GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Development of a cassava core collection based on single nucleotide polymorphism markers

E.J. Oliveira[1], C.F. Ferreira[1], V.S. Santos[1] and G.A.F. Oliveira[2]

[1]Embrapa Mandioca e Fruticultura, Cruz das Almas, BA, Brasil
[2]Universidade Federal do Recôncavo da Bahia, Cruz das Almas, BA, Brasil

Corresponding author: E.J. Oliveira
E-mail: eder.oliveira@embrapa.br

Genet. Mol. Res. 13 (3): 6472-6485 (2014)
Received July 23, 2013
Accepted January 13, 2014
Published August 25, 2014
DOI http://dx.doi.org/10.4238/2014.August.25.11

ABSTRACT. Single nucleotide polymorphism (SNP) markers were used in the largest cassava (*Manihot esculenta* Crantz) germplasm collection from Brazil to develop core collections based on the maximization strategy. Subsets with 61, 64, 84, 128, 256, and 384 cassava accessions were selected and named PoHEU, MST64, PoRAN, MST128, MST256, and MST384, respectively. All the 798 alleles identified by 402 SNP markers in the entire collection were captured in all core collections. Only small alterations in the diversity parameters were observed for the different core collections compared with the complete collection. Because of the optimal adjustment of the validation parameters representative of the complete collection, the absence of genotypes with high genetic similarity and the maximization of the genetic distances between accessions of the PoHEU core collection, which contained 4.7% of the accessions of the complete collection, maximized the genetic conservation of this important cassava collection. Furthermore, the development of this core collection will allow concentrated efforts toward future characterization and agronomic evaluation of accessions to maximize the diversity and genetic gains in

cassava breeding programs.

Key words: Breeding; Genetic diversity; Germplasm; Core subset; *Manihot esculenta* Crantz; Molecular markers

INTRODUCTION

Cassava (*Manihot esculenta* Crantz) is a species that produces starchy roots with high adaptability to different environments, and it is considered to be a basic food for more than 800 million people, most of whom are in Sub-Saharan Africa (Lebot, 2009). In Brazil, especially Northeast region, cassava plays an important role in food security because of its ability to grow under adverse environmental conditions (poor soils and low water availability) where other crops cannot be cultivated. In contrast, in mid-south Brazil, cassava has a strong industrial appeal. Although widely cultivated in tropical and subtropical regions of Africa, Asia, and Latin America, cassavas are native to South America.

One of the largest cassava germplasm collections in Latin America belongs to Embrapa Mandioca e Fruticultura (Bahia, Brazil), with more than 1300 accessions maintained *in vivo*. This variability is represented mostly by creole varieties that were selected naturally or by producers. Because cassava is an allogamous crop, its propagation is mainly carried out vegetatively through cuttings. Therefore, a cassava germplasm may be maintained vegetatively *in situ* and *ex situ* in the field and *in vitro* in the laboratory or by botanical seeds. However, improved cassava varieties and landraces that are maintained in the Cassava Germplasm Bank (CGB) at Embrapa are almost exclusively kept as *in vitro* plantlets or in clonal forms in the field.

The cassava germplasm collections are important to the preservation of genetic variability, allowing the development of new varieties with drought tolerance, disease resistance, and better starch quality and yield. According to Belaj et al. (2012), germplasm banks are responsible for maintaining accessions, documentation, evaluation, and making the information regarding the genetic resources of the species available for effective use in plant breeding. However, maintaining a species that propagates vegetatively, such as cassava, is very laborious, thus limiting the size of the collection. Furthermore, there is a gap between the diversity of the cassava germplasm collection and its effective use in the development of new varieties.

In general, the high number of accessions, the high maintenance cost, and the lack of complete information regarding the diversity of the accessions hinder successful use of the genetic potentials of most germplasm collections (Brown, 1989a,b; van Hintum et al., 2000; Bhattacharjee et al., 2012). However, the management of large germplasm collections may be improved by the use of sub-samples that represent the maximum variability of the species, which are also known as core collections.

The significant advances in the evaluation of genetic diversity in germplasm were followed by the development of different methodological approaches to guide the development of core collections (Schoen and Brown, 1993; Franco et al., 2005, 2006). The maximization strategy (M) that maximizes the number of alleles in each locus (Schoen and Brown, 1993) has been used in many studies to maintain the different alleles of a collection, eliminating redundancy and capturing most of the genetic diversity for a restricted number of accessions.

Core collections have been developed for many species (McKhann et al., 2004; Franco et al., 2006; Richards et al., 2009; Belaj et al., 2012); however, the first cassava core collection was

developed by the International Center for Tropical Agriculture (CIAT) with approximately 630 accessions (Hershey et al., 1994). Recently, Bhattacharjee et al. (2012) developed a cassava core collection that included 22.6% of the 1890 accessions that were maintained in the germplasm bank at the International Institute of Tropical Agriculture (IITA) using 40 morpho-agronomic descriptors.

