Theor Appl Genet (2014) 127:1423–1440
DOI 10.1007/s00122-014-2309-8

ORIGINAL PAPER

Potential of SNP markers for the characterization of Brazilian cassava germplasm

Eder Jorge de Oliveira · Cláudia Fortes Ferreira · Vanderlei da Silva Santos · Onildo Nunes de Jesus · Gilmara Alvarenga Fachardo Oliveira · Maiane Suzarte da Silva

Received: 7 February 2013 / Accepted: 31 March 2014 / Published online: 16 April 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract

Key message **High-throughput markers, such as SNPs, along with different methodologies were used to evaluate the applicability of the Bayesian approach and the multivariate analysis in structuring the genetic diversity in cassavas.**

Abstract The objective of the present work was to evaluate the diversity and genetic structure of the largest cassava germplasm bank in Brazil. Complementary methodological approaches such as discriminant analysis of principal components (DAPC), Bayesian analysis and molecular analysis of variance (AMOVA) were used to understand the structure and diversity of 1,280 accessions genotyped using 402 single nucleotide polymorphism markers. The genetic diversity (0.327) and the average observed heterozygosity (0.322) were high considering the bi-allelic markers. In terms of population, the presence of a complex genetic structure was observed indicating the formation of 30 clusters by DAPC and 34 clusters by Bayesian analysis. Both methodologies presented difficulties and controversies in terms of the allocation of some accessions to specific clusters. However, the clusters suggested by the DAPC analysis seemed to be more consistent for presenting higher probability of allocation of the accessions within the clusters. Prior information related to breeding patterns and geographic origins of the accessions were not sufficient for providing clear differentiation between the clusters according to the AMOVA analysis. In contrast, the F_{ST} was maximized when considering the clusters suggested by the Bayesian and DAPC analyses. The high frequency of germplasm exchange between producers and the subsequent alteration of the name of the same material may be one of the causes of the low association between genetic diversity and geographic origin. The results of this study may benefit cassava germplasm conservation programs, and contribute to the maximization of genetic gains in breeding programs.

Communicated by A. J. Bervillé.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-014-2309-8) contains supplementary material, which is available to authorized users.

E. J. de Oliveira (✉) · C. F. Ferreira · V. da Silva Santos ·
O. N. de Jesus
Embrapa Cassava and Fruits, Rua da Embrapa s/n,
Cruz das Almas, BA, Brazil
e-mail: eder.oliveira@embrapa.br

C. F. Ferreira
e-mail: claudia.ferreira@embrapa.br

V. da Silva Santos
e-mail: vanderlei.silva-santos@embrapa.br

O. N. de Jesus
e-mail: onildo.nunes@embrapa.br

G. A. F. Oliveira · M. S. da Silva
Federal University of Recôncavo da Bahia,
Cruz das Almas Campus, Cruz das Almas, BA, Brazil
e-mail: gfachardo@yahoo.com.br

M. S. da Silva
e-mail: maisuzarte@yahoo.com.br

Introduction

Cassava (*Manihot esculenta* Crantz) is known for its drought tolerance and stable productivity even when cultivated in soils of low fertility. Cassava is a species that produces starchy roots with high adaptability and is considered a major basic food for more than 800 million people worldwide, especially in sub-Saharan Africa (Lebot 2009). In Brazil, mostly in the northeast region, cassava plays a key role in food security based on its ability to grow in the presence of adverse conditions (poor soils and low water availability) that are lethal to other crops. In addition, in the mid-southern region of Brazil, cassava plants have a strong industrial appeal (Dixon et al. 2003).

Cultivated in tropical and subtropical regions of Asia, Latin America and Africa, cassava production reached 230 million tons in 2010, whereas Brazilian production was responsible for more than 10 % of the worldwide cassava production (FAO 2012). However, there are many contrasts between the different cassava-producing countries, even though one of the main issues in common between African countries and different regions of Brazil is the low average yield (average of 10.2 and 13.7 t ha^{-1} in African countries and Brazil, respectively). In many situations, these yields may be six times lower than the potential yield of the crop (Lebot 2009).

Cassavas originated in the Americas and were transferred by the Portuguese to the rest of the world, particularly to the African continent, in the sixteenth century. In Africa and Latin America, cassavas have been cultivated and produced for many centuries by small farmers, which has resulted in a large number of local cassava varieties (Nweke et al. 2002).

Cassava is an allogamous species propagated predominantly by stakes. Therefore, the cassava germplasm may be preserved in situ and *ex situ* in the field and in vitro in the laboratory or by botanical seeds. However, improved cassava varieties and landraces are almost exclusively preserved as in vitro plantlets or in the field as clones. One of the greatest Latin American cassava germplasm collections belongs to Embrapa Cassava and Fruits (Cruz das Almas, Brazil), which maintains more than 1,300 accessions in vivo. This variability is represented mostly by landraces selected naturally or artificially by producers.

This collection is important because it preserves the genetic variability of cassava plants, which can be used in the development of new drought-tolerant, disease-resistant, high-quality and high-yield varieties. Furthermore, the analysis of the genetic relationships present in different cassava accessions is an important endeavor for the development of breeding programs, as these studies provide information about the genetic diversity and stratification of breeding populations. In addition, the precise evaluation of the genetic diversity of cassava plants contributes information regarding the (a) genetic variability of cultivars, (b) identification of parental combinations that allow the development of segregating progenies with maximum genetic variability, and (c) introgression of desirable genes present in the germplasm available.

However, one of the challenges of germplasm banks especially those for orphan crops such as cassava, is the ability to measure, register and compare the genetic variability in a large number of accessions (de Bang et al. 2011). This is especially valid for the agronomic characterization of cassavas in the field, where the evaluation of a large number of genotypes is extremely time consuming, costly and highly influenced by environmental effects. Therefore, until now, most studies published regarding genetic diversity in cassavas have used only a small sample of plant collections.

Kawuki et al. (2011) evaluated the phenotypic variability of the cassava germplasm available in six countries in Africa (Uganda, Kenya, Tanzania, Rwanda, Democratic Republic of Congo and Madagascar) for 29 qualitative and four quantitative characteristics (dry matter content, harvest index, leaf retention and root cortex thickness). The results indicated a limited power of discrimination of the cassava accessions based on the qualitative descriptors, even though the quantitative descriptors showed sufficient variation for implementing desired breeding schemes. Therefore, the characterization based on morphological descriptors is important; however, the use of morphological descriptors has been ineffective for the complete understanding of the genetic variability present in large cassava germplasm collections (Benesi et al. 2010; Kawuki et al. 2011), and also in annual (Cieslarová et al. 2012) and perennial crops (Prakash et al. 1996). At the same time, while quantitative descriptors may contribute to the study of genetic variation, the need for experiments with replicates and analyses carried out on cassavas grown in different environments under many years of cultivation becomes an extremely costly and laborious process that has been limited to only part of the germplasm collection.

In contrast, molecular markers facilitate the measurement of diversity based on a large number of neutral loci without the influence of the environment and with high discrimination power. In diversity studies, a large number of molecular markers have been used including restriction fragment length polymorphisms, amplified fragment length polymorphisms (AFLPs), random amplified polymorphic DNA (RAPDs), diversity array technology, simple sequence repeats or microsatellites (SSR), and a combination of markers and single nucleotide polymorphisms (SNPs) (Olsen 2004; Kawuki et al. 2009; Ferguson et al. 2012). However, advances in molecular technologies have made SNPs an attractive option for high-throughput genotyping due to the relatively low cost per data point, the high abundance of SNPs in the genome, the locus specificity and codominance of SNPs, the potential for high-throughput analysis and the low genotyping error rate (Rafalski 2002; Schlotterer 2004; Chagné et al. 2007). Recently, Ferguson et al. (2012) identified 2,954 putative SNPs from which 1,190 were technically and biologically validated for use in cassavas.

Bayesian clustering methods have been used to identify genetic clusters under an explicit population genetics model (Pritchard et al. 2000; Jombart et al. 2010; Arnaud et al. 2011; Semagn et al. 2012). Several advantages can be achieved using Bayesian methods, especially assigning admixed individuals to population clusters, since the prior information could assist the calculation of ancestry. As for the Bayesian analysis as implemented in the Structure v.2.3 (Pritchard et al. 2000) program, there are powerful analytical tools to identify the genetic structure within different sets of data. However, this approach assumes that the populations are in Hardy–Weinberg equilibrium and that there is linkage equilibrium between the loci, i.e., prerequisites that are frequently violated in natural populations and in germplasm accessions. Therefore, this approach relies on assumptions such as the type of population subdivision, which can restrict their applicability. Moreover, it requires considerable computational time when estimating large datasets. In contrast, multivariate analyses such as discriminant analysis of principal components (DAPC), is a new methodological approach which has the ability to identify genetic structures in very large datasets within negligible computational time, and the absence of any assumption about the population genetic model (Jombart et al. 2010).

Although studies of genetic variation in cassavas have provided valuable information about the genetic diversity, population structure and origin of the species, most of the cassava germplasm has not yet been genotyped. Therefore, the main objective of this study was to investigate the population structure and the relationship patterns of cassava accessions from the cassava germplasm bank from Embrapa Cassava and Fruits in Brazil for use in breeding programs. Other objectives of our study were to evaluate the applicability of the Bayesian approach and multivariate systems in structuring the genetic diversity in cassavas. The availability of this information will facilitate the design of efficient strategies to use in cassava breeding programs and to target future collection sites to enable better use of the genetic resources of the species.

Materials and methods

Plant material

One thousand, two hundred and eighty accessions from various ecosystems in Brazil, Colombia, Venezuela and Nigeria were evaluated from the cassava germplasm bank at Embrapa Cassava and Fruits (Cruz das Almas, Brazil). Although passport information was restricted, basic information pertaining to the collection sites at the national level, was available. However, several of the accessions did not have complete passport data due to non-standardized collection practices in some countries, and stake submissions by producers without including any basic passport information (Supplement 1).

