Scientia Agricola
http://dx.doi.org/10.1590/0103-9016-2014-0043

Genotypic variation of traits related to quality of cassava roots using affinity propagation algorithm

Eder Jorge de Oliveira[1]*, Fernanda Alves Santana[2], Luciana Alves de Oliveira[1], Vanderlei da Silva Santos[1]

[1]Embrapa Cassava and Fruits, R. Embrapa, s/n, C.P. 007 – 44380-000 – Cruz das Almas, BA – Brazil.
[2]Federal University of Bahia Reconcavo, R. Rui Barbosa, 710, Centro – 44380-000 – Cruz das Almas, BA – Brazil.
*Corresponding author <eder.oliveira@embrapa.br>

Edited by: Antonio Augusto Franco Garcia

Received February 02, 2014
Accepted July 07, 2014

ABSTRACT: The conservation, sustainable evaluation and use of cassava (*Manihot esculenta* Crantz) genetic resources are essential to the development of new commercial varieties. This study aimed to evaluate the quality of cassava roots and to estimate genetic variation and clustering in cassava germplasm using the Affinity Propagation algorithm (AP), which is based on the concept of "message passing" between data points. AP finds "exemplars" of each group and members of the input set representative of clusters. The genotypic data of 474 cassava accessions were evaluated over a period of two years for starch yield (StYi), root dry matter (DMC), amylose content (AML), and the level of cyanogenic compounds (CyC). The AP algorithm enabled the formation of nine diversity groups, whose number reflects the high genetic diversity of this germplasm. A high homogeneity of genetic distances was observed within all the groups, except for two groups in which there was a partial overlap caused mainly by a high variation of the CyC trait. In addition, no relationship between the genetic structure and CyC (sweet and bitter cassava) was observed. Analysis of variance of the nine clusters confirmed the presence of differences between the groups. Thus, the results of this study can be used in future breeding programs (hybridization or selection) to introduce new genetic variability into commercial cultivars to avoid problems related to low genetic variation and to improve the quality of cassava roots.

Keywords: germplasm, amylose content, cyanogenic compounds, yield

Introduction

Cassava (*Manihot esculenta* Crantz) is a crop that is widely cultivated in many tropical countries in Africa, Latin America and Asia (Ceballos et al., 2007). Being adapted to low soil fertility and marginal, irregular rain conditions and having relatively stable productivity and flexibility in harvesting, cassava has great potential both as a secure source of food as well as a tool for reducing poverty due to its use not only in traditional agriculture but also in high productivity industrial production systems that are highly technical.

Cassava is from South America, more specifically from the Upper Amazon Basin (Olsen, 2004). Therefore, most of the genetic diversity currently used worldwide comes from Brazil. Cassava germplasm has several traits potentially useful to commercial varieties, such as differential starch characteristics (Ceballos et al., 2007), disease resistance (Raji et al., 2007) and root quality (Chávez et al., 2005). In general, breeders believe that most cassava varieties have low root productivity and low competitiveness as compared with improved varieties. However, this belief must be examined via an intensive evaluation of the data on cassava germplasm regarding its various agronomic and economic interest attributes.

As these data are not fully available, the use of the genetic resources of *M. esculenta* in cassava breeding programs has been limited. Furthermore, while some progress has been achieved in traditional evaluations of agronomic, morphological and molecular traits (Benesi et al., 2010; Duraisamy et al., 2011), these evaluations do not necessarily reflect the diversity associated with the quality of cassava roots that is basic to their use, whether for *in natura* consumption or industrial purposes. One of the factors that has hindered the development of new cassava varieties is the lack of information about root quality as well as details of the amount of genetic diversity in Brazilian germplasm.

Considering that both the evaluation of the genetic resources of cassava in Brazil and the precise evaluation of genetic variation are crucial in the development of optimal management strategies for sustainable conservation and for using germplasm to generate new varieties, this study aimed to evaluate the quality of the roots of germplasm accessions and to estimate the genetic variation to be used in cassava breeding programs.

Materials and Methods

Plant material

Four-hundred and seventy four germplasm accessions belonging to the Cassava Germplasm Bank (CGB) from Embrapa Cassava and Fruits (Cruz das Almas, in the state of Bahia, Brazil, 12° 48' S; 36° 06' W; 225 m a.s.l.), originating from several ecosystems in Brazil, Colombia, Venezuela and Nigeria were evaluated. This database consists of landraces and improved varieties resulting from conventional breeding procedures, such as crossing, and a selection of landraces identified by farmers or research institutions.

Experimental design

Two field trials were carried out over two years (2011 and 2012) in Cruz das Almas. A randomized block design with three replications and ten plants per plot was used in 2011, and an augmented block design was used in 2012, with the number of accessions evenly distributed in ten blocks with ten plants per plot. As experimental checks, not-yet-recommended improved clones (9624-09, 98150-06, and 9824-09) as well as landraces (Cigana and Eucalipto) and recommended varieties (BRS Aipim Brasil, BRS Dourada, BRS Tapioqueira, BRS Caipira, BRS Verdinha and BRS Gema de Ovo) were used.

The planting was carried out at the beginning of the region's rainy season (May-July) using 15-20-cm stem cuttings in a single row. The spacing was 0.9 m between rows and 0.8 m between plants, and all the recommended cultivation practices for cassava were followed. The plants were harvested 11 months after planting.

Traits evaluated

Dry matter content of the roots (DMC): this was obtained using 100-g root samples previously washed under running water, cut into pieces, peeled and cut into quarters, which were crushed to form a homogeneous mass. Then, the samples were dried in an oven with forced air circulation at 60 °C for 48 h until constant weight was achieved.

Starch yield (StYi): this was obtained by subtracting 4.65 % from the DMC, which corresponds to the ash content, protein content, and lipids and fibers. Next, the starch content was multiplied by the average yield of the accession to obtain the *StYi* in t ha^{-1}.

