Tree Genetics & Genomes (2009) 5:1–10
DOI 10.1007/s11295-008-0163-z

ORIGINAL PAPER

Molecular characterization of an international cacao collection using microsatellite markers

Dapeng Zhang · Sue Mischke · Elizabeth S. Johnson · Wilbert Phillips-Mora · Lyndel Meinhardt

Received: 31 December 2006 / Revised: 7 April 2008 / Accepted: 18 April 2008 / Published online: 24 June 2008

Abstract Plant germplasm collections invariably contain varying levels of genetic redundancy, which hinders the efficient conservation and utilization of plant germplasm. Reduction of genetic redundancies is an essential step to improve the accuracy and efficiency of genebank management. The present study targeted the assessment of genetic redundancy and genetic structure in an international cacao (*Theobroma cacao* L.) collection maintained in Costa Rica. A total of 688 cacao accessions maintained in this collection were genotyped with 15 simple sequence repeat (SSR) loci, using a capillary electrophoresis genotyping system. The SSR markers provided a high resolution among the accessions. Thirty-six synonymously labeled sets, involving 135 accessions were identified based on the matching of multilocus SSR profiles. After the elimination of synonymous sets, the level of redundancy caused by closely related accessions in the collection was assessed using a simulated sampling scheme that compared allelic diversity in different sample sizes. The result of the simulation suggested that a random sample of 113 accessions could capture 90% of the total allelic diversity in this collection. Principal Coordinate Analysis revealed that the Trinitario hybrids from Costa Rica shared a high similarity among groups as well as among individual accessions. The analysis of the genetic structure illustrated that the within-country/within-region difference accounted for 84.6% of the total molecular variation whereas the among-country/among-region difference accounted for 15.4%. The Brazilian germplasm contributed most to this collection in terms of total alleles and private alleles. The intercountry/interregion relationship by cluster analysis largely agreed with the geographical origin of each germplasm group and supported the hypothesis that the Upper Amazon region is the center of diversity for cacao. The results of the present study indicated that the CATIE International Cacao Collection contains a high level of genetic redundancy. It should be possible to rationalize this collection by reducing redundancy and ensuring optimal representation of the genetic diversity from distinct germplasm groups. The results also demonstrated that SSR markers, together with the statistical tools for individual identification and redundancy assessment, are technically practical and sufficiently informative to assist the management of a tropical plant germplasm collection.

Keywords *Theobroma cacao* L. · Cacao · Genetic diversity · Population structure · Conservation · SSR molecular markers · Genebank · CATIE · Costa Rica

Communicated by D. Grattapaglia

D. Zhang (✉) · S. Mischke · E. S. Johnson · L. Meinhardt
Sustainable Perennial Crops Laboratory, Plant Sciences Institute, Beltsville Agricultural Research Center, USDA/ARS, Bldg. 1,
Beltsville, MD 20705, USA
e-mail: ZhangD@ba.ars.usda.gov

W. Phillips-Mora
Laboratorio de Biotecnologia, Tropical Agricultural Research and Higher Education Center (CATIE),
P.O. Box 7170, Turrialba, Costa Rica

Introduction

Cacao is an important tropical crop native to South America (Cuatrecasas 1964; Young 1994 Dias 2001). The Upper Amazon is generally believed to be the center of origin of cacao because the greatest morphological diversity is observed in this region (Cheesman 1944; Bartley 2005). The species comprises a large number of highly morpho-

logically variable and mutually interfertile populations. A cacao pod may contain numerous seeds but the seeds are recalcitrant (Cheesman 1944; Cuatrecasas 1964). As a result, cacao germplasm collections must be maintained as clonally propagated living trees. A dozen major cacao germplasm collections in tropical regions of the world serve as germplasm repositories.

As is the case with many other plant germplasm collections, cacao collections invariably contain duplicate accessions, both within and between genebanks. Records of accessions of global holdings comprise 27,800 clone names, 14,000 of which are synonyms, which represent 13,800 separate clones (International Cacao Germplasm Database, http://www.icgd.rdg.ac.uk). This high rate of redundancy not only hinders the efficient conservation of these collections but also hampers the effectiveness of germplasm evaluation and utilization. Comprehensive assessment of individual identity and population structure have been identified as high priority tasks for cacao genebank management (Turnbull et al. 2004).

Molecular markers such as random amplified polymorphic DNA and amplified fragment length polymorphism have sufficient discriminatory power to distinguish clones, but these tools often fail to reach clear conclusions in the identification of duplicates and mislabeled genotypes. These markers do not identify duplicates by exact match of banding pattern. Rather, assessment is by similarity (or distance) estimation. "Identical" clones are declared when the similarity reaches certain threshold values (Christopher et al. 1999; Perry et al. 1998; Sounigo et al. 2001). Since the development of simple sequence repeats (SSR) markers in cacao (Lanaud et al. 1999), SSR-based DNA fingerprinting has been increasingly applied in cacao germplasm characterization (Lanaud et al. 2001; Motamayor et al. 2002; Saunders et al. 2004; Cryer et al. 2006; Schnell et al. 2005; Takrama et al. 2005). To date, however, reports have been scarce on large scale application of SSR markers to identify duplicate clones and assess the structure of genetic diversity in cacao genebanks.