Genotyping

The SNP markers used are available at the Cassava Genome Database (http://cassava.igs.umaryland.edu/cgi-bin/index.cgi). The SNPs were identified using a bacterial artificial chromosome (BAC)-based fingerprint map of an inbred cultivar of cassava, whose end of BAC clones were sequenced and low-copy sequences throughout the genome, were selected. The genotyping of 354 selected SNPs from genetic loci and 48 other loci based on the physical map of cassava (Supplement 2) was conducted by using a MassArray system (Sequenom iPLEXassay, San Diego, USA). Polymerase chain reaction (PCR) was conducted by MassExtend (Sequenom), where 15 ng of genomic DNA was isolated from leaf samples using the CTAB protocol (Doyle and Doyle 1990) with modifications. Quantification was performed in 1.0 % (p/v) agarose gels stained with ethidium bromide (1.0 mg L^{-1}) using a series of concentrations of Lambda phage (Invitrogen) as the standard. Locus-specific PCR and primer detection were designed using the MassARRAY Assay Design 3.0 software (Sequenom, San Diego, USA).

DNA samples were amplified based on a multiplex reaction, and PCR products were used in a single-base extension reaction for each locus. The resulting products were desalted and transferred to a 384-element SpectroCHIP array. Alleles were discriminated by mass spectrometry (Sequenom, San Diego, USA). Only samples and SNPs with less than 10 % of data lost were analyzed.

Diversity analysis

The following information was obtained from the estimate of the allele frequencies of the SNPs: (1) observed heterozygosity (Ho), calculated directly from the sample by the genotypic frequencies, (2) expected heterozygosity (He), calculated according to Nei (1973): $\mathrm{He} = 1 - \sum p_{ij}^2$, where p_{ij} is the frequency of the jth allele for the ith locus, (3) polymorphism information content (PIC) according to Botstein et al. (1980) $\mathrm{PIC} = 1 - \left(\sum_{i=1}^{k} p_i^2\right) - \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} 2p_i^2 p_j^2$, where k is the number of alleles and p_i and p_j are the frequencies of the i and j alleles, respectively. All of the estimates were calculated by using the PowerMarker v.3.25 software (Liu and Muse 2005).

Linkage disequilibrium (LD) was estimated as the correlation coefficient (r^2) between all pairs of SNPs using the genetics package for the 2.11 version of the R program (R Development Core Team 2010).

Population structure analysis

Three complementary approaches were used to evaluate the structure of the genetic diversity between accessions from the cassava germplasm: clustering based on the DAPC, a Bayesian method and molecular analysis of variance (AMOVA).

Discriminant analysis of principal components

The DAPC available in the adegenet package for the 2.11 version of the R program (R Development Core Team 2010) was used for the definition of the clusters of the cassava hybrids and the parentals because this technique does not require pre-defined genetic groups (Jombart et al. 2010). Repeated clusters with the *K-means* method and the Bayesian information criteria (BIC) were used to define the number of groups where the *K* with the lowest BIC value represented the most probable number of groups for the set of data analyzed. However, in the presence of genetic structure in cline (a particular type of continuous variation) or of the hierarchical form, the BIC values can be reduced after the identification of the real *K* value. Therefore, a reduction of the BIC values was visually analyzed to identify the *K* value where the BIC values were slightly reduced (Jombart et al. 2010). *K* values were tested from 1 to 50 with 10 runs for each *K*. After setting the number of groups, the axes of the principal components analysis that explained more than 90 % of the total variance of the data, were kept.

To avoid over-fitting during the discrimination of the groups by the DAPC, the ideal number of principal components was estimated using the optim.a.score function in the adegenet package. The *a*-score measures the proportion of successful reassignments of the DAPC in comparison to the k-means cluster (observed discrimination) and the random cluster (random discrimination). The *a*-score is calculated as (Pt–Pr), where Pt is the probability of re-attribution using the real cluster, and Pr is the probability of re-attribution for the clusters permuted randomly. An *a*-score close to one (1) means that the DAPC solution is highly discriminatory and stable, whereas low values (close to zero) indicate low power of discrimination of the accessions or instability of results.

Bayesian analysis

The genetic structure of the cassava accessions was analyzed using Structure v.2.3 software (Pritchard et al. 2000). The number of clusters was inferred using five independent runs with 50,000 burn-ins and 30,000 MCMC (Monte Carlo Markov Chain) permutations after the burn-ins using the admixture ancestry model, non-linked loci and correlated allele frequencies with *K* varying from 2 to 50.

The Structure v.2.3 algorithm can be effective to infer the correct number of clusters in a data set that presents some relationship of isolation by distance. In the case where the dispersion patterns between the populations are not homogeneous, the estimated log probability of the data does not provide a correct estimate of the number of clusters. Therefore, the ΔK by Evanno et al. (2005) was estimated to verify whether the inferred number of clusters presented consistent and reliable results. These methods are *ad hoc* statistics that assist the researcher in finding the ideal value for *K*. For this reason, ΔK statistics are used because they are based on the rate of variation between the repeated values of *K* to infer the highest level of structure based on the data set (Evanno et al. 2005). In this case, the Structure software assumes that the genotype of each individual at each locus is totally unknown.

After the value of *K* is defined, the averages of the traditional estimates of genetic differentiation based on F_{ST} of the five independent runs were used to evaluate the consistency of the clusters in terms of genetic differentiation.

Analysis of molecular variance

Analysis of molecular variance (AMOVA) was performed by decomposition of the principal components into different hierarchical levels: (a) type of accession (improved variety or landrace), (b) origin of accessions [Brazil (midwest region, north, northeast and south regions), Colombia, Nigeria and Venezuela], and (c) the theoretical cluster obtained by the DAPC and Bayesian analysis. These analyses were performed using GenAlEx 6.1 software (Peakall and Smouse 2006).

Results

Genetic diversity

Of the 402 SNPs analyzed, six were monomorphic in the germplasm evaluated (MH008_I10, MH125_P15, 6282, 7945, 31679 e 36127) and therefore removed from the subsequent analysis leaving 396 SNPs for our analysis. The average Ho for all loci was 0.322 and varied from 0.001 to 0.999. For approximately 87 % of the cases, the Ho estimates were lower than 0.50. In contrast, the genetic diversity (He) varied from 0.002 to 0.500 with an average He of 0.327. In this case, approximately 42 % of the SNPs presented an He value >0.45 (Fig. 1 and Supplement 2).

The 7622, 1167, 7622, 1167, 19501, 35279, 23076, 35735 and MH076_J24 loci showed an excess of heterozygotic plants and although these loci did not show signs of linkage disequilibrium, they could be associated with the process of domestication of the species over the years.

The PIC values varied from 0.002 to 0.375 with an average PIC value of 0.262 (Fig. 1 and Supplement 2). However, approximately 48 % of the SNPs presented with PIC estimates over 0.30, which qualified these SNPs as

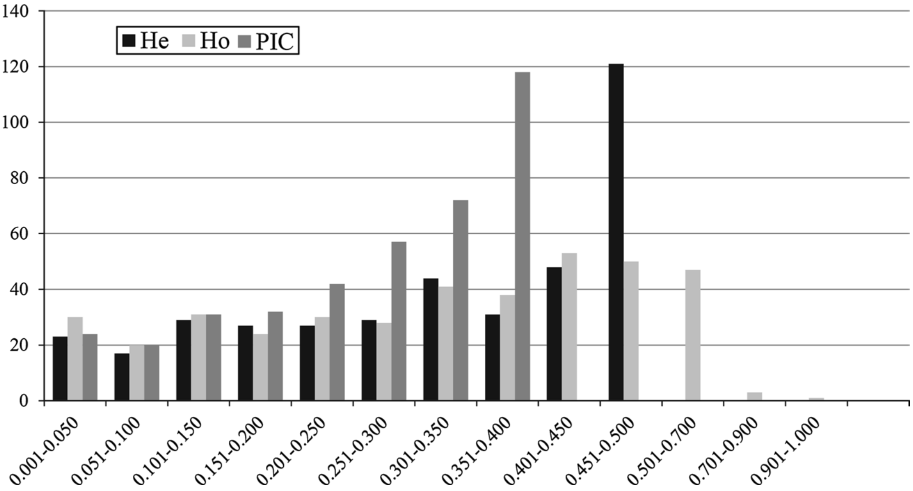

Fig. 1 Distribution of the frequencies of the Ho, genetic diversity (He) and polymorphism information content by the assessment of the cassava germplasm using SNP markers

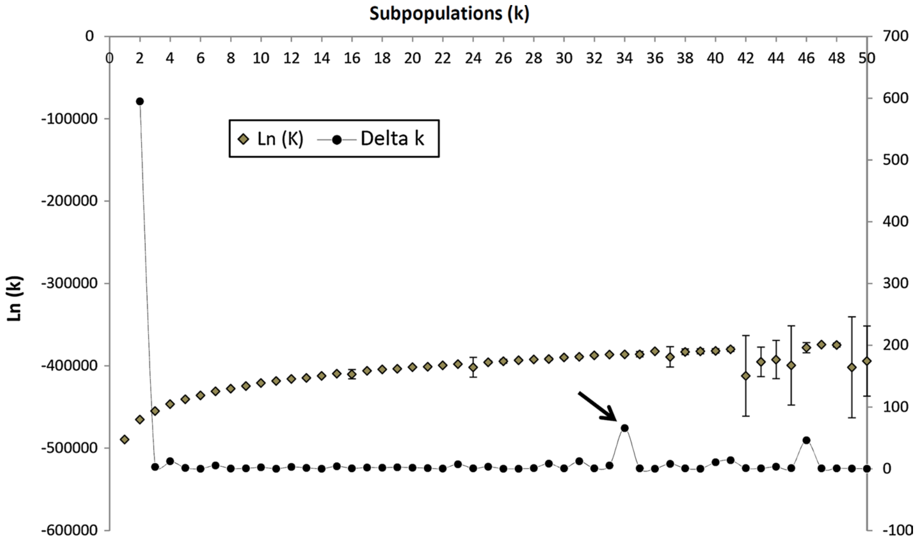

Fig. 2 Likelihood-log from the Structure v.2.3 software with *K* varying from 1 to 50. The values of Δ(*K*) as the criteria for the determination of the number of groups, are presented by the *arrow*

candidates for use in genetic variability studies or for the detection of polymorphisms associated with breeding because these SNPs were relatively more informative.

