Plant Mol Biol Rep
DOI 10.1007/s11105-016-0999-6

ORIGINAL PAPER

Establishment of a Core Collection of Traditional Cuban *Theobroma cacao* Plants for Conservation and Utilization Purposes

Igor Bidot Martínez[1] · **Marlyn Valdés de la Cruz**[2] · **Manuel Riera Nelson**[1] · **Pierre Bertin**[3]

Abstract Presently, *Theobroma cacao* L. (cacao) in Cuba is mainly cultivated in the eastern region where plantations comprise a mixture of clonal varieties, hybrids, progeny of Trinidad Selected Hybrids, and traditional—also known as ancient—cacao. The ancient genetic resources, probably the plants most closely related to the original introductions, are endangered by their progressive replacement by modern clones. To promote the conservation and utilization of these genetic resources, a representative sample of 537 traditional Cuban cacao plants was used to develop a core collection. Core collections based on 15 simple sequence repeat (SSR) markers were generated using five different sampling algorithms: random sampling, simulated annealing, stepwise clustering with random sampling, the M strategy, and maximum genetic diversity. The five core collections were designed to capture 95 % of the SSR alleles in the complete collection. The genetic, morphological, and geographical diversity of each core collection was compared with that of the entire collection. The entire collection contained 139 alleles, including 104 rare ones, with the 95 % allelic coverage threshold achieved with 133 alleles. The core collection generated by the maximum genetic diversity algorithm had the lowest number of accessions (185), the highest mean genetic distance (0.486), the lowest morphological character redundancy, and the highest diversity as assessed by the mean Shannon-Weaver diversity index (0.757). This core collection can thus serve as the basis of future improvement programs based on local genetic resources.

Keywords Genetic diversity · Morphological diversity · Relic cacao · SSR markers · Sampling strategies

Introduction

Many germplasm collections of crop species and their relatives are currently too large to achieve their main goals: collection, maintenance, evaluation, and utilization of genetic diversity. The large sizes of these collections are the main reason why such germplasm is poorly exploited and information is lacking for economically important traits, thereby preventing selection of the most appropriate parents for breeding programs (van Hintum et al. 2000; Studnicki et al. 2013).

A viable method to overcome these difficulties entails the development of a core collection (Glaszmann et al. 2010; Çalişkan 2012); that is, a sampling of accessions that represents the genetic diversity of the entire collection with the lowest possible level of redundancy (Frankel and Brown 1984; Brown 1989). The construction of such a reduced set may be desirable because of economic, space, and other considerations (Spooner et al. 2005).

A core collection is a key element in the development of conservation or evaluation projects (Engelmann et al. 2007). The reduced number of selected accessions allows for conservation of a crop species at a relatively low cost compared with the complete collection (Ramanatha and Hodgkin 2002). The availability of a core collection can increase the use of germplasm collections in genetic improvement programs, accelerate the search for alleles or beneficial

✉ Pierre Bertin
pierre.bertin@uclouvain.be

[1] Facultad Agroforestal de Montaña, Universidad de Guantánamo, Guantánamo, Cuba

[2] Facultad de Biología, Universidad de La Habana, Havana, Cuba

[3] Earth and Life Institute-Agronomy (ELI-A), Université catholique de Louvain, Croix du Sud 2, L7.05.11, 1348 Louvain-la-Neuve, Belgium

Published online: 13 August 2016

adaptations, and aid the acquisition of information about a species' genetic diversity (Engels and Visser 2003; Glaszmann et al. 2010; Çalişkan 2012).

Selection of an appropriate sampling strategy is an important prerequisite to the development of a core collection of a size suitable to adequately represent the genetic spectrum and to maximize the capture of genetic diversity in available crop collections (Çalişkan 2012). Because of the importance of such a step to the creation of a core collection, several different selection methods have been developed. These methods use different approaches to select accessions that are based on morphological or genetic information, dendrograms, or distance matrices (Schoen and Brown 1993; Hu et al. 2000; Marita et al. 2000; Liu and Muse 2005).

In cacao, several core collections have been established from germplasm collections. For example, 90 % of the genetic diversity of the 154 plants in a collection maintained at the Tropical Agriculture Research Station in Mayaguez, Puerto Rico, was captured using a 37-plant (24.03 %) core collection (Irish et al. 2010), representing various types available for commercial production (Zhang et al. 2006). Such another is the collection of 668 cacao plants from the Centro Agronómico Tropical de Investigación y Enseñanza, Costa Rica, in which 90 and 95 % of the genetic diversity was captured with 113 (16.92 %) and 245 (36.68 %) accessions, respectively (Zhang et al. 2009).

An alternative way is the establishment of a core collection that captures not only the variation but also contains genotypes with useful characters. An example in cacao is the International Cocoa GeneBank, Trinidad, where a core collection of genotypes has been elaborated that captures the range of variability and contains useful fruit and seed morphological characteristics and resistance to diseases (Sounigo et al. 2006).

In our previous investigation (Bidot Martínez et al. 2015) aiming to study and capture the genetic diversity of ancient Cuban cacao or traditional Cuban cacao plants, several expeditions were carried out throughout the entire country. The traditional Cuban cacao plants are probably the closest to the ones originally introduced in Cuba in 1540 and between 1791 and 1803 (Núñez 2010). A total of 537 plants obtained from these surveys where morphologically and genetically analyzed using 33 morphological descriptors and 15 simple sequence repeat (SSR) markers, respectively. The ex situ conservation of a collection with such a high number of plants would be complex and expensive, with this complexity exacerbated by the difficulty of maintaining a perennial, relatively large, woody species. In addition, this number of plants is not manageable for future manipulation of their characteristics for improvement purposes.

In light of this situation, the objective of the present study was to compare five selection algorithms to generate the most efficient core collection in terms of genetic and phenotypic diversity of traditional Cuban cacao for the ex situ conservation and utilization.

Materials and Methods

Plant Materials

A total of 537 traditional Cuban cacao plants were sampled from local farms in a previously published work (Bidot Martínez et al. 2015), throughout the country wherever old plantations and isolated trees were encountered, between May 2009 and April 2012. Only the oldest plants (approximately 60–80 years old according to their owners), presumably corresponding to individuals genetically closest to the plants originally introduced onto the island, were collected. New plantations comprising clonal varieties, hybrid cacao, and/or progeny of Trinidad Selected Hybrids were discarded, as these varieties are already maintained in gene banks and details on their genetic and phenotypic states are available(Turnbull and Hadley 2011; CIRAD 2013; USDA-ARS 2013).

Core Collection Methods

DNA isolation, PCR amplification and SSR analysis were performed in a previous published work (Bidot Martínez et al. 2015) using 15 microsatellite markers (Lanaud et al. 1999) recommended as international standards for cacao genetic characterization (Saunders et al. 2004). The markers were evenly distributed in the genome (in nine of the ten linkage groups of cacao), and no linkage disequilibrium or null alleles have been described (Brown et al. 2008). Five methods based on SSR data were used to develop core collections from the complete collection of 537 traditional cacao plants: random sampling, the M strategy, maximum genetic diversity, stepwise clustering with random sampling, and simulated annealing. All the core collection algorithms were programmed in R statistical language v3.01 (R Core Team 2013).

For random sampling (Escribano et al. 2008), the accessions were randomly selected from the complete collection without replacement. This strategy assumed no prior knowledge of the original collection.

The M strategy (maximization or marker allele richness) (Schoen and Brown 1993) maximizes the number of alleles at each marker by eliminating redundant accessions. An initial subgroup of plants was taken at random from the complete collection; one accession was then eliminated so that a second subgroup with the highest possible number of alleles was generated. In the next step, the accession among those not taken in the initial subgroup and that produced the highest increase in the number of alleles of the second subgroup was selected to create a third subgroup. This process was repeated until the number of retained alleles did not increase further.