Cordeiro et al. (1995) provided the general criteria that should be used to develop a Brazilian core collection. However, until now, no core collection has been developed for cassavas using single nucleotide polymorphism (SNP) markers. Because core collections are essentially dynamic because of the addition of new accessions and the availability of additional information (agronomic and/or molecular), the development of different core collections is justified. Furthermore, because the establishment of a cassava core collection may facilitate the evaluation and use of its genetic diversity in genetic breeding, this study aimed to develop a core collection that was representative of the diversity preserved in one of the largest cassava germplasm collections in Latin America at Embrapa Mandioca e Fruticultura.

MATERIAL AND METHODS

Plant materials

One thousand two hundred eighty accessions from the CGB at Embrapa Mandioca e Fruticultura (Cruz das Almas, BA, Brazil), obtained from multiple ecosystems in Brazil, Colombia, Venezuela, and Nigeria, were assayed. This germplasm bank consists of landraces and improved varieties obtained from conventional breeding methods such as crosses and selections, as well as from the selection of landraces with a high-yield potential identified by producers or research institutions.

DNA extraction

DNA was extracted from young cassava leaves following the protocol described by Doyle and Doyle (1990). A 1.0% agarose gel (w/v) was used to estimate DNA concentrations by comparing the fluorescent signal from DNA that was stained with 1.0 mg/mL ethidium bromide to the signals from a dilution series of commercial Lambda DNA (Invitrogen, Carlsbad, CA, USA) of known concentration.

Molecular characterization of cassava accessions by SNP markers

The genotyping of 354 SNP markers from genic regions and another 48 markers derived from the cassava physical map was carried out by the MassArray system (Sequenom iPLEXassay, San Diego, CA, USA). Polymerase chain reaction (PCR) was conducted according to MassExtend (Sequenom) using 15 ng genomic DNA. Locus-specific PCR primers were designed using the MassARRAY Assay Design 3.0 software (Sequenom).

The DNA samples were amplified by a multiplex reaction, and the PCR products were used for a one-base extension reaction for each specific locus. The products were desalinized and transferred to a 384-element SpectroCHIP array. The alleles were discriminated by mass spectrometry (Sequenom). Only the samples and SNPs that contained less than 10% unknown data were analyzed.

Genetic diversity of the entire collection

The cassava germplasm diversity was assayed according to the following parameters: total number of alleles (N_A), polymorphism information content (PIC), shared allele distance (SAD), observed heterozygosity (H_O), and expected heterozygosity (H_E) or genetic diversity. The PowerMarker v 3.25 (Liu and Muse, 2005) software was used for analysis.

Construction of the core collection

Two different algorithms based on the M strategy were used to develop the core collection. The standard M strategy described by Schoen and Brown (1993) was employed using the MSTRAT software (Gouesnard et al., 2001). Four main subsets of different sizes were assayed, including 64 (5.0% of the entire set), 128 (10.0%), 256 (15.0%), and 384 (20.0%) accessions. Hereafter, these collections will be named MST64, MST128, MST256, and MST384, respectively. For each sample size, 50 independent replications and 100 interactions were created, and the core collection with the highest Shannon diversity index was selected to make up the germplasm subset for each collection size.

The M strategy with random and advanced selection with heuristic search, as proposed by Kim et al. (2007), was carried out using the PowerCore v1.0 software. The M strategy with random and heuristic searches selects the most diverse accessions to represent the entire set of alleles from the SNPs in the entire collection. Through this method, the final size of a core collection is unknown *a priori* and depends on the levels of variability and redundancy in the collection. These collections are named PoRAN and PoHEU for the M strategy with random search and advanced heuristic search, respectively.

Validation of the representativeness of the core collections

Genetic diversity measures (N_A, PIC, H_O, H_E, and SAD) were estimated separately for each of the core collections as well as for the entire collection. The diversity estimates of the different core collections were compared with that of the entire cassava germplasm.

Molecular analysis of variance (AMOVA)

AMOVA was carried out to compare the diversity of the accessions that represent the different core collections: MST64, MST128, MST256, MST384, PoRAN, and PoHEU. These analyses were performed using the GenAlEx 6.1 software (Peakall and Smouse, 2006). The genetic structure of the core collections in comparison with the entire collection was inferred based on the genetic relationships between the accessions, which were evaluated by the principal component analysis (PCA) of the 1280 cassava accessions using the genetic distance matrix of the shared alleles of the SNP markers. The PCA analysis was performed with the Genealx 6.1 software (Peakall and Smouse, 2006).

RESULTS

Development of core collections and comparison with the entire collection

Four cassava core collections with 5, 10, 15, and 20% (64, 128, 256, and 384 acces-

sions, respectively) of the total number of accessions were constructed from the original collection using the MSTRAT software. The core collection obtained by the PowerCore software using the M strategy with random selection was composed of 84 accessions (6.6%), whereas that obtained by the advanced M strategy with heuristic search was composed of 61 accessions (4.7%).

All alleles detected in the cassava germplasm were maintained in the different core collections (Table 1). PIC estimates were slightly higher for the MST64, MST128, PoRAN, and PoHEU collections than in the entire collection, although all collections presented the same variation in the PIC values for the different SNP loci.

Table 1. Diversity parameters for the different cassava core collections and the entire Cassava Germplasm Bank (CGB) at Embrapa Mandioca e Fruticultura.