The cacao collection at CATIE is one of the two international cacao germplasm collections in the world. The other is the International Cocoa Genebank, Trinidad, which is curated by the University of West Indies, Trinidad. The CATIE collection was initiated in 1944 to promote the exchange of germplasm of tropical crops. At the time this study was initiated, this collection maintained 745 clones or accessions from Central America, Mexico, South America, the Caribbean, Asia, and Africa. Some of these accessions, such as the criollo genotypes, are not found in other collections. The collection has been an important source for resistance to frosty pod and Phythophtora pod rot diseases. In 1978, the collection was catalogued by the International Board for Plant Genetic Resources IBPGR (now Bioversity International) as one of the two "International Cacao Collections", and since 2004, it is under the auspices of the Food and Agriculture Organization and covered by an international treaty for the protection of plant genetic resources (Phillips-Mora et al. 2006). As with most other cacao germplasm collections, passport data for this collection is incomplete. Some primary and secondary contributors of germplasm were unable to guarantee the authenticity of the material supplied. This is considered a common cause of the introduction of mislabeled accessions into cacao collections.

In this paper, we report a study in which 15 SSR loci were used to characterize the CATIE collection. Our first objective was to identify duplicate accessions and reduce redundancy in this collection, and our second objective was to assess level and organization of genetic diversity in this collection and provide baseline information to explore potential measures of genebank rationalization. We show that SSR markers, together with the statistical tools for individual identification and redundancy assessment, are technically practical and sufficiently informative to assist in the management of tropical plant germplasm collections. We believe that this example of using molecular marker technology for genebank management should have wide implications for other tropical perennial crops. This study is a part of the international collaborative project on DNA fingerprinting of cacao germplasm in the Americas.

Materials and methods

Plant material and DNA isolation

A total of 688 cacao accessions were used in the present study. The samples used for DNA fingerprinting profiles included leaves of various ages collected from individual cacao accessions held at CATIE's International Cacao Collection in Turrialba, Costa Rica. Each sampled branch was tagged for potential revisiting. Both accession name and DNA extraction number were used to label each sample. DNA was extracted according to the instruction of DNeasy Plant System (Qiagen Inc., Valencia, CA, USA). Modifications were made to cope with the high level of endogenous phenolics in the cacao leaf samples (Saunders et al. 2004; Zhang et al. 2006).

SSR analysis

DNA amplification used primer sets with sequences previously described (Lanaud et al. 1999; Risterucci et al. 2000; Saunders et al. 2004). Primers were synthesized by Proligo (Boulder, CO, USA), and forward primers were 5′-

labeled using WellRED fluorescent dyes (Beckman Coulter Inc., Fullerton, CA, USA). Polymerase chain reaction (PCR) was performed as described in Saunders et al (2004), using commercial hot-start PCR supermixes that had been fortified with an additional 30 U of the respective hot-start Taq DNA polymerase (Invitrogen Platinum Taq, Carlsbad, CA, USA; Eppendorf HotMaster Taq, Brinkman, Westbury, NY, USA) added to each milliliter of the supermix. The amplified PCR products were separated by capillary electrophoresis as previously described (Saunders et al. 2004) using a CEQ 8000 genetic analysis system (Beckman Coulter Inc.). Data analysis was performed using the CEQ 8000 Fragment Analysis software version 7.0.55 according to manufacturer's recommendations (Beckman Coulter Inc). SSR fragment sizes were automatically calculated to two decimal places by the CEQ 8000 Genetic Analysis System. Allele calling was performed using the CEQ 8000 binning wizard software (CEQ 8000 software version 7.0.55, Beckman Coulter Inc.).

Identification of duplicates

Duplicates were identified by using pair-wise comparisons among all 688 individuals based on their multilocus SSR profile. The computer program GenAlEx 6 (Peakall and Smouse 2006) was used for genotype matching. Accessions with different names that were fully matched at 15 loci were declared duplicates or synonymously mislabeled accessions. Statistical rigor was also assessed for match declaration to determine whether two individuals may share the same multilocus genotype by chance (Waits et al. 2001). Probability of identity (PID) between siblings (PID-sib) was estimated using GIMLET V.1.3.2 (Valière 2002). The identified synonymous sets were excluded in the subsequent analysis of genetic diversity and population structure.

Assessment of genetic redundancy caused by closely related accessions

In addition to the duplicates, we also assessed the level of genetic redundancy caused by closely related accessions. The assessment was carried out by measuring allelic richness against a given number of individuals, following the sampling method of maximization strategy (Schoen and Brown 1993). The procedure was originally designed for the development of germplasm core collections implemented in the computer program MSTRAT (Gouesnard et al. 2001). For each simulated sampling, Shannon's diversity index was used to represent the sampled diversity.

To further illustrate the redundant contribution from the Trinitario hybrids from Costa Rica, pair-wise Euclidian distances were computed for the 221 accessions in ARF, PMCT, CC and UF groups using the program of GenAlEx 6 (Peakall and Smouse). The pairwise distances were then presented by Principle Coordinates Analysis (PCO) using the same program.