The premise of the maximum genetic diversity algorithm (Marita et al. 2000) is that the genetic variability of the core collection should correspond to the distribution of genetic

distances in the population of interest. For this method, a matrix of genetic distances (1 − proportion of shared alleles) (Bowcock et al. 1994) was constructed with the adegenet package (Jombart 2008; Jombart and Ahmed 2011). A rank was assigned to each accession based on its genetic distance relative to the other accessions; the value of this rank ranged from 1, indicating the highest difference, to a value equal to the number of accessions, corresponding to the highest similarity. Starting from a randomly chosen accession, the accession with the lowest rank relative to the other previously chosen accessions was selected. Accessions with the lowest sums of ranks—relative to the accessions included in the core collection—and not falling within a certain distance of any previously selected accession were ultimately chosen. This latter distance equaled the number of traditional Cuban cacao plants in the complete collection (537) minus the number of accessions included in the core collection, divided by the number of accessions included in the core collection.

For stepwise clustering with random sampling (Hu et al. 2000), a dendrogram based on the unweighted pair-group method with arithmetic averages clustering technique was generated with the phangorn package (Schliep 2011) from a genetic distance matrix (1 − proportion of shared alleles) (Bowcock et al. 1994) created with adegenet (Jombart 2008; Jombart and Ahmed 2011). An accession was selected at random from each lower-order level subgroup to constitute the core collection. A new dendrogram was then generated, and the process was repeated until a subset of the desired size was obtained.

In the simulated annealing algorithm (Liu and Muse 2005), each accession was given a weight based on its allele number and randomly added to a subgroup of accessions. The subgroup was compared with another subgroup formed by exchanging the accession with the lowest number of alleles with a non-selected accession. The subgroup with the highest number of alleles was selected to restart the cycle.

Core collections with size ranging from 3 to 536 accessions selected, were developed using each of these five methods. Ten repetitions of each method were carried out to select the corresponding core collection with the highest number of alleles. When several repetitions resulted in collections with the same number of alleles, the collection with the highest Nei's gene diversity index was selected. The amount of genetic redundancy caused by closely related accessions was estimated for each of the five core collections by plotting allelic richness vs. the number of selected accessions.

The municipality-level geographical distribution of plants in each developed core collection was compared with that of the complete collection.

Core Collection Evaluation

To compare the performance of the different core collection development methods, the smallest collection from each method that retained at least 95 % of the SSR alleles of the traditional Cuban cacao plants was used.

The following parameters among core collections were compared: number of selected accessions, total number of alleles, number of rare alleles (i.e., those with a frequency lower than 0.05), and mean genetic distance (1 − proportion of shared alleles).

Allelic coverage (CV)—the percentage of alleles retained in the core collection (Kim et al. 2007)—was calculated according to the formula:

$$CV = (1 - PN)*100,$$

where PN is the proportion of non-informative alleles (PN), that is, the proportion of alleles not present in the core collection (Franco et al. 2006). PN was calculated as:

$$PN = 1 - |A_{cs}| / |A_{col}|,$$

where A_{cs} and A_{col} are the number of alleles in core and complete collections, respectively.

The relationship between PN and CV is inverse: maximizing one variable minimizes the other.

The frequency distribution of alleles in each core collection was compared with that of the complete collection of Cuban plants using the Chi-square test. Expected (H_e) and observed (H_o) heterozygosities, and the number of alleles were compared at the locus level using the Friedman rank sum test as implemented in the agricolae package (de Mendiburu 2013). A post hoc Dunnett's test in the asbio package (Aho et al. 2013) was used to compare the different subsets generated from the complete collection.

Principal coordinate analysis (PCO) based on microsatellite markers was performed using the ape package (Paradis et al. 2004) to compare the distribution of accessions in the core collection with that of the complete collection of traditional Cuban cacao plants. A two-dimensional plot was generated based on the first two PCO components.

The phenotypic diversity of all 537 traditional Cuban cacao plants and plants of the core collections was evaluated with 33 cacao morphological descriptors (Table 1) (Engels et al. 1980; Bekele and Butler 2000) in a previous published work (Bidot Martínez et al. 2016). Only qualitative descriptors were used because the traditional Cuban cacao plants are very old plants and do not bear the appropriate number of flowers and fruits (between 10 and 30) for an accurate quantitative evaluation. The frequency of each category of the 33 morphological descriptors and the Shannon-Weaver diversity index (Shannon and Weaver 1949) were calculated.

For selection of core collections, several diversity and representation parameters were taken into consideration. In particular, selection was based on the following criteria: (1) a low number of accessions, generally corresponding to 5–20 % of the complete collection (van

Table 1 Evaluated morphological descriptors (Engels et al. 1980; Bekele and Butler 2000) in the complete and core collections of traditional Cuban cacao plants

Organ	Morphological descriptor	Abbreviation	Categories
Seed	Transversal section	SeTs	1, round; 2, intermediate; 3, flat
	Cotyledon color	SeCC	1, white; 2, light purple; 3, intensely purple
	Pulp color	SePC	1, white; 2, light yellow; 3, intensely yellow
	Pulp flavor	SePf	1, sweet; 2, slightly acidic; 3, acid
Fruit	Length/width ratio	FrLwr	1, L/W > 2; 2, L/W < 2
	Immature fruit color	FrIc	0, white; 1, light green; 2, intermediate green; 3, intensely green; 4, light red; 5, intermediate red; 6, intensely red
	Mature fruit color	FrMc	1, yellow; 2, orange
	Anthocyanin in ridge of the immature fruit	FrAri	1, absent; 2, present
	Anthocyanin in furrow of the immature fruit	FrAfi	1, absent; 2, present
	Anthocyanin in ridge of the mature fruit	FrArm	1, absent; 2, present
	Anthocyanin in furrow of the mature fruit	FrAfm	1, absent; 2, present
	Ridge number and distribution	FrRnr	1, 10 regular ridge; 2, 5 pairs of ridges; 3, 5 ridge
	Furrow depth	FrFd	1, superficial; 2, median; 3, deep
	Surface rugosity	FrSr	0, absent; 1, slight; 2, intermediate; 3, intense
	Apex form	FrAf	0, round; 1, slightly obtuse; 3, obtuse; 5, slightly acute; 7, pointed; 9, acuminate
	Basal constriction	FrBc	0, absent; 1, slight; 2, intermediate; 3, strong; 4, very strong
	Mesocarp hardness	Frmh	1, soft; 2, intermediate; 3, hard
	General form[a]	FrGf	1, Amelonado; 2, Cundeamor; 3, Angoleta; 4, Calabacillo; 5, Criollo; 6, Pentagona
Flower	Staminode orientation	FlSto	1, vertical; 2, convergent; 3, divergent
	Staminode color	FlStc	1, white; 2, light purple; 3, intensely purple
	Sepal color	FlSec	1, white; 2, light purple; 3, intensely purple; 4, purple in rim, white in center
	Sepal union	FlSeu	1, joined; 2, separated
	Sepal orientation	FlSeo	1, horizontal; 2, pendant
	Petal parts balance	FlPpb	1, Shell = Taenial segment + Ligule; 2, Shell = Taenial segment = 1/2 Ligule; 3, Taenial segment < Shell < Ligule; 4, Shell = Taenial segment = Ligule)
	Ligule orientation	FlLo	1, horizontal; 2, pendant
	Guide lines	FlGl	1, absent; 2, present
	Petal yellow-coloration intensity	FlPyc	1, white; 2, light yellow; 3, intensely yellow
	Petal red-coloration intensity	FlPrc	1, white; 2, light red; 3, intensely red
	Stamen filament color intensity	FlSfc	1, white; 2, light purple; 3, intensely purple
	Pedicel color	FlPc	1, green; 2, light purple; 3, intensely purple
	Closed flower color	FlCfc	1, white; 2, light purple; 3, intensely purple
Flush leaves	Green-coloration intensity	FLGc	1, absent; 2, light green; 3, intermediate green; 4, dark green
	Red-coloration intensity	FLRc	1, absent; 2, light red; 3, intermediate red; 4, intensely red

[a] This antique cacao fruit classification has been maintained because it is still presently used by researchers and farmers in Cuba

Hintum et al. 2000; Le Cunff et al. 2008; Studnicki et al. 2013); (2) high allelic richness; (3) a low proportion of non-informative alleles and higher allele coverage; (4) high genetic distance among accessions (Thachuk et al. 2009); (5) high number of rare alleles; (6) no significant difference (according to the Chi-square test) in the frequency distribution of alleles of at least 95 % of loci between the core and complete collection; and (7) no significant difference in variability parameters (H_o and H_e) between the core and the complete collection according to Friedman rank sum and Dunnett's tests (Brown 1989; Grenier et al. 2000; Franco et al. 2005; Escribano et al. 2008).