Measures of LD were calculated for each of pair of SNPs. The overall average r^2 for all pairwise comparisons was 0.016 (range from 0.00 to 0.84) and the number of SNP pairs showing evidence of LD > 0.25 was low (0.21 % of total) (Supplement 3).

Estimate of the structure

The mixed model of the Structure v.2.3 software was analyzed with the number of groups (*K*) varying from 1 to 50 independent repetitions for each group. The most adequate *K* value to describe the structure of the cassava germplasm bank based on the Pritchard criteria (Pritchard et al. 2007) is the one with the least likelihood value after reaching the LnP(D) plateau. However, as illustrated in Fig. 2, a clearly defined plateau was not observed in this set of cassava accessions. In addition, the Evanno et al. (2005) method, which uses more formal criteria taking into consideration the most abrupt rupture in the slope of the distribution of the LnP(D) values, was used for the definition of the most adequate *K* value for structuring the cassava germplasm. Figure 2 shows an abrupt drop in the curve with Δ(*K*) equal to 2 and other two peaks of Δ(*K*) with $K = 34$ and 46. The broad genetic variability based on morphological observation indicates that $K = 2$ is not sufficient to describe the maximum variability of the cassava germplasm. Therefore, the next highest Δ(*K*) peak (=34), was used for the subsequent cluster analysis.

Considering that the choice of the adequate number of groups to represent the genetic variability is an important step to infer about the germplasm preserved in the cassava germplasm bank, the DAPC method was used for the formation of clusters. In DAPC analysis, 200 principal components

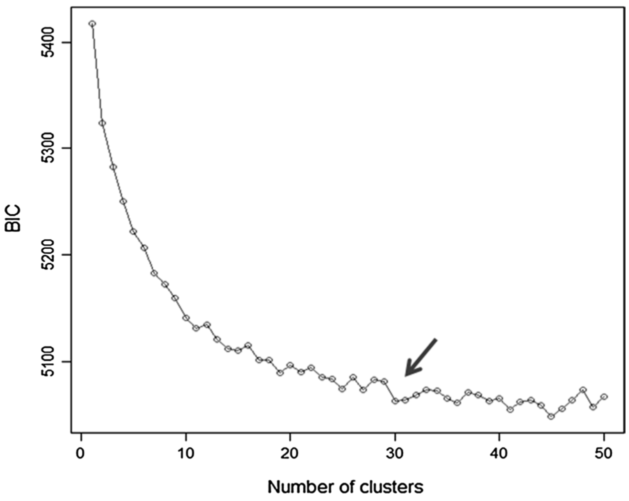

Fig. 3 Distribution of the differences between expected and Ho for the analysis of 1,280 cassava accessions by 402 SNP markers

(PCs) were maintained in the analysis in the preliminary transformation of data, which explained 91 % of the total genetic variation in the set. The definition of the number of PCs required for the analysis is related to the power of reduction of the dimension of the data. In general, many studies have indicated the use of components that retain more than 80 % of the genetic variance. In the context of DAPC, the equilibrium point should be determined between the power of discrimination of the clusters and the stability of the attributions of the genotypes in each cluster. Therefore, the analysis with 200 PCs guarantees high statistical power in evaluating the genetic structure of the cassava germplasm.

Figure 3 indicates the presence of hierarchical structure in the cassava germplasm bank at Embrapa Cassava and Fruits and illustrates the selection process of the ideal number of groups. According to Jombart et al. (2010) the best BIC is indicated by an elbow in the curve of BIC values matching with the smallest BIC and the “true” number of cluster. However, considering that it was not so clear in Fig. 3, but the most abrupt initial drop and resumption of the curve according to the BIC began with $K = 30$, this K was used to represent the variability of the cassava germplasm by the DAPC. Moreover, even with $K = 30$ not having the lowest BIC, it was observed that the BIC values from $K = 30$ tend to be quite homogeneous.

Clustering of molecular diversity

Initially, clusters were formed considering a priori information about the classification of the cassava accessions. To allow for a more uniform cluster, classification on two and nine groups, were used for the breeding pattern and geographic distributions of the collection sites of the accessions, respectively; however, the accessions were clustered in increasing order of the accessions according to the proportion of shared genome in each population (q_K) (Fig. 4). In both cases, there is great discrepancy in the pattern of the clusters proposed by the Structure v.2.3 software in comparison to the one based on the a priori classification of the germplasm. For example, considering two groups for the breeding pattern, both groups (green and red—Fig. 4) have improved cassava accessions and landraces. Likewise, with nine groups according to their origin, there is not a clear differentiation of the accessions based on the collection sites (Fig. 4b).

Even though, $K = 2$ is not sufficient to describe the cassava variability, the Structure v.2.3 software was used to understand the structuring for these two groups. We observed that these two clusters group together landraces and improved cassava and accessions from different geographic origins (Fig. 5). The proportion of landraces and improved accessions was very similar in both groups, 66 and 34 %, respectively, indicating no predominance of a given pattern at any group.

Considering $K = 34$, for the genetic distribution of the accessions, a pattern of clustering was also not observed even based on the breeding pattern of the cassava accessions (Fig. 5). Only five groups were formed exclusively by landraces (clusters 3, 5, 16, 19 and 31); the remaining clusters showed a range of 1–20 improved accessions per cluster. Therefore, the relationship of groups according to breeding pattern presented a broad overlap and also indicated likely admixture of gene pools. Moreover, with $K = 34$, the degree of admixture is higher in improved accessions (29.8 %) compared with the landraces (21.8 %). Likewise, considering the geographic origins, only clusters 14 and 30 presented accessions belong to just one region (Brazil northeast). For the remaining clusters we could not find a reasonable level of structuring according to their geographical classes.

Considering an arbitrary cut-off of ancestry at 71 % for the arrangement of the genotypes in the groups, 603 accessions (47 %) were attributed to one of the 34 groups (Table 1). A large number of plants (677) seem to have ancestry in more than one cluster with values of q_K lower than 71 % for 30 of the 34 groups. Therefore, the results of the population structure allows the assertion that the cassava germplasm analyzed contains a great amount of genetic diversity derived from sometime in the past where complex hybrids were bred from crosses between completely different genetic backgrounds (Fig. 5). A clear relationship among accessions with similar mixed ancestry and breeding pattern was not observed because among the accessions with q_K lower than 71 %, 88 and 589 improved accessions and landraces were observed, respectively. Considering the q_K higher than 71 %, a quite similar proportion was identified (70 improved accessions and 533 landraces).

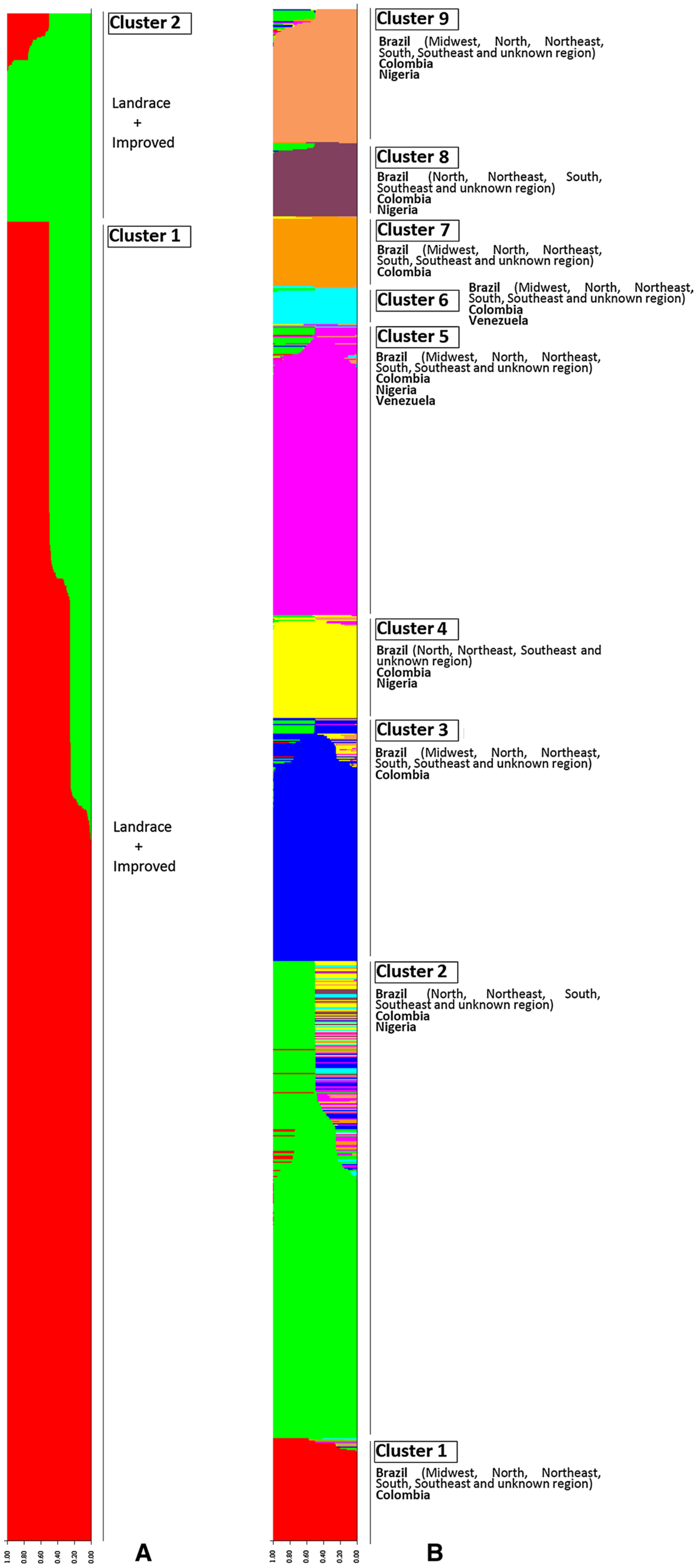

Fig. 4 Results from the attribution of the genetic structure analysis of the 1,280 accessions from the cassava germplasm. *Bar plot* from the Structure v.2.3 software, using a priori information about: (**a**) the breeding pattern, i.e., traditional versus improved accessions, and (**b**) geographical origins. Each individual is represented by a *vertical line* divided into *K* colored segments with length proportional to individual coefficient of participation in the *K* clusters or probability of attribution (q_K) for each cluster