Analysis of genetic diversity

After the exclusion of synonymous sets, the accessions with clear record of introduction, including 548 accessions (five accessions with unknown origin were not used in the diversity analysis), was classified into 12 groups based on their geographical origin, including Brazil, Central America, Colombia, Ecuador, Mexico, Peru, Trinidad, and French Guiana. The Trinitario hybrids from Costa Rica, PMCT, CC, UF, and ARF, were each treated as an independent group in order to illustrate the interrelationship among these groups. Summary statistics for measuring genetic diversity were computed using the program PowerMarker (Liu and Muse 2005). These estimations included mean number of alleles per locus, observed heterozygosity and gene diversity (expected heterozygosity), and within-population inbreeding coefficient. Deficiency of heterozygosity for microsatellite loci was checked using Wilcoxon test (Cornuet and Luikart 1996).

Pair-wise genetic distances (Nei et al. 1983) among the 12 groups were calculated using PowerMarker and the dendrogram was produced using the software TreeView Version 1.6.6. (Page 1996). To understand the hierarchical structure of the molecular variation in this collection, the program of analysis of molecular variance (AMOVA; Excoffier et al. 1992) implemented in the software of Arlequin 3.0 (Excoffier et al. 2005) was used for computation. The total molecular variance was partitioned into the components of among-group and among-individual/within-group. The significance of Φ statistics (Excoffier et al. 1992) was tested by permutation, with the probability of nondifferentiation, for 1,000 randomizations.

Results

Identification of duplicates groups

Individual genotype matching (pair-wise comparisons) based on multilocus SSR profiles identified 36 synonymous sets, involving 135 accessions (Table 1). Within each group, the individuals shared identically sized alleles in all 15 loci and, thus, met our definition of mislabeling. The size of the synonymously mislabeled group ranged from two to 36 clones per group. The rate of mislabeling varies greatly among the accession groups. The rate of duplicates in the introduced germplasm groups (i.e., RIM, EET, and SGU) was higher than that of the local accession groups (i.e., UF, CC, PMCT, and Criollo). In total, the duplicated accessions

Table 1 Thirty-six synonymous groups (involving 135 accessions) within the CATIE cocoa collection identified by microsatellite DNA analysis

Group	Accession
1	Amelonado_15
	Amelonado_22
2	Amelonado 13
	Catongo Blanco
3	ARF-3
	4-D
4	ARF-1
	ARF-4
	ARF-9
5	CC-38
	CC-46
6	CC_41
	CC-47
	CC-49
8	CC-213
	CC-215
9	CC-30
	CC-80
10	CC-256
	EEG-29
11	CC-232
	EET-48
7	CCN-10(A)
	CCN-10(B)
	CCN-16
	CCN-51
12	Criollo-11
	Criollo-22
13	Criollo-14
	Criollo-50
14	CU-1
	CU-2
15	EET-183
	EET-283
16	EET-228
	EET-408
17	EET-62
	EET-67
	EET-94
	EET-96
	EET-162
18	EET-236
	EET-381
	P-8
	P-22
	RIM-2
	RIM-6
	RIM-8
	RIM-9
	RIM-13
	RIM-15
	RIM-19
	RIM-21
	RIM-23
	RIM-24
	RIM-34
	RIM-39
	RIM-41
	RIM-43
	RIM-48
	RIM-52
	RIM-56
	RIM-71
	RIM-75
	RIM-76
	RIM-78
	RIM-105
	RIM-106
	RIM-117
	UF-667_set_D
	UF-676
	UF-677a
	UF-677b
19	GU-147N
	GU-171N
20	GS-36
	ICS-29
	ML-106
	OC-77
	P-7p
	Papayo-K20
21	PMCT-11
	PMCT-29
22	PMCT-15
	PMCT-6
23	PMCT-10
	PMCT-18
24	PMCT-30
	PMCT-32
25	PMCT-71
	ARF-5
26	RB-43
	RB-46
	RB-41
27	SGU-2
	SGU-60
	SGU-71
	SGU-72
28	SGU-53
	SGU-67
	SGU-75
	SGU-90
	SGU-93
	SGU-94
	SGU-104
29	SIAL-325
	SIC-6
30	SIAL-163
	SIAL-339
	SIAL-98
	SIC-2
	Laranjo_156
	Para
	YLV-F
31	SIC-1
	SIC-7
32	SIC-329
	SIC-433
	SIC-802
	YLV-E
33	UF-11
	UF-12
	UF-10A
	UF-168
	ICS-39
	EQX-100
	GS-46
34	UF-242
	UF-713
	ICS-4
	UF-705
35	Yamada_LVD
	YLV-A
36	5-E
	ARF-10

Accessions in same synonymous group shared identical multilocus SSR profiles

accounted for approximately 19.6% of the clones genotyped from this collection. The clones in the 36 synonymous groups, together with the clones lacking clear passport data, were excluded in subsequent analyses of genetic diversity and population structure.

Assessment of genetic redundancy

The simulation between sample size and diversity representation revealed a high level of redundancy in the CATIE collection. The maximum diversity, as measured by Shannon diversity index, is 126 averaged over ten simulated measurements (Fig. 1). The relationship between sample size and allelic diversity shows that 90% of the diversity can be captured with a sample of 113 accessions, whereas 95% of the diversity can be captured with a sample of 245 accessions. Continued increasing in sample size will only result to a negligible increase of diversity.