Statistical Analysis Software

All packages used for core collection development, genetic and morphological diversity analyses, distance calculations, and PCO were from the R statistical language v3.01 (R Core Team 2013).

Results

Selection of Plants for Core Collections

The 185 plants selected with the maximum genetic diversity algorithms constitute the more representative core collection of the genetic and phenotypic diversity of traditional Cuban cacao.

The 95 % allele threshold from the complete sample of 537 plants was achieved using each of the five analyzed algorithms with 185 to 386 plants as follows: 343 from random selection, 386 from simulated annealing, 312 from stepwise clustering with random sampling, 297 from the M strategy, and 185 from the maximum genetic diversity algorithm. These numbers corresponded to 63.9, 71.9, 58.1, 55.3, and 34.5 % of the complete collection, respectively, and indicated that some redundancy existed within the complete collection. On the basis of this minimal redundancy criterion, the maximum genetic diversity algorithm was considered to be the most efficient method.

Capture of 100 % of alleles required the selection of 454 plants using random sampling, 507 with the simulated annealing algorithm (Liu and Muse 2005), 446 from stepwise clustering with random sampling (Hu et al. 2000), 465 with the M strategy (Schoen and Brown 1993), and 390 with the maximum genetic diversity algorithm (Marita et al. 2000). In all cases, more than 100 additional plants were thus necessary to retain 100 % rather than 95 % of alleles. Therefore the core collections possessing 95 % of allelic diversity were selected for subsequent analyses (Fig. 1).

With one exception, all five core collections contained accessions from every Cuban municipality in which traditional Cuban cacao had been sampled. The exception was the core collection developed using the M strategy, which lacked any plants collected from the municipality of Manicaragua in the central Cuban province of Villa Clara. Nevertheless, it is necessary to consider that in this municipality the sample size was very small, with only two plants located (Table 2).

Core Collection Evaluation

The 95 % allelic diversity threshold required 133 of the 139 alleles in the complete collection. This number represented an allelic coverage (CV) of 95.68 %. In each core collection, 98 rare alleles were captured out of a total of 104 in the complete collection (Table 3).

The consistency of the five core collections across the ten repetitions was the highest with the stepwise clustering with random sampling algorithm (100.00 % of the plants selected were present in all ten repetitions) and with the simulated annealing algorithm (94.10 %). Satisfactory results were obtained with the maximum genetic diversity algorithm (73.95 %) but not with the random sampling method and the M strategy (0.56 and 0.37 %, respectively).

Among the generated core collections, the collection created using the maximum genetic diversity algorithm (Marita et al. 2000) had the highest mean genetic distance among accessions (0.486) and the smallest number of accessions (185).

Allele distributions of the three core collections generated by random sampling, stepwise clustering with random sampling and the M strategy were not significantly different from that of the complete collection (according to the Chi-square test). Conversely, distributions of five loci (33.33 %) were significantly different between the core collection obtained from the simulated annealing algorithm and the complete collection. In the case of the maximum genetic diversity algorithm, significant differences were observed for all loci.

Generally, statistically significant differences in observed and expected heterozygosities were not found between the complete collection and the core collections identified with the five algorithms. The only exception was the core collection that was derived with the maximum genetic diversity algorithm in which the expected heterozygosity (0.555) was significantly ($p < 0.05$) higher than that of the complete collection (0.419) (Table 3).

The PCO captured only a small percentage of the total variance. Nevertheless, the distribution of plants was similar among the five core collections and covered most of the genetic diversity of the entire collection of traditional Cuban cacao plants (Fig. 2). There was one exception: the simulated annealing algorithm selected plants that were restricted to the left side and center of the PCO plot. With respect to each tested algorithm except the maximum genetic diversity method, numerous plants overlapped in the PCO plot; in the case of the maximum genetic diversity algorithm, only a few plants

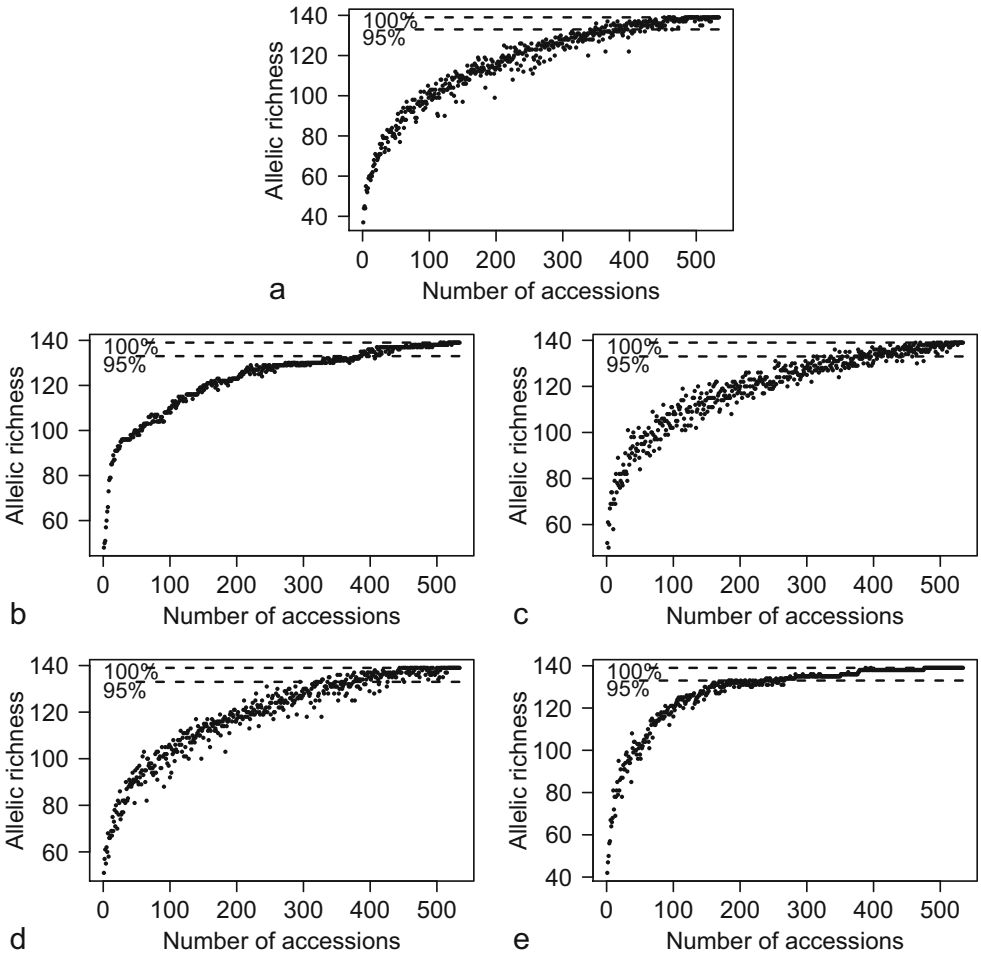

Fig. 1 Relationship between core collection allelic richness and the number of accessions chosen by different selection algorithms. **a** Random selection; **b** simulated annealing (Liu and Muse 2005); **c** stepwise clustering with random sampling (Hu et al. 2000); **d** M strategy (Schoen and Brown 1993); **e** maximum genetic diversity algorithm (Marita et al. 2000). *Dashed lines* indicate 95 and 100 % allelic richness levels relative to the 537 ancient Cuban cacao plants in the complete collection

Table 2 Geographical distribution of plants selected for core collections having 95 % allelic richness

