Tree Genetics & Genomes (2008) 4:821–831
DOI 10.1007/s11295-008-0155-z

ORIGINAL PAPER

Genetic diversity and structure of farm and GenBank accessions of cacao (*Theobroma cacao* L.) in Cameroon revealed by microsatellite markers

Ives Bruno M. Efombagn · Juan C. Motamayor · Olivier Sounigo · Albertus B. Eskes · Salomon Nyassé · Christian Cilas · Ray Schnell · Maria J. Manzanares-Dauleux · Maria Kolesnikova-Allen

Received: 26 November 2007 / Revised: 6 March 2008 / Accepted: 26 March 2008 / Published online: 11 June 2008

Abstract The genetic diversity of 400 accessions collected in cacao farms, 95 GenBank, and 31 reference accessions was analyzed using the 12 microsatellite markers. The GenBank and reference accessions were subdivided into 12 accession groups (AG) that belong to the traditional cacao genetic groups (GG) Lower Amazon Forastero (LA), Upper Amazon Forastero (UA), Trinitario, and Criollo (Cr). The 12-microsatellite loci revealed a total of 125 alleles, 113 of which were present in the farm accession group (FA). The within and between group variation for all AGs accounted respectively for 81% and 19% of the total molecular variation. The average F_{is} for the FA was 0.15 suggesting a moderate level of inbreeding. Significant differences for the level of gene diversity were found between the farm (0.50), GenBank (0.42 to 0.62), and reference (0.10 to 0.60) AGs. Genetic differentiation among AGs was variable with F_{st} values varying between 0.14 and 0.57 for the different AGs. Analysis using a Bayesian model-based method showed the existence of a high level of admixture for the farm accessions group. The LA genes were most represented in the FA (54%), followed by UA (33%) and Cr (7%). The genes of LA were also the most represented in the GenBank (48%), followed by UA (24%) and Cr (14%). Only 14% and 6% of the genes of the GenBank and farm accessions, respectively, could not be attributed to any of the reference GGs. The results suggest the predominating presence of LA genes in the Cameroon farm accessions and a high level of admixture, with apparent presence of genes of more than three GGs in most accessions. The traditional Trinitario types appear to have almost disappeared from farmers fields. The admixture must be the result of hybridization and recombination of these genes from the different GGs in seed gardens and in farmers' fields. The use of selected farm accessions will depend on the GG that it belongs to and also on their level of heterozygosity. Further implications of the results for breeding and for introduction of new germplasm into the Cameroon GenBank are discussed.

Keywords Genetic diversity · Tree improvement

Communicated by D. Grattapaglia

I. B. M. Efombagn (✉) · S. Nyassé
Institute of Agricultural Research for Development (IRAD),
P.O. Box 2067 or 2123, Yaoundé, Cameroon
e-mail: efombagn@yahoo.fr

J. C. Motamayor · R. Schnell · M. Kolesnikova-Allen
MARS, Inc, c/o USDA–ARS–SHRS,
13601 Old Cutler Road,
Miami, FL 33158, USA

O. Sounigo · A. B. Eskes · C. Cilas
CIRAD, UPR31,
34398 Montpellier, France

M. J. Manzanares-Dauleux
Agrocampus Rennes APBV, UMR INRA,
BP35327, 35653 Le Rheu, France

M. Kolesnikova-Allen
Central Biotechnology Laboratory,
International Institute of Tropical Agriculture (IITA),
Ibadan, Nigeria

Introduction

Cacao (*Theobroma cacao* L.) is a perennial tree belonging to the family *Malvaceae* (Alverson et al. 1999). It is a diploid preferentially allogamous tropical species (Cope 1984). The upper Amazon (UPA) region of south American rainforest was hypothesized earlier to be the center of genetic diversity of cacao (Cheesman 1944; Cuatrecasas 1964),

which is now supported by molecular evidences (Motamayor et al. 2002). Traditionally, cacao has been subdivided into four genetic groups: Upper Amazon Forastero (UA), Lower Amazon Forastero (LA), Criollo (Cr, anciently cultivated by the Indians in Central America), and Trinitario (Tr, hybrids between LA and Criollo types). Africa is the main cacao producing continent, with approximately 70% of the world's production. Cameroon produces around 4% of the world cacao crop.

While the first attempts to cultivate cacao in Asia and Oceania were based on the introduction of Criollo varieties, the first material introduced in West Africa was of LA origin (Bartley 2005). The LA type (Amelonado) was introduced into Cameroon towards 1892 from Sao Tomé where it was imported by Spanish in the fourteenth century (Champaud 1966). In Cameroon, the first recorded attempt to grow the cacao is connected with the establishment in 1876 of 13 plants (probably originating from Trinidad) in the Royal Botanic Gardens located at the city of Limbé in south western part of the country. Later, introductions of supposedly LA and Tr cacao types from Sao Tomé or from Fernando Po into Cameroon were made by the locals, as occurred in the other parts of west Africa (Preuss 1901; Bartley 2005). Cacao diversity in Cameroon has been enlarged by further introductions from Trinidad, Grenada, Costa Rica, France, and West Africa in such a way that the country established one of the largest collection of cacao cultivars that existed at the beginning of the twentieth century (Bartley 2005). Field surveys carried out in 1950s revealed LA and Tr types (Braudeau et al. 1952; Braudeau and Divaret 1955). The first step of the breeding process was the selection of clones in the 1950s from the dominating local Trinitario populations and these were identified as selections from the Nkoemvone (SNK) accessions (research station, southern Cameroon). The local cacao germplasm based in research collections has been broadened in the 1950s and 1960s mainly by the introduction of UA types. This was followed by the selection of hybrid varieties, mainly crosses between local Trinitario and introduced UA accessions. In the 1970s and 1980s, hybrids selected for high yield potential and precocity were reproduced in biclonal seed gardens and distributed to cacao producing areas. Seed gardens were supposed to produce true hybrid varieties, but due to incomplete self-incompatibility of the female parental clones, much of the seeds produced may in fact have resulted rather from the selfing than from the out-crossing (Lanaud et al. 1987). Besides hybrid seeds, throughout the cacao cultivation history in Cameroon, farmers have used seeds from their own plantation or from neighbors as the main source of planting materials. Such practice could possibly lead to the inbreeding as it can be expected that most of such seeds is produced by selfing. Considering the different germplasm introductions and the different ways by which farmers have obtained their planting materials in the past, it became important to capture the range and type of the genetic diversity of cacao found today in farmers' field in Cameroon. A better understanding of the diversity in farmers' fields will help to guide the possible use of selected farm accessions in breeding.