Amylose content (AML): the starch extraction was performed manually using a 500-g root sample cut into pieces and ground in a blender with a non-cutting helix (1:1 ratio water / root) and then filtered through a 150-mesh sieve. The starch suspension was kept in a cold chamber at 5 °C for 12 h. The supernatant was then discarded, and the decanted starch was washed with 95 % ethanol and dried in an oven with forced air circulation at 40 °C for 48 h. The dried starch was analyzed with regard to its amylose content according to ISO (2005) protocol. The starch sample was dispersed in 95 % ethanol gelatinized with sodium hydroxide and acidified with acetic acid. After the addition of the iodine solution, the blue complex formed was quantitated by spectrophotometry at 620 nm (Biospectro, model SP 220).

Cyanogenic compounds (CyC): The determination of CyC, especially free cyanide, α-hydroxynitrile and cyanogenic glycosides, present in the samples was performed by an extraction of these compounds followed by a subsequent reaction with chloramine-T and Isonicotinate/1,3 dimethyl barbiturate and spectrophotometric determination at 605 nm. Linamarase enzyme was used, which is extracted from the bark of the cassava cortex according to Cooke (1979), to release the glycosidic cyanide.

Estimation of genotypic values

A combined analysis of different experimental designs was carried out, using statistical models for incorporating block designs. In this case, all the blocks were set to represent random effects, while the design effects were treated as fixed by adjusting the experiments to a randomized complete block, and at another level for the experiments, to an incomplete block design. In this case, the model set the block effects within each design type.

The linear mixed model that was used to describe the data was $y = Xb + Zg + Wp + e$ (Henderson, 1984), where y is the data vector, b is the vector of fixed effects associated to the mean and block effect, g is the vector of random genetic effects, p is the vector of the random effects of the plots, e is the vector of random errors, and X, Z and W are the incidence matrices that are associated with the unknown parameters b, g and p, respectively, to the y data vector.

The mixed model methodology allows for the estimation of b using a generalized least squares procedure and for g and p using BLUP (Best Linear Unbiased Prediction), which predicts the random genetic effects and uncorrelated random effects not included in the model (Henderson, 1984).

Once the quality traits in cassava roots had been correlated with each other and their statistical dependence taken into account when analyzing multivariate data, Pearson's correlation coefficients among those traits were estimated.

Genetic diversity and clustering

The cassava accessions' genotypic values, obtained by BLUP, were used to calculate the genetic distance matrix using the n*egDistMat* () function of the APCluster package (Bodenhofer et al., 2011) from the R program version 3.0.1 (R Development Core Team, Vienna, AT). The negative Euclidean distances were calculated, in which the *negDistMat* () function provides the following variants to compute the distance $d(x; y)$ between two accessions .$x = (x_1; . . .; x_n)$ and $y = (y_1; . . .; y_n)$, where:

$$d(x, y) = -\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The affinity propagation method (AP) was used to promote the clustering of cassava accessions. The data were subjected to 100 independent runs to verify the consistency of selecting the exemplars and clustering analysis. This procedure identifies a set of centers (exemplars) from the data set, taking into account each accession as a network node, and recursively transmits real-valued messages along the edges of the network until a good set of exemplars and their corresponding cluster emerges. At any time, the magnitude of each message reflects the current affinity that a particular accession has to be chosen as a new exemplar of the group (Frey and Dueck, 2007).

The messages exchanged between data points can be from two types: "responsibility" $r(i,k)$ and "availability" $a(i,k)$ (Sakellariou et al., 2012). "Responsibility" reflects the accumulated evidence of how appropriate point k is to serve as an exemplar for point i, considering other potential exemplars for this same point. In contrast, "availability" reflects the accumulated evidence of how appropriate it would be for point i to choose point k as its exemplar, taking into account the other points at which point k may be an exemplar. Initially, availabilities are set to zero. The AP was implemented as follows:

$$r(i,k) \leftarrow s(i,k) - \max_{k':k' \neq k} \{a(i,k') + s(i,k')\}$$

$$a(i,k) \leftarrow \min\left\{0, r(k,k) + \sum_{i':i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\}$$

where the matrix $s(i,k)$ is the similarity between the two nodes i and k, and the diagonal of this matrix represents the preferences for each node. The AP algorithm is iterated until a good set of exemplars emerges from the equation above. Each node i can then be assigned to the exemplar k which maximizes the sum $a(i,k) + r(i,k)$, and if $i = k$, then i is an exemplar. AP can be applied to problems where the similarities are neither symmetric nor do they satisfy triangle inequality (Frey and Dueck, 2007).

The k-means method that has similar properties for grouping data (partition analysis) was used to compare the relative stability of clusters obtained by the AP method. The genotypic values of cassava accessions were used to obtain 100 independent runs for clustering using the *k-means* () function from the R program version 3.0.1 (R Development Core Team, Vienna, AT). A plot of the total within-groups sums of squares against the number of clusters was used to find the best number of cluster.

Results and Discussion

Trait correlation

Pearson's correlations among the quality traits of cassava roots were overall low to intermediate (Table 1). The correlations among cassava germplasm for the traits AML, DMC, CyC and StYi indicated that selection of populations expressing multiple desirable quality traits is possible by employing this germplasm. In contrast, the demands for cassava roots (for food or industrial purposes) have rapidly increased in recent times, and high AML is a new trait that should be improved by cassava breeders. Therefore, the important negative correlation between DMC and AML (-0.41) could be an issue once the ideal cassava had the highest AML and DMC levels. Even with low to intermediate correlations, the statistical dependence of these quality traits makes multivariate analysis an appropriate approach for analyzing the data.

Table 1 – Pearson's correlations among dry matter content (DMC), starch yield (StYi), amylose content (AML) and cyanogenic compounds (CyC) evaluated in 474 cassava accessions.

	AML	DMC	CyC
DMC	-0.41		
CyC	-0.16	0.04	
StYi	0.03	0.12	-0.11

Number of groups

Unlike other clustering methods, the strategy implemented by AP does not require a prior estimate of the number of groups. Instead, the AP method defines the number of exemplars in the analysis that are representative of the sample. AP has as input a similarity value $s(i,k)$ for each data point k, where the data with the highest $s(i,k)$ values are selected as the cluster exemplar. These values are known as "preferences". In this case, the number of exemplars identified, which refers to the number of groups, was influenced by both the genetic distance input and the message-passing procedure (Frey and Dueck, 2007).