Region	Mountain range	Province	Municipality	Number of plants (proportion with regard to traditional Cuban plants (%))									
				Complete population	Random sampling (Escribano et al. 2008)		Simulated annealing algorithm (Liu and Muse 2005)		Stepwise clustering with random sampling (Hu et al. 2000)		M strategy (Schoen and Brown 1993)		Maximum genetic diversity algorithm (Marita et al. 2000)	
Oriental	Nipe-Sagua-Baracoa	Guantánamo	Baracoa	383	246	(64.23)	282	(73.63)	227	(59.27)	211	(55.09)	128	(33.42)
	Sierra Maestra	Granma	Buey Arriba	27	18	(66.67)	10	(37.04)	16	(59.26)	16	(59.26)	6	(22.22)
			Campechuela	10	6	(60.00)	10	(100.00)	5	(50.00)	8	(80.00)	8	(80.00)
		Santiago de Cuba	Santiago de Cuba	9	4	(44.44)	4	(44.44)	1	(11.11)	7	(77.78)	1	(11.11)
			Songo-La Maya	73	48	(65.75)	53	(72.60)	42	(57.53)	39	(53.43)	20	(27.40)
Central	Macizo de Guamuhaya	Sancti Spíritus	Trinidad	10	7	(70.00)	5	(50.00)	8	(80.00)	7	(70.00)	6	(60.00)
		Cienfuegos	Cumanayagua	23	13	(56.52)	21	(91.30)	11	(47.83)	9	(39.13)	14	(60.87)
		Villa Clara	Manicaragua	2	1	(50.00)	1	(50.00)	2	(100.00)	0	(0.00)	2	(100.00)

Table 3 Comparison of the five generated core collections and the complete population of 537 traditional Cuban cacao plants

	Complete population	Random sampling (Escribano et al. 2008)	Simulated annealing algorithm (Liu and Muse 2005)	Stepwise clustering with random sampling (Hu et al. 2000)	M strategy (Schoen and Brown 1993)	Maximum genetic diversity algorithm (Marita et al. 2000)
Number of accessions selected	537	343	386	312	297	185
Number of alleles	139a	133a	133a	133a	133a	133a
Allele coverage (%)	–	95.68	95.68	95.68	95.68	95.68
Number of rare alleles (frequency <0.05)	104	98	98	98	98	98
Mean genetic distance (1 − proportion of shared alleles)	0.367	0.365	0.398	0.377	0.364	0.486
Percentage of loci with no significant difference in frequency distribution of alleles between the core and complete collections (chi squared test)	–	100.00	66.67	100.00	100.00	0.00
Observed heterozygosity	0.370a	0.381a	0.456a	0.371a	0.380a	0.468a
Expected heterozygosity	0.419a	0.420a	0.475a	0.427a	0.422a	0.555b

Different letters indicate statistically significant differences according to Friedman rank sum and Dunnett's tests ($p < 0.05$)

that genetically clustered together were selected, thereby allowing for a homogeneous distribution of selected plants.

Morphological diversity was maintained in all five developed core collections, with all morphological descriptor categories retained from the complete collection. In each core collection, however, a high level of redundancy in morphological characteristics was present, with a high number of plants falling into the same morphological descriptor categories. The maximum genetic diversity algorithm generated the core collection with the lowest redundancy; this collection had the smallest number of selected plants in 96 out of the 100 morphological descriptor categories observed in the traditional Cuban cacao plants (Fig. 3).

The mean Shannon-Weaver diversity index value for all morphological descriptors in the complete collection of traditional Cuban cacao plants, 0.75, was similar to that of each of the five core collections: 0.75 (random selection) and 0.76 (simulated annealing, stepwise clustering with random sampling, M strategy, and maximum genetic diversity). The core collection obtained using the maximum genetic diversity algorithm exhibited higher Shannon-Weaver index values for 18 morphological characters than did the complete collection. In addition, the Shannon-Weaver index value of 13 of these characters was higher in this core collection than in the others, revealing the higher morphological diversity of the subset obtained using the maximum genetic diversity algorithm (Table 4).

Discussion

The present study allowed to determine the most efficient core collection method in terms of genetic and phenotypic diversity of traditional Cuban cacao for the ex situ conservation and utilization, among five selected algorithms. An ideal collection without accessional redundancy must show a linear relationship between allelic richness and the number of plants (Zhang et al. 2009). On the contrary, the asymptotic curve in the core collections selected with the five algorithms indicates redundancy in the complete traditional Cuban cacao collection (Fig. 1).

The smallest number of selected accessions, 185 (or 34.5 % of the complete collection), was obtained with the maximum genetic diversity algorithm; this number was approximately half that selected with the other methods (Table 3). The reason for this lower number may be a consequence of the fact that methods based on genetic distances, such as the maximum genetic diversity algorithm, sample relatively small and large numbers of accessions from homogeneous and heterogeneous groups, respectively. This approach thus leads to the selection of a high number of alleles in a few plants (Jansen and van Hintum 2007).

By definition, 95 % of the alleles in the complete collection were conserved, as well as 94.23 % of rare alleles (Table 3). The conservation and increased frequency of rare alleles is very important for their conservation and use—especially in poorly characterized populations—because these alleles may confer or regulate important agronomic features (Li et al. 2002; Garkava-Gustavsson et al. 2005). This situation is true for the traditional Cuban cacao plants, in which 74.82 % of the alleles (104 out of 139) are rare. Although a high number of alleles was found at each locus (an average of 9.27), two alleles at each locus always represented more than 80 % of the total allelic frequency. Consequently, most of the remaining ones had frequencies lower than 5 % and could be considered as rare alleles (Bidot Martínez et al. 2015).

The percentage of accessions (34.5–71.9 %) selected by each algorithm necessary to capture 95 % of the alleles was higher than the 10 % usually considered as appropriate for

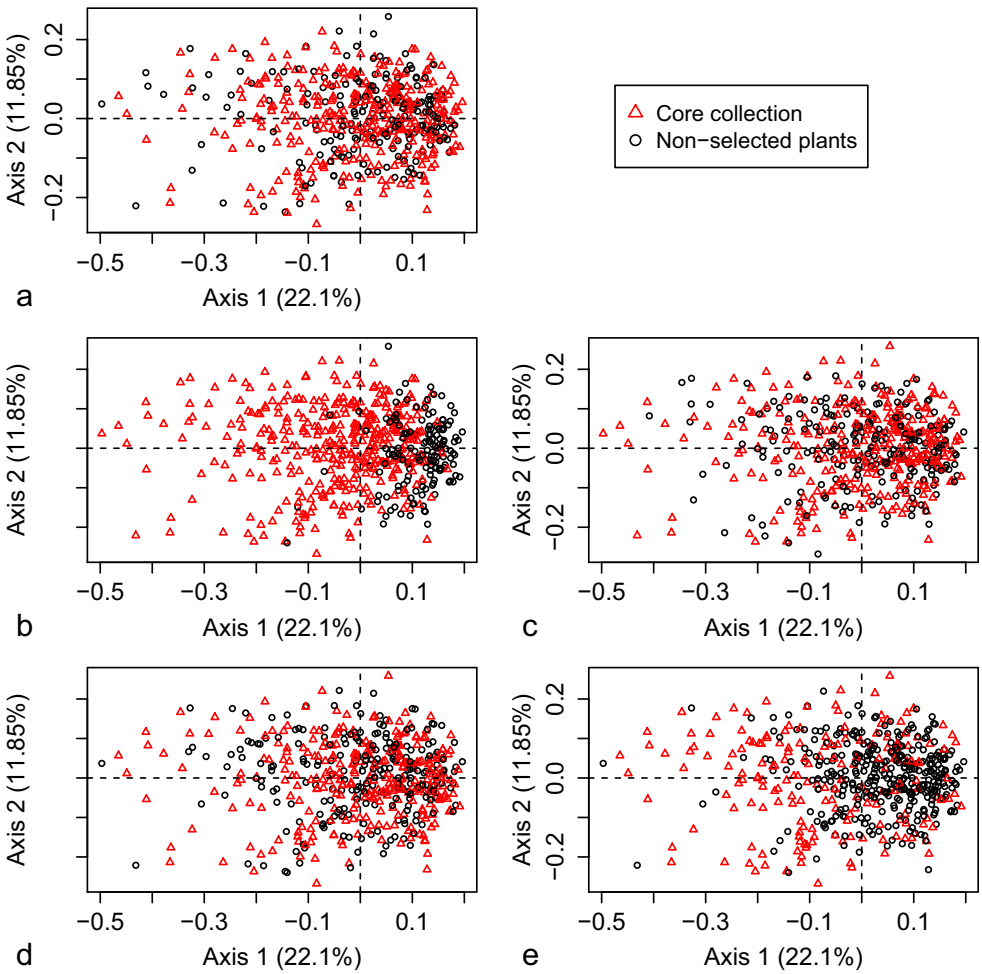

Fig. 2 Plot of the first two principal components from principal coordinate analysis of the 537 traditional Cuban cacao plants (*circles* and *triangles*) and the five core collections with 95 % allelic richness (*triangles*) that were derived using different selection algorithms. **a** Random selection; **b** simulated annealing (Liu and Muse 2005); **c** stepwise clustering with random sampling (Hu et al. 2000); **d** M strategy (Schoen and Brown 1993); **e** maximum genetic diversity algorithm (Marita et al. 2000). The variation captured by the two axes refers to that of the complete collection