An example of genetic redundancy in this collection was illustrated by the large number of Trinitario hybrids, as presented by the Principle Coordinates Analysis (Fig. 2). The plot shows that there is little genetic difference among the four hybrids groups. Their contribution to the overall allelic diversity overlapped both at population and individual level.

Analysis of genetic diversity

A total of 231 alleles were identified in the 548 accessions with distinctive SSR genotypes, with the mean of 14.2 alleles/locus. Within each of the geographical groups, the number of alleles ranged from 2.1 in the French Guiana to 9.7 in Brazil group (Table 2). The private allelic richness was much higher in the introduced germplasm groups from South America. The two groups from Brazil and Ecuador alone accounted for 56% of the total number of private alleles, although these two groups only contributed 18.4% of the total accessions used in the diversity analysis. In contrast, the four hybrids groups from Costa Rica (ARF, PMCT, CC, and UF) comprised 40% of the total accessions but only two private alleles were detected in these four groups (Table 2).

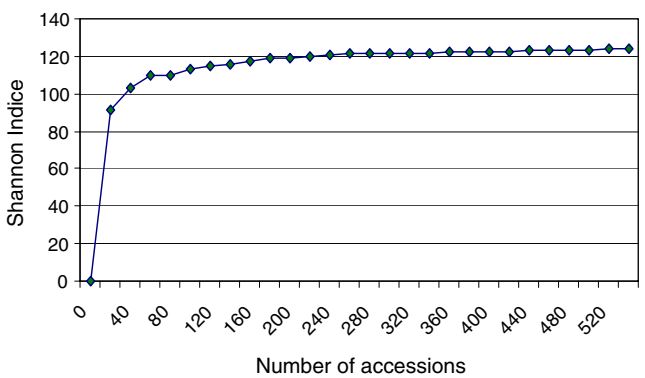

Fig. 1 Simulated relationship between sample size and genetic diversity (measured by Shannon's diversity index) in the CATIE cacao collection

The mean gene diversity was 0.51 and the observed heterozygosity was 0.46 in the 12 germplasm groups (Table 2). Heterozygosity deficiency was detected in the group of French Guiana (H_e=0.30, H_o=0.13) and the group of Brazil (H_e=0.60; H_o=0.34) by Wilcoxon test (Table 2). In concurrence with the result of heterozygosity deficiency, significant positive inbreeding coefficients were found in French Guiana (f=0.571, P<0.01) and Brazil (f=0.445, P< 0.01), respectively. Heterozygosity deficiencies were not observed in the other ten groups which all had nonsignificant inbreeding coefficient (Table 2).

The results of AMOVA showed significant genetic variation both within and among the accession groups (Table 3). The within-group variation accounted for 84.6% of the total molecular variance, whereas the inter-accession group accounted for the 15.4% (Table 3). The contribution to the intergroup variability by each germplasm group was also reflected by the mean genetic distances among these groups (Table 4). In general, the germplasm groups introduced from South America made more contribution to the intergroup variability than the local groups. The group from French Guiana had the highest mean genetic distance (D=0.848, followed by the group from Peru (D=0.309), Colombia (D=0.282), and Ecuador (D=0.255). The mean distance of the four Trinitario hybrids groups (ARF, CC, PMCT, and UF) ranged from 0.161 to 0.204 (Table 4).

Clustering of the 12 accession groups based on Nei's genetic distance (Nei et al. 1983) resulted in a dendrogram that largely agreed with the known geographical origin of these groups (Fig. 3). Accessions from Mexico, Central America, and Trinidad were closely grouped together. The Trinitario hybrids (ARF, PMCT, CC, and UF) all fall between the Mexico–Central America–Trinidad group and the South American groups, which reflected their hybrid nature. Among the 12 accession groups, the French Guiana wild cacao had the most unique position in the dendrogram.

Discussion

Plant genebank curators are increasingly challenged to conserve broader genetic diversity without a proportional increase in operational funding. Genetic redundancy, due to duplicates and closely related accessions, is a common problem in the management of plant genebanks (Engels and Visser 2003). A high rate of redundancy in cacao collections not only increases the cost of genebank management, it also hampers the potential exploitation of