Considering that it was possible to deduce the definition of groups based on the input value of "preference", the formation of nine diversity groups was observed when using a genetic input distance of 0.35, which was calculated on the basis of the genotypic data from the four quality traits from cassava roots analyzed in this study (Figure 1). Table 2 shows the classification of the 474 cassava accessions according to their cluster. In contrast, the K-means clustering suggested that the 6-cluster solution was a good fit to the data based on the bend in the graph related to total within-groups sums of squares against the number of clusters (data not shown). This difference in the indication of the optimal number of clusters between the methods is due to AP's tendency to discover the more representative objects in the dataset, once it performs clustering without additional parameters.

The determination of the optimal or acceptable number of groups is an important factor in determining the reliability of the groups. This process basically involves the establishment of the criteria to be used to separate groups of two or more accessions whose genetic distance must be lower within the groups when compared with the overall average distance and whose distance between groups is greater than the distance within the groups involved in the analysis (Brown-Guedira et al., 2000).

In general, the AP method allowed for the creation of a large number of groups, which certainly reflects the high genetic diversity of cassava germplasm in Brazil. In principle, this statement could contradict what would be expected from a species that is predominantly vegetatively or asexually propagated, in which case, a great reduction in genetic diversity would be expected over time, due to the accumulation of systemic pathogens and preferences for more vigorous and adapted varieties

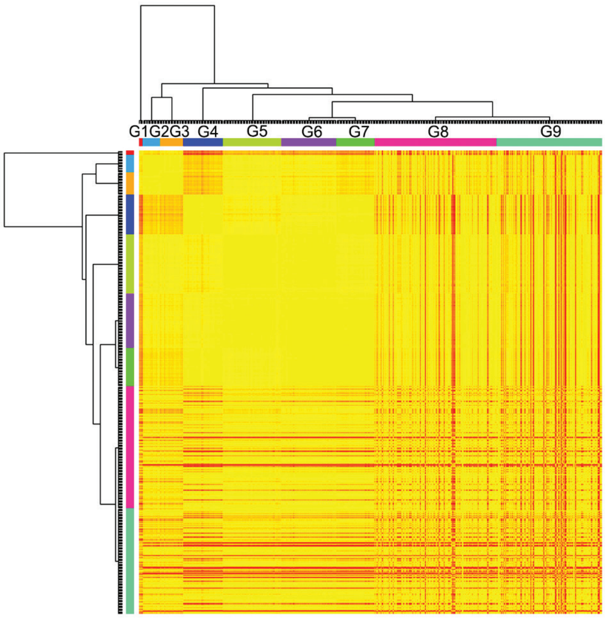

Figure 1 – Heatmap and hierarchical clustering of the genetic distance on 474 cassava accessions based on quality-related traits of the roots. Yellow color indicates a low similarity between accessions, while orange indicates a high similarity.

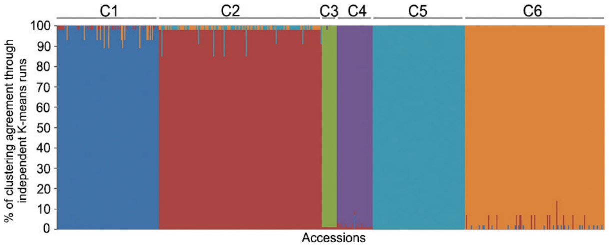

Figure 2 – Distribution of cassava accessions in six clusters formed by the K-means method and percentage of clustering agreement for each accession in 100 K-means analyzes. The predominant colors represent the differences among clusters.

with a higher capacity to produce stem cuttings. Moreover, throughout the evolution of cassava, its frequent outcrossings allowed for the production of a large group of spontaneous seeds, which, under natural and directed selection by conventional farmers, were selected to produce new varieties and have thereby maintained a high level of genetic diversity, which is still maintained in many traditional communities. In fact, considering the high levels of outcrossing and sexual propagation, cassava accessions are predominantly maintained by vegetative propagation, allowing for high heterozygosity and extensive plasticity in the expression of phenotypic characteristics, such as those relating to the quality of the roots that have been observed in the present study.

Comparison between K-means and AP

AP and K-means analyses were carried out to determine the performance of both algorithms in terms of effectiveness and accuracy when analyzing agronomic data from the quality of cassava root. Considering six diversity groups by the K-means method, the groups formed are inconsistent for 16 % of the independent analysis, i.e., cassava accessions were incorrectly allocated into different clusters (Figure 2). Conversely, for all analyzes performed using the AP algorithm, the selection was carried out using the same individuals as exemplars, as well as the same accessions in the nine different clusters.

A hypothesis to explain this inconsistency in the cluster analysis may be the fact that K-means revealed a weakness in the determination of the initial exemplar (center point), which is obtained randomly. Therefore, if the initial exemplar is not appropriate, then the cluster will be not be maximal. Consequently, the K-means algorithm could show a high error rate and may not provide the best cluster results (Refianti et al., 2012). In contrast, the AP simultaneously considers all accessions as possible exemplars (center points) where each message is sent to reflect the latest interest held by each data point to be able to select another data point as their exemplar. Through this message-passing process, an algorithm tests the possibility of all data to become the center of the cluster, and then each accession has the same opportunity to become an exemplar.

Frey and Dueck (2007) compared the reconstruction errors for the AP and K-center clustering for each number of clusters, aiming to detect putative exons comprising genes from mouse chromosome 1. The AP method achieved higher true positive rates (TPR), especially at low false-positive rates (FPR), with regard to addressing the question of how well these methods perform in detecting bona fide gene segments. According to Frey and Dueck (2007), at an FPR rate of 3 %, AP achieved a TPR rate of 39 %, whereas the best K-center clustering result was 17 %.