Different molecular techniques have been used during the past decade to assess genetic diversity in cacao populations. The large genetic diversity of cacao in the native and producing areas around the world has been confirmed by analyses using isozymes and random amplified polymorphic DNA markers (Russel et al. 1993; Sounigo et al. 2005), or a restriction fragment length polymorphism (RFLP; Laurent et al. 1993, 1994; N'Goran et al. 2000; Lerceteau et al. 1997). From the end of the 1990s onwards, microsatellite-based DNA fingerprinting has been increasingly used for the diversity assessment (Lanaud et al. 1999; Zhang et al. 2006) and for the investigations into the origin and dispersal of cacao (Motamayor et al. 2002, 2003). Recently, an agreement was reached that a set of standard simple sequence repeat (SSR) primers should be used to characterize *T. cacao* germplasm collections (Saunders et al. 2004).

In the current paper, results are reported on a study in which 12 SSR loci were used to fingerprint 400 cacao accessions collected in farms from the main cacao growing areas of Cameroon and 126 accessions from the local and international collections. The latter accessions include reference genotypes that are supposedly at the origin of the Cameroon farm and GenBank accessions. The objective was to assess the genetic diversity and population structure in farmers and breeders' germplasm and to study their relationship with the reference genotypes. The questions addressed are the following:

(a) Do the farmers' and breeders' accessions in Cameroon harbor distinct genetic populations?
(b) What is the level of genetic diversity in farmers' and breeders' populations?
(c) What are the lessons to be learned for cacao breeding purposes?

Materials and methods

Plant material

The 526 cacao genotypes included in the study consisted of three main groups: 400 trees from cacao farms (FA) in Cameroon, 95 clones from the GenBank of the Institute of Agricultural Research for Development (IRAD, Cameroon), and 31 reference genotypes that are not present

in the Cameroon GenBank (Table 1). The farm accessions were collected in the different cacao growing areas of the country. These accessions represent the broad geographic range of cacao cultivated in Cameroon (Fig. 1). Selection was made by the farmers in the presence of the breeder. The criteria used by the farmers were the incidence of the *Phytophthora* pod rot (Ppr) disease and the high yield potential observed in each selected accession over several years of production. Furthermore, the selected accessions were identified by the farmers as belonging to the traditional cacao cultivars or to the hybrid cultivars distributed from seed gardens from the 1970s onward. The GenBank and reference genotypes were subdivided into 12 accession groups (AG). According to the traditional classification of cacao, the AGs were identified as belonging to the four main genetic groups (GGs) of the cacao species: Upper Amazon Forastero (UA), Lower Amazon Forastero (LA, containing the Amelonado type), Criollo (Cr), Trinitario (Tr), or as hybrid types between these groups. As shown in Table 1, the GenBank accessions include the following AGs: Imperial College Selection (ICS, selected in Trinidad), SNK (local selections made in farms in the 1950s with numbers below 600), the "SNK600" series (clones selected in Tr×Tr and Tr×UA crosses), T clones (selected in Ghana in the 1950s in crosses among three UA origins: Iquitos Mixed Callabacillo (IMC), Nanay (NA), and Parinari (PA)) and UPA clones (selected in crosses among T clones in Côte d'Ivoire or Cameroon). The reference AGs used in this study are composed of four UA origins, collected by Pound (1943) in the center of diversity of cacao (NA, PA, IMC, Scavina (SCA)), a LA (Amelonado) and a Criollo origins (Table 1). These reference AGs were chosen such as to represent GGs that are known to have participated directly, or as ancestors, in the composition of the FA and in the GenBank AGs in Cameroon. SSRs profiles of the reference AGs were obtained from the Sub Horticultural Research Station (SHRS) of the United States Department of Agriculture (USDA), Miami, USA.

PCR and capillary electrophoresis

DNA was extracted from the leaf material of cacao accessions employing an improved semiautomated rapid

Table 1 Farm, GenBank, and reference accessions used in the study

Type of accessions	Accessions group (AG) code	Genetic group (GG)	Country/region of origin	Number of accessions
Farm selections	FA	Various	Cameroon	400
Cameroon GenBank accessions	ICS	Tr	Trinidad	7
	SNK[a]	Tr	Cameroon	50
	SNK600-Tr[b]	ICS×SNK	Cameroon	11
	SNK600-Tr×UA	Tr×UA	Cameroon	17
	T[c]	UA×UA hybrids	Trinidad (selected in Ghana)	5
	UPA[d]	UA×UA hybrids	Ghana (selected in Côte d'Ivoire or in Cameroon)	3
Reference accessions (from outside Cameroon)	NA[e]=NA 30; 32; 33; 34; 35	UA	Peru	5
	PA=PA 7; 35; 70; 107; 121; 300	UA	Peru	6
	IMC=IMC 11; 47; 60; 67; 76	UA	Peru	5
	SCA=SCA 5; 6; 9; 11; 12	UA	Peru	5
	Am=Oram 1 & 2; SIAL70; SIC 19 & 806	LA (Amelonado)	Brazil	5
	Cr=SJU1, BEN 5 & 1, LIB1, RAN	Criollo	BEN=Venezuela LIB=Nicaragua RAN=Mexique SJU1=Ecuador	5
Total				526

The accession groups (AGs) from the GenBank and reference genotypes correspond each to one of the four known genetic groups (GG) used in the study

UA Upper Amazon Forastero, *LA* Lower Amazon Forastero (Amelonado); *Tr* Trinitario, *Cr* Criollo, *PA* Parinari, *NA* Nanay, *IMC* Iquitos Mixed Callabacillo, *SCA* Scavina

[a]Trinitario clones selected from farmers fields in the South of Cameroon in the 1960s and transferred to the GenBank of the IRAD Nkoemvone Research Station (Southern Cameroon)

