Conserv Genet (2008) 9:327–337
DOI 10.1007/s10592-007-9345-8

RESEARCH ARTICLE

Microsatellite variation and population structure in the ''Refractario'' cacao of Ecuador

Dapeng Zhang · Michel Boccara · Lambert Motilal · David R. Butler · Pathmanathan Umaharan · Sue Mischke · Lyndel Meinhardt

Received: 30 December 2006 / Accepted: 8 May 2007 / Published online: 5 June 2007

Abstract Utilization of germplasm for crop improvement is often hampered by absence of information regarding origin, genetic identity and genealogical relationships of germplasm groups or populations. Molecular marker technology offers an efficient tool to verify or reconstruct passport data. Using a high-throughput genotyping system with 15 microsatellite loci, we fingerprinted 482 accessions in 48 putative half-sib families of Refractario cacao (a group of germplasm collected from nine farms in Ecuador). Based on the multilocus profiles, a Bayesian method for individual assignment was applied to verify membership in each half-sib family. Multivariate statistical analysis showed that the Refractario genetic profile was different from other groups tested, except for the ''Nacional'' cacao from the coastal valley of Ecuador. Hierarchical partitioning of genetic variance in the Refractario cacao showed that 76% of the variation was contributed by intra-family difference, whereas the inter-family and inter-farm difference accounted for 15 and 9% of total variance, respectively. All three sources of variation were highly significant ($P < 0.01$). Cluster and Principal Coordinates Analyses revealed a population sub-structure in Refractario, which was also highly heterozygous, suggesting hybridization derived from Nacional cacao and multiple other parental varieties, which all shared a similar genetic background. The improved understanding of identities and structure in Refractario cacao will contribute to more efficient conservation and use of this germplasm group in cacao breeding.

Keywords DNA fingerprinting · Ecuador · Germplasm · Genetic diversity · Population structure · South America · *Theobroma cacao* L.

D. Zhang (✉) · S. Mischke · L. Meinhardt
USDA/ARS, Beltsville Agricultural Research Center, Sustainable Perennial Crops Laboratory, Plant Sciences Institute, 10300 Baltimore Avenue, Bldg. 001 Rm. 223, Beltsville, MD 20705, USA
e-mail: ZhangD@ba.ars.usda.gov

M. Boccara · L. Motilal · D. R. Butler
Cacao Research Unit, The University of the West Indies, St. Augustine, Trinidad and Tobago, West Indies

M. Boccara
Centre de coopération internationale en recherche agronomique pour le développement, Montpellier Cedex 5, France

P. Umaharan
Department of Life Science, The University of the West Indies, St. Augustine, Trinidad and Tobago, West Indies

Introduction

Efficient utilization of plant germplasm held in genebanks is largely dependent on the availability and accuracy of the passport data and other related information. The absence of detailed knowledge regarding the origin, genetic identity, relationships among individual progenies and population structure has hampered the potential exploitation of germplasm in crop improvement, and has been the case with cacao germplasm. Cacao is native to the South American rainforest with its putative centre of diversity located in the upper Amazon region of Peru, Ecuador and Colombia (Cuatrecasas 1964; Cheesman 1944; Bartley 2005). The species comprises a large number of highly morphologically variable populations, which can be crossed with each other (Cheesman 1944; Pound 1945; Bartley 2005). The majority of the cacao germplasm held in the International Cacao Genebank, Trinidad (ICG,T) was collected in the 1930s and the passport data were

either incomplete or had many ambiguities (Lockwood and End 1993; Motilal and Butler 2003). Cacao germplasm is usually clonally propagated and maintained as living trees in germplasm collections, because the seeds do not remain viable for much longer than a week after harvesting (Coe and Coe 1996). Managing large collections of cacao germplasm is operationally challenging. Errors of documentation commonly occur during the transportation, propagation or maintenance of material, when germplasm was exchanged or otherwise obtained at different times, resulting in a large numbers of trees with unconfirmed identities (Motilal and Butler 2003; Turnbull et al. 2004).

A substantial amount of work has been reported on the use of molecular markers for cacao germplasm management (Engels 1986; Laurent et al. 1993, 1994; Lerceteau et al. 1997; N'Goran et al. 2000; Sounigo et al. 1996, 2001). Since the development of simple sequence repeat (SSR) markers for cacao (Lanaud et al. 1999), SSR-based DNA fingerprinting has been increasingly applied in cacao germplasm characterization (Charters and Wilkinson 2000; Lanaud et al. 2001; Motamayor et al. 2002, 2003; Saunders et al. 2001, 2004; Cryer et al. 2006; Schnell et al. 2005; Takrama et al. 2005). However, application of molecular markers for large-scale genotyping of genebank collection and comprehensive characterization of whole germplasm groups has been scarce.