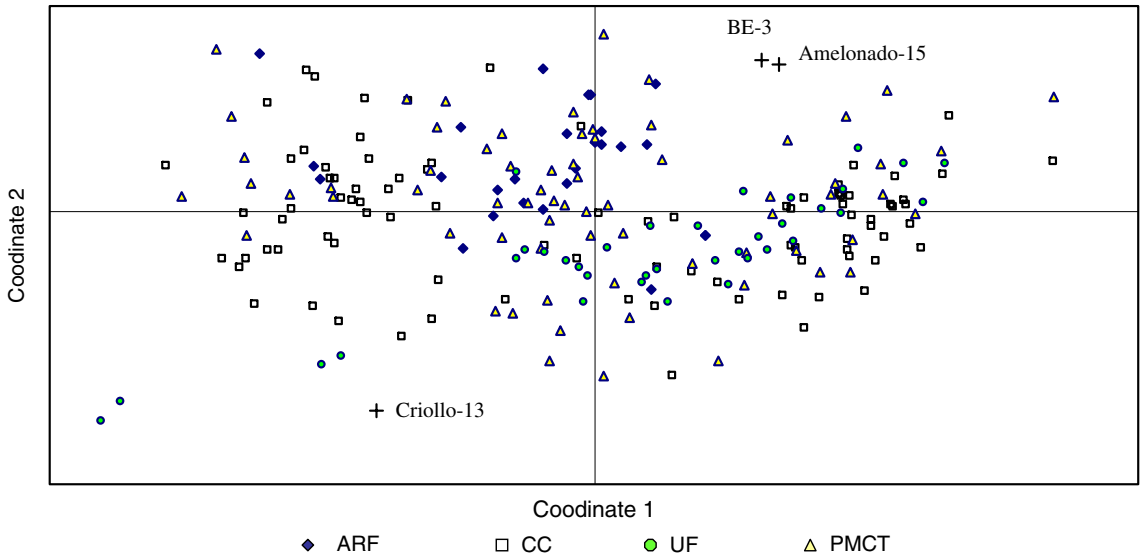

Fig. 2 PCO plot of 221 cocoa accessions belonging to the ARF, CC, PMCT, and UF accession groups. Clone Criollo-13, BE-3, and Amenolado-15 were included as external controls (first axis=34.7% of total information and the second=20.8%)

germplasm for crop improvement. For many tropical tree species, the demand for comprehensive assessments of genetic redundancy and diversity is particularly urgent because their seeds are often recalcitrant and the germplasm can only be maintained in a field genebank.

In cacao germplasm collections, large efforts have been made to identify these redundant accessions using morphological data (Motilal and Butler 2003; Bekele et al. 2006). However the results were often not conclusive due to the plasticity and seasonality of the morphological characteristics. Molecular analysis can complement traditional approaches for identifying duplicates. The development of simple sequence repeats markers in cacao (Lanaud et al. 1999) have significantly enhanced the capacity of individual identification of cacao germplasm. This technique allows unambiguous verification of duplicates by a matching of multilocus DNA profiles between the potential duplicates (Cryer et al. 2006; Takrama et al. 2005; Zhang et al. 2006). In our previous papers, we have demonstrated the effectiveness of using SSR markers for the identification of synonymously mislabeled accessions and that the use of 15 SSR loci for cacao individual identification is statistically rigorous (Zhang et al. 2006; Johnson et al. 2007). In the present study, we identified 135 duplicate accessions which accounts for 19.6% of the assessed accessions. The benefit of using molecular characterization for cacao genebank management, therefore, is highly significant. However, there are exceptional cases in which

Table 2 Diversity parameters for the 12 germplasm groups (including a total of 548 accessions)

Accession group	*N*	*K*	H_e	H_o	Wilcoxon test	Private allele	Inbreeding coefficient (*f*)
French Guiana	19	2.1	0.30	0.13	0.002	0	0.571
Colombia	12	5.7	0.70	0.68	0.163	3	0.076
Trinidad	40	6.5	0.63	0.60	0.210	2	0.061
Mexico	31	6.9	0.63	0.65	0.281	0	0.000
ARF	26	7.4	0.72	0.79	0.190	1	0.000
PMCT	69	7.7	0.69	0.66	0.134	0	0.053
CC	89	8.9	0.71	0.66	0.221	0	0.065
UF	37	4.8	0.61	0.69	0.381	1	0.000
Central America	90	7.9	0.65	0.52	0.117	2	0.200
Peru	34	8.1	0.71	0.54	0.104	2	0.257
Ecuador	53	8.9	0.76	0.66	0.135	5	0.144
Brazil	48	9.7	0.60	0.34	0.011	9	0.445
Mean	36.5	5.6	0.51	0.46			0.125

K=average number of alleles per locus, Wilcoxon test=*P*-value of Wilcoxon test for heterozygosity deficiency (Cornuet and Luikart 1996) using stepwise mutational model, and *f*=within-population inbreeding coefficient

N Number of accessions in the group, H_e expected heterozygosity (mean gene diversity), H_o observed heterozygosity

Table 3 Analysis of molecular variance for SSR variation among and within 12 germplasm groups held in the CATIE International Cocoa Collection

Source	*df*	SSD	MSD	Variance component	% Total[a]	*P* value[b]
Among populations	11	1097.9	99.82	2.00	15.4	0.01
Within populations	536	6048.8	11.29	11.29	84.6	
Brazil	48	617.0	12.85			
Trinidad	40	397.3	9.93			
C. America	90	1049.1	11.66			
Colombia	12	131.0	10.92			
Ecuador	53	683.3	12.89			
F. Guiana	19	131.6	6.93			
Mexico	31	289.7	9.35			
Peru	34	453.1	13.33			
ARF	26	252.6	9.72			
PMCT	69	750.2	10.87			
CC	89	998.1	11.21			
UF	37	295.6	7.99			
Total	547	7146.8	111.1			

SSD Sum of squared deviations, *MSD* Mean squared deviations
[a] Percent of total molecular variance
[b] Probability of obtaining a larger component estimate. Number of permutations = 1,000

closely related clones are indistinguishable based on the 15 loci. For example, point mutations that may cause phenotypic change, i.e., the change of pod or seed color, are often associated with few mutations. Therefore, morphological examination remains an important tool that can play a complementary role in the identification of duplicates in cacao germplasm. Currently, all these identified duplicate groups are being verified by morphological comparison using a field guide (Johnson et al. 2007).