[b]So-called SNK600 series of clones, they were selected on-station in UA×Tr and in Tr×Tr crosses

[c]Cloned genotypes derived from F1 crosses between UA parents (NA, PA, and IMC) made in Trinidad in the 1940s

[d]Cloned genotypes derived from crosses between T clones selected in Côte d'Ivoire or Cameroon

[e]The reference accessions of the study are listed within each accessions group (AG)

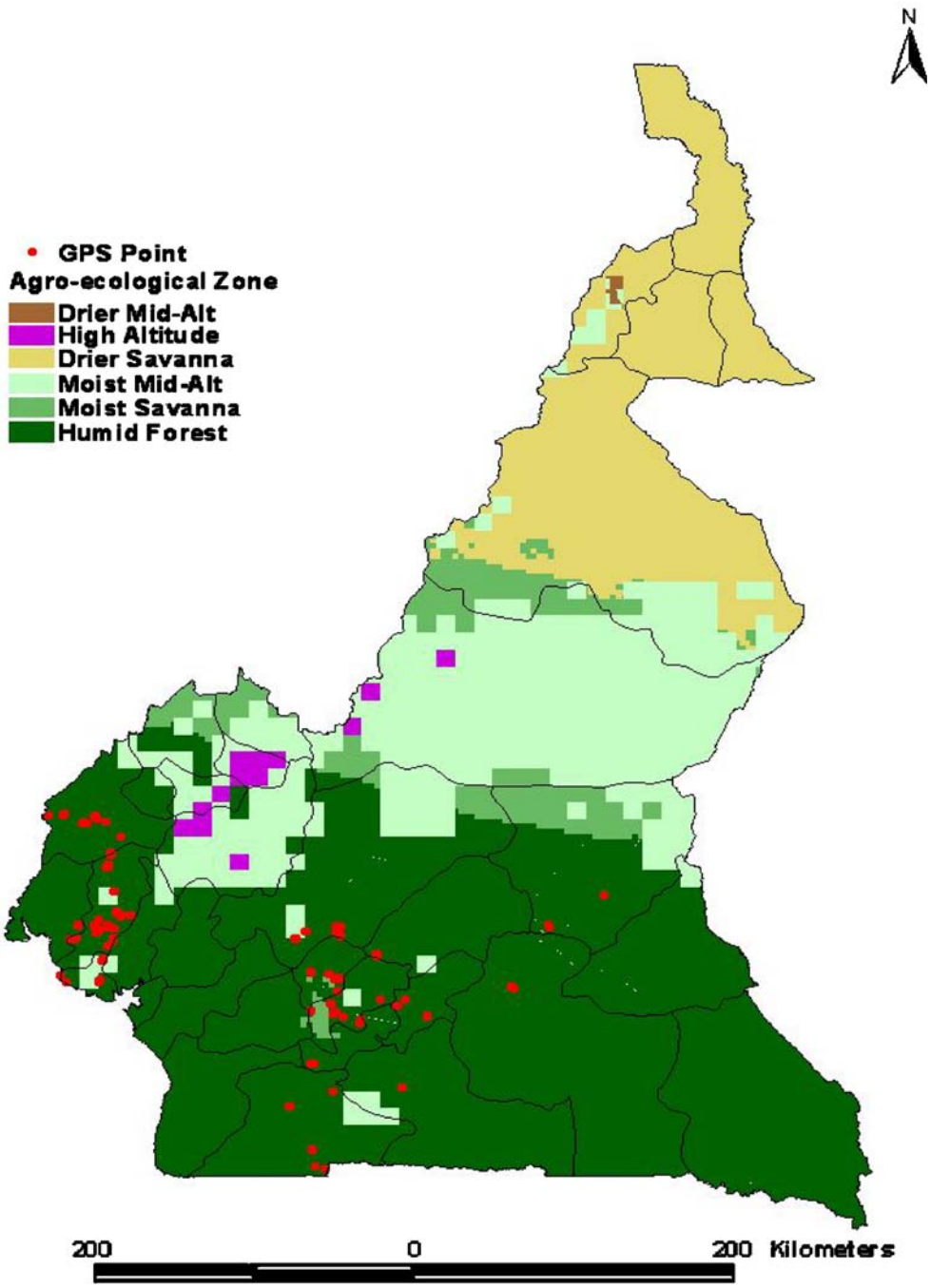

Fig. 1 Map of Cameroon (cacao agroecological zone) showing the sampling sites (*red* GPS points) of cacao germplasm

method of extracting genomic DNA for molecular markers' analysis in cacao (Bhattacharjee et al. 2004). A set of 12 SSR primers was used for genotyping. These primers of the study have previously been described by Lanaud et al. (1999) and Saunders et al. (2004). Characteristics such as nucleotide repeats, allele size, and number of alleles per locus are given in Table 2. Polymerase chain reactions (PCR) were carried out in a 5-µl reaction mix containing 2.5 ng (1 µl) of template DNA, 1 µl of 55× PCR buffer (10 mM Tris-HCl pH 8.3, 50 mM Kcl), 1 µl of 25 mM MgCl2, 0.5 µl of each forward and reverse primers, 0.2 µl of 10 mM dNTPs (dATP, dCTP, dGTP, dTTP), and 0.1 µl of 5 U Taq polymerase (Bioline, UK). Amplifications were carried out in a Gradient Cycler PTC 200 (MJ Research, USA). PCR cycling conditions were as follows: 5 min initial denaturation at 94°C, 35 cycles of amplification at 94°C for 30 s, 51°C annealing for 1 min, and 72°C for 1 min. This was followed by further primer extension at 72°C for 7 min. PCR reactions were performed on an MJ Research Dyad 96-wells PCR.