Parameter[1]	Entire collection	MST64	MST128	MST256	MST384	PoRAN	PoHEU
N_A	798	798	798	798	798	798	798
PIC	0.258	0.267	0.262	0.258	0.258	0.270	0.263
H_O	0.318	0.299	0.307	0.311	0.321	0.292	0.297
H_E	0.322	0.330	0.325	0.323	0.323	0.334	0.325
SAD	0.260	0.284	0.267	0.255	0.260	0.274	0.289

[1]N_A = total number of alleles; PIC = polymorphism information content; H_O = observed heterozygosity; H_E = expected heterozygosity; and SAD = shared allele distance.

The H_O estimates were decreased for all core collections except MST384. In contrast, H_E estimates were slightly increased for the MST64, MST128, PoRAN, and PoHEU collections compared with the entire collection.

With regard to the genetic distances of the accessions in the core collections that were calculated based on the distances of shared alleles, all of the collections except MST256 and MST384 had slightly increased genetic distances between accessions compared with the entire collection. In addition to showing a variation of 0.136 to 0.402, the PoHEU collection had the highest average SAD value (0.289 compared with 0.26 for the entire collection), whereas the remaining collections contained accessions with genetic distances that were close to 0.01. In fact, the genetic distances evaluated based on the shared allele distances varied from 0.081 to 0.402 for MST64 and from 0.136 to 0.402 for PoHEU, and the remaining core collections showed variations that ranged from 0.010 to 0.402 (Figure 1).

When the breeding pattern of the cassava accessions was evaluated based on the classification of landraces and improved varieties, there was a proportional distribution of the genotypes for all core collections relative to the entire collection (Figure 2). Collection MST128 showed the same proportions in the distribution of accessions as the entire collection (12% of improved varieties and 88% landraces). A similar observation was made when the geographic origins of the cassava accessions were considered, although the MST64 and PoHEU collections had a decreased representation of the accessions from northeast Brazil and an increased representation of Brazilian accessions of unknown origin. In contrast, the MST128 collection had a reduced representation of accessions from the Northeast Brazil and an increased number of accessions from the North Brazil (Figure 3).

Only 17 accessions were common to all six core collections. Another 17 were common to five core collections. However, most accessions (388) were present in only one of the core collections (Figure 4).

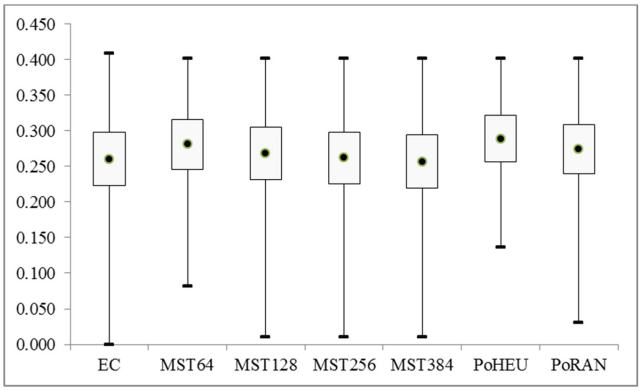

Figure 1. Boxplot of the distribution of the genetic distance of shared alleles for the different core collections containing a total of 1280 accessions from the cassava germplasm that were genotyped with single nucleotide polymorphism markers.

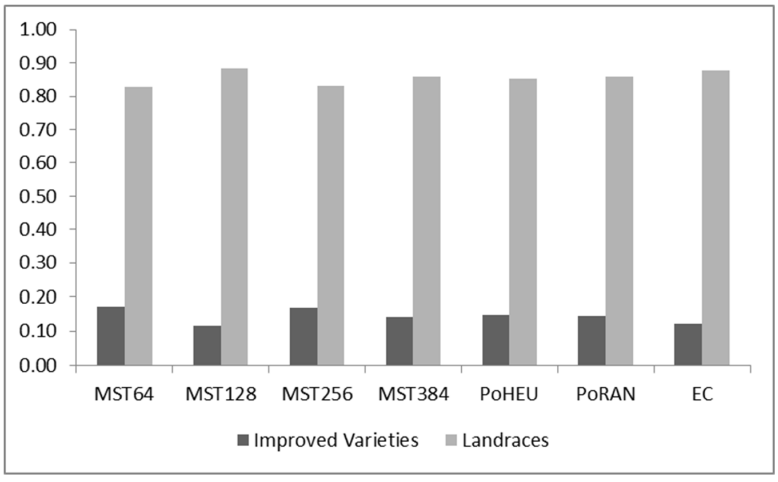

Figure 2. Distribution of cassava accessions according to the breeding pattern (landraces and improved varieties) in the entire collection (EC) of the Cassava Germplasm Bank at Embrapa Mandioca e Fruticultura and in the core collections.

Differentiation of the core collections

AMOVA, which assessed the genetic variation between and within the different collections, showed that the most significant differences in the molecular variance of the SNPs were almost entirely within the core collections (99.70%). Only 0.30% of the molecular variance was attributed to differences between core collections (), suggesting that, in general, the collections were similar in terms of the allelic variance of the SNPs. Although the percentages of genetic variation attributed to each core collection were quite similar (variations between 14.28 and 19.78%), the MST64 and PoHEU collections presented higher genetic differentiation compared with the remaining collections, with values of 17.97 and 19.78%, respectively.