Considering that: i) the lack of knowledge of the true genetic structure of cassava accessions evaluated on the basis of root quality traits, ii) the presence of strong inconsistencies in the clusters formed by the K-means method, which has been supported by other authors, and iii) the low reliability and repeatability of the clusters formed by the K-means method, the genetic diversity analysis was performed based on the AP algorithm only.

Genetic diversity grouping from the AP method

According to the hierarchical clustering (Figure 1), there was great homogeneity in the genetic distances within all the groups, except for Groups 8 and 9. In the latter two groups, there was partial overlap. The hierarchical clustering using a heatmap form was based on the matrix of the genetic distance (negative Euclidean distance) (Figure 1). Therefore, the heatmap classified the accessions, taking into account the distance profile of

Table 2 – List of cassava accessions belonging to the nine clusters formed by affinity propagation analysis.

Cluster	Accessions
1	BGM0045, BGM0082, BGM0083, BGM0128, BGM0204, BGM0211, BGM0248, BGM0250, BGM0271, BGM0273, BGM0279, BGM0295, BGM0312, BGM0319, BGM0331, BGM0356, BGM0360, BGM0509, BGM0583, BGM0624, BGM0631, BGM0791, BGM0847, BGM0885, BGM0887, BGM0928, BGM0930, BGM0947, BGM1024, BGM1053, BGM1123, BGM1131, BGM1354, BGM1452, BGM1608, BGM1650, BGM1667, BGM1669, BGM1672, BGM1719 and BRS Dourada
2	BGM0010, 2BGM0120, BGM0165, BGM0337, BGM0426, BGM0544, BGM0823, BGM0892, BGM0937, BGM1027, BGM1085, BGM1165, BGM1282, BGM1387, BGM1393, BGM1598, BGM1624 and BGM1626
3	BRS Tapioqueira, BRS Caipira, Aipim Brasil, BGM0075, BGM0207, BGM0214, BGM0263, BGM0303, BGM0338, BGM0409, BGM0440, BGM0447, BGM0477, BGM0550, BGM0576, BGM0598, BGM0601, BGM0636, BGM0649, BGM0807, BGM0846, BGM0856, BGM0894, BGM1026, BGM1067, BGM1108, BGM1203, BGM1257, BGM1281, BGM1347, BGM1366, BGM1398, BGM1406, BGM1409, BGM1410, BGM1432, BGM1447, BGM1481, BGM1494, BGM1510, BGM1550, BGM1552, BGM1567, BGM1581, BGM1622, BGM1636, BGM1640, BGM1643, BGM1713, BGM1716, BGM1721, BGM1771, BGM1816, BGM1850, BGM1969, BGM2019,
4	BGM0104, BGM0175, BGM0184, BGM0249, BGM0361, BGM0365, BGM0382, BGM0547, BGM0567, BGM0667, BGM0767, BGM0837, BGM0878, BGM1124, BGM1183, BGM1191, BGM1484, BGM1523, BGM1645, BGM1681, BGM1942, MandimBM and MandimBP
5	BRS Verdinha, AmarelaSC, BGM0023, BGM0025, BGM0057, BGM0093, BGM0097, BGM0103, BGM0160, BGM0188, BGM0255, BGM0316, BGM0378, BGM0390, BGM0394, BGM0419, BGM0473, BGM0540, BGM0574, BGM0579, BGM0778, BGM0822, BGM1037, BGM1118, BGM1125, BGM1186, BGM1200, BGM1245, BGM1273, BGM1345, BGM1348, BGM1370, BGM1371, BGM1403, BGM1408, BGM1417, BGM1420, BGM1508, BGM1524, BGM1533, BGM1534, BGM1547, BGM1559, BGM1595, BGM1623, BGM1678, BGM1701, BGM1707, BGM1723, BGM1726, BGM1729, BGM1730, BGM1811, BGM1867, BGM1876, BGM1884, BGM2018, BGM2020, FilhaPreta and PretaSC,
6	BGM0660, BGM1184, BGM119 and BGM1957
7	BGM0016, BGM0051, BGM0145, BGM0155, BGM0166, BGM0179, BGM0182, BGM0201, BGM0297, BGM0445, BGM0456, BGM0499, BGM0575, BGM0600, BGM0932, BGM1206, BGM1211, BGM1218, BGM1222, BGM1236, BGM1270, BGM1287, BGM1291, BGM1321, BGM1487, BGM1489, BGM1513, BGM1515, BGM1569, BGM1579, BGM1607, BGM1610, BGM1621, BGM1632, BGM1648, BGM1666, BGM1685, BGM1814 and BGM1865
8	9624-09, 98150-06, 9824-09, BGM0011, BGM0018, BGM0027, BGM0033, BGM0039, BGM0041, BGM0049, BGM0116, BGM0127, BGM0135, BGM0136, BGM0144, BGM0146, BGM0148, BGM0149, BGM0162, BGM0163, BGM0189, BGM0196, BGM0236, BGM0254, BGM0276, BGM0330, BGM0332, BGM0341, BGM0343, BGM0399, BGM0405, BGM0406, BGM0408, BGM0425, BGM0428, BGM0432, BGM0433, BGM0434, BGM0436, BGM0452, BGM0465, BGM0469, BGM0472, BGM0492, BGM0501, BGM0539, BGM0543, BGM0552, BGM0556, BGM0562, BGM0587, BGM0640, BGM0678, BGM0702, BGM0728, BGM0800, BGM0810, BGM0815, BGM0820, BGM0824, BGM0845, BGM0868, BGM0873, BGM0876, BGM0877, BGM0890, BGM0895, BGM0901, BGM0924, BGM1010, BGM1042, BGM1043, BGM1050, BGM1057, BGM1073, BGM1078, BGM1100, BGM1163, BGM1174, BGM1175, BGM1177, BGM1178, BGM1179, BGM1189, BGM1202, BGM1219, BGM1277, BGM1283, BGM1333, BGM1361, BGM1362, BGM1416, BGM1423, BGM1440, BGM1476, BGM1495, BGM1518, BGM1526, BGM1549, BGM1560, BGM1562, BGM1589, BGM1590, BGM1597, BGM1602, BGM1616, BGM1619, BGM1620, BGM1671, BGM1677, BGM1682, BGM1691, BGM1692, BGM1693, BGM1696, BGM1736, BGM1741, BGM1817, BGM1824, BGM1956, BGM2038, BGM2042, BGM2043, BGM2044 and Cigana
9	BGM0089, BGM0098, BGM0143, BGM0153, BGM0164, BGM0177, BGM0217, BGM0298, BGM0307, BGM0324, BGM0367, BGM0375, BGM0376, BGM0384, BGM0443, BGM0455, BGM0464, BGM0471, BGM0497, BGM0532, BGM0541, BGM0542, BGM0546, BGM0590, BGM0591, BGM0611, BGM0620, BGM0623, BGM0626, BGM0630, BGM0638, BGM0642, BGM0656, BGM0677, BGM0706, BGM0733, BGM0745, BGM0752, BGM0776, BGM0783, BGM0785, BGM0788, BGM0817, BGM0818, BGM0821, BGM0826, BGM0834, BGM0867, BGM0889, BGM0896, BGM0908, BGM0940, BGM0943, BGM0956, BGM0991, BGM1012, BGM1077, BGM1116, BGM1122, BGM1136, BGM1137, BGM1138, BGM1153, BGM1171, BGM1185, BGM1193, BGM1194, BGM1195, BGM1197, BGM1198, BGM1209, BGM1210, BGM1311, BGM1342, BGM1357, BGM1359, BGM1363, BGM1367, BGM1407, BGM1413, BGM1418, BGM1450, BGM1457, BGM1464, BGM1477, BGM1482, BGM1490, BGM1491, BGM1517, BGM1535, BGM1539, BGM1546, BGM1561, BGM1576, BGM1586, BGM1593, BGM1611, BGM1614, BGM1617, BGM1702, BGM1750, BGM1751, BGM1767, BGM1822, BGM1830, BGM1832, BRS Gema de Ovo and Eucalipto