A volume of 0.5 µl of diluted PCR products (1:19) were mixed with 9.5 µl of formamide (PE-Applied Biosystems)

Table 2 Data summary for the 12 microsatellite markers across the 526 farm, GenBank, and reference's AGs

Primer Name	Nucleotide repeats structure	No of alleles	Size range (bp)	GD	Private allelic richness	PIC
mTcCIR 3	$(ct)_{20}(ta)_{21}$	15	206–247	0.69	3.32	0.72
mTcCIR 6	$(tg)_{7}(ga)_{13}$	9	222–246	0.17	2.64	0.56
mTcCIR 9	$(ct)_{8}n_{15}(ct)_{5}$	12	254–295	0.36	4.56	0.33
mTcCIR 12	$(cata)_{4}n_{18}(tg)_{16}$	11	187–252	0.60	1.87	0.72
mTcCIR 15	$(tc)_{19}$	13	230–254	0.53	2.0	0.65
mTcCIR 17	$(gt)_{7}n_{4}(ga)_{12}$	4	269–310	0.23	0.79	0.22
mTcCIR 18	$(ga)_{12}$	8	331–354	0.50	1.31	0.50
mTcCIR 19	$(ct)_{28}$	13	348–377	0.50	3.16	0.50
mTcCIR 21	$(tc)_{11}n_{5}(ca)_{12}$	10	140–169	0.48	2.08	0.41
mTcCIR 24	$(ag)_{13}$	7	184–200	0.21	3.83	0.12
mTcCIR 25	$(ct)_{21}$	13	128–164	0.53	1.75	0.62
mTcCIR 26	$(tc)_{9}c(ct)_{4}tt$	10	282–311	0.60	2.01	0.59
All loci		125	–	0.45	2.44	0.50

PIC follows the definition of Hearne et al. (1992)
PIC Polymorphism information content, *GD* gene diversity

and 0.5 μl ROX-labeled GeneScan-500 size standard (PE-Applied Biosystems). DNA fragments were denatured at 95°C for 5 min and size fractioned using capillary electrophoresis on an ABI 3100 automatic DNA sequencer (PE-Applied Biosystems). The GENESCAN 3.7 software (PE-Applied Biosystems) was applied to size the peak patterns, using the internal ROX 500 size standard and GENEMAPPER 3.5 (PE-Applied Biosystems) for allele calling.

Diversity analysis

Basic statistics related to the assessment of the genetic diversity were computed. Due to significant variation in sampling size between farm, GenBank, and reference AGs, the average gene diversity (GD = expected heterozygosity), allelic richness, and private allelic (Leberg 2002) were estimated using the rarefaction approach implemented in the software HP-RARE 1.0 (Kalinowski 2005). Rarefaction approach is used to standardize the mean number of alleles per locus or the number of private alleles for a given AG, based on the number of genes (alleles) present in the smallest sample (AG) of accessions present in the study. The informativeness (polymorphism information content, PIC) of the 12 SSRs were estimated using the CERVUS software (Marshall et al. 1998). The PIC values were estimated following the definition of Hearne et al. (1992).

Analysis of population structure

In order to assess the structure of genetic diversity within and among the AGs, we used four complementary approaches: *F* statistics, a principal coordinates analysis (PCoA), a Bayesian model-based clustering method, and an analysis of molecular variance (AMOVA). Considering the different cacao groups, F_{st} and F_{is} (Weir and Cockerham 1984) were also computed using the FSTAT software (Goudet 1995) as measures of the genetic diversity within and among AGs. Dissimilarity matrixes were computed and PCoA was performed on them using the Darwin 5.0 software (Perrier et al. 2003).

The Bayesian model-based clustering method of Pritchard et al. (2000), as executed by the software STRUCTURE 2.2. (http://www.pritch.bsd.uchicago.edu), was also used. This method assumes that each cacao genotype in the whole sample may result from the admixture of an unknown number of differentiated ancestral populations (the reference AGs in our study), with membership coefficients totaling 1. A cluster assignment is defined based on the membership assignment probability. STRUCTURE estimates the proportion of the genome of each individual having an ancestry in each subcluster and applies a posterior probability of the data Pr(X/K), where X represents the data. Preliminary studies performed at USDA found a strong genetic differentiation (F_{sts} over 0.25) among the reference AG (unpublished results). Given the known strong genetic structure of the a priori population information existing in our AGs, we used the admixture option on the prior population information model with unlinked loci and correlated allele frequencies. We ran two simulations where farm and GenBank AGs were respectively compared to the reference AGs. Based on the number of AGs compared in each simulation, we used respectively K=7 (the FA and the six reference AGs) and K=12 (the six GenBank and the six reference AGs), with ten replicate runs for each K value, a burning period length of 10^5, and a postburning simulation length of 5×10^5. The POPFLAG option was used to update allelic frequencies in order to

assign the genes present in the farm and GenBank accession groups to each of the reference AG.

The genetic structure of the cacao AGs was further investigated by an AMOVA using ARLEQUIN 3.1 (Excoffier et al. 2005). Tolerance was set to 5% of the missing data per locus.

Results

Overall SSR diversity

All microsatellite loci were polymorphic and met the assumptions of independence (no linkage between pairs of loci). A total of 125 alleles were detected at 12 microsatellite markers across the 526 accessions from the farm, GenBank, and reference AGs (Table 2). The number of alleles per locus ranged from four (mTcCIR 17) to 15 (mTcCIR 3), with an average of ten alleles across the 12 loci. To determine the private allelic richness with the rarefaction method, the standardization was performed through the sampling of the three gene copies since the smallest AG were composed of three accessions. The average private allelic richness was 2.44 per locus (Table 2). Informativeness of the 12 loci (PIC) was highly variable (0.12–0.72) with the eight loci showing the PIC values≥0.50.

Molecular diversity analysis in farm and GenBank

Although a total of 125 alleles were identified in all farm and GenBank accessions, the allelic richness differed among the AGs (Table 3). The highest number of alleles was found in the FA group (113 alleles) and the lowest in the Cr reference group (14 alleles). The average gene diversity and allelic richness (calculated according to the rarefaction approach), were highest for UPA, PA, T, and IMC and lowest for Am and Cr. The average F_{is} values varied considerably among the different AGs (0.01 to 0.26), suggesting the highly variable inbreeding levels. The highest F_{is} values were obtained for Am and followed by FA, Cr, and SNK.

The average gene diversity and allelic richness were quite high for the FA. The private allelic richness was highest for SCA (12.51), followed by IMC (5.32) and PA (2.77), with a value of 2.03 in the FA (Table 3). The estimate for the genetic differentiation (Weir and Cockerham 1984) among all the AGs varied considerably, with F_{st} values varying from 0.14 to 0.57 for the different AGs. The significant population differentiation detected by contingency table was supported by the AMOVA's permutation result (Excoffier et al. 2005) with between-group variance being highly significant ($p<0.001$). About 17.5% of the total molecular variance was due to the differences between AGs whereas 82.5% of the variance was partitioned among all the 526 accessions of the study.