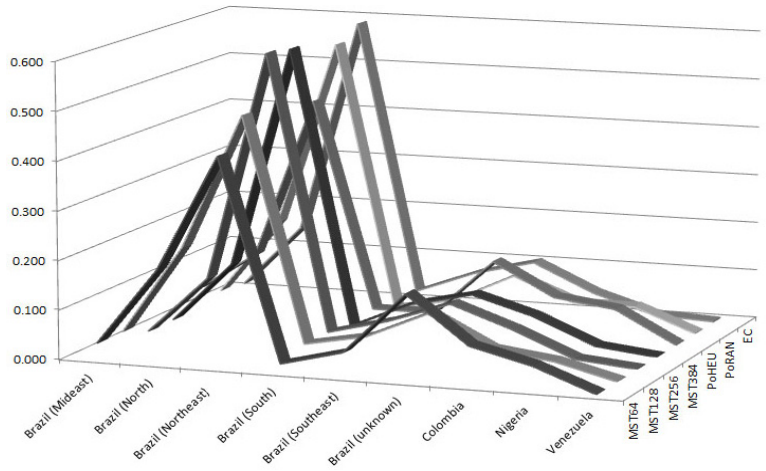

Figure 3. Distribution of cassava accessions according to their geographic origin in the entire collection (EC) of the Cassava Germplasm Bank (CGB) at Embrapa Mandioca e Fruticultura and in the different core collections.

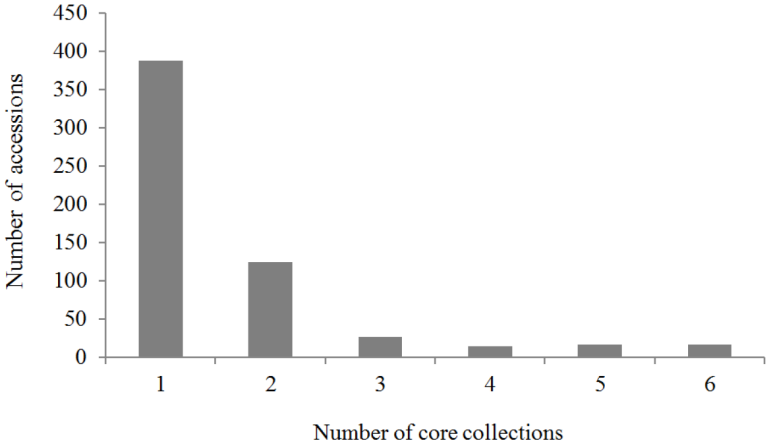

Figure 4. Distribution of accessions selected according to their presence in one or more cassava core collections established using SNP molecular markers.

Table 2. Molecular analysis of variance (AMOVA) of the different cassava core collections obtained by the maximization (M) strategy with 5.0% (MST64), 10.0% (MST128), 15.0% (MST256), and 20.0% (MST384) of accessions from the total set and with random (PoRAN) and advanced heuristic (PoHEU) searches.

Source of variation	Degrees of freedom	Average squares	Percent of variation (%)
Between collections	5	239.225	0.30
Within collections	971	166.155	99.70
MST64	63	212.72	19.78
MST128	127	177.64	16.52
MST256	255	157.86	14.68
MST384	383	153.61	14.28
PoHEU	60	193.2	17.97
PoRAN	83	177.07	16.47
Total	976		100.00

Validation of the core collections based on PCA

PCA was performed to validate the different core collections. The results showed that the distribution of the core collections and the entire collection may be represented by the first two principal components, which were responsible for 60.78% of the total genetic variance (Figure 5).