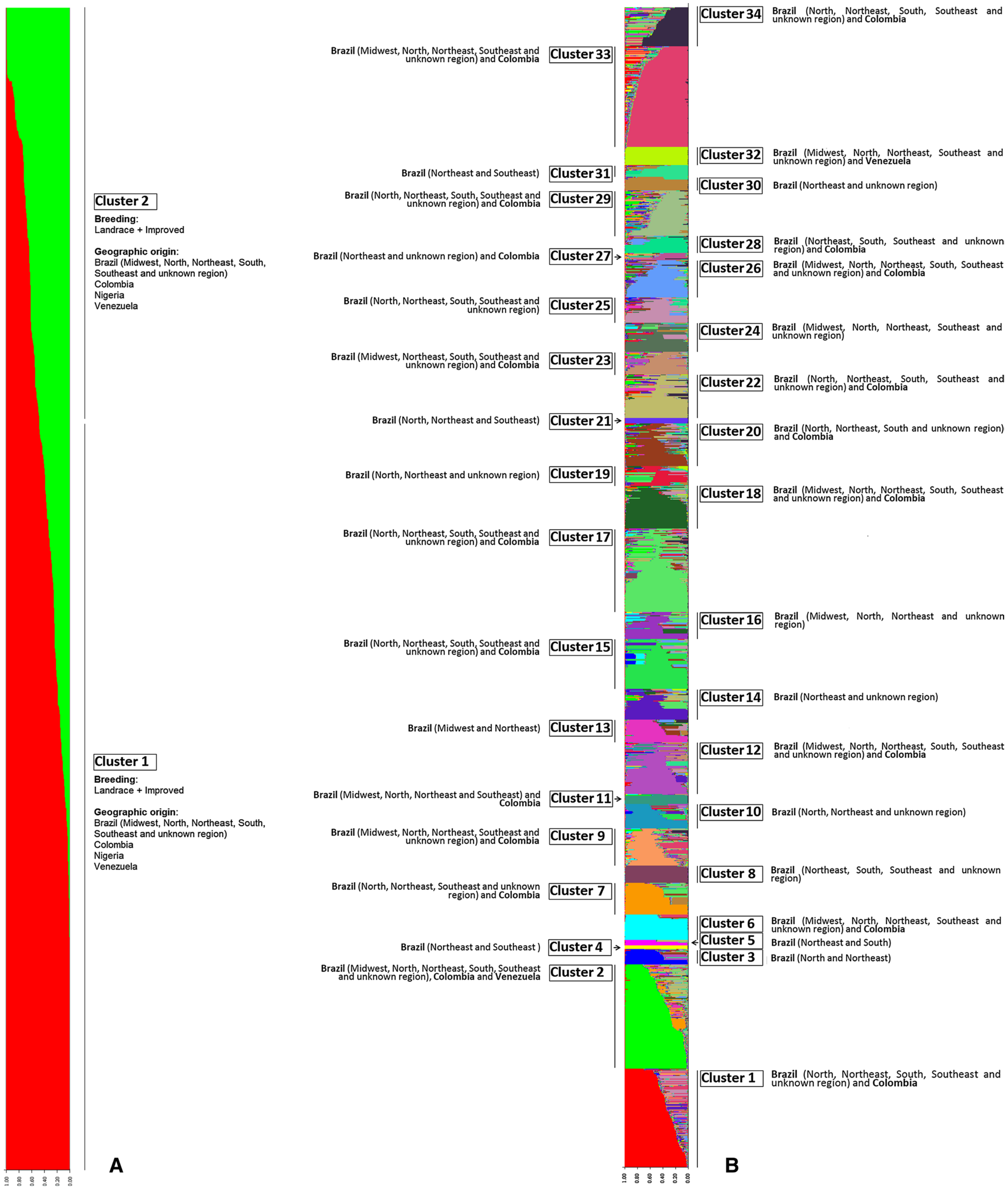

Fig. 5 Results from the attribution of the analysis of the genetic structure of 1,280 accessions from the cassava germplasm. *Bar plot* from the Structure v.2.3 software using $K = 2$ and 34, assuming 2 and 34 different population and each and each one represented by *different colors*. Each individual is represented as a *vertical line* divided into K colored segments, with length proportional to the individual coefficient of participation in the K clusters or probability of attribution (q_K) for each cluster

Table 1 Ancestry pattern for 34 groups of cassava accessions by the Bayesian analysis ($K = 34$)

Cluster	Ancestry (q_K %)				
	>0.91	0.71–0.90	0.51–0.70	0.31–0.50	<0.30
1	18	40	42	7	2
2	38	23	29	22	3
3	5	0	9	2	1
4	4	0	0	0	0
5	3	1	0	2	0
6	23	1	2	2	0
7	11	2	18	3	1
8	18	1	0	0	0
9	0	6	15	17	3
10	9	1	9	8	0
11	9	0	0	2	0
12	8	18	19	12	0
13	7	0	13	5	0
14	12	0	7	9	6
15	23	1	23	8	0
16	6	0	9	14	1
17	32	12	15	32	1
18	27	5	9	6	0
19	0	0	14	8	0
20	10	2	15	17	3
21	6	0	0	0	0
22	20	1	12	14	1
23	6	3	8	5	3
24	11	0	10	9	2
25	8	0	13	6	1
26	11	8	14	7	1
27	0	0	2	6	0
28	9	0	7	2	1
29	0	8	21	18	3
30	12	0	2	1	0
31	3	5	1	3	1
32	20	0	0	0	0
33	35	51	14	10	1
34	1	9	11	19	3

For the DAPC assessment, 30 diversity groups were identified (Fig. 6) on the basis of the SNP analysis. DAPC searches for a reduced internal space in which the accessions are better discriminated in pre-defined groups. One way to assess the quality of this discrimination is to analyze the re-attribution of the individuals to their previous groups where the success in this re-attribution is a sign of high discriminatory power. By using the same threshold in the Bayesian analysis (71 %), nine (clusters 1, 2, 9, 10, 21, 22, 24, 25 and 26) of the 30 clusters formed by the DAPC analysis were accurate in allocating genotypes to the appropriate groups (*a*-score = 1.00) (Table 2). Eleven other clusters (5, 8, 11, 12, 13, 15, 16, 17, 18, 27 and 30) presented high reliability in the attribution of the cassava accessions (*a*-score = 0.81–0.99), while five clusters (3, 14, 19, 23 and 28) presented average reliability in the attribution of the cassava accessions (*a*-score = 0.50–0.65). However, five other clusters (4, 6, 7, 20 and 29) presented low reliability and stability in the clustering of the cassava accessions (*a*-score = 0.00–0.24) (Table 2).

In general, there seemed to be some inconsistency between the groups identified by the Bayesian analysis and the DAPC analysis (Table 3). Complete coincidence was observed in the clustering pattern between both methodologies for clusters 4, 8, 15, 21 and 32 from the Bayesian analysis (Fig. 5) and clusters 25, 22, 16, 8 and 9 from the DAPC analysis (Fig. 6 and Table 3). In contrast, the cassava accessions allocated in clusters 2, 17, 20, 22, 24, 29 and 34 from the Bayesian analysis, were divided into at least eight clusters in the DAPC analysis (Table 3).

Differentiation of the population structure defined a priori and posteriori

Using a priori information regarding the classification of the cassava accessions in improved varieties and landraces, the AMOVA was analyzed, and results demonstrated that the most significant differences in the molecular variance of the SNPs exist within the groups of classification (51.61 %) and in the individuals themselves (48.27 %). Only 0.12 % of the molecular variance was attributed to differences between the improved genotypes and the landraces (Table 4). Similarly, regarding the classification based on the breeding patterns, when the germplasm was classified a priori on the basis of their geographical origin, results also demonstrated that most of the molecular variation exists within the groups and individuals (50.60 and 48.27 %, respectively). Only 1.13 % of the variation appeared to be caused by differences in the population structure for this grouping pattern. When the groups identified by the Structure and DAPC methods, most variability was found among individuals (91 %), with low variability between and within groups.