In addition to duplicates, genetic redundancies caused by closely related accessions can also be systematically assessed using multilocus SSR fingerprinting data. The sampling method of maximization strategy (Schoen and Brown 1993) can assess the level of genetic redundancy in a collection by plotting allelic richness against different sample sizes. An ideal collection with zero redundancy then should have a straight linear relationship between sample size and allelic diversity. On the other hand, a curved linear relationship (Fig. 1) indicates redundancies in the collection. An example of the high rate of redundancy was illustrated by the relationship among the large number of breeding lines selected from various Trinitario hybrid families (Fig. 2). Included in these hybrid breeding lines are the accessions in the groups ARF, CC, PMCT, and UF. The present study showed that the intergroup variation among these groups was small, as illustrated by the PCO analysis (Fig. 2). Moreover, few private alleles were found within these groups (Table 2). The result, thus, demonstrates an overrepresentation of these hybrid breeding lines in the CATIE collection for the purpose of long-term conservation. Rationalization is needed in order to improve the conservation efficiency of this collection. A refined sampling, based on both agronomic traits and molecular data, would be necessary to optimize the representation both in genotype diversity and in allelic diversity. One obvious step of rationalization should be to reduce the large number of Trinitario hybrids maintained in this collection. The saved space then can be used to introduce new germplasm that can fill the diversity gap in this collection.

The results obtained in this study also provide critical baseline information for the development of cacao core collection. Core collections refer to a limited set of accessions derived from an existing germplasm collection chosen to represent the genetic spectrum in the whole collection (Frankel and Brown 1984). The purpose of a

Table 4 Genetic distances (Nei et al. 1983) among the 12 germplasm groups held in CATIE

	ARF	Brazil	Trinidad	Central America	Colombia	Ecuador	French Guiana	Mexico	Peru	PMCT	CC	UF
ARF	0.000											
Brazil	0.124	0.000										
Trinidad	0.146	0.167	0.000									
C. America	0.170	0.157	0.056	0.000								
Colombia	0.223	0.282	0.300	0.266	0.000							
Ecuador	0.195	0.287	0.221	0.296	0.261	0.000						
F. Guiana	1.101	1.146	1.159	1.082	1.460	1.286	0.000					
Mexico	0.176	0.201	0.088	0.049	0.328	0.339	1.001	0.000				
Peru	0.201	0.331	0.422	0.450	0.316	0.308	0.990	0.473	0.000			
PMCT	0.070	0.138	0.089	0.089	0.204	0.164	1.110	0.083	0.273	0.000		
CC	0.160	0.143	0.095	0.042	0.291	0.296	1.150	0.044	0.429	0.094	0.000	
UF	0.159	0.207	0.097	0.116	0.306	0.176	1.232	0.102	0.448	0.094	0.120	0.000
Mean distance	0.182	0.212	0.189	0.185	0.282	0.255	0.848	0.192	0.309	0.161	0.191	0.204

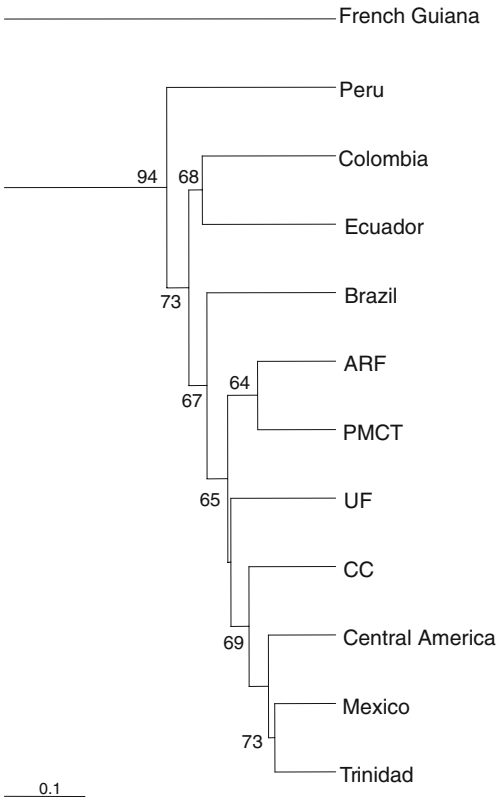

Fig. 3 Clustering of the 12 cocoa accession groups held in the CATIE International Cacao Genebank. The dendrogram was based on Nei's genetic distance with 1,000 permutations. *Numbers* near nodes in the tree are percent bootstrap support for the node. Only bootstrap values higher than 50% are shown