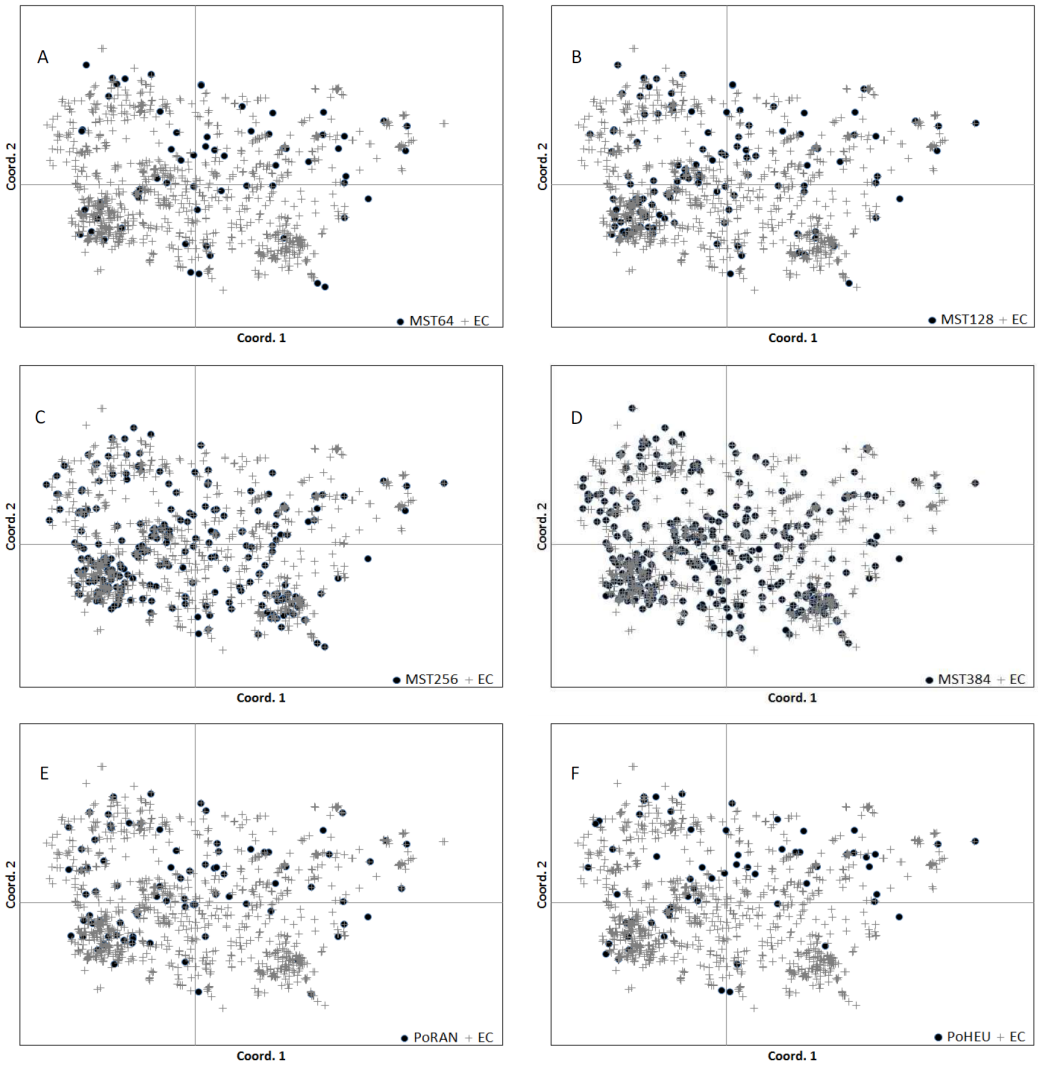

Figure 5. Principal component analysis using the shared-allele distances for the entire collection and the different cassava core collections: **A.** M strategy with 5.0% selection from the entire collection (MST64); **B.** M strategy with 10.0% selection from the entire collection (MST128); **C.** M strategy with 15.0% selection from the entire collection (MST256); **D.** M strategy with 20.0% selection from the entire collection (MST384); **E.** M strategy with random selection of accessions (PoRAN); and **F.** M strategy with advanced selection using a heuristic search (PoHEU).

There is a good distribution of the cassava accessions in the complete collection based on the first two principal components, with a greater concentration of accessions in the lower left quadrant (gray cross in Figure 5) than in other quadrants. In general, all core collections were highly representative of the entire collection (dark circles in Figure 5). The MST384 collection enabled the selection of highly representative accessions from the entire collection. In contrast, the smaller collections (MST64, PoRAN, and PoHEU) enabled the selection of accessions from all quadrants of the PCA graph, which effectively contributed to the representation of the cassava collection.

DISCUSSION

Representativeness of the cassava core collection

According to Cordeiro et al. (1995), three basic criteria should be followed to select cassava accessions to constitute a core collection of the species: a) classification based on the breeding pattern (landraces or improved material), b) agro-ecological origin, and c) important agronomic characteristics, especially for breeders. Considering the availability of complete agronomic evaluations for all of the accessions of the CGB at Embrapa Mandioca e Fruticultura, the first two criteria were then carefully observed when the core collections were formed.

Even though the core collections showed small deviations in the ratio of cassava accessions from the entire collection based on the breeding patterns and geographical origins of the accessions, two accessions from Venezuela in the complete collection were not included in any of the core collections. These two accessions had a SNP-based genetic profile that was very similar to that of the Brazilian landraces. Considering the large flow of propagation material between Brazil and Venezuela, especially from the Amazon, it is expected that these accessions have gene pools similar to those from Brazil; therefore, their inclusion in the core collections was not prioritized.

In general, the accessions from all geographical regions in the entire collection were represented in the different core collections (Figure 3). Of the CGB at Embrapa Mandioca e Fruticultura approximately 94% of the accessions originated from Brazil, and the remaining accessions were from Colombia, Venezuela, and Nigeria. According to Fu (2012), germplasm collections might not represent the gene pools of the species very well because of possible biases in the geographical representation and the random sampling of the germplasm. However, because cassavas are native to Brazil (Olsen and Schaal, 2001), this bias in the conservation of the largest part of the national germplasm is justified because these accessions likely represent the genetic variability that is distributed in many germplasm banks of the species.