each accession in relation to the others, where the reddish color refers to a larger divergence. Group 1 consisted of 40 germplasm accessions (mostly landraces) and the sweet variety, BRS Dourada. According to the boxplot of this group (Figure 3), the most striking traits of this group were its low level of cyanogenic compounds, moderate starch yield and dry matter content.

In a study conducted in Nigeria with landraces and improved varieties for 18 traits (agronomic and quality of cassava roots), Raji et al. (2007) observed that landraces showed better root quality and superior agronomic characteristics compared to improved varieties as well as lower cyanogenic compounds, with 40.0 mg kg^{-1} being found for the landrace 'Isunikankiyan' and 128.6 mg kg^{-1} for an improved cultivar (TMC30572). In general these values were higher than those reported in this study; however, they confirm that the preference for cassava with a wide range of cyanogenic compounds is quite common, particularly among small farmers, despite a general preference for sweet varieties.

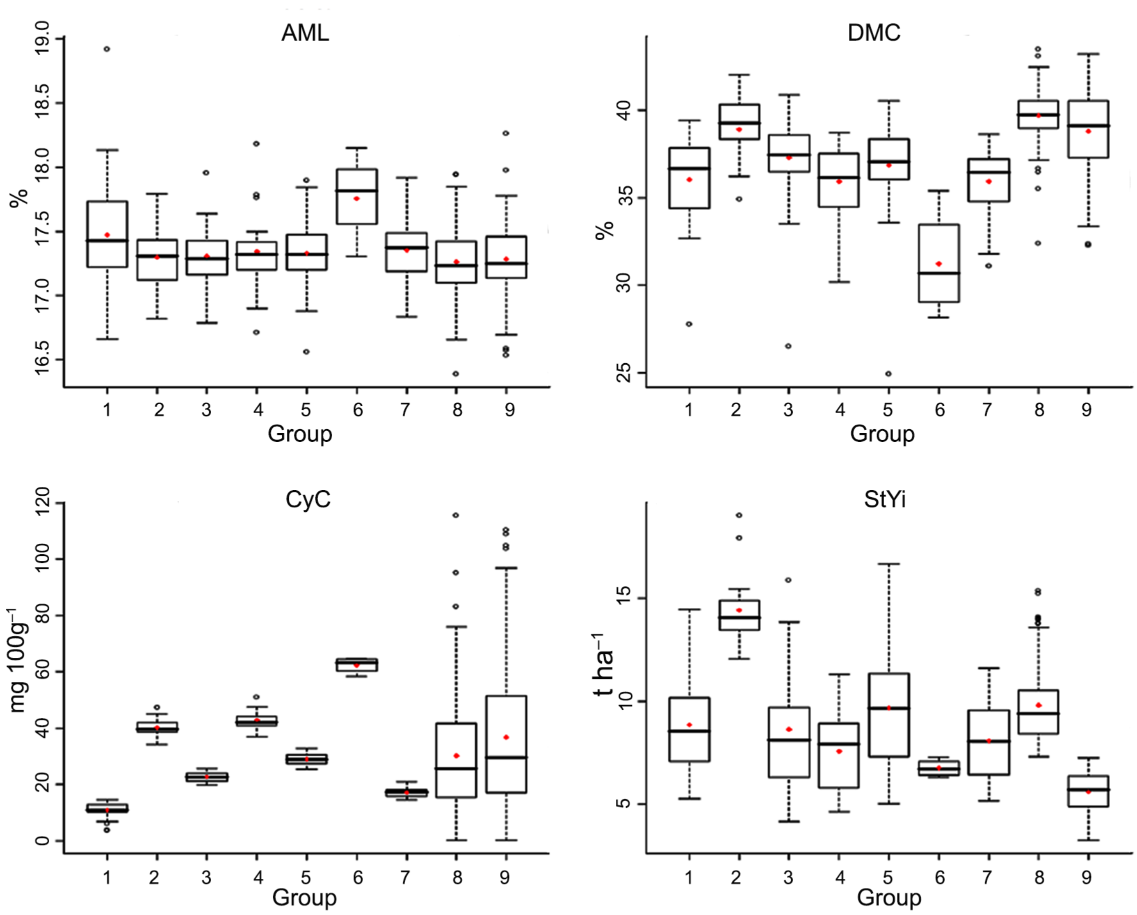

Figure 3 – Boxplot analysis of traits: amylose (AML), dry matter content (DMC), cyanogenic compounds (CyC) and starch productivity (StYi) among the nine clusters based on affinity propagation analysis.