Genetic distance-based diversity analysis

The first plan of the PCoA shows that the UA, LA (Amelonado), and Cr accessions are most widely separated from each other (Fig. 2). The farm accessions are distributed in between the putative parental populations made up by the GenBank and reference AGs. As expected (Motamayor et al. 2002), the Tr AGs (ICS and SNK) are in between the locations of the Cr and Am genotypes. The SNK accessions (Trinitario selected in Cameroon) are

Table 3 Diversity statistics for the different accession groups (AGs) analyzed

Main AGs	AG code	*N*	GD	F_{is}	A*t*	A/L	Allelic richness	Private allelic richness	F_{st}
Farm	FA	400	0.50	0.15	113	9.41	31.3	2.03	0.18
GenBank	SNK	50	0.42	0.13	50	3.91	25.5	0.94	0.23
	SNK600-Tr	11	0.43	0.12	38	3.16	27.1	0.01	0.23
	SNK600-Tr×UA	17	0.45	0.07	51	4.25	28.8	0.54	0.22
	T	5	0.55	0.08	41	3.41	33.8	0.58	0.22
	UPA	3	0.62	0.08	36	3.0	36.0	0.64	0.14
	ICS	7	0.52	0.05	27	2.25	25.5	0.01	0.25
Reference	NA	5	0.47	0.07	32	2.66	27.9	1.73	0.30
	PA	6	0.60	0.01	43	3.58	34.2	2.77	0.25
	SCA	5	0.42	0.20	33	2.75	27.0	12.51	0.42
	IMC	5	0.57	0.07	36	3.0	31.4	5.32	0.27
	Cr	5	0.10	0.13	14	1.16	13.6	0.95	0.57
	Am	5	0.20	0.26	19	1.58	17.8	0.81	0.30
Total or mean		526	0.55	–	125		27.6	2.4	–

N Number of accessions analyzed, *GD* gene diversity, *At* total number of alleles, *A/L* mean number of alleles per locus

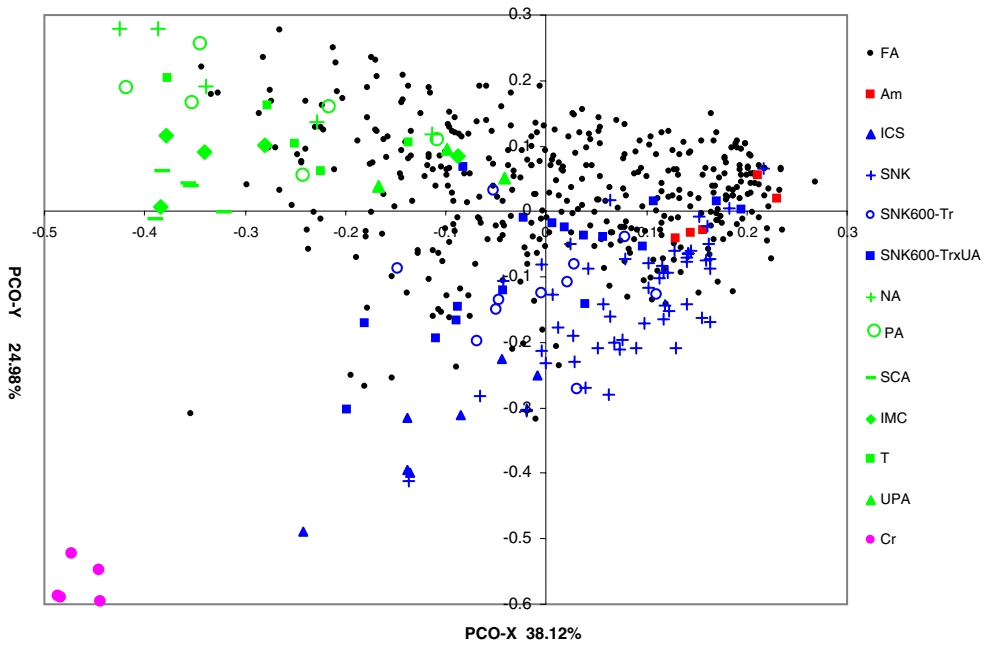

Fig. 2 Principal coordinates analysis based the distances among individuals of farm, GenBank (SNK, SNK600-Tr, SNK600-Tr×UA, ICS, UPA, and T) and reference AGs (Am, NA, PA, SCA, IMC, Cr)

genetically nearer to Am than the ICS accessions (Trinitario from Trinidad).

Population structure analysis

Using the a priori information related to the predefined AGs, two simulations were achieved with the STRUCTURE Program. In the first simulation (Fig. 3a), each GenBank accession was associated with six probabilities corresponding to the six reference AGs, showing the degree to which the genotypes were related to each reference AG. In the second simulation (Fig. 3b), each of the farm accessions was associated with the same reference AGs. Given the high rate of admixture observed in farm and GenBank material (Fig. 3), we calculated the proportions of the genome (posterior probabilities) of the reference AGs that according to STRUCTURE program was associated with the FA and GenBank AGs (Table 4).