The geographical criteria for the classification of the cassava accessions were quite robust, especially because the environmental factors that influence the phenotypic expression of many characteristics were considered, whereas the selection within each extract (geographic origin) increases the chances of maximizing the genetic variability that is retained in the core collection because of the concentration of genic combinations that were adaptive to environmental conditions in the different extracts (Cordeiro et al., 2000). However, using the origins of cassava accessions to define the extracts that should make up the core collections may not be the most adequate criterion because the frequent exchange of propagative materials among producers leads to incorrect inferences. Therefore, genotypes with different names that were

collected in a certain region might have originated in another region, or genotypes with the same name might be completely different. Given these considerations, the use of more objective criteria, such as genotypic data from molecular markers, may increase the accuracy in defining the cassava core collections.

Selection of core collections and comparison with the entire collection

The development of a cassava core collection that is capable of representing the maximum variability in a smaller set of accessions is highly desirable for conservation and crop breeding. In other crops (Escribano et al., 2008; Miranda et al., 2010) and cassava (Hershey et al., 1994), it has been shown that a low intensity of sampling (varying from 4.7 to 20% of an entire collection) was enough to sufficiently represent the genetic diversity that is found in the entire collection.

Previous evidence suggested that maintaining allelic and genetic diversities in core collections is an inherent characteristic of the M strategy and nested selection methods (Schoen and Brown, 1993; Marita et al., 2002) when molecular markers are used. However, capturing all alleles in core collections is not common to all species and data sets, considering that only 93% of alleles were represented in the core collection of corn (Todorovska et al., 2005) and only 98% of alleles were in the core collection of wheat (Balfourier et al., 2007). Core collections need to eliminate redundancies of accessions while maintaining the genetic diversity and important genotypic and phenotypic characteristics of the crop. Therefore, regardless of the collection size, this study demonstrated a high allelic representation (100%) for all SNP markers relative to the entire cassava collection. Similar results for the maintenance of all alleles in a germplasm were also reported by Agrama et al. (2009) for a mini rice core collection composed of 12% of an entire collection using 70 microsatellite markers.

In general, small alterations in the genetic parameters were observed in the different core collections relative to the entire germplasm. However, the core collections that showed the best adjustment of the validation parameters to represent the entire collection and the maximization of the genetic distances between the accessions were PoHEU and MST64. Therefore, the methods implemented by PowerCore and MSTRAT software were very similar considering the selection of the cassava core collections. According to Franco et al. (2006), the M strategy is the most powerful function for the selection of accessions with great allelic diversity and for eliminating redundancies from non-informative alleles that appear because of co-ancestry. In fact, the M strategy (Schoen and Brown, 1993) can effectively select the accessions of a core collection and minimize the probability that any existing allele in the entire collection is missing from the core collection. The implementation of this strategy in the MSTRAT program (Gouesnard et al., 2001) allows the alleles to be selected in an interactive manner, achieving high diversity by the allelic richness criteria and large sum-of-squares values of the variables analyzed. In contrast, the PowerCore algorithm utilizes a heuristic search based on the A* search concept, which is defined by the evaluation criteria of variable coverage (the ratio between the values in the core collection and those in the entire collection and the average of all variables) (Kim et al., 2007).

The PoHEU and MST64 collections differentiated the genotypes and had a low frequency of accessions with short genetic distances (close to zero). In fact, the genetic distances increased by an average of 7.93 and 11.01% for MST64 and PoHEU, respectively. Similar observations

were made in a cassava core collection that was based on morpho-agronomic descriptors, where the core collection retained the phenotypic diversity that was present in the entire germplasm, showing an increment of 15% in the average Gower distance between the accessions (Bhattacharjee et al., 2012). It is possible that this average increase in the genetic distances of these core collections occurred because of the elimination of accessions with high genetic similarity.

In general, a core collection should have 10% of the original collection size and represent approximately 70% of the genetic diversity of the original collection (Brown, 1989a,b). However, different percentages might be acceptable because the establishment of a core collection depends on the size of the original collection, the quality of the data collected for the characterization and evaluation of the stratification of the original collection, and the sampling strategies (Cochran, 1977). A good core collection should incorporate the maximum diversity of the species with minimal redundancy and the smallest size possible to facilitate its management and use in the development of cultivars (Brown, 1989a).

In species such as cassavas, where propagation is mainly asexual *in vitro*, the high costs of accession maintenance in the field and the high vulnerability and losses due to adverse environmental conditions are the main characteristics that differentiate such species from those whose propagative materials can be stored as seeds. Therefore, given the lowest intensity of sampling among all core collections (4.7%) and its genetic diversity between accessions, the PoHEU collection (established by PowerCore) may be the appropriate choice for the practical applications involving cassava genetic conservation. Other studies showed that the heuristic algorithm is capable of effectively reducing the number of accessions in germplasm collections while maintaining an almost complete proportion of the diversity in phenotypic and molecular characteristics (Kim et al., 2007; Agrama et al., 2009). The heuristic algorithm enabled the formation of a rice core collection containing only 1% of the entire germplasm in comparison with approximately 10% when proportional core collection and random core collection methods were applied (Chung et al., 2009). Similar to the cassava core collection, the heuristic algorithm is capable of reducing the maximum number of accessions without significant losses in the parameters that define the core collection.