The global F_{ST} estimates of 0.20 for the improved genotypes and 0.15 for the improved landraces confirm the lack of difference between the clusters formed a priori based on information about the breeding pattern of the accessions (Table 5). Considering the geographic origin of the accessions as a priori information for the classification of the genotypes, the F_{ST} values varied from 0.01 to 0.53, whereas the accessions from Brazil (midwest, northeast, north and southeast regions) and Nigeria presented very similar F_{ST} values (0.18–0.25). Furthermore, there was less differentiation for the accessions obtained from the southern region of

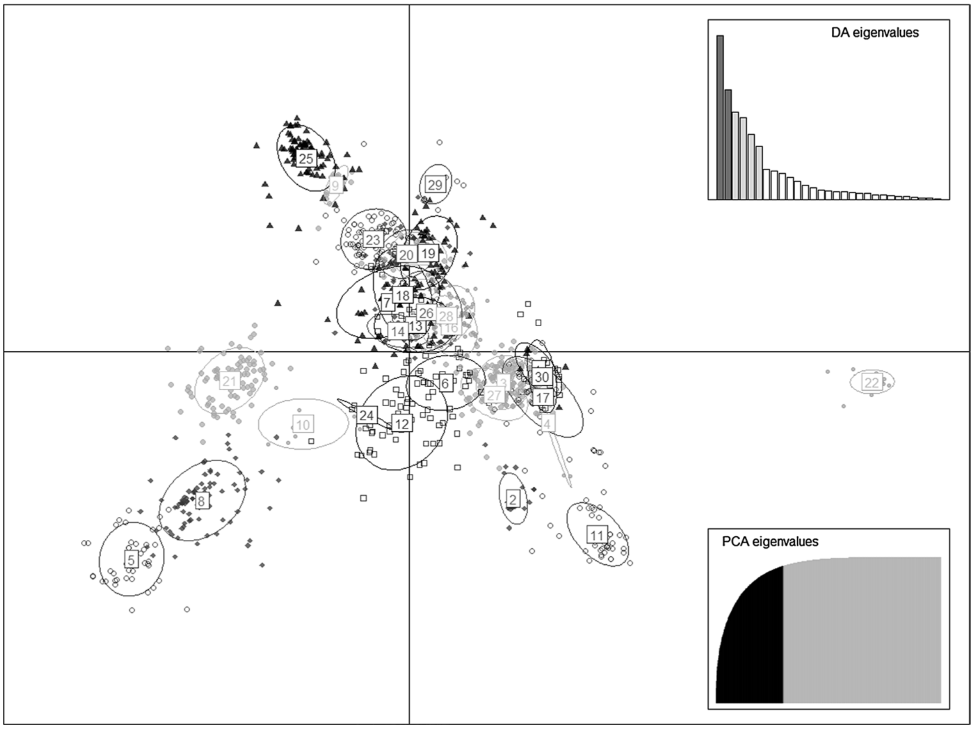

Fig. 6 *Dispersion plots* of the first and second principal component of the DAPC based on the analysis of 1,280 cassava accessions using 402 SNP markers. The clusters, represented by *numbers* and *different colors*, represent the analyzed accessions

Brazil ($F_{ST} = 0.13$) and Venezuela ($F_{ST} = 0.01$), whereas the accessions from Colombia showed greater variation in relation to the remaining genotypes ($F_{ST} = 0.39$). Therefore, both patterns of a priori clustering suggest a limited and barely significant differentiation between the subpopulations and reveal a lack of clearly defined clusters.

Considering the cluster formed by the DAPC and Bayesian analyses, a greater differentiation of the groups with F_{ST} values varying from 0.28 to 0.58 for DAPC and from 0.29 to 0.58 in the Bayesian analysis was observed; both of these methods produced average estimates of 0.46. Therefore, the clusters based on a posteriori information can be more useful in classifying the cassava germplasm by maximizing the variation among groups and minimizing the variation within groups.

Discussion

Genetic diversity identified by SNP markers

The 396 SNP markers were able to identify genetic diversity in terms of the PIC value varying from 0.002 to 0.375 with an average PIC value of 0.262. The high number of polymorphic SNPs is consistent with the mode of reproduction (open pollination), genetic breeding system (selection of landraces) and the level of genetic variation in *M. esculenta* Crantz. Recently, Ferguson et al. (2012) identified 12 % of monomorphic SNPs when analyzing 53 cassava accessions. Furthermore, the average values of Ho, He and PIC were 0.35, 0.36 and 0.28, respectively; these values are similar to the ones presented in this work where 1,280 accessions were assessed.

The diversity reported in the current study is limited in comparison to other types of markers such as microsatellites, where PIC values vary between 0.09 and 0.86 (Fregene et al. 2003; Raji et al. 2009b; de Bang et al. 2011) even when analyzing a single sample from the cassava germplasm. The bi-allelic nature of the SNP markers explains this difference. In this case, the distribution of both alleles in the most extreme case is 50–50, and, consequently, the values of PIC and He could not be higher than 0.5. In this case, it is possible that a higher number of SNP markers are necessary to have the same level of discrimination in comparison to multiallelic markers such as microsatellites. In fact, Chao et al. (2009) reported that in wheat, the PIC values based on SNP markers were about three times lower than the ones based on microsatellite markers. However, if the objective is to study the intraspecific variation such as the case of the cassava germplasm, SNPs may be more appropriate than microsatellite markers due to their high distribution in the genome and the high degree of automation in the genotyping process.

Another important aspect to consider in genetic analysis is that the number of loci to be assessed is not absolute depending on the objective of the study, the extent of information gained from the marker and the genetic relationship between the individuals under study (the more diverse, the

Table 2 Probability of attributions of the 1,280 cassava accessions in different groups based on the DA of principal components ($K = 30$)

Cluster	Probability of attribution (*a*-score)				
	>0.91	0.71–0.90	0.51–0.70	0.31–0.50	<0.30
1	7	1	0	0	0
2	27	0	0	0	0
3	36	70	45	7	0
4	0	0	4	1	0
5	41	2	1	0	0
6	2	7	16	6	0
7	1	3	6	14	0
8	80	5	1	0	0
9	20	0	0	0	0
10	8	1	0	0	0
11	56	0	0	1	0
12	34	10	3	0	0
13	13	13	1	5	0
14	3	6	3	3	0
15	42	0	0	1	0
16	60	9	3	7	0
17	39	4	2	3	0
18	17	1	0	1	0
19	3	27	18	10	1
20	1	2	19	25	2
21	84	1	0	0	0
22	19	0	0	0	0
23	27	28	20	14	0
24	3	0	0	0	0
25	84	1	0	0	0
26	19	0	0	0	0
27	30	3	3	2	0
28	13	5	9	5	0
29	0	1	14	1	0
30	28	3	3	0	0

fewer number of markers are needed) (Kawuki et al. 2009). In the present study, the 396 polymorphic SNPs were able to detect a high level of genetic structure in the cassava germplasm and therefore seemed sufficient for this type of study.

In terms of r^2, the overall LD measured in our sample was very low (0.016). However, the estimates of LD in the current study are based on a limited number of SNP markers, none positioned on a cassava consensus map. Therefore, our estimates do not provide a high resolution view of the distribution of LD across the cassava genome. A higher density of markers would provide for a more accurate calculation of LD. Nevertheless, the number of SNPs used to understand the genetic variability of cassava seems to be non-redundant once the level of markers with LD > 0.25 was low.

Genetic structure

The main challenges in the analysis of any set of genetic data are related to the identification and separation of homogeneous groups and the identification of quantitative data that support the presence of these clusters (Patterson et al. 2006). Using SNP markers, the level of genetic differentiation, population structure and relationship patterns between a diversified set of the cassava germplasm were investigated using a model based on the analysis of population structure and DAPC. These different methods showed many possible diversity groups, although it was not possible to establish a relationship between the clusters formed based on SNPs and agronomic information of economic importance such as fresh root yield, dry matter content, disease resistance, and adaptation to specific environments due to the lack of information for all of the 1,280 cassava accessions. However, results from other species demonstrate the lack of a clear grouping pattern of the germplasm based on phenotypic data alone (Xia et al. 2004, 2005; Semagn et al. 2012).

Cassava germplasm collections in Brazil are mainly constituted by landraces, and characterization studies have focused on the analysis of morphological descriptors and in the content of cyanogenic compounds in the roots to classify the genotypes as bitter or sweet (Fukuda and Alves 1987). However, the use of morphological descriptors seems insufficient to allow a clear distinction between the cassava accessions due to the limited capacity of discrimination (Benesi et al. 2010; Kawuki et al. 2011). The simplicity of use, color, central lobe shape and growth habit characteristics, enable a clear distinction between many cassava accessions and therefore are used by producers to identify cassava varieties. However, although the characteristics of color may play an important role in the differentiation of cassava varieties, genotypes that are morphologically similar may be classified erroneously (Elías et al. 2001).

Furthermore, the lack of detailed records about the pedigree of the cassava accessions preserved at the Embrapa Cassava and Fruits and the lack of complete passport information with precise georeferenced data increase the difficulty in understanding the genetic variation of the germplasm of this species. In contrast, the molecular characterization of the accessions of the collection using SNP markers provides a powerful alternative to determine the accurate representation of the genetic diversity of the species, to identify gaps in the collection that may be filled and to better understand the complexities of the relationships between cultivars and landraces despite limited passport information.

In general, some characteristics intrinsic to cassavas present difficulties for the development of new cultivars such

Table 3 Distribution of genotypes present in the group defined by the Bayesian analysis in the cluster formed by the DAPC

DAPC Clusters	Clusters of the Bayesian analysis																
	1[a]	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	16.7	3.3
2	–	–	–	–	–	85.7	–	–	–	–	–	–	–	–	–	10.0	–
3	–	84.4	–	–	33.3	–	2.9	–	–	–	–	–	16.0	23.5	–	–	3.3
4	–	–	–	–	–	–	–	–	–	–	–	–	12.0	–	–	–	–
5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
6	1.8	5.2	5.9	–	–	–	–	–	9.8	–	–	–	–	2.9	–	–	3.3
7	2.8	–	–	–	–	–	–	–	–	–	–	–	4.0	–	–	–	–
8	15.6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
9	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
10	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
11	–	1.7	–	–	–	–	–	–	–	–	–	–	–	–	–	–	57.6
12	–	–	–	–	–	14.3	–	–	82.9	–	–	–	–	–	–	6.7	–
13	–	0.9	–	–	–	–	–	–	2.4	–	–	–	–	–	–	–	2.2
14	0.9	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
15	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
16	–	–	–	–	–	–	–	–	2.4	63.0	–	1.8	–	–	100.0	–	–
17	–	1.7	–	–	–	–	2.9	–	–	–	–	–	–	14.7	–	–	28.3
18	–	0.9	5.9	–	–	–	–	–	–	–	–	–	–	–	–	–	–
19	0.9	–	–	–	–	–	–	–	–	–	–	–	4.0	–	–	–	–
20	–	–	17.7	–	–	–	–	–	–	–	–	–	4.0	–	–	–	–
21	78.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
22	–	–	–	–	–	–	–	100.0	–	–	–	–	–	–	–	–	–
23	–	–	41.2	–	66.7	–	–	–	–	22.2	–	–	–	55.9	–	3.3	–
24	–	0.9	–	–	–	–	–	–	–	–	9.1	–	–	–	–	–	–
25	–	–	23.5	100.0	–	–	–	–	2.4	11.1	90.9	98.3	–	–	–	–	–
26	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	63.3	–
27	–	0.9	–	–	–	–	–	–	–	–	–	–	4.0	–	–	–	1.1
28	–	2.6	–	–	–	–	2.9	–	–	–	–	–	56.0	2.9	–	–	–
29	–	–	5.9	–	–	–	–	–	–	3.7	–	–	–	–	–	–	–
30	–	0.9	–	–	–	–	91.4	–	–	–	–	–	–	–	–	–	1.1