In Group 2, 18 accessions of the germplasm were grouped and their main distinguishing traits were high starch yield and dry matter content. Moreover, a concentration of cyanogenic compounds below 40 mg kg^{-1} was observed. Therefore, in addition to being of industrial interest due to their high starch yield and dry matter content, there is also the possibility of using germplasm accessions from this group for use as either parental varieties or directly for fresh consumption because of the low level of cyanogenic compounds.

Starch yield is a crucial point to be considered when choosing the variety to be used by a large scale production system. Because of its use in the starch industry, it is necessary to maximize production per hectare efficiency. Therefore, it is important to consider the productive potential of the variety and its starch content, which can be translated into starch yield. If we consider the Brazilian national average yield to be approximately 13.0 t ha^{-1} (IBGE, 2013), which is associated with an average dry matter content of 35 %, it results in an average yield of 4.5 t ha^{-1} of starch. Therefore, the use of the germplasm accessions of Group 2 may contribute to an increased production of more than 3 times the current Brazilian national average starch yield.

Group 3 consisted of 53 germplasm accessions together with the improved varieties BRS Tapioqueira, BRS Caipira and BRS Aipim Brazil, while Group 4 included 23 germplasm accessions, and Group 5, 59 germplasm accessions plus the variety BRS Verdinha. The close relationship found between BRS Tapioqueira and BRS Caipira was based on the origin of these genotypes, which share a common parentage. In general, it was observed that the traits which distinguished these groups the most were the dry matter content, cyanogenic compounds and starch yield (Figure 3). Additionally, the amylose content of these groups was quite homogeneous.

Group 6 consisted of only four germplasm accessions and each is characterized by their higher content of amylose and cyanogenic compounds (above 60 mg kg^{-1}) and lower dry matter content and starch yield. Recently, breeding programs have focused not only on developing cultivars with high yield and phenotypic stability but also on adding root qualities that best meet the different needs of the starch industry. In this regard, the function-al properties of starch (viscosity, solubility, and swelling index) change according to its amylose content. Cassava starch, composed exclusively of amylopectin, has several advantages for commercial purposes, like waxy maize starch (Ceballos et al., 2007). In contrast to maize, cassava starch with its high amylose content is important in the food industry for the development of products with lower digestibility. In general, the amylose content in cassava roots ranges from 13.6 to 25.0 % (Rickard et al., 1991; Defloor et al., 1998; Moorthy, 2004; Ceballos et al., 2007). In the present study, the genotypic values of the amylose content ranged from 16.39 to 18.92 %, and this variation was from17.30 to 18.15 % in Group 6 (Figure 3). Therefore, the genotypes in Group 6 can be used in crosses to increase the amylose content in recurrent selection programs.

Group 7 consisted of 39 germplasm accessions characterized by the grouping of individuals with medium amylose content, dry matter and starch yield, but with low cyanogenic compound content. In contrast, the largest group (Group 8) was composed of 121 germplasm accessions, along with clones 9624-09, 98150-06, 9824-09 and Cigana landrace. A highlighted trait of this group was the elevated content of dry matter in the roots, reaching almost 44 % in some accessions (Figure 3). However, even with elevated amounts of dry matter, the average starch yield for this group was considered to be medium, which is due possibly, to the lower average yield of fresh roots. In addition, there was high variation in the cyanogenic compounds, which certainly was not a predominant trait for the grouping of these accessions.

Group 9 consisted of 106 germplasm accessions, including one improved variety (BRS Gema de Ovo) and one landrace (Eucalipto). The most noticeable trait of this group was the high average of dry matter content and the higher range for this variable in addition to a wide range of cyanogenic compounds and lower starch yield.

According to Jansz and Uluwaduge (1997), based on the cyanogenic compounds in cassava roots, cassava can be divided into three classes: low toxicity or sweet cassava (< 50 mg kg^{-1}), medium toxicity (between 50 and 100 mg kg^{-1}), and high toxicity or bitter cassava (> 100 mg kg^{-1}). Thus, both Groups 8 and 9 clustered accessions were classified into the three toxicity classes mentioned above, and there was no relation between the genetic structure and the cyanogenic compounds (sweet and bitter cassava). Overall, these results confirm the observations of other authors indicating that this relationship is weak because of the polygenic nature of this trait (Benesi et al., 2010).

Group Significance

The analysis of variance of the nine groups identified by the AP method indicated the presence of at least one difference between the groups ($p < 0.001$) for all four traits evaluated (Table 3), confirming the data from the boxplot (Figure 3). Furthermore, a multivariate analysis of variance (MANOVA) of the nine groups of cassava germplasm diversity against the four quantitative variables produced a Wilks' lambda mean of 0.22 ($F_{32,\ 1619} = 25.00$, $p < 0.001$). Therefore, considering the quality traits of the root, the nine groups identified by AP cluster analysis are consistently different.

This information has important implications for the conservation of germplasm collections, since the limited financial resources invested in most gene banks' routine activities make the curators prioritize the activities of germplasm characterization and evaluation. In such cases, evaluations can therefore be conducted in accessions from different groups. Moreover, cassava germplasm classification based on the quality of the roots may contribute to the selection of accessions to be used in cassava breeding programs, especially by optimizing opportunities for transgressive segregation from crosses between genotypes belonging to different groups with wide divergence, wherein there is a greater likelihood that the unrelated genotypes belonging to different clusters may contribute unique and desirable alleles from different loci (Beer et al., 1993).