The international core collection formed by CIAT is composed of 630 of the 5500 accessions of the germplasm (11.45%). Compared with this study, a greater percentage of accessions was likely selected for the core collection at CIAT because of the use of many types of information, such as diversity of origin, geographic diversity, isozyme patterns, morphological descriptors, and agronomical descriptors. In this study, because 4.7% of the cassava collection can represent the entire cassava germplasm at Embrapa Mandioca e Fruticultura without a significant loss of information and drastic alterations in the genetic diversity parameters, some redundancy/genetic similarity of the cassava accessions is likely maintained in this collection. In contrast, one might speculate that the bi-allelic nature of the SNP markers limited the detection of certain variability parameters, such as the PIC (maximum value of 0.50 for each locus). Usually, molecular markers offer more information about genetic diversity. When comparing SNP markers with microsatellite markers in corn, Yang et al. (2011) reported low estimates for H_E and PIC using SNPs. Moreover, Van Inghelandt et al. (2010) showed that although similar inferences could be made regarding the structure and diversity of the heterotic groups in corn using either SNPs or microsatellites, the number of markers needed to obtain similar estimates of genetic diversity was 7 times greater for SNPs than for microsatellites, and the modified Roger's distance was 11 times greater for SNPs than for microsatellites.

Applications in conservation, characterization, and use of cassava genetic resources

According to Chavarriaga-Aguirre et al. (1999), the annual cost of maintaining a cassava germplasm in the field is USD 17.09 per accession, and the cost rises to USD 26.22 per accession under *in vitro* conditions. Measures that aim to optimize the conservation and use of cassava germplasm are necessary to avoid losses and to guarantee the security of these genetic resources. Therefore, to facilitate the management and use of the cassava germplasm, a core collection composed of 61 accessions of the germplasm was established by the M strategy with advanced selection and a heuristic search (Kim et al., 2007), considering the representation of the cassava accessions based on their breeding patterns and geographic origins.

The proposal to concentrate efforts in the characterization and evaluation of the germplasm in a small but representative core collection plays an extremely important role in improving the access of breeders and other users to cassava genetic resources. This is especially important in the case of cassava, where the increase in the number of accessions stored in germplasm banks is not usually accompanied by an increase in funds for research and collection maintenance. Consequently, the use of such genetic resources has been limited because of insufficient characterization and evaluation. Therefore, the core collections may potentially contribute to the efforts in the characterization and agronomic evaluation of important attributes in these accessions.

By maximizing the genetic diversity in a reduced number of genotypes, the PoHEU collection may facilitate studies on the variability and correlation of morphological and agronomical characteristics. An in-depth evaluation of this core collection may also be useful to choose the ideal parents to use in cassava breeding programs and to develop quantitative trait loci maps. Cassava genetic breeding programs are particularly interested in using genetic resources for gene and allele discovery. Such genes and alleles, which are linked to important biological processes, can be discovered via association mapping and genomic selection strategies, and many such projects are already in progress.

The approach developed in this study may be effectively applied to develop core collections for cassava and other species. However, the ideal set of accessions that comprise the core collection must be dynamic and revised periodically when new accessions are incorporated into the collection or when other characteristics become available. The importance of a dynamic core collection is reflected in this study, in which we found that only 17 accessions were common to all six cassava core collections and that the remaining accessions were different in each core collection. This finding indicates the need to include complementary criteria in the formation of the core collections, such as morphological and agronomic characteristics and information about genetic structure and pedigree. In general, core collections in other species (McKhann et al., 2004; Balfourier et al., 2007; Richards et al., 2009) have shown that the genetic diversity of a core collection may be maximized when a set of specific molecular or phenotypic characteristics are simultaneously used.

Although not all phenotypic data are available for the entire cassava germplasm, we suggest that the cassava core collection formed with SNP markers may effectively serve as a reference to research activities related to the conservation and use of these genetic resources. Breeding programs may prioritize the characterization and evaluation of this core collection in their search for desirable characteristics, such as those associated with biotic and abiotic resistance and root and yield characteristics. Small core collections that are focused on specific

characteristics may be part of the strategy to increase the use of cassava genetic resources in breeding programs.

ACKNOWLEDGMENTS

Research supported by Empresa Brasileira de Pesquisa Agropecuária (Embrapa) and Banco Interamericano de Desenvolvimento (BID), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERENCES

Agrama HA, Yan WG, Lee F and Fjellstrom R (2009). Genetic assessment of a mini-core subset developed from the USDA rice GenBank. *Crop Sci.* 49: 1336-1346.

Balfourier F, Roussel V, Strelchenko P, Exbrayat-Vinson F, et al. (2007). A worldwide bread wheat core collection arrayed in a 384-well plate. *Theor. Appl. Genet.* 114: 1265-1275.

Belaj A, Dominguez-García MC, Atienza SG and Urdíroz NM (2012). Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genet. Genomes* 8: 365-378.

Bhattacharjee R, Dumet D, Ilona P and Folarin S (2012). Establishment of a cassava (*Manihot esculenta* Crantz) core collection based on agro-morphological descriptors. *Plant Genet. Resour.* 10: 119-127.