core collections of cultivated plants (Igartua et al. 1998; Agrama et al. 2009; Zorić et al. 2012) and the 5–20 % proposed for core collections in general (van Hintum et al. 2000; Le Cunff et al. 2008; Studnicki et al. 2013). It was also higher than in the core collection of the International Cocoa GeneBank, Trinidad (14.25 %, Motilal et al. 2013), in the Tropical Agriculture Research Station in Mayaguez, Puerto Rico (24.03 %, Irish et al. 2010), but similar to the core collection of the Centro Agronómico Tropical de Investigación y Enseñanza, Costa Rica to capture the 95 % of the genetic diversity (36.68 %, Zhang et al. 2009). Nevertheless, these percentages are merely reference values for most plant collections. The size of a core collection depends on the species, the collection objective, the characteristics of the population to be conserved, the selection strategy adopted, and the size of the complete collection to be sampled (van Hintum et al. 2002).

Several examples of core collections generated from small collections exist in which more than 20 % of plants from the complete collections have been selected. Examples include the lingonberry (*Vaccinium vitis-idaea* L.) core collection at the Swedish University of Agricultural Sciences with 69 accessions (31.51 % of the complete collection) (Garkava-Gustavsson et al. 2005), the mulberry (*Morus* spp.) collection at the National Mulberry GeneBank of the Sericultural Research Institute, Chinese Academy of Agricultural Sciences, with 21 accessions (28.77 %) (Lin et al. 2011), and the collection of *Solanum pimpinellifolium* L., the ancestor of cultivated tomato (*Solanum lycopersicum* L.), at the World Vegetable Center with 34 accessions (39.47 %) (Rao et al. 2012).

Retention of less than 20 % of the total number of accessions in a core collection is more easily reached in larger-sized collections. For instance, a grape (*Vitis vinifera* L.) core collection derived from 2262 accessions in the Vassal collection in France (Le Cunff et al. 2008) captured 100 % of alleles with 326 (14.41 %) accessions. This phenomenon may be due to the presence of genetically similar plants in large collections leading to higher redundancy, which allows a higher percentage of alleles to be retained with a smaller proportion of plants.

The larger number of plants needed to preserve 95 % of the alleles of traditional Cuban cacao plants (Table 3) may be due to their broad genetic base, which includes seven of the ten genetic groups described by Motamayor et al. (2008)—mainly Criollo, Amelonado, and Marañón with lower proportions of

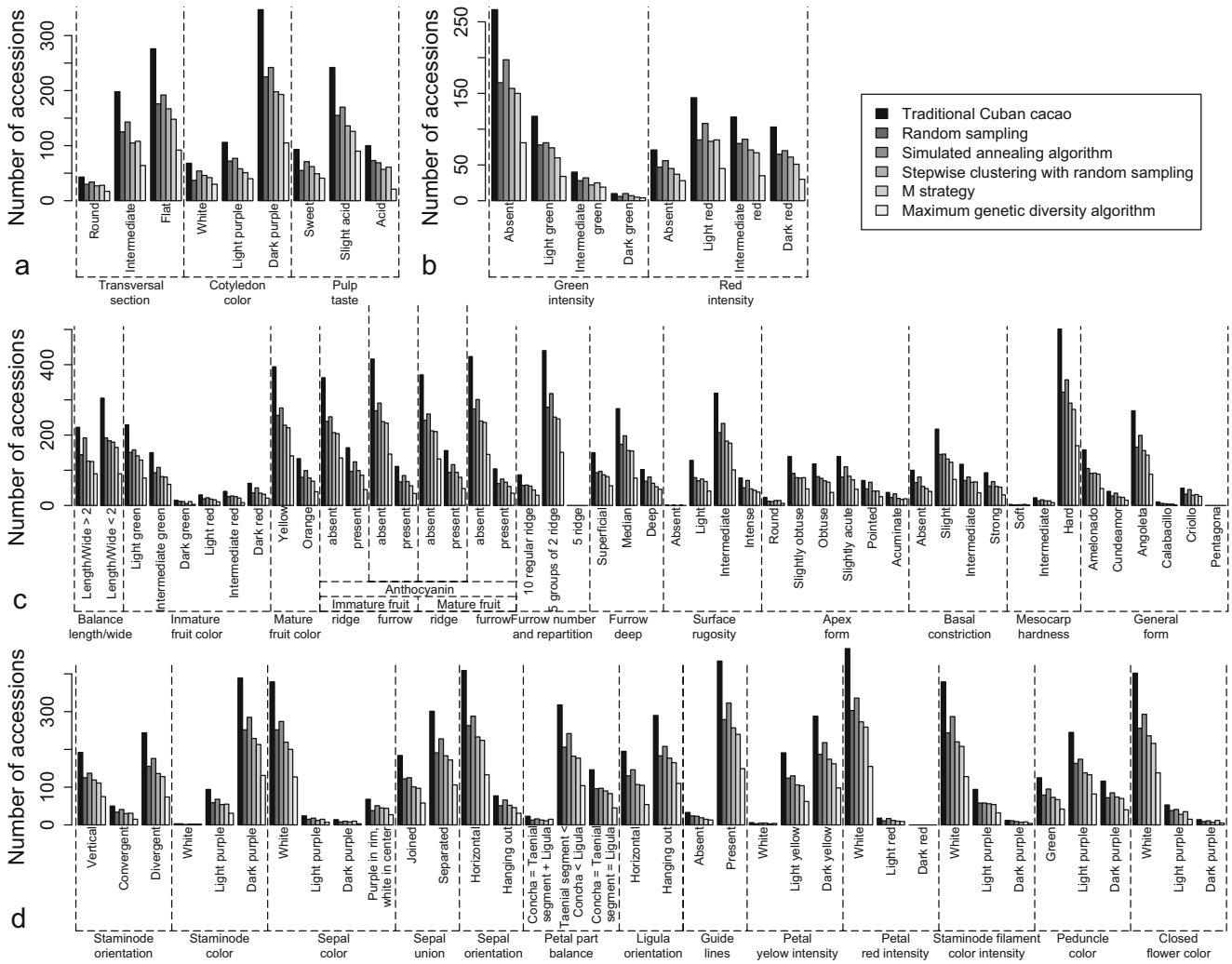

Fig. 3 Number of accessions in the complete collection of 537 traditional Cuban cacao plants and the core collections that were associated with each morphological character state. **a** Seed, **b** flushed leaf, **c** fruit, and **d** flower characters

Contamana, Nanay, Nacional, and Iquitos (Bidot Martínez et al. 2015). In addition, traditional Cuban cacao plants, even though cultivated, have rarely been selected and crossed except for sporadic selections possibly carried out by farmers. This absence of selection could have caused a higher allelic richness in Cuban cacao compared with improved varieties. Thus, a larger-sized core collection is needed to capture most of the alleles (Zhao et al. 2010).

Compared with the original collection, a higher mean genetic distance is expected among plants in a core collection. This situation is a consequence of the fact that a core collection encompasses the diversity of the complete collection with fewer redundant accessions (Franco et al. 2005; Thachuk et al. 2009). Larger genetic distances were indeed obtained among accessions in the five core collections developed in this study. The core collection generated using the maximum genetic diversity algorithm was characterized by a higher mean genetic distance among a lower number of plants, with a consequently lower number of redundant accessions. Other core collections, such as the Olive Worldwide Germplasm Bank Marrakech olive tree (*Olea europea* L.) collection (El Bakkali et al. 2013), have also been shown to have higher mean genetic distances than those of the original complete collection.