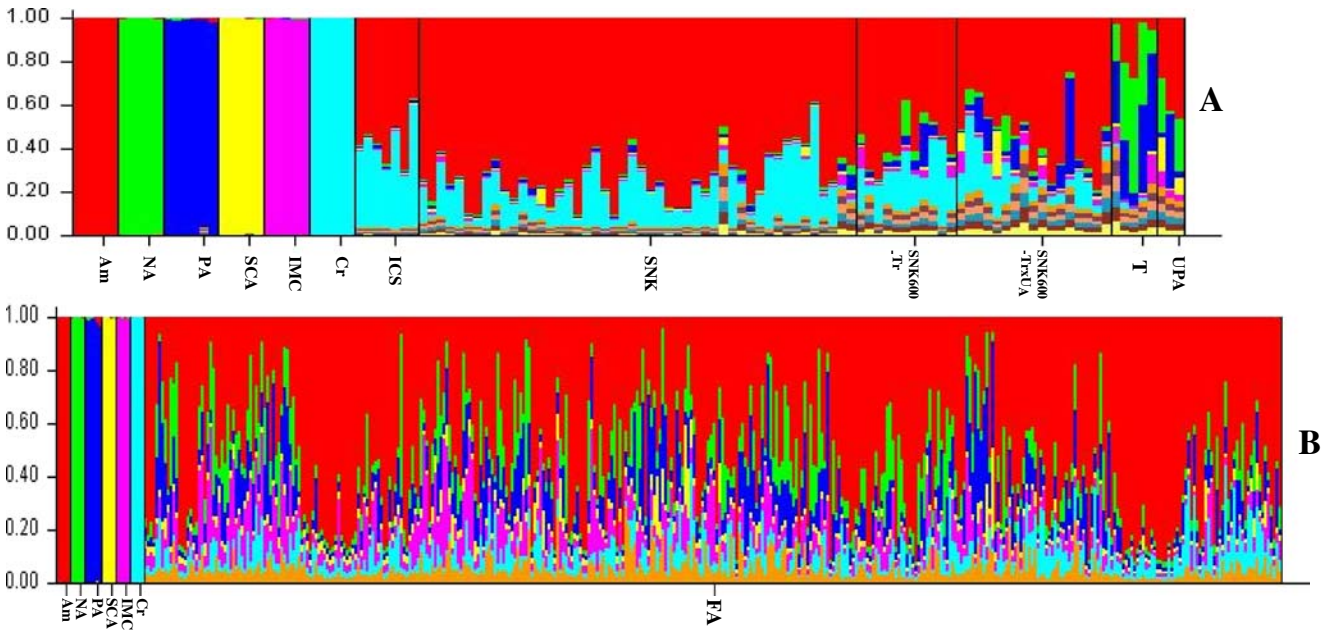

Fig. 3 Diversity structure depictured for 400 farm, 95 GenBank (SNK, SNK600-Tr, SNK600-Tr×UA, T clones, and UPA) and 31 reference (Am, NA, PA, SCA, IMC, ICS, and Cr) accessions according to the STRUCTURE program (Pritchard et al. 2000). Reference accession groups are compared to GenBank accessions in **(a)** and to the farms' accessions in **(b)** by assigning different *colors* for each reference AG. *Colors* that are different from the ones used for the reference AGs represent genes of undetermined origin

Table 4 Gene proportion of GenBank and farm accessions found in reference GGs calculated with the STRUCTURE program

GenBank and farm Ags		Reference GGs						Undetermined	Total
		LA (Am)	UA				Cr		
			NA	PA	SCA	IMC			
GenBank	ICS	0.56	0.01	0.01	0.00	0.01	0.37	0.05	1
	SNK	0.73	0.01	0.01	0.01	0.01	0.16	0.05	1
	SNK600–Tr	0.57	0.04	0.04	0.01	0.03	0.19	0.10	1
	SNK600–Trx UA	0.54	0.03	0.07	0.03	0.03	0.12	0.15	1
	T	0.12	0.31	0.30	0.01	0.05	0.01	0.20	1
	UPA	0.39	0.17	0.13	0.08	0.04	0.00	0.19	1
	Mean	0.48	0.24				0.14	0.14	1
Farm		0.54	0.10	0.11	0.03	0.09	0.07	0.06	1

Proportions were expressed as the posterior probability value of all individuals of each of the AGs (depicted in Fig. 3)

High proportions of the genes in the GenBank and farm AGs (86% and 94%, respectively) were associated with the reference AGs. The GenBank AGs were most associated with the Am reference genome (48%), followed by the UA (24%) and Cr (14%) genomes. The Am genome was highly associated with SNK (73%), SNK600-Tr (57%), SNK600-TrxUA (54%), ICS (56%), and less so with UPA (39%) and T (12%) (Table 4). The UA genome was most associated with T (67%) and UPA (42%), and the T group was mostly with NA (31%) and PA (30%). The Cr genome was most associated with ICS (37%) and less so with the three SNK groups. The farm accession group was also most associated with the Am genome (48%), followed by the UA (33%) and Cr (7%) genomes. Within the UA groups, the PA (11%), NA (10%), and IMC (9%) genomes were most represented. The contribution of the SCA genome was very small (3%).

Discussion

All the SSRs markers used in this study, which were developed by Lanaud et al. (1999), appeared to be highly polymorphic (Table 2). Eight of the 12 markers contributed most to the overall genetic diversity. In total, 125 alleles were detected. The large majority of these (113) were present in the FA, suggesting that the FA harbor most of the genetic richness that is present in the GenBank and reference populations. It shows the close relationship of the farm population with the GenBank and reference AGs. An exception is to be made for the genomes of the SCA and IMC reference AGs that have high private allelic richness values and appear to be less represented in the FA.

The overall genetic diversity was high both in farm and GenBank materials. The large variability in farm accessions can be explained by the large variation of cacao introduced in Cameroon at the beginning of the twentieth century, as reported by Bartley (2005), and also to the introduction of UA germplasm in the 1950s and its subsequent use in the cacao breeding program. The rarefaction approach showed that the FA had a private allelic richness of 2.03 in the smallest sample of individuals. This value was higher than those of the GenBank's AGs (SNK, ICS, T, UPA) that have been used in the local breeding program for the release of hybrid cacao varieties. Therefore, the farm accessions may also harbor some genetic diversity that is not present in the current GenBanks.

The genetic diversity for the FA (estimated by GD values) was similar to that of the GenBank and reference accessions, except for the Cr and Am which had as expected very low GD values. The average level of F_{is} (0.15) for the FA was higher than that of the GenBank accessions, suggesting an inbreeding effect in the farm populations. This can be explained by the presence of naturally inbred LA types in the farm populations and also by the way farmers have obtained seedlings for new plantings (mostly by using open-pollinated seeds that is expected to be partly self-pollinated).