Brown AHD (1989a). Core collections: a practical approach to genetic resources management. *Genome* 31: 818-824.

Brown AHD (1989b). The Case for Core Collections. In: The Use of Plant Genetic Resources. In: 156 (Brown AHD, Frankel OH, Marshall DR and Williams JT, eds.). Cambridge University Press, Cambridge, 136-156.

Chavarriaga-Aguirre P, Maya MM, Tohme J and Duque MC (1999). Using microsatellites, isozymes and AFLPs to evaluate genetic diversity and redundancy in the cassava core collection and to assess the usefulness of DNA-based markers to maintain germplasm collections. *Mol. Breed.* 5: 263-273.

Chung HK, Kim KW, Chung JW, Lee JR, et al. (2009). Development of a core set from a large rice collection using a modified heuristic algorithm to retain maximum diversity. *J. Integr. Plant Biol.* 51: 1116-1125.

Cochran WG (1977). Sampling Techniques. 3rd edn. John Wiley and Sons, New York.

Cordeiro CMT, Morales EAV, Ferreira P and Rocha DMS (1995). Towards a Brazilian Core Collection of Cassava. In: Core Collections of Plant Genetic Resources (Hodgkin T, Brown AHD, van Hintum TJL and Morales EAV, eds.). John Wiley and Sons, Chichester, 155-168.

Cordeiro CMT, Abadie T, Burle ML and Rocha DMS (2000). A Coleção Nuclear de Mandioca no Brasil. Embrapa Recursos Genéticos e Biotecnologia, Brasília.

Doyle JJ and Doyle JL (1990). Isolation of plant DNA from fresh tissue. *Focus* 12: 13-15.

Escribano P, Viruel MA and Hormaza JI (2008). Comparison of different methods to construct a core germplasm collection in woody perennial species with simple sequence repeat markers. A case study in cherimoya (*Annona cherimola*, Annonaceae), an underutilized subtropical fruit tree species. *Ann. Appl. Biol.* 153: 25-32.

Franco J, Crossa J, Taba S and Shands H (2005). A sampling strategy for conserving genetic diversity when forming core subsets. *Crop Sci.* 45: 1035-1044.

Franco J, Crossa J, Warburton ML and Taba S (2006). Sampling strategies for conserving maize diversity when forming core subsets using genetic markers. *Crop Sci.* 46: 854-864.

Fu YB (2012). Genetic structure in a core subset of cultivated barley germplasm. *Crop Sci.* 52: 1195-1208.

Gouesnard B, Bataillon TM, Decoux G, Rozale C, et al. (2001). MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. *J. Hered.* 92: 93-94.

Hershey C, Iglesias C, Iwanaga M and Tohme J (1994). Definition of a Core Collection for Cassava. Report for the First Meeting of the International Network for Cassava Genetic Resources (1992), Cali.

Kim KW, Chung HK, Cho GT, Ma KH, et al. (2007). PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* 23: 2155-2162.

Lebot V (2009). Tropical Root and Tuber Crops, Cassava, Sweet Potato, Yams and Aroids. Crop Production Science in Horticulture 17. CABI, Wallingford.

Liu K and Muse SV (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21: 2128-2129.

Marita JM, Rodriguez JM and Nienhuis J (2002). Development of an algorithm identifying maximally diverse core collections. *Genet Resour. Crop Evol.* 47: 515-526.

McKhann HI, Camilleri C, Berard A, Bataillon T, et al. (2004). Nested core collections maximizing genetic diversity in *Arabidopsis thaliana*. *Plant J.* 38: 193-202.

Miranda C, Urrestarazu J, Santesteban LG and Royo JB (2010). Genetic diversity and structure in a collection of ancient Spanish pear cultivars assessed by microsatellite markers. *J. Am. Soc. Hortic. Sci.* 135: 428-437.

Olsen K and Schaal B (2001). Microsatellite variation in cassava (*Manihot esculenta*, Euphorbiaceae) and its wild relatives: further evidence for a southern Amazonian origin of domestication. *Am. J. Bot.* 88: 131-142.

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

Richards CM, Volk GM, Reeves PA and Reilley AA (2009). Selection of stratified core sets representing wild apple (*Malus sieversii*). *J. Am. Soc. Hortic. Sci.* 134: 228-235.

Schoen DJ and Brown AHD (1993). Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. *Proc. Natl. Acad. Sci. U. S. A.* 90: 10623-10627.

Todorovska E, Abumhadi N, Kamenarova K and Jeleva D (2005). Biotechnological approaches for cereal crops improvement. Part II: use of molecular markers in cereal breeding. *Biotechnol. Equip.* 19: 91-104.

van Hintum TJL, Brown AHD, Spillane C and Hodgkin T (2000). Core Collections of Plant Genetic Resources. International Plant Genetic Resources Institute, Rome.

Van Inghelandt D, Melchinger AE, Lebreton C and Stich B (2010). Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theor. Appl. Genet.* 120: 1289-1299.

Yang X, Xu Y, Shah T, Li H, et al. (2011). Comparison of SSRs and SNPs in assessment of genetic relatedness in maize. *Genetica* 139: 1045-1054.