The maximum genetic diversity algorithm was unable to meet one criterion, namely, the requirement of no significant difference in the frequency distribution of alleles of at least 95 % of loci between the complete collection and the core collection (Table 3). This result is not detrimental to conserving genetic diversity. This core collection had the smallest size, containing at least 100 fewer plants than the other core collections. This reduced number of selected accessions constituted a more manageable and economical core collection. With such a reduced number of plants, however, it was difficult to obtain an allele distribution similar to that of the complete traditional Cuban cacao collection. The resulting

Table 4 Shannon-Weaver diversity index (Shannon and Weaver 1949) values for morphological descriptors of 537 traditional Cuban cacao plants and five derived core collections

Organ	Morphological descriptor	Shannon-Weaver Diversity Index					
		Complete population	Random sampling (Escribano et al. 2008)	Simulated annealing algorithm (Liu and Muse 2005)	Stepwise clustering with random sampling (Hu et al. 2000)	M strategy (Schoen and Brown 1993)	Maximum genetic diversity algorithm (Marita et al. 2000)
Seed	Transversal section	0.91	0.92	0.93	0.91	0.94	0.93
	Cotyledon color	0.86	0.84	0.89	0.88	0.86	0.95
	Pulp color	0.00	0.00	0.00	0.00	0.00	0.00
	Pulp flavor	0.99	1.00	1.00	1.01	1.01	0.94
Fruit	Length/wide ratio	0.68	0.68	0.69	0.68	0.68	0.69
	Immature fruit color	1.43	1.43	1.45	1.40	1.43	1.35
	Mature fruit color	0.57	0.55	0.58	0.57	0.55	0.52
	Anthocyanin in ridge of the immature fruit	0.62	0.60	0.63	0.63	0.61	0.56
	Anthocyanin in furrow of the immature fruit	0.52	0.50	0.53	0.53	0.49	0.49
	Anthocyanin in ridge of the mature fruit	0.61	0.59	0.62	0.62	0.59	0.58
	Anthocyanin in furrow of the mature fruit	0.50	0.48	0.50	0.52	0.48	0.49
	Ridge number and distribution	0.45	0.46	0.43	0.47	0.43	0.44
	Furrow depth	1.02	1.02	1.02	1.03	1.00	1.07
	Surface rugosity	0.94	0.92	0.94	0.94	0.94	0.99
	Apex form	1.63	1.62	1.63	1.63	1.63	1.65
	Basal constriction	1.32	1.31	1.34	1.31	1.31	1.32
	Mesocarp hardness	0.21	0.18	0.21	0.22	0.24	0.22
	General form	1.20	1.21	1.21	1.19	1.21	1.22
Flower	Staminode orientation	0.95	0.96	0.96	0.96	0.97	0.94
	Staminode color	0.53	0.54	0.51	0.53	0.55	0.55
	Sepal color	0.72	0.67	0.73	0.73	0.80	0.70
	Sepal union	0.66	0.67	0.65	0.65	0.65	0.65
	Sepal orientation	0.44	0.44	0.48	0.48	0.46	0.49
	Petal parts balance	0.78	0.77	0.76	0.79	0.77	0.86
	Ligule orientation	0.67	0.68	0.68	0.66	0.67	0.63
	Guide lines	0.26	0.28	0.24	0.25	0.22	0.28
	Petal yellow-coloration intensity	0.73	0.72	0.73	0.74	0.72	0.77
	Petal red-coloration intensity	0.16	0.15	0.19	0.18	0.16	0.21
	Stamen filament color intensity	0.60	0.62	0.56	0.61	0.63	0.60
	Pedicel color	1.04	1.03	1.05	1.05	1.05	1.04
		0.48	0.51	0.50	0.45	0.57	0.45

Table 4 (continued)

Organ	Morphological descriptor	Shannon-Weaver Diversity Index					
		Complete population	Random sampling (Escribano et al. 2008)	Simulated annealing algorithm (Liu and Muse 2005)	Stepwise clustering with random sampling (Hu et al. 2000)	M strategy (Schoen and Brown 1993)	Maximum genetic diversity algorithm (Marita et al. 2000)
	Closed flower color						
Flush leaf	Green-coloration intensity	0.96	0.98	0.99	0.97	0.96	1.03
	Red-coloration intensity	1.36	1.36	1.36	1.36	1.34	1.37
	Mean	0.75	0.75	0.76	0.76	0.76	0.76

distribution was non-uniform: at each locus, two alleles together constituted more than 80 % of the alleles present, with the remaining alleles, many of them rare (less than 5 % frequency), having lower frequencies.

With respect to heterozygosity, the highest H_e was observed in the core collection developed using the maximum genetic diversity algorithm (Table 3). Although such differences between the core and the complete collection should be avoided, increases in H_o and H_e have been found in core collections such as in the *V. vinifera* collection in France (Le Cunff et al. 2008) and the core collection in the International Cocoa GeneBank, Trinidad (Motilal et al. 2013). This could be because the maximum genetic diversity algorithm maximizes the genetic distance among the selected plants, thus allowing to conserve as much alleles than the other algorithms with only about half the number of plants. As a consequence, there is an increase in heterozygosity among the reduced number of plants selected.

In the present work, the selection of the plants for the traditional Cuban cacao core collection was based on the 15 microsatellite markers recommended as international standards for cacao genetic characterization (Lanaud et al. 1999; Saunders et al. 2004) and qualitative morphological descriptors. These 15 microsatellite markers are present in nine of the ten cacao linkage groups and no linkage disequilibrium or null alleles have been described (Brown et al. 2008). They have been used for the creation of the cacao core collections in CATIE, Costa Rica (Zhang et al. 2009) and USDA-ARS Tropical Agriculture Research Station, Mayaguez, Puerto Rico (Irish et al. 2010), using in both cases the M strategy (Schoen and Brown 1993). This is one of the most used algorithms for the selection of core collections in several species, included cacao, because of its efficiency and also because it is one of the first described. However, new algorithms have been described more recently that could be more efficient than the M strategy, as those selected and used in the present work.

The microsatellites have been the most widely used molecular markers in studies of cacao genetic diversity because they are co-dominant, highly frequent, randomly distributed throughout the genome, highly repeatable, and relatively few expensive (Guiltinan 2007; Kumar et al. 2009). However, SNPs have been included more recently in the genetic analysis of cacao but not yet in the selection of core collection (Livingstone et al. 2015). An alternative way should be to consider economical and agronomical morphological descriptors, in addition to the molecular markers.

The consistency of each of the five core collections across the ten repetitions was the highest with the stepwise clustering with random sampling algorithm (100 % of the plants selected were present in all ten repetitions) and with the simulated annealing algorithm (94.10 %). Satisfactory results were obtained with the maximum genetic diversity algorithm (73.95 %) but not with the random sampling method and the M strategy (0.56 and 0.37 %, respectively).

A comparison of morphological characteristics and Shannon-Weaver diversity index values between core and complete collections indicated that phenotypic diversity was preserved in the five core collections. The five core collections were morphologically very similar, differing primarily in the number of selected plants. The maximum genetic diversity algorithm yielded the collection with the lowest morphological redundancy because it contained the lowest number of selected accessions. This core collection also included the highest number of descriptors with Shannon-Weaver diversity index values higher than those of the complete collection; it also contained the most descriptors with the highest values of this index. This core collection thus featured the highest morphological diversity and the lowest number of plants (Fig. 3; Table 4).

The morphological descriptors evaluated in this work are broadly used in Cuba and are based on the work of Engels et al. (1980). Two additional morphological descriptors were used for fruit classification and the color of fruit, because they are used by researchers in Cuba.

Morphological diversity and characters of interest should also be considered for the creation of a cacao core collection. The Criollo genetic group, for instance, is highly homozygous and shows little genetic diversity. However, it is interesting for its fine flavor (Irish et al. 2010). Among the Cuban traditional cacao plants, several individuals with white cotyledons have been identified. This character linked to the fine flavor of cacao should be retained among the plants that constitute the core collection.