DAPC Clusters	Clusters of the Bayesian analysis																
	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3	–	–	–	–	4.2	–	–	–	–	12.5	–	76.0	–	–	–	–	4.7
4	–	–	–	–	4.2	–	–	–	–	–	–	–	–	–	–	–	–
5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	39.6	–
6	2.1	–	2.1	–	4.2	8.0	–	–	–	–	–	8.0	–	–	–	–	9.3
7	2.1	–	12.8	–	2.1	–	6.3	7.1	2.4	12.5	21.1	–	–	15.4	–	–	–
8	–	–	–	100.0	–	–	–	–	–	–	–	–	–	–	–	56.8	–
9	–	–	–	–	–	–	–	–	–	–	–	–	–	–	100.0	–	–
10	–	–	–	–	–	–	–	–	–	12.5	–	–	–	–	–	0.9	16.3
11	–	–	–	–	–	–	–	–	–	–	–	4.0	–	–	–	–	–
12	–	–	–	–	–	–	–	–	–	25.0	–	–	–	–	–	2.7	4.7
13	–	–	4.3	–	4.2	4.0	3.1	–	–	12.5	–	–	–	–	–	–	48.8
14	–	–	2.1	–	–	–	34.4	–	2.4	–	–	2.0	–	–	–	–	–

Table 3 continued

DAPC Clusters	Clusters of the Bayesian analysis																
	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
15	87.2	9.1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
16	–	18.2	–	–	–	–	–	–	–	–	–	2.0	–	–	–	–	–
17	–	27.3	–	–	4.2	–	12.5	–	–	12.5	–	2.0	–	–	–	–	–
18	–	–	–	–	–	64.0	–	–	–	–	–	–	–	–	–	–	2.3
19	–	13.6	40.4	–	2.1	–	–	71.4	17.1	–	15.8	–	6.7	23.1	–	–	–
20	–	22.7	17.0	–	–	–	21.9	17.9	7.3	–	47.4	4.0	13.3	30.8	–	–	–
21	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
22	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
23	2.1	9.1	10.6	–	–	–	15.6	3.6	68.3	–	–	–	13.3	30.8	–	–	9.3
24	–	–	–	–	–	–	–	–	–	12.5	–	–	–	–	–	–	–
25	–	–	–	–	–	24.0	–	–	2.4	–	–	–	–	–	–	–	–
26	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
27	–	–	–	–	72.9	–	–	–	–	–	–	–	–	–	–	–	–
28	6.4	–	8.5	–	2.1	–	3.1	–	–	–	5.3	2.0	–	–	–	–	4.7
29	–	–	2.1	–	–	–	3.1	–	–	–	10.5	–	66.7	–	–	–	–
30	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

[a] Percentage of distribution (%) of cassava accessions in the 30 DAPC groups. For example, 1.8, 2.8, 15.6, 0.9, 0.9 and 78.0 % of the accessions grouped in cluster 1 of the Bayesian analysis were allocated in clusters 6, 7, 8, 14, 19 and 21 of the DACP, respectively

Table 4 AMOVA considering: (a) two groups: improved varieties and landraces; (b) nine groups with respective geographic origins: Brazil (midwest, northeast, north, southeast regions and south, unknown region), Colombia, Nigeria and Venezuela; (c) thirty-four groups according to structure; and (d) thirty groups according to DAPC

Source of variation	Improved varieties and landraces			Geographic origins		
	Degree of freedom	Mean square	% of variation	Degree of freedom	Mean square	% of variation
Between populations	1	376.49	0.12	8	480.98	1.13
Within populations	1,278	138.08	51.61	1,271	136.11	50.60
Within individuals	1,280	128.94	48.27	1,280	128.94	48.27
Source of variation	Structure groups			DAPC groups		
	Degree of freedom	Mean square	% of variation	Degree of freedom	Mean square	% of variation
Between populations	33	173.286	0.01	29	192.161	0.02
Within populations	1,246	72.800	0.08	1,250	72.684	0.07
Within individuals	1,280	62.630	0.91	1,280	62.630	0.91

as long productive cycles, high level of genetic load in the varieties used in crosses, lack of clearly defined cassava populations hindering the efficient modification of allelic frequencies and the highly heterozygotic nature of the crop such that the dominance effects largely contribute to the performance of the genotypes under evaluation, which has led to much effort being put into the production of seeds in controlled crosses. However, the need for the quick expansion of the cassava crop has led breeders to register some landraces as varieties after a series of evaluations of the productive potential and stability in different cassava-producing regions. Therefore, one of the hypotheses that may explain the lack of clusters with clear distinction between improved cassava varieties and landraces in Brazil is the fact that many landraces are considered improved genotypes, and at the same time, many cultivars were unduly introduced into the germplasm banks at many institutions.

Previous studies carried out with cassavas were not able to identify any relationship between the genetic structure estimated with AFLP markers and the classification of the accessions in sweet and bitter cassavas, which is most likely due to the polygenic nature of cyanogenic compounds, which is strongly influenced by genetic and environmental effects (Benesi et al. 2010). In addition, when a small portion of the

Table 5 Genetic differentiation (F_{ST}) of cassava accessions according to a priori information (breeding pattern and geographic origin) and a posteriori information based on the DAPC and Bayesian analysis using the Structure v.2.3. software

A priori cluster				A posteriori cluster							
Breeding pattern	F_{ST}	Geographic origin	F_{ST}	$K = 30$ (DACP)				K = 34 (Bayesian analysis)			
				Cluster	F_{ST}	Cluster	F_{ST}	Cluster	F_{ST}	Cluster	F_{ST}
Improved	0.20	Brazil/midwest	0.19	1	0.39	18	0.54	1	0.41	18	0.49
Landraces	0.15	Brazil/northeast	0.18	2	0.54	19	0.36	2	0.42	19	0.31
		Brazil/northeast	0.18	3	0.47	20	0.54	3	0.49	20	0.44
		Brazil/southeast	0.23	4	0.53	21	0.39	4	0.48	21	0.53
		Brazil/south	0.13	5	0.52	22	0.41	5	0.42	22	0.52
		Brazil/unknown	0.53	6	0.28	23	0.47	6	0.49	23	0.55
		Colombia	0.39	7	0.51	24	0.53	7	0.49	24	0.58
		Nigeria	0.25	8	0.45	25	0.40	8	0.50	25	0.51
		Venezuela	0.01	9	0.52	26	0.46	9	0.43	26	0.35
				10	0.36	27	0.36	10	0.50	27	0.29
				11	0.53	28	0.39	11	0.53	28	0.40
				12	0.38	29	0.39	12	0.53	29	0.29
				13	0.55	30	0.52	13	0.46	30	0.54
				14	0.40			14	0.48	31	0.35
				15	0.54			15	0.46	32	0.54
				16	0.42			16	0.48	33	0.46
				17	0.58			17	0.51	34	0.47

cassava germplasm (94 accessions) at Embrapa Cassava and Fruits was assessed, Carvalho and Schaal (2001) did not identify a high correlation in the clusters formed from the analysis using morphological descriptors and RAPD markers and SSR-primed PCR; this could be related to the low genetic variability of the accessions analyzed. Furthermore, this small sample size may not have reflected the real cassava diversity.

Fregene et al. (2003) argue that from the time of the dispersion of cassavas to the rest of the world, there were many founding events with concomitant effects resulting in the reduction of diversity and an increase in genetic differentiation. In fact, the reproduction system (asexual) of the species itself may favor reduction of the genetic diversity over time due to systemic accumulation of pathogens and preference of cassava producers for specific types of varieties with characteristics of agronomic importance such as content of cyanogenic compounds, (sweet and bitter cassavas), maturation period, root quality for fresh consumption and processing, resistance to diseases, drought tolerance and fitness for intercropping with other crops; these qualities are especially important for the semi-arid regions in the northeast region of Brazil. However, the heterozygotic state and the allogamous nature of cassavas has led to a large number of seeds being produced in commercial farms, which in turn results in voluntary seedlings with a completely different genotypic constitution from the original landraces. If there is any competitive advantage of these new genotypes, natural and human selection may produce new varieties, which artificially maintains a high level of genetic diversity (Doyle et al. 2001). Furthermore, the accumulation of somatic mutations and the transmission of these mutations to subsequent generations by vegetative propagation is a multiplication factor of genetic polymorphisms (Elías et al. 2001; Sardos et al. 2008). All of these influences may act in favor of maintaining and creating genetic variability in cassavas as proven by the high inter- and intravarietal diversity observed in the areas of cassava producers (Elías et al. 2000).

The population structure analysis of the Brazilian cassava germplasm using SNP markers showed that regardless of the methodology used to determine the diversity, most clusters presented a high percentage of plants (53 and 22 % in the Bayesian and DAPC analysis, respectively) with ancestry in more than one cluster. This reflects the large genetic variability of the species, which could be due to the generation of voluntary seedlings in producer areas and a significant exchange of germplasm, which certainly occurred during the long process of domestication of cassavas in Brazil when cassava producers exchanged germplasms.

Bayesian analysis and multivariate ordination

The Bayesian and DAPC analysis presented consistent results regarding the detection of a complex genetic structure between the cassava accessions using SNP markers. However, there is discrepancy in the indication of the number of diversity groups, which could be related to the approach of the method. According to Jombart et al. (2010), the type of population structure influences the precision of the method as to the definition of the number of clusters, whereas the inferences in structured populations in the island model are more precise than in continuous populations, which seems to be the case for the cassava germplasm. Nevertheless, even when the real K value is not identified, the number of clusters usually inferred is very close to the real value. However, discrepancies as to the number of clusters identified with these methodologies seem to be common. In a recent study analyzing weed beets, Arnaud et al. (2011) found the adequate number of groups required to represent the genetic variability of this species to be $K = 3$ based on Bayesian analysis and $K = 5$ based on DAPC analysis.