core collection is to facilitate the use of germplasm by providing a preliminary look at the diversity available in the larger collection (Brown 1989). Firstly, the information regarding the level of genetic redundancy will help to determine the sampling intensity for the construction of core collection. The reported sampling intensities vary among species, ranging from 5% to 20% of the total number of accessions (Brown 1989; Schoen and Brown 1993; van Hintum 1999; van Hintum et al. 2000). Our result shows that a sample intensity of 19.3% (113 accessions) would be sufficient to capture 90% of the total allelic diversity. Secondly, the multilocus maker data generated for this collection can be used for marker-assisted sampling of a core collection. In the past, core collections were often developed based on morphological agronomic and other passport data, because genotype data of the entire collection is usually not available. Two sampling strategies, the H and M strategies, were proposed for the construction of core collections based on marker revealed diversity (Schoen and Brown 1993). The H strategy seeks to maximize the diversity in the core collection by sampling accessions from groups in proportion to their within-group genetic diversity. In this strategy, the germplasm will be partitioned into clusters or groups based on a traits, location, or other criteria. Alternatively, the M strategy (or maximization strategy) examines all possible core collections without relying upon stratified sampling. The core collection that maximizes the allelic richness can then be singled out (Schoen and Brown 1993). A sampling algorithm has been developed to implement the maximization strategy for the construction core collections using cacao as an example (Marita et al. 2000). This algorithm sets a sampling intensity for the core collection then maximizes diversity in the core collection by ensuring that only dissonantly related accessions are sampled. The MSTRAT software (Gouesnard et al. 2001) extended the M strategy to include qualitative and quantitative data coding for morphological and agronomic variables. Therefore MSTRAT allows users to quantify diversity both based on molecular markers and morphological and agronomic traits. Since allelic diversity is often uncorrelated with diversity in morphological and agronomic traits in crop germplasm, the inclusion of both complementary measurements is essential for the construction of a cacao core collection. The process of verification and compilation of morphological and agronomic data is still ongoing for the CATIE collection. This data will be combined with the molecular data to develop a core collection for this genebank.

Genetic redundancy is just one of the problems regarding the issue of genetic identity in cacao germplasm. Quite often, accessions in the genebank are received from other collections. The transfer process between cacao collections is frequently subject to errors in identification. Differences in labeling occur across germplasm collections as well as within collections. To fully confirm whether a given accession is mislabeled, we need to compare the SSR profile of this accession with the corresponding original tree. Reference SSR profiles are being developed for each original tree. Then the reference SSR profiles will be compared with the putatively mislabeled accessions. Corrections will then be made as necessary (Turnbull et al. 2004; Boccara and Zhang 2006).

Despite the high rate of redundancy in the CATIE collection, the present study also showed that the CATIE collection holds a high level of genetic diversity in terms of allelic richness. This high level of allelic richness was due to the diverse genetic groups, ranging from Upper Amazon to Mexico, that comprise this collection (Engels 1986, Lockwood and End 1993; Phillips-Mora et al. 2006). The largest allelic contribution, especially the contribution of private alleles, from the Brazilian group suggested the unique genetic composition in the Brazilian germplasm. It is well known that the Brazilian Amazon harbors a diverse range of cacao populations (Bartley 2005; Dias 2001; Sereno et al. 2006). However, so far, little comparative diversity analysis between Brazil and the upper Amazon

region has been carried out. Further investigation is needed to verify the level of allelic diversity in different regions and river systems, which would have significant implications for both in situ and ex situ conservation of cacao genetic diversity.

Acknowledgements We wish to thank Emily Leamy, Elizabeth Gingold, and Sarah Gingold for assisting with the microsatellite genotyping and Antonio Mora, Carlos Astorga, and Ulrike Krauss for their contributions in germplasm identification and sampling. We also wish to thank Lambert Motilal, Ainong Shi and two anonymous reviewers for reviewing the manuscript.

References

Bartley BGD (2005) The genetic diversity of cacao and its utilization. CABI, Wallingford

Bekele F, Bekele I, Butler D, Bidaisee G (2006) Patterns of morphological variation in a sample of cacao (*Theobroma Cacao* L.) germplasm from the International Cocoa Genebank, Trinidad. Genet Resour Crop Evol 53:933–948

Boccara M, Zhang D (2006) Progress in resolving identity issues among the Parinari accessions held in Trinidad: the contribution of the collaborative USDA/CRU project. In: Annual report for 2005. The University of the West Indies, St. Augustine, Trinidad and Tobago

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

Cheesman EE (1944) Notes on the nomenclature, classification and possible relationships of cocoa population. Trop Agric 21:144–159

Christopher Y, Mooleedhar V, Bekele F, Hosein F (1999) Verification of accession in the ICG, T using botanical descriptors and RAPD analysis. In: Annual report 1998. The University of the West Indies, St. Augustine, Trinidad and Tobago, pp pp 15–18

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144:2001–2014

Cryer NC, Fenn MGE, Turnbull CJ, Wilkinson MJ (2006) Allelic size standards and reference genotypes to unify international cocoa (*Theobroma cacao* L.) microsatellite data. Genet Resour Crop Evol 53:1643–1652 http://dx.doi.org/10.1007/s10722-005-1286-9

Cuatrecasas J (1964) Cacao and its allies. A taxonomic revision of the genus Theobroma. Contr U S Natl Herb 35:375–614 Smithsonian Institution, Washington, DC

Dias LAS (2001) Origin and distribution of *Theobroma cacao* L: A new scenario. In: Dias LAS (ed) Genetic improvement of cacao. Available at: http://ecoport.org/ep?SearchType=earticleView&earticleId=197&page=-2

Engels JMM (1986) The systematic description of cacao clones and its significance for taxonomy and plant breeding. Ph.D. Thesis. Agricultural University Wageningen, The Netherlands