Selection of a core collection that fulfills all genetic criteria is impracticable because of the inter-relationships of the evaluated parameters (El Bakkali et al. 2013). Considering all analyzed criteria, the best results were obtained with the maximum genetic diversity algorithm (Marita et al. 2000). The obtained core collection had the lowest number of plants (185) and the highest mean genetic distance (0.486).

All generated core collections contained 133 alleles, of which 98 were rare and 4 % were non-informative, with 95.68 % allelic coverage and complete representation of all morphological characteristics. The only statistically significant differences with respect to the complete collection were in H_o, H_e, the allele frequency distribution, and, evidently, the number of selected plants (Table 3). The maximum genetic diversity algorithm yielded the core collection with the lowest number of selected plants, the lowest redundancy in morphological descriptors (Fig. 3), and the highest phenotypic diversity as evaluated by the Shannon-Weaver diversity index (Table 4). This collection was also highly representative of the geographical distribution of traditional Cuban cacao plants, comprising plants collected in every municipality in Cuba were this cacao was found (Table 2). In summary, this algorithm allows to decrease the geographical, morphological, and genetic redundancy of the complete traditional Cuban cacao collection, without diminishing its diversity.

The most representative core collection of traditional Cuban cacao was selected with the maximum genetic diversity algorithm. A similar result was obtained by Gustavsson (2008) who compared two algorithms to generate a lingonberry core collection based on RAPD molecular markers. Other comparisons of different core collection algorithms where made by Escribano et al. (2008) and Studnicki et al. (2013) without including, however, the maximum genetic diversity algorithm.

Nowadays, the selection of plants for a core collection of cacao (Zhang et al. 2009; Irish et al. 2010; Motilal et al. 2013) or other species (Garkava-Gustavsson et al. 2005; Escribano et al. 2008; Lin et al. 2011) is based almost exclusively on molecular markers. Few core collections are based on morphological descriptors (Studnicki et al. 2013). In the present work, both SSR genetic diversity and qualitative morphological diversity were captured for the constitution of the traditional Cuban cacao core collection. This combined approach is not common and could be used in the selection of core collections of cacao or other species.

The present study represents the first analysis of a cacao plant collection carried out in Cuba to select a core collection for ex situ conservation and future use. Although in situ conservation is an easier, less-expensive option only requiring plants to be located and identified, it carries a high risk of losing genetic material due to factors such as replacement and tree death. As an alternative, ex situ conservation in a location similar to the original site is widely used to safeguard important populations in danger of extinction or replacement. An ex situ strategy also allows easy, immediate access to conserved material for research or production (Normah et al. 2013). Compared with the complete collection of traditional Cuban cacao plants, the core collection generated in the present study can serve as the basis of a more efficient selection program (Engels and Visser 2003). This core collection is more economical and should accelerate the efficient exploitation of adaptations and beneficial alleles in traditional Cuban cacao plants for use in genetic improvement programs (Reeves et al. 2012).

Conclusions

Using each of the five selection algorithms, at least 95 % of alleles and 94.23 % of rare alleles in the complete traditional Cuban cacao plant collection were captured in a reduced number of plants. The core collection generated using the maximum genetic diversity algorithm had the lowest number of accessions (185) and the highest mean genetic distance (0.486). This collection also featured the lowest morphological redundancy and the highest phenotypic diversity as assessed by its mean Shannon-Weaver diversity index value (0.757).

On the basis of the present results, we recommend the use of the maximum genetic diversity algorithm described by Marita et al. (2000) to select traditional Cuban cacao plants for conservation and future use in the genetic improvement of Cuban cacao. Also, this algorithm could be used for the selection of core collections of cacao and other species. A program of genetic improvement based on local genetic resources, potentially more adapted to Cuban conditions, should be more efficient (with respect to adaptation) than programs based on high-yielding introductions.

Acknowledgments We thank the Commission de Coopération au Développement of Belgium (ARES-CCD) for financial support. We are also grateful to farmers, researchers, and technicians of the IIAB and the staff of the Cuban provincial agricultural delegations for their cooperation.

References

Agrama HA, Yan WG, Lee F, Fjellstrom R, Chen MH, Jia M, McClung A (2009) Genetic assessment of a mini-core subset developed from the USDA rice genebank. Crop Sci 49:1336–1346

Aho K, Winston V, Roberts D (2013) Asbio: a collection of statistical tools for biologists. Version 0.4-11

Bekele F, Butler DR (2000) Proposed short list of cocoa descriptors for characterisation. Paper presented at the Working procedures for cocoa germplasm evaluation and selection. In: Eskes AB, Engels JMM, Lass RA (eds) Proceedings of the CFC/ICCO/IPGRI Project Workshop. IPGRI, Montpellier, pp 41–48

Bidot Martínez I, Riera Nelson M, Flamand M-C, Bertin P (2015) Genetic diversity and population structure of anciently introduced Cuban cacao plants. Genet Resour Crop Evol 62:67–84. doi:10.1007/s10722-014-0136-z

Bidot Martínez I, Valdez de la Cruz M, Riera Nelson M, Bertin P (2016) Morphological characterization of traditional cacao (*Theobroma cacao* L.) plants in Cuba. Genet Resour Crop Evol. doi: 10.1007/s10722-015-0333-4

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457. doi:10.1038/368455a0

Brown AHD (1989) Core collection: a practical approach to genetic resources management. Genome 31:818–824

Brown JS, Sautter RT, Olano CT, Borrone JW, Kuhn DN, Motamayor JC, Schnell RJ (2008) A composite linkage map from three crosses between commercial clones of cacao, Theobroma cacao L. Trop Plant Biol 1:120–130

Çalişkan M (ed) (2012) Genetic diversity in plants. InTech, Croatia

CIRAD (2013) CocoaGenDB. http://cocoagendb.cirad.fr/index.html. Accessed 18 April 2013.

R Core Team (2013) R: A language and environment for statistical computing. Version 3.01

de Mendiburu F (2013) Agricolae: statistical procedures for agricultural research. Version 1:1–4

El Bakkali A, Haouane H, Moukhli A, Costes E, Van Damme P, Khadari B (2013) Construction of core collections suitable for association mapping to optimize use of mediterranean olive (*Olea europaea* L.) genetic resources. PLoS One 8(5):e61265. doi:10.1371/journal.pone.0061265

Engelmann F, Dulloo ME, Astorga C, Dussert S, Anthony F (eds) (2007) Complementary strategies for *ex situ* conservation of coffee (*Coffea arabica* L.) genetic resources. A case study in CATIE, Costa Rica, Topical reviews in Agricultural Biodiverstity. Biodiversity International, Rome

Engels JMM, Visser L (eds) (2003) A guide to effective management of germplasm collections. IPGRI handbooks for genebanks, vol 6, Rome.

Engels JMM, Bartley BGD, Enriquez GA (1980) Cacao descriptors, their stage and modus operandi. Turrialba 30(2):209–218

Escribano P, Viruel MA, 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 underutilised subtropical fruit tree species. Ann Appl Biol 153:25–32

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

Frankel OH, Brown AHD (1984) Current plant genetic resources—a critical appraisal. In: Chopra VL, Joshi BC, Sharma RP, Bansal HC (eds) Genetics: new frontiers, vol 4, Proceedings of the XV International Congress of Genetic. Oxford and IBH, New Delhi, pp 1–11

Garkava-Gustavsson L, Persson HA, Nybom H, Rumpunen K, Gustavsson BA, Bartish IV (2005) RAPD-based analysis of genetic diversity and selection of lingonberry (Vaccinium vitis-idaea L.) material for ex situ conservation. Genet Resour Crop Evol 52:723–735

Glaszmann JC, Kilian B, Upadhyaya HD, Varshney RK (2010) Accessing genetic diversity for crop improvement. Curr Opin Plant Biol 13:1–7. doi:10.1016/j.pbi.2010.01.004

Grenier C, Deu M, Kresovich S, Bramel-Cox PJ, Hamon P (2000) Assessment of genetic diversity in three subsets constituted from the ICRISAT sorghum collection using random vs non-random sampling procedures B. Using molecular markers. Theor Appl Genet 101:197–202