Some discrepancies were also observed regarding the clustering of the cassava accessions using both methodologies. This was especially true for accessions with low ancestry in a certain cluster. However, the cluster suggested by the DAPC analysis seemed to be more consistent for presenting greater probability of the allocation of the accessions within the clusters (*a*-score) (Table 3). In addition, the analysis of the population structure of the cassava germplasm involves samples from genotypes from different genetic origins in different and unknown proportions, which leads to linkage disequilibrium between non-linked loci (Ersoz et al. 2007). Consequently, this may increase the rate of false-positives, which are statistically associated to the characteristics analyzed without actually being involved in the phenotypic variation. Furthermore, Jombart (2008) have suggested that the cluster calculated by Bayesian analysis may be inappropriate when the populations are structured by the cline. In these cases, the DAPC may identify different genetic structures including clines without having to meet the assumptions of the Bayesian approaches.

Semagn et al. (2012) evaluated the diversity in corn germplasm at CIMMYT and the relationships between the elite genotypes using SNP markers and different methodological approaches for the classification of the genotypes [models based on structural analysis, principal components analysis, cluster neighbor-joining and discriminating analysis (DA)]. All of these approaches indicated the presence of three major clusters, which is in agreement with the pedigree information. However, the cluster analysis showed low agreement with the other methods as to the attribution of genotypes to their respective groups, but according to these authors, the clustering pattern of the principal components of the models based in the analysis of the structure and DA was more reliable in the attribution of genotypes to specific clusters.

The clear definition of the number of clusters in the cassava germplasm using the DAPC strategy consists of a methodological advance, and until now, only hierarchical methods or analysis of principal components have been used for genetic studies in cassava. Therefore, the formation of clusters has always been performed empirically without defined criteria for the classification and ordination of genotypes.

This type of analysis is more advantageous in comparison to other classical approaches such as Bayesian approaches that are implemented in the Structure software (Pritchard et al. 2000) for the following reasons: (a) exploratory characters do not require adaptations to the genetic structure such as Hardy–Weinberg equilibrium or the absence of linkage disequilibrium, (b) PCA is not computationally intensive and can be used for a large number of data sets, and (c) PCA allows one to approach complex questions such as the identification of adaptation and the association of genetic variability to specific environmental conditions (Jombart et al. 2009). However, PCA does not allow one to infer the formation of groups of diversity and thus requires an a priori definition to study the population structure. Furthermore, PCA is not appropriate for providing a clear image of the variation between populations. These shortcomings can be overcome by DA, which defines a model in which synthetic variables with the partitioned genetic variation within and between clusters are created to maximize genetic variation within clusters (Jombart et al. 2010).

Alterations in the molecular variance

Partitioning of AMOVA demonstrates variations within and between the germplasm components. AMOVA analysis showed that most of the diversity in the cassava accessions exists within the populations and individuals. Only 0.12 and 1.13 % of the molecular variance is due to differentiation of the germplasm based on the a priori classification of breeding patterns and geographic origins, respectively. In contrast, the theoretical cluster formed by the DAPC and the Bayesian analysis allowed the maximization of the genetic differentiation between the clusters (average $F_{ST} = 0.46$). Comments of the same nature related to low genetic differentiation with a priori information were observed in other crops. In corn, Semagn et al. (2012) reported low F_{ST} (0.012) values between two heterotic groups defined a priori based on the knowledge of the combination capacity in field experiments. However, considering the defined population structure based on a posteriori analysis using SNP markers, the F_{ST} values were much higher contributing to the maximization of differences between the groups ($F_{ST} = 0.113$ for $K = 2$ and 0.118 for $K = 3$).

A greater differentiation of the groups (F_{ST} ranging from 0.28 to 0.58 for DAPC and from 0.29 to 0.58 in the Bayesian analysis), considering the clusters based on a posteriori information, were observed. Even though bi-allelic markers such as SNPs were used, the estimation of genetic differentiation of the Brazilian cassava germplasm is higher than those reported from Africa and the Americas ($F_{ST} = 0.09$—Fregene et al. 2003; $F_{ST} = 0.12$—Lokko et al. 2006; $F_{ST} = 0.03$—Montero-Rojas et al. 2011), with estimations based on microsatellite markers. Futhermore, Fregene et al. (2003) reported a large variation in the genetic differentiation of the germplasm due to its origin, once estimations higher than F_{ST} (0.27) were observed between accessions from Nigeria and Guatemala; genotypes with different evolutionary histories. However, the use of different markers and germplasm accessions may lead to discrepancies in the results as to the estimation of genetic differentiation between populations. Using AFLP and microsatellites to analyze a collection of African cassava landraces and elite cultivars, Raji et al. (2009a) found a high F_{ST} in the landrace population for SSR (0.746) and AFLP markers (0.656). This high genetic differentiation between landraces and elite cultivars suggest that geographical or regional variation could be responsible for most of the genetic differentiation observed.

In general, considering only the effects of the molecular markers, some works have indicated greater power of detection of genetic differentiation between populations using SNPs in comparison to microsatellite markers. In citrus, Garcia-Lor et al. (2013) investigated the average molecular differentiation between 'true citrus fruit trees' using three types of markers (SSRs, SNPs, indels). The F_{ST} estimates between the eight basic taxa confirmed the high degree of stratification in differentiated taxa. However, the levels of diversity revealed by the three types of markers were quite different (0.596, 0.644 and 0.392 for indel, SNP and microsatellites markers, respectively). In *Populous tremula* the genetic differentiation across populations using 206 SNPs was 0.017, which is very close to estimates based on microsatellite data ($F_{ST} = 0.015$) (Hall et al. 2007; Ma et al. 2010).

Many authors have tried to establish an association between geographic origin and genetic divergence in cassavas; however, the results seem to be quite contradictory. Devi et al. (2009) did not find any pattern of distribution of cassava accessions according to different eco-geographic regions of collection, indicating that the geographic and genetic diversities were not related. Observations of the same nature were reported by Ferguson et al. (2012) when 53 cassava accessions were analyzed using SNP markers, and no structuralization was observed based on the geographic origin, which was particularly evident in the accessions from the neotropical regions and Africa; however, substructures between germplasms from southeast, southwest and central and western Africa, were observed. This indicates a strong international exchange of germplasm whose geographic origins need to be preserved to understand the genetic variation of the species and how the genotype flow occurred in different countries. The similarity of Asian and African cassava accessions with the American germplasm especially from Latin America reflects the movement of germplasm from South America to these regions (Kawuki et al. 2009). Benesi et al. (2010) demonstrated the potential use of AFLP markers in grouping most cassava accessions according to their geographic region. However, the low number of accessions used (78) and the regional adaptations within the accessions certainly contributed to this type of structure.

Many studies using microsatellite markers are in agreement as to the weak relationship between the diversity and the genetic origin of cassavas. Using 67 SSR loci and 283 cassava landraces from Africa (Tanzania and Nigeria) and the neotropical region of Brazil (Brazil, Colombia, Peru, Venezuela, Guatemala, Mexico and Argentina), Fregene et al. (2003) reported low levels of differentiation between the samples from each country; however, the authors observed enough differentiation to reveal an accentuated substructure of the African landraces to allow the separation of the African from the neotropical landraces.

The known natural and artificial selection of cassava seedlings generated spontaneously in agricultural farms leads to the establishment of new cassava landraces (Elías et al. 2000, 2001). Furthermore, another agricultural practice that increases genetic diversity is the exchange of planting material between producers (Elías et al. 2000). This activity certainly occurred and still occurs in many Brazilian states due to (a) different cycles of the expansion of the crop and (b) a lack of propagative material in certain regions due to diseases or even stakes due to intense drought, which often occurs in the northeast region of Brazil and forces producers to purchase propagative material from other regions. Consequently, name changes of the imported variety are common, which leads to confusion of synonyms of the same material or leads to a single genotype with different names or different genotypes with the same name. This practice makes the attribution of the geographic origin of the material very difficult when collections occur. Therefore, the intense movement of stakes in the national territory and changes in names of landraces and varieties may contribute to fact that the clusters formed by the Bayesian and DAPC analyses do not show any correlation with the geographic origin of the accessions from the germplasm at Embrapa Cassava and Fruits.

Future perspectives

The studies elucidating the characterization of this crop need to be intensified especially considering the fact that the germplasm is the basic raw material used by the breeder to introduce new genes and important economic and agronomic attributes for the development of new cassava cultivars for better growth potential. The discovery of the high genetic diversity and structure of the cassava germplasm will be extremely useful in genetic breeding because it will allow for the following: (1) assisting the selection of parental combinations for crosses and development of segregating populations aimed at the development of genotypes with maximum genetic variability for genetic mapping or other types of selection, (2) describing heterotic groups, (3) determining the level of genetic variability in certain subgroups of selected germplasms for specific characteristics, and (4) estimating the possible loss of genetic diversity during programs of selection or germplasm conservation.

Depending on the objective of the genetic breeding program, to select the best parents for crosses and to attribute genotypes to certain heterotic groups, breeders use different data, especially phenotypic and pedigree information. However, considering that such data are not available for all accessions from the germplasm bank at Embrapa Cassava and Fruits, the SNP markers is the only information available for the classification of the cassava germplasm into heterotic groups.

In the near future, the acquisition of information regarding phenotypic data on characteristics of interest in the entire cassava germplasm may lead to a better classification of this germplasm. Together with the structure of the cassava germplasm bank in 30 diversity groups based on DACP analysis (Fig. 6), this phenotypic and molecular information will open new opportunities for the exploration of the heterotic effects between the groups formed. In this case, progenitors with high genetic diversity within groups may be used in diallelic crosses aimed at the implementation of backcross selection or even for maximizing the potential to develop superior populations for clonal competition tests.