Engels JMM, Visser L (2003) A guide to effective management of germplasm collections. IPGRI handbooks for Genebanks No. 6. IPGRI, Rome, Italy

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491

Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50

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. Oxford and IBH, New Delhi, pp 1–11

Gouesnard B, Bataillon T, Decoux G, Rozale C, Schoen DJ, David JL (2001) MSTRAT: an algorithm for building gerplasm core collections by maximizing allelic or phenotypic richness. J Hered 92:93–94

Johnson SE, Mora A, Schnell RJ (2007) Field guide efficacy in the identification of reallocated clonally propagated accessions of cacao. Genet Resour Crop Evol 54:1301–1313

Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A, Lagoda PJL (1999) Isolation and characterization of microsatellites in *Theobroma cacao* L. Mol Ecol 8:2141–2143

Lanaud C, Motamayor JC, Risterucci AM (2001) Implications of new insight into the genetic structure of *Theobroma cacao* L. for breeding strategies. In: Proceedings of the International Workshop on New Technologies for Cocoa Breeding, Kota Kinabalu, Malaysia, London: Ingenic Press, 89–107. http://www.personal.psu.edu/users/a/o/aoa113/ingenic/documents/communications/meetings/past/2000INGENIC.pdf (30 Dec 2005)

Liu J, Muse S (2005) PowerMarker: new genetic data analysis software v3.23. Released 2/1/2005. Free program distributed by author, available from http://www.powermarker.net

Lockwood C, End M (1993) History, technique and future needs for cacao collection. In: Proceedings of the international workshop on conservation, characterization and utilization of cocoa genetic resources in the 21st century. Port-of-Spain, Trinidad and Tobago: The University of the West Indies, Cocoa Research Unit, 1–14

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, Risterucci AM, Lopez PA, Ortiz CF, Moreno A, Lanaud C (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. Heredity 89:380–386

Motilal L, Butler D (2003) Verification of identities in global cacao germplasm collections. Genet Resour Crop Evol 50:799–807

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. J Mol Evol 19:153–170

Page RDM (1996) TreeVIEW: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12:357–358

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

Perry MD, Davey MR, Power JB, Lowe KC, Bligh HFJ, Roach PS, Jones C (1998) DNA isolation and AFLP genetic fingerprinting of *Theobroma cacao* (L.). Plant Mol Biol Report 16:49–59

Phillips-Mora W, Mora A, Johnson ES, Astorga C (2006) Recent efforts to improve the genetic and physical conditions of the international cacao collection at CATIE. Proceedings of the 15th COPAL International Cocoa Research Conference, San Jose, Costa Rica, Oct. 9–14, 2006 pp 611–623

Risterucci AM, Grivet L, N'Goran JAK, Pieretti I, Flament MH, Lanaud C (2000) A high-density linkage map of *Theobroma cacao* L. Theor Appl Genet 101:948–955

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

Schnell RJ, Olano CT, Brown JS, Meerow AW, Cervantes-Martinez C, Nagai C, Motamayor JC (2005) Retrospective determination of the parental population of superior cacao (*Theobroma cacao* L.) seedlings and association of microsatellite alleles with productivity. J Am Soc Hortic Sci 130:181–190

Schoen DJ, Brown AHD (1993) Conservation of allelic richness in wild corp relatives is aided by assessment of genetic markers. Proc Natl Acad Sci USA 22:10623–10627

Sereno ML, Albuquerque PSB, Vencovsky R, Figueira A (2006) Genetic diversity and natural population structure of cacao (*Theobroma cacao* L.) from the Brazilian Amazon evaluated by microsatellite markers. Conserv Genet 7:13–24

Sounigo O, Christopher Y, Bekele F, Mooleedhar V, Hosein F (2001) The detection of mislabelled trees in the International Cocoa Genebank, Trinidad (ICG,T) and options for a global strategy for identification of accessions. In: Bekele F et al (ed) Proc Int Workshop on new technologies and cocoa breeding. Kota Kinabalu, Sabah, Malaysia, 16–17 Oct. 2000. INGENIC, London, pp 34–39

Takrama JF, Cervantes-Martinez CT, Phillips-Mora W, Brown JS, Motamayor JC, Schnell RJ (2005) Determination of off-types in a cacao breeding program using microsatellites. INGENIC Newsl 10:2–8

Turnbull CJ, Butler DR, Cryer NC, Zhang D, Lanaud C, Daymond AJ, Ford CS, Wilkinson MJ, Hadley P (2004) Tackling mislabelling in cocoa germplasm collections. INGENIC Newsl 9:8–11

Valière N (2002) Gimlet, a computer program for analysing genetic individual identification data. Mol Ecol Notes 2:377–379

van Hintum TJL (1999) The general methodology for creating a core collection. In: Johnson RC, Hodgkin T (eds) Core collections for today and tomorrow. International Plant Genetic Resources Institute, Rome, pp 10–17

van Hintum TJL, Brown AHD, Spillane C, Hodgkin T (2000) Core collections of plant genetic resources. IPGRI Tech. Bull. 3. International Plant Genetic Resources Institute, Rome

Waits LP, Luikart G, Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. Mol Ecol 10:249–256

Young AM (1994) The chocolate tree. Smithsonian Institution, Washington and London

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