The GD values (expected heterozygosity) found in our study for UA and LA GGs are comparable to values previously found in other studies, using the microsatellites (Motamayor et al. 2002) or RFLP (N'Goran et al. 2000). However, Tr has usually presented higher GD values than the UA or LA GGs, whereas in our study, two AGs of Tr origin (SNK and SNK600-Tr) had lower GD values (0.42 and 0.43, respectively) than the UA groups, such as PA (0.60), IMC (0.57), and T (0.55). This is likely caused by the origin of the SNK and SNK600-Tr accessions, descending from the natural mating in farmers' fields among Tr and LA ancestral genotypes. We found a significant variation for the degree of differentiation among the AGs (F_{st}=0.14–0.57). The highest value is obtained by Cr, showing that this GG is most widely differentiated from the other AGs.

The genetic distance-based analysis revealed the associations between the genetic diversity in farm, GenBank, and reference AGs (Fig. 2). Most of the farm accessions were located in between the LA and UA reference accessions or in between LA and Tr accessions. Others appear to be so closed to the Am, Tr, or UA groups that they may in fact belong to these GGs. None of the FA was near to Cr. This may mean that genotypes containing many Criollo genes have been selected against in the history of cacao cultivation in Cameroon. Such would be expected due to the high susceptibility of Criollo to pests (mirids) and diseases such as *Phytophthora* pod rot and cacao dieback. The location of the three UPA accessions in between UA and Amelonado suggests that these accessions may in fact not be pure UA genotypes. This is confirmed by the high proportion (0.39) of Amelonado genes in the UPA accessions (Table 4). Two of the accessions identified as SNK600-Tr×UA are very close to the Am reference accessions and two SNK600-Tr accessions are in between the Tr and UA accessions, suggesting that these groups may also contain a few mislabeled accessions.

Population structure analysis using the STRUCTURE program showed that 96% of the genetic diversity in FA can be explained by the genetic diversity in the reference AGs (Table 4). This shows that the reference AGs are the main genetic groups involved in the evolution of the farmers populations in Cameroon. Among the reference AGs, the genome of Am has the largest impact on the FA (54%) followed by the AGs of UA origin (33%). Within UA, the most important contributions come from the PA, NA, and IMC. Despite the use of the clones SCA6 and SCA12 in the seed gardens in Cameroon, the contribution of the SCA genome in the FA is very small (3%). This may be partly explained by the fact that the original SCA6 introduced into Cameroon is a mislabeled clone (Risterucci et al. 2001).

The participation of the reference genomes in the FA is partly explained by the presence of a small proportion of rather pure Am and UA types in the farmers' fields (Figs. 2 and 3b). However, the impact of the reference genomes is mainly apparent in the high proportion of accessions with the admixture of genes from more than two reference AGs (Am, one or more UA AGs, and Cr). The admixture must be due to hybridization (in seed gardens) and to substantial natural recombination among different GGs in the farmers' fields. Farmers tend to use seeds issued from open pollination in their plantations for new plantings (replacement of dead cacao trees, extension or the creation of new cacao farms). In many cases, the same farm may therefore contain first or second generation hybrid cacao cultivars, selfings, and backcrosses to local genotypes causing the substantial recombination of genes from different origins. Visual analysis of Fig. 3 suggests that at least 50% of the FA is of hybrid origin (UA×Am or UA×Tr crosses).

In the GenBank populations, the degree of admixture of genes from different reference AGs (Fig. 3a) tallied with the known origins of these AGs. The SNK accessions showed to possess less Cr genes than the ICS accessions. This is in line with the putative origin of the local Tr population, including the Am progenitors introduced in the country from Sao Tomé (Champaud 1966) that were widely planted in cacao farms. Eleven out of the 95 GenBank accessions contained a high proportion (above 0.20) of undetermined genes; these would need to be studied with more detail as they may be mislabeled accessions of complex origin.

Knowledge of the genetic diversity and population structure of germplasm collections is an important foundation for crop improvement. In the past, breeding work in Cameroon has involved local cacao material collected in farmers' fields, including the SNK accessions selected for their high yield potential. The next step was to establish seed gardens by crossing local selections with introduced material, aiming mainly at high yield potential. The selection methods adopted from 2003 onwards include a participatory approach, involving farmers in the identification of superior trees in their plantations and in the selection of new varieties in on-farm trials. The main objectives of the current improvement program are to select new hybrid and clone varieties with resistance to diseases and pests (mainly Ppr and mirids) and to start also looking for varieties with improved quality, to add value to the product. Furthermore, Cameroon intends to introduce more germplasm to widen the useful genetic diversity in the GenBank which is currently very small.

The following considerations are made with regards to the impact of the findings reported here for current cacao breeding program in Cameroon. Firstly, the large genetic variation and substantial private allelic richness detected among the farm accessions appear to be in support of the participatory selection approach adopted in Cameroon. The possibility to select for resistance to Ppr among the farm accessions has already been demonstrated by Efombagn et al. (2007), who found that 9% of the FA are resistant to Ppr compared to only 5% for the local GenBank accessions. The accessions in farmers' fields should equally be valuable material when selecting for the other important traits. Secondly, the molecular results will be very useful to decide on how to use promising FA in further breeding. If an accession is relatively homozygous and belongs to rather pure GGs (Am, Tr, UA), it is suitable to be used as a parent for creation of new uniform hybrid varieties (of UA×LA or UA×Tr type). On the other hand, if the accession is very heterozygous and of hybrid origin, it will be more suitable for the selection of new clones. Thirdly, Fig. 3b shows that many FA contain an important proportion of Tr genes (combination of Am and Cr genes).

This would be in favor of the selection for quality traits related to the Tr germplasm, such as fruity flavor. Fourthly, the information generated on the genetic diversity provides good guidance for the choice of useful new germplasm to be introduced into the cacao GenBank in Cameroon. The following materials should receive priority: (a) materials from underrepresented GGs with high value for the cacao breeding in Cameroon (e.g., containing genes for Ppr resistance, quality, and yield), (b) accessions of the already available AGs that are currently underrepresented and that have high private allelic richness (such as PA, IMC, and SCA), (c) other unique GGs that have been identified in cacao and that are not yet present in the collection (e.g., clones from French Guiana, Ucayali, and the LCT-EEN and Chalmers collections), and (d) after the evaluation of the farm accessions, the best ones should be included into the GenBank for long-term conservation.

