghelandt, D., Melchinger, A. E., Lebreton, C., & Stich, B. (2010). Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theoretical and Applied Genetics*, *120*(7), 1289-1299. DOI: 10.1007/s00122-009-1256-2

Zaher, H., Boulouha, B., Baaziz, M., Sikaoui, L., Gaboun, F., & Udupa, S. M. (2011). Morphological and genetic diversity in olive (*Olea europaea* subsp. *europaea* L.) clones and varieties. *Plant Omics*, *4*(7): 370-376. DOI: 10.1007/s00122-003-1350-9

Wang, L., Jiao, S., Jiang, Y., Yan, H., Su, D., Sun, G., ... Sun. L. (2013). Genetic diversity in parent lines of sweet sorghum based on agronomical traits and SSR markers. *Field Crop Research*, *149*, 11-19. DOI 10.1016/j.fcr.2013.04.013

Wu, G. A., Terol, J., Ibanez, V., López-García, A., Pérez-Román, E., … Curk, F., 2018. Genomics of the origin and evolution of Citrus. *Nature*, *554*(7692), 311. DOI: 10.1038/nature254