Considerations for breeding

With the increasing number of improved genotypes and germplasm accessions used in cassava breeding programs, the use of ordering algorithms and the classification of genetic variability have gained prominence in the pre-breeding actions or in the parental preparation for use in crosses. In general, the use of multivariate algorithms that allows for the simultaneous analysis of multiple agronomic characteristics, regardless of the data set (morphological, agronomic, biochemical or molecular), is widely employed for germplasm classification, ordering the genetic variability for a large number of accessions, or for the analysis of the genetic relationship between improved genotypes (Mohammadi and Prasanna et al., 2003).

Among the various multivariate algorithms, cluster analysis, principal component analysis (PCA), principal coordinates analysis (PCoA), and multidimensional scaling (MDS) are commonly employed and seem to be particularly useful in plants (Mohammadi and Prasanna et

Table 3 – Analysis of variance of the nine groups of genetic diversity based on the evaluation of root quality traits in 474 cassava germplasm and varieties.

Sources of variation	DF	Mean square			
		AML	DMC	CyC	StYi
Cluster	8	0.26**	144.66**	4865.94**	235.68**
Error	465	0.08	4.07	305.21	4.07

DF: degree of freedom, **significant ($p < 0.01$), AML: amylose content, DMC: dry matter content of the roots, CyC: Cyanogenic compounds, StYi: starch yield.

al., 2003). Moreover, despite being relatively unknown in plant breeding, perhaps because of its fairly recent development, the AP analysis method has properties for dealing with multiple traits that are very interesting.

The high potential of the AP algorithm for data clustering has been demonstrated in a number of areas of knowledge, from human face image analysis to gene expression in many organisms (Frey and Dueck, 2007; Sumedha and Weigt, 2007; Borile et al., 2011), including plants (Kiddle et al., 2010). In general, the AP algorithm effectively reveals the hierarchical grouping structure present in the various types of data sets.

Unlike other clustering methods such as K-means, the choice of initial exemplar is not a step that undermines the groups when using the AP method because all accessions are potential exemplars to be tested. Therefore, the AP algorithm tends to produce a low error rate as compared with the K-means, as a consequence of their high robustness and invariance for the assignment of accessions within each cluster (Refianti et al., 2012).

In cassava, the information obtained from the AP cluster analysis indicated a wide variation, especially for traits such as cyanogenic compounds, dry matter content, and starch yield, which, thereby, provides extensive scope for improving this crop through hybridization and selection. Studies assessing the quality of cassava roots and the genetic variation of large numbers of germplasm accessions have not been previously undertaken in Brazil. This hampers engagement with the current policy of sustainable agricultural systems, in which it is necessary to use the components of diversity in a proper way and avoid medium and long term diversity reduction. Additionally, what is observed as regards Brazilian cassava nowadays, especially in agricultural systems that use intensive technologies, is the use of a very limited number of cassava cultivars, which certainly share common ancestors and, therefore, reduce the allelic diversity available to the production system.

Brazil, being the center of cassava's origin and diversity must characterize and evaluate its genetic resources appropriately so that they can be effectively used for developing new cultivars. Thus, breeding by hybridization and selection among accessions from the groups established in this study may contribute to the introduction of new genetic diversity to help avoid problems related to limited genetic variation and may contribute to an improvement in the quality of cassava roots.

Future perspectives

Estimating genetic variation among cassava accessions is a prerequisite for efficient germplasm management and its use in breeding programs. Agronomic traits, especially those related to root quality are very important for grouping genetic resources and are also essential to the improvement of existing varieties by introducing novel genetic variation. However, clusters formed by agronomic traits were not in accordance with molecular marker data (Esmaeilzadeh et al., 2005; Garcia et al., 2007; Barakat et al., 2013).

Possible reasons for the lack of correlation between molecular and morpho-agronomic variation could be the wide genome coverage of molecular markers, including coding and non-coding regions, and by the fact that molecular markers are less subject to artificial selection in comparison with morpho-agronomic markers (Semagn, 2002). Therefore, the contradiction of the results from these types of data indicates that germplasm clustering and selection for crossing in breeding programs should not rely on a single measurement. Hereafter, these cassava accessions will be genotyped using high-throughput approaches, such as genotyping by sequencing (GBS), for future studies focusing on germplasm characterization aiming to carry out a full analysis of genetic diversity (morpho-agronomic and molecular data).

Moreover, in general, clustering methods such as AP produce and maintain very distinct groups of accessions, with a reduced level of genetic variation within each group, but with high genetic variability among clusters. In theory, these clusters could be used to obtain a maximum heterotic response when hybrids are produced by crossing genotypes from different groups. Therefore, potential heterotic groups based on the quality of cassava roots and AP algorithm can be used in crosses to test the robustness of these clusters. In addition, clusters containing many accessions, i.e. Clusters 3, 5, 8 and 9 (Table 2), can be subdivided in heterotic groups based on other agronomic plant traits, such as stem growth habit, plant height, first branching height, plant shape, number of storage roots/plant, starch content, harvest index and resistance to postharvest deterioration.

These potential heterotic groups in cassava could help breeders to meet the changing needs of modern agriculture, which lead farmers, for economic reasons, to accept only a few of the highest yielding cultivars with good root quality. Thus, we hope that our focus on germplasm preservation and characterization using different approaches could ensure that farmers have access to the best varieties of cassava.

Acknowledgements

The authors thank Bahia State Foundation for Research Support (FAPESB) and the Brazilian National Council for Scientific and Technological Development (CNPq) for their financial assistance and scholarship support.

References

Barakat, M.N.; Al-Doss, A.A.; Elshafei, A.A.; Ghazy, A.I, Moustafa, K.A. 2013. Assessment of genetic diversity among wheat doubled haploid plants using TRAP markers and morpho-agronomic traits. Australasian Journal of Crop Science 7: 104-111.

Beer, S.C.; Goffreda, J.; Phillips, T.D.; Murphy, J.P.; Sorrells, M.E. 1993. Assessment of genetic variation in *Avena sterilis* using morphological traits, isozymes, and RFLPs. Crop Science 33: 1386-1393.