Guiltinan MJ (2007) Cacao. In: Pua EC, Davey MR (eds) Biotechnology in agriculture and forestry, vol 60, Transgenic Crops V. Springer-Verlag, Berlin Heidelberg, pp 497–518

Gustavsson L (2008) Genetic diversity in fruit and berry crops estimated with molecular markers. PhD thesis, Swedish University of Agricultural Sciences, Alnarp

Hu J, Zhu J, Xu HM (2000) Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. Theor Appl Genet 101:264–268

Igartua E, Gracia MP, Lasa JM, Medina B, Molina-Cano JL, Montoya JL, Romagosa I (1998) The Spanish barley core collection. Genet Resour Crop Evol 45:475–481

Irish BM, Goenaga R, Zhang D, Schnell R, Brown JS, Motamayor JC (2010) Microsatellite fingerprinting of the USDA-ARS Tropical Agriculture Research Station cacao (*Theobroma cacao* L.) germplasm collection. Crop Sci 50:656–667

Jansen J, van Hintum T (2007) Genetic distance sampling: a novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. Theor Appl Genet 114:421–428

Jombart T (2008) Adegenet: a R package for the multivariate analysis of genetic markers. Bioinform 24:1403–1405. doi:10.1093/bioinformatics/btn129

Jombart T, Ahmed I (2011) Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinform 27:3070–3071. doi:10.1093/bioinformatics/btr521

Kim K-W, Chung H-K, Cho G-T, Ma K-H, Chandrabalan D, Gwag J-G, Kim T-S, Cho E-G, Park Y-J (2007) PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. Bioinform 23(16):2155–2162

Kumar P, Gupta VK, Misra AK, Modi DR, Pandey BK (2009) Potential of molecular markers in plant biotechnology. Plant Omics J 2(4):141–162

Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A (1999) Isolation and characterization of microsatellites in *Theobroma cacao* L. Mol Ecol 8:2141–2145

Le Cunff L, Fournier-Level A, Laucou V, Vezzulli S, Lacombe T, Adam-Blondon A-F, Boursiquot J-M, This P (2008) Construction of nested genetic core collections to optimize the exploitation of natural diversity in *Vitis vinifera* L. subsp. sativa. BMC Plant Biol 8(31):1–12. doi:10.1186/1471-2229-8-31

Li Q, Xu Z, He T (2002) *Ex situ* genetic conservation of endangered *Vatica guangxiensis* (Dipterocarpaceae) in China. Biol Conserv 106:151–156

Lin Z, Wei Z, Jun C, Yong H, Xing S, Li L, Sheng Q (2011) Analysis of genetic diversity and construction of core collection of local mulberry varieties from Shanxi province based on ISSR marker. Afr J Biotechnol 10(40):7756–7765

Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinform 21:2128–2129

Livingstone D, Royaert S, Stack C, Mockaitis K, May G, Farmer A, Saski C, Schnell R, Kuhn D, Motamayor JC (2015) Making a chocolate chip: development and evaluation of a 6K SNP array for *Theobroma cacao*. DNA Res:1–15. doi:10.1093/dnares/dsv009

Marita JM, Rodriguez JM, Nienhuis J (2000) Development of an algorithm identifying maximally diverse core collections. Genet Resour Crop Evol 47:515–526

Motamayor JC, Lachenaud P, da Silva JW, Loor R, Kuhn DN, Brown JS, Schnell RJ (2008) Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L). PLoS One 3(10):e3311

Motilal LA, Zhang D, Mischke S, Meinhardt LW, Umaharan P (2013) Microsatellite-aided detection of genetic redundancy improves management of the International Cocoa Genebank, Trinidad. Tree Genet Genomes 9:1395–1411

Normah MN, Malik SK, Chaudhury R, Salma I, Makeen MA (2013) Conservation of tropical fruit genetic resources. In: Normah MN (ed) Conservation of tropical plant species. Springer Science + Business Media, New York, pp 137–170

Núñez N (2010) El cacao y el chocolate en Cuba, 2nd edn. Fundación Fernando Ortíz, La Habana

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinform Appl Note 20(2):289–290. doi:10.1093/bioinformatics/btg412

Ramanatha V, Hodgkin T (2002) Genetic diversity and conservation and utilization of plant genetic resources. Plant Cell Tissue Organ Cult 68:1–19

Rao ES, Kadirvel P, Symonds RC, Geethanjali S, Ebert AW (2012) Using SSR markers to map genetic diversity and population structure of *Solanum pimpinellifolium* for development of a core collection. Plant Genet Resour: Charact Util 10(1):38–48

Reeves PA, Panella LW, Richards CM (2012) Retention of agronomically important variation in germplasm core collections: implications for allele mining. Theor Appl Genet 124(6):1155–1171. doi:10.1007/s00122-011-1776-4

Saunders JA, Mischke S, Leamy EA, Hemeida AA (2004) Selection of international molecular standards for DNA fingerprinting of *Theobroma cacao*. Theor Appl Genet 110:41–47

Schliep KP (2011) phangorn: phylogenetic analysis in R. Bioinform 27(4):592–593

Schoen DJ, 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

Shannon CE, Weaver W (1949) The mathematical theory of communication. Urbana Press, University of Illinois, Champaign, IL

Sounigo O, Bekele FL, Iwaro AD, Thévenin JM, Bidaisee G, Umaharan R, Sankar A, Sukha DA, Motilal LA, Butler DR, Eskes A (2006) Description of cocoa clones proposed for the CFC/ICCO/IPGRI project collection. In: Eskes A, Yoel E (eds) Global approaches to cocoa germplasm utilization and conservation: Final report of the CFC/ICCO/IPGRI project on "Cocoa germplasm utilization and conservation: A global approach" (1998–2004). Amsterdam, pp 67–81

Spooner D, van Treuren R, de Vicente MC (2005) Molecular markers for genebank management. IPGRI Technical Bulletin, vol 10. IPGRI, Rome

Studnicki M, Mądry W, Schmidt J (2013) Comparing the efficiency of sampling strategies to establish a representative in the phenotypic-based genetic diversity core collection of orchardgrass (*Dactylis glomerata* L.). Czech J Genet Plant Breed 2013(1):36–47

Thachuk C, Crossa J, Franco J, Dreisigacker S, Warburton M, Davenport GF (2009) Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. BMC Bioinform 10:243

Turnbull CJ, Hadley P. International Cocoa Germplasm Database (ICGD) (2011) NYSE Liffe/CRA Ltd./University of Reading, UK. http://www.icgd.reading.ac.uk Accessed 12/7/2013

USDA-ARS (2013) Germplasm Resources Information Network - (GRIN). National Germplasm Resources Laboratory. http://www.ars-grin.gov/. Accessed 18 April 2013.

van Hintum TJL, Brown AHD, Spillane C, Hodgkin T (2000) Core collections of plant genetic resources. IPGRI Technical Bulletins No. 3. International Plant Genetic Resources Institute, Rome

van Hintum TJL, Sackville Hamilton NR, Engels JMM, van Treuren R (2002) Accession management strategies: splitting and lumping. In: Engels JMM, Ramanatha Rao V, Brown AHD, Jackson MT (eds) Managing plant genetic diversity., pp 113–120

Zhang D, Mischke S, Goenaga R, Hemeida AA, Saunders JA (2006) Accuracy and reliability of high-throughput microsatellite genotyping for cacao clone identification. Crop Sci 46:2084–2092. doi:10.2135/cropsci2006.01.0004

Zhang D, Mischke S, Johnson ES, Phillips-Mora W, Meinhardt L (2009) Molecular characterization of an international cacao collection using microsatellite markers. Tree Genet Genomes 5:1–10

Zhao W, Cho GT, Ma KH, Chung JW, Gwag JG, Park YJ (2010) Development of an allele mining set in rice using a heuristic algorithm and SSR genotype data with least redundancy for the post-genomic era. Mol Breed 26:639–651

Zorić M, Dodig D, Kobiljski B, Quarrie S, Barnes J (2012) Population structure in a wheat core collection and genomic loci associated with yield under contrasting environments. Genetica 140:259–275