Acknowledgements The authors thank the Institute of Agricultural Research for Development (IRAD), Cameroon which authorized the publication of this paper. The study was partially financed by the West Africa Cacao Diversity Project (IITA/USAID), the CFC/ICCO/Bioversity project titled 'Cacao Productivity and Quality Improvement, a Participatory Approach' and MARS, Inc. We thank Nanette Langevin (USDA–ARS–SHRS, Miami, FL, USA), Sunday Taiwo (IITA-CBL, Ibadan, Nigeria), François Edoa, Innocent Badjeck, Essomo Ngomba, and K. Daniel Vefonge (IRAD, Cameroon) for their technical support in the present work.

References

Alverson WS, Whitlock BA, Nyffeler R, Bayer C, Baum DA (1999) Phylogeny of the core Malvales: evidence from ndhF sequence data. Am J Bot 86:1474–1486

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

Bhattacharjee R, Kolesnikova-Allen M, Aikpokpodion P, Taiwo S, Ingelbrecht I (2004) A semi-automated rapid method of extracting genomic DNA for molecular marker analysis in cocoa, *Theobroma cacao* L. Plant Mol Biol Report 22:435a–435h

Braudeau J, Divaret P (1955) Rapport Annuel de la Station Expérimentale du Centre (IRA), Nkoemvone. In: Rapport Annuel de l'Inspection Générale de l'Agriculture, Territoire du Cameroun, pp 60–102

Braudeau J, Grimaldi E, Lavabre E (1952) Rapport Annuel de la Station du Cacaoyer de Nkoemvone (IRA). In: Rapport Annuel du Service de l'Agriculture, Territoire du Cameroun, pp 56–12

Champaud J (1966) L'économie cacaoyère au Cameroun. Cah ORSTOM Sér Sci hum III(3):105–124

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

Cope FW (1984) Cacao *Theobroma cacao* (Sterculiaceae). In: Simmonds NW (ed) Evolution of crops plants. Longman, London, pp 285–289

Cuatrecasas J (1964) Cacao and its allies: a taxonomic revision of the genus *Theobroma*. Contrib US Natl Herb 35:379–614

Efombagn MIB, Nyassé S, Sounigo O, Kolesnikova-Allen M, Eskes AB (2007) Participatory cocoa selection in Cameroon: *phytophthora* pod rot resistant accessions identified in farmers' field. Crop Prot 26(10):1467–1473

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

Goudet J (1995) FSTAT 1.2: a computer program to calculate F-statistics. J Heredity 86:485–486

Hearne CM, Ghosh S, Todd JA (1992) Microsatellites for linkage analysis of genetic traits. Trends Ecol Evol 8:288–294

Kalinowski ST (2005) HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. Mol Ecol Notes 5:187–189

Lanaud C (1987) Nouvelles données sur la biologie du cacaoyer (*Theobroma cacao* L.): diversité des populations, système d'incompatibilité, haploids spontanées. Leurs conséquences pour l'amélioration génétique de cette espèce. D. Sc. Thesis, Université de Paris Sud, Centre d'Orsay, France

Lanaud C, Risterucci AM, Piretti I, Falque M, Bouet A, Lagoda PJL (1999) Isolation and characterization of microsatellites in *Theobroma cacao* L. Mol Ecol 8:2141–2152

Laurent V, Risterucci AM, Lanaud C (1993) Variability for nuclear ribosomal genes within (*Theobroma cacao*). Heredity 71: 96–103

Laurent V, Risterucci AM, Lanaud C (1994) Genetic diversity in cocoa revealed by cDNA probes. Theor Appl Genet 88:193–198

Leberg PL (2002) Estimating allelic richness: effects on sample sizes and bottlenecks. Mol Ecol 11:2445–2449

Lerceteau E, Robert T, Pétiard V, Crouzillat D (1997) Evaluation of the extent genetic variability among *Theobroma cacao* L. accessions using RAPD and RFLP markers. Theor Appl Genet 95:10–19

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. Mol Ecol 7:639–655

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

Motamayor JC, Risterucci AM, Heath M, Lanaud C (2003) Cocoa domestication II. Progenitor germplasm of the *Trinitario cacao* cultivar. Heredity 91:322–330

N'Goran JAK, Laurent V, Risterucci AM, Lanaud C (2000) Comparative genetic diversities studies of *Theobroma cacao* L. using RFLP and RAPD markers. Heredity 73:589–597

Perrier X, Flori A, Bonnot F (2003) Data analysis methods. In: Hamon P, Seguin M, Perrier X, Glazmann JC (eds) Genetic diversity of cultivated tropical plants. Enfield, Montpellier, pp 43–76

Preuss P (1901) Expedition nach central und Sudamerika 1899/900. Verlag des Kolonial-Wirtschaftlichen Komitees, Berlin

Pound JF (1943) Cacao and witches broom diseases (*Marasmius pernicious*). Report on a recent visit to the Amazon territory of Peru, September 1942–February 1943. Government Printer, Port of Spain, Trinité and Tobago

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

Risterucci AM, Eskes AB, Fargeasm D, Motamayor JC, Lanaud C (2001) Use of microsatellite markers for germplasm identity analysis in cocoa. In: Bekele F, End M, Eskes A (eds) Proceedings of the international workshop on new technologies in cocoa breeding, INGENIC, 16–17 October 2000. Kota, Malaysia, pp 25–33

Russell JR, Hosein F, Johnson E, Waugh R, Powell W (1993) Genetic differentiation of cocoa (*Theobroma cacao* L.) populations revealed by RAPD analysis. Mol Ecol 2:89–97

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

Sounigo O, Umaharan R, Christopher Y, Sankar A, Ramdahin S (2005) Assessing the genetic diversity in the International Cocoa Genebank, Trinidad (ICG,T) using isozyme electrophoresis and RAPD. Genet Resour Crop Evol 52: 1111–1120

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. Evolution 38:1358–1370

Zhang D, Arevalo-Gardini E, Mischke S, Zuniga-Cerdanes L, Barreto-Chavez A, Adriazola del Aguila J (2006) Genetic diversity and structure of managed and semi-natural populations of cocoa (*Theobroma cacao*) in the Huallaga and Ucayali Valleys of Peru. Ann Bot 98:647–655