Benesi, I.R.M.; Labuschagne, M.T.; Herselman, L.; Mahungu, N. 2010. Ethnobotany, morphology and genotyping of cassava germplasm from Malawi. Journal of Biological Sciences 10: 616-623.

Bodenhofer, U.; Kothmeier, A.; Hochreiter, S. 2011. APCluster: an R package for affinity propagation clustering. Bioinformatics Applications Note 27: 2463-2464.

Borile, C.; Labarre, M.; Franz, S.; Sola, C.; Refrégier, G. 2011. Using affinity propagation for identifying subspecies among clonal organisms: lessons from *M. tuberculosis*. BMC Bioinformatics 12: 224.

Brown-Guedira, G.L.; Thompson, J.A.; Nelson, R.L.; Warburton, M.L. 2000. Evaluation of genetic diversity of soybean introductions and North American ancestors using RAPD and SSR markers. Crop Science 40: 815-823.

Ceballos, H.; Sánchez, T.; Morante, N.; Fregene, M.; Dufour, D.; Smith, A.M.; Denyer, K.; Pérez, J.C.; Calle, F.; Mestres, C. 2007. Discovery of an amylose: free starch mutant in cassava (*Manihot esculenta* Crantz). Journal of Agricultural and Food Chemistry 55: 7469-7476.

Chávez, A.L.; Sánchez, T.; Jaramillo, G.; Bedoya, J.M.; Echeverry, J.; Bolaños, E.A.; Ceballos, H.; Iglesias, C.A. 2005. Variation of quality traits in cassava roots evaluated in landraces and improved clones. Euphytica 143: 125-133.

Cooke, R.D. 1979. Enzymatic assay for determining the cyanide content of cassava and cassava products. Centro International de Agricultura Tropical, Cali, Colombia.

Defloor, I.; Dehing, I.; Delcour, J.A. 1998. Physico-chemical properties of cassava starch. Starch 50: 58-64.

Duraisamy, R.; Rathinasamy, S.A.; Natesan, S.; Muthurajan, R.; Ramineni, J.J.; Karuppusamy, N.; Lakshmanan, P.; Chokkappan, P.; Gandhi, K. 2011. Starch content and cassava mosaic disease genetic diversity with relation to yield in South Indian cassava (*Manihot esculenta* Crantz) germplasm. Journal of Crop Science and Biotechnology 14: 179-189.

Esmaeilzadeh, M.M.; Trethowan, R.M.; William, H.M.; Rezai, A.; Arzani, A.; Mirlohi, A.F. 2005. Assessment of genetic diversity in bread wheat genotypes for tolerance to drought using AFLPs and agronomic traits. Euphytica 141: 147-156.

Frey, B.J.; Dueck, D. 2007. Clustering by passing messages between data points. Science 315: 972-976.

Garcia, M.V.; Balatti, P.A.; Arturi, M.J. 2007. Genetic variability in natural population of *Paspalum dilatatum* Poir. analyzed by means of morphological traits and molecular markers. Genetic Resource and Crop Evolution 54: 935-946.

Henderson, C.R. 1984. Applications of Linear Models in Animal Breeding. University of Guelph, Guelph, Canada.

Instituto Brasileiro de Geografia e Estatística [IBGE]. 2013. Systematic survey of agricultural production = Levantamento sistemático da produção agropecuária. Available at: http://www.ibge.gov.br/home/estatistica/indicadores/agropecuaria/IBGE/ [Accessed Nov 14, 2013].

International Organization for Standarization [ISO]. 2005. Norme ISO 6647 (F). Riz – Determination de la Teneur em Amylose. ISO, Geneva, Switzerland.

Jansz, E.R.; Uluwaduge, D.I. 1997. Biochemical aspects of cassava (*Manihot esculenta* Crantz) with special emphasis on cyanogenic glucosides: a review. Journal of the National Science Council of Sri Lanka 25: 1-24.

Kiddle, S.J.; Windram, O.P.F.; Mchattie, S.; Mead, A.; Beynon, J.; Buchanan-Wollaston, V.; Denby, K.J.; Mukherjee, S. 2010. Temporal clustering by affinity propagation reveals transcriptional modules in *Arabidopsis thaliana*. Bioinformatics 26: 355-362.

Mohammadi, S.A.; Prasanna, B.M. 2003. Analysis of genetic diversity in crop plants: salient statistical tools and considerations. Crop Science 43: 1235-1248.

Moorthy, S.N. 2004. Tropical sources of starch. p. 321-359. In: Eliasson, A.C., ed. Starch in food. CRC Press, Boca Raton, FL, USA.

Olsen, K.M. 2004. SNPs, SSRs and inferences on cassava's origin. Plant Molecular Biology 56: 517-526.

Raji, A.A.; Ladeinde, T.A.O.; Dixon, A.G.O. 2007. Agronomic traits and tuber quality attributes of farmer grown cassava landraces in Nigeria. Journal of Tropical Agriculture 45: 9-13.

Refianti, R.; Mutiara, A.B.; Juarna, A.; Ikhsan, S.N. 2012. Analysis and implementation of algorithm clustering affinity propagation and k-means at data student based on GPA and duration of bachelor-thesis completion. Journal of Theoretical and Applied Information Technology 35: 69-76.

Rickard, J.E.; Asaoke, M.; Blanshard, J.M.V. 1991. The physicochemical properties of cassava starch. Tropical Science 31: 189-207.

Sakellariou, A.; Sanoudou, D.; Spyrou, G. 2012. Combining multiple hypothesis testing and affinity propagation clustering leads to accurate, robust and sample size independent classification on gene expression data. BMC Bioinformatics 13: 270.

Semagn, K. 2002. Genetic relationships among ten endod types as revealed by a combination of morphological, RAPD, and AFLP markers. Hereditas 137: 149-156.

Sumedha, M.L.; Weigt, M. 2007. Clustering by soft-constraint affinity propagation: applications to gene-expression data. Bioinformatics 23: 2708-2715.