Author contributions E.J.O., C.F.F., V.S.S. and O.N.J. conceived, designed and analyzed the data of the experiments. G.A.F.O. and M.S.S. performed the experiments.

Acknowledgments The authors would like to thank Empresa Brasileira de Pesquisa Agropecuária (Embrapa) for the financial support. The authors gratefully acknowledge the support through fellowships from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical standards We declare that the experiments comply with the current Brazilian laws in which they were performed. In addition, we declare that the authors do not have any competing financial interests, including funding support, employment and personal financial interests.

References

Arnaud J-F, Cuguen J, Fénart S (2011) Metapopulation structure and fine-scaled genetic structuring in crop-wild hybrid weed beets. Heredity 107:395–404

Benesi IRM, Labuschagne MT, Herselman L, Mahungu N (2010) Ethnobotany, morphology and genotyping of cassava germplasm from Malawi. J Biol Sci 10:616–623

Botstein D, White RL, Skolmick H, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphism. Am J Med Genet 32:314–331

Carvalho LJCB, Schaal BA (2001) Assessing genetic diversity in the cassava (*Manihot esculenta* Crantz) germplasm collection in Brazil using PCR-based markers. Euphytica 120:133–142

Chagné D, Batley J, Edwards D, Forster JW (2007) Single nucleotide polymorphisms genotyping in plants. In: Oraguzie N, Rikkerink E, Gardiner S, De Silva H (eds) Association mapping in plants. Springer, New York, pp 77–94

Chao S, Zhang W, Akhunov E, Sherman J, Ma Y, Luo M, Dubcovsky J (2009) Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. Mol Breed 23:23–33

Cieslarová J, Hýbl M, Griga M, Smýkal P (2012) Molecular analysis of temporal genetic structuring in pea (*Pisum sativum* L.) cultivars bred in the Czech Republic and in former Czechoslovakia since the Mid-20th century. Czech J Genet Plant Breed 48:61–73

de Bang TC, Raji AA, Ingelbrecht IL (2011) A multiplex microsatellite marker kit for diversity assessment of large cassava (*Manihot esculenta* Crantz) germplasm collections. Plant Mol Biol Rep 29:655–662

Devi AKB, Gin G, Rahul Y (2009) Assessment of genetic diversity in cassava (*Manihot esculenta* Crantz) germplasm. J Root Crops 35:108–111

Dixon AGO, Bandyopadhyay R, Coyne D, Ferguson M, Shaun R, Ferris B, Hanna R, Hughes J, Ingelbrecht I, Legg J, Mahungu N, Manyong VMD, Neuenschwander P, Whyte J, Hartmann P, Ortiz R (2003) Cassava: from poor farmers's crop to pacesetter of African rural development. Chron Hortic 43:8–15

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus 12:13–15

Doyle M, Pujol B, Elias M (2001) Ecology and genetics of populations of cassava managed by Amazonian Amerindians: or, how to maximize the benefits and minimize the costs of outcrossing in a vegetatively propagated crop. In: Fauquet C, Taylor NN (eds) Proc 5th int scientific meeting of the cassava biotechnology network. Danforth Plant Science, Missouri

Elías M, Panaudà O, Robertà T (2000) Assessment of genetic variability in a traditional cassava (*Manihot esculenta* Crantz) farming system, using AFLP markers. Heredity 85:219–230

Elías M, Penet L, Vindry P, Mckey D, Panaud O, Robert T (2001) Unmanaged sexual reproduction and the dynamics of genetic diversity of a vegetatively propagated crop plant, cassava (*Manihot esculenta* Crantz), in a traditional farming system. Mol Ecol 10:1895–1907

Ersoz ES, Yu J, Buckler ES (2007) Applications of linkage disequilibrium and association mapping in crop plants. In: Varshney RK, Tuberosa R (eds) Genomics-assisted crop improvement. Springer, Dordrecht, pp 97–119

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. Mol Ecol 14:2611–2620

FAO (2012) FAOSTAT database. FAO, Italy (Accessed 26 June 2012)

Ferguson ME, Hearne SJ, Close TJ, Wanamaker S, Moskal WA, Town CD, Young J, Marri PR, Rabbi IY, Villiers EP (2012) Identification, validation and high-throughput genotyping of transcribed gene SNPs in cassava. Theor Appl Genet 124:685–695

Fregene MA, Suarez M, Mkumbira J, Kulembeka H, Ndedya E, Kulaya A, Mitchel S, Gullberg U, Rosling H, Dixon AG, Dean R, Kresovich S (2003) Simple sequence repeat marker diversity in cassava *landraces*: genetic diversity and differentiation in an asexually propagated crop. Theor Appl Genet 107:1083–1093

Fukuda WMG, Alves AAC (1987) Banco ativo de germoplasma de mandioca do Centro Nacional de Mandioca e Fruticultura. Rev Bras Mandioca 6:65–97

Garcia-Lor A, Curk F, Snoussi-Trifa H, Morillon R, Ancillo G, Luro F, Navarro L, Ollitrault P (2013) A nuclear phylogenetic analysis: SNPs, indels and SSRs deliver new insights into the relationships in the 'true citrus fruit trees' group (Citrinae, Rutaceae) and the origin of cultivated species. Ann Bot 111:1–19

Hall D, Luquez V, Garcia VM, St Onge KR, Jansson S, Ingvarsson PK (2007) Adaptive population differentiation in phenology across a latitudinal gradient in European aspen (*Populus tremula* L.): a comparison of neutral markers, candidate genes and phenotypic traits. Evolution 61:2849–2860

Jombart T (2008) Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405

Jombart T, Pontier D, Dufour AB (2009) Genetic markers in the playground of multivariate analysis. Heredity 102:330–341

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:94

Kawuki RS, Ferguson M, Labuschagne M, Herselman L, Kim DJ (2009) Identification, characterisation and application of single nucleotide polymorphisms for diversity assessment in cassava (*Manihot esculenta* Crantz). Mol Breed 23:669–684

Kawuki RS, Ferguson M, Labuschagne MT, Herselman L, Orone J, Ralimanana I, Bidiaka M, Lukombo S, Kanyange MC, Gashaka G, Mkamilo G, Gethih J, Obiero H (2011) Variation in qualitative and quantitative traits of cassava germplasm from selected national breeding programmes in sub-Saharan Africa. Field Crops Res 122:151–156

Lebot V (2009) Tropical root and tuber crops: cassava, sweet potato, yams and aroids. Crop Production Science in Horticulture. CABI, Wallingford, UK

Liu K, Muse S (2005) PowerMarker: integrated analysis environment for genetic marker data. Bioinformatics 21:2128–2129

Lokko Y, Dixon A, Offei S, Danquah E, Fregene M (2006) Assessment of genetic diversity among African cassava *Manihot esculenta* Grantz accessions resistant to the cassava mosaic virus disease using SSR markers. Genet Resour Crop Evol 53:1441–1453

Ma X-F, Hall D, Onge KRS, Jansson S, Ingvarsson PK (2010) Genetic differentiation, clinal variation and phenotypic associations with growth cessation across the *Populous tremula* photoperiodic pathway. Genetics 186:1033–1044

Montero-Rojas M, Correa AM, Siritunga D (2011) Molecular differentiation and diversity of cassava (Manihot esculenta) taken from 162 locations across Puerto Rico and assessed with microsatellite markers. AoB Plants plr010 doi:10.1093/aobpla/plr010

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323

Nweke FI, Spencer DSC, Lynam JK (2002) The cassava transformation. Africa's best kept secret. Michigan State University, East Lansing

Olsen KM (2004) SNPs, SSRs and inferences on cassava's origin. Plant Mol Biol 56:517–526

Patterson N, Price AL, Reich D (2006) Population structure and Eigen analysis. PLoS Genet 2:e190

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in excel. Population genetic software for teaching and research. Mol Ecol Notes 6:288–295

Prakash CS, He G, Jarret RL (1996) DNA marker-based study of genetic relatedness in United States sweet potato cultivars. J Am Soc Hortic Sci 121:1059–1062

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Pritchard JK, Wen X, Falush D (2007) Documentation for structure software: version 2.2. http://pritch.bsd.ichicago.edu/software. Accessed 12 April 2012

R Development Core Team (2010). R: A language and environment for statistical computing, reference index version 2.12.1. R foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0 http://www.R-project.org

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100

Raji AAJ, Fawole I, Gedil M, Dixon AGO (2009a) Genetic differentiation analysis of African cassava (*Manihot esculenta)* landraces and elite germplasm using amplified fragment length polymorphism and simple sequence repeat markers. Ann Appl Biol 155:187–199

Raji AAJ, Anderson JV, Kolade OA, Ugwu CD, Dixon AGO, Ingelbrecht IL (2009b) Gene-based microsatellites for cassava (*Manihot esculenta* Crantz): prevalence, polymorphisms, and cross-taxa utility. BMC Plant Biol 9:118

Sardos J, Mackey E, Duval MF, Malapa R, Noyer JL, Lebot V (2008) Evolution of cassava (*Manihot esculenta* Crantz) after recent introduction into a South Pacific Island system: the contribution of sex to the diversification of a clonally propagated crop. Genome 51:912–921

Schlotterer C (2004) The evolution of molecular markers—just a matter of fashion? Nat Rev Genet 5:63–69

Semagn K, Magorokosho C, Vivek BS, Makumbi D, Beyene Y, Mugo S, Prasanna BM, Warburton ML (2012) Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. BMC Genom 13:113

Xia XC, Reif JC, Hoisington DA, Melchinger AE, Frisch M, Warburton ML (2004) Genetic diversity among CIMMYT maize inbred lines investigated with SSR markers: I. Lowland tropical maize. Crop Sci 44:2230–2237

Xia XC, Reif JC, Melchinger AE, Frisch M, Hoisington DA, Beck D, Pixley K, Warburton ML (2005) Genetic diversity among CIMMYT maize inbred lines investigated with SSR markers: II. Subtropical, tropical mid altitude, and highland maize inbred lines and their relationships with elite US and European maize. Crop Sci 45:2573–2582