''Refractario'' cacao originated from a large group of germplasm selected during the 1920s from the coastal valley of Ecuador for it's potential resistance to witches' broom disease (Pound 1938, 1943). Uninfected trees (or those with slight infection) were selected around 1923, and seedlings from fruits of these trees were raised in nurseries and subjected to natural infection. The plants that passed the nursery screening were then established on various farms in Ecuador. The Refractario germplasm in the ICG,T refers to progeny from fruit collected in 1937 from trees on farms in Ecuador with an absence of infection by witches' broom disease [caused by *Crinipellis perniciosa* (Stahel) Singer] (Pound 1938, 1943; Bartley 2001). In the 1938 report, the collector, Dr. F. J. Pound stated that fruits were collected from ''some'' 80 trees, each being a seedling progeny of a ''Refractario'' type that was 5–10 years old. Seeds from these fruits were bulked and sent to Barbados for quarantine purposes where the seedlings were raised (Pound 1938, 1943). The Refractario accession group was subsequently transferred to Trinidad in the form of budwood and was planted in Marper Farm in Plum Mitan, Manzanilla and Trinidad. Later, the ICG,T was established in the 1980s at the University Cacao Research Station, Centeno under the management of the Cacao Research Unit (CRU) of The University of the West Indies (Kennedy and Mooleedhar 1993). The entire set of surviving Refractario accessions was re-propagated and planted with about 16 replicated trees in each plot.

Today, the Refractario group accounts for roughly one quarter of the 2,300 accessions of cacao germplasm held in ICG,T. However, this accession group has so far not been widely used for cacao breeding. This is largely due to the absence of detailed knowledge regarding their origins, genetic composition and the relationships among the progenies. Until a few years ago, there was paucity of information regarding the nomenclature of these progenies (Bartley 2000). The locations where the uninfected ''Refractario'' trees were growing were never clearly given. Furthermore, there were insufficient details on the actual genetic composition of the progenies from which the seeds were collected, or their relationship with other germplasm groups held in the genebank (Bartley 2000, 2001).

In this paper, we report a study in which 15 SSR loci were used to characterize the Refractario accessions. Our objectives were to (1) identify mislabelled accessions in the Refractario group; (2) quantify the genetic diversity within and among different farms (Hacienadas); and (3) assess the genetic relationship between Refractario and other cacao germplasm groups. This study is a part of the International collaborative project on DNA fingerprinting of cacao germplasm in the Americas. The resultant information will improve our understanding about the South American cacao gene pool, and facilitate the efficient use of the Ecuadorian germplasm for cacao genetic improvement.

Materials and methods

Plant material and DNA isolation

For this experiment, a total of 482 accessions, representing 48 half-sib families were used, most of which came from Marper Farm in Trinidad where most of the original trees are still present. The name of the 48 half-sib families, as well as the nine farms from which they were originally collected, are listed in Table 1. Leaf samples of variable ages were collected from each accession and each sampled branch was tagged for potential revisiting. In several cases, putative duplicate trees from a different field and plot were sampled, which resulted in two or three samples for these accessions. Therefore, each sample was labelled with both accession name and DNA extraction number and kept as an individual sample for profiling.

The study also included a group of ''control'' accessions comprised of:

- Twenty-seven Parinari accessions and 22 ICS accessions from the San Juan Estate and Marper Farm in Trinidad.

Table 1 Refractario accessions and their home families assigned by Bayesian clustering method

Farm (Hacienda)	Accession series	Number of half-sib families	Number of genotyped progenies	Number of correctly assigned progenies	% Error rate[a]
Amalia	AM	2	71	46	35.2
Balao	B	17	57	39	31.6
Clementina	CL	6	59	47	20.3
Clementina mixed[c]	CLM	Unknown	18	11	Unknown
Javilla	JA	6	76	58	23.7
La Paz	LP	5	52	39	25.0
L[b]	LX	1	14	9	35.7
Moquique	MOQ	6	75	53	29.3
San Juan	SJ	2	33	22	33.3
Santa Lucia	SLA, SLC	2	27	24	11.1
Total		48	482	348	27.3

The threshold of assignment probability was $P = 0.90$

[a] The computation of mean error rate does not include the group of ''Clementina mixed''

[b] There was no information regarding the source of the Hacienda ''L'' in Pound's collecting report (Pound 1938, 1945). Bartley (2000, 2001) expressed doubt about the exact meaning of ''L''

[c] Accessions labelled as ''Clementina mixed'' were originally collected from the Clementina farm. These accessions were named separately because the number of half-sib families in this group was unknown. The result of assignment test shown that 11 of the 18 accessions used in the present study can be assigned to a single family, whereas the rest seven accessions could not be decided. Therefore, these 11 accessions were counted as one family and were merged with the Clementina group in the subsequent analysis

- Eleven additional international clones from the collections at ICG,T and Centro Agronómico Tropical de Investigación y Enseñanza, Costa Rica (CATIE): three Lower-Amazon Forasteros (BE-3, Amelonado-15, Amelonado-22), one Criollo (Criollo 13), one French Guiana wild cacao (GU 102/A), six hybrids from Nacional cacao (NAL 1, NAL 2, NAL 3, NAL 4, UF 705 and EET 96).
- Four Nacional accessions (Las Brisas 20, La Gloria 24, CCAT 11/19 and EB 04/02) from INIAP (Instituto Nacional de Investigaciones AgroPecuarias) Ecuador.

DNA was extracted at CRU following the protocol of Kobayashi et al. (1998) and quantified with ethidium-staining in 1% agarose gels. Aliquots of 50 µl were prepared and shipped to the USDA Beltsville Agricultural Research Center.

SSR analysis

Amplification of microsatellite loci used 15 primers with sequences previously described (Lanaud et al. 1999; Risterucci et al. 2000; Saunders et al. 2004). These 15 loci have been suggested as a standardized SSR primers to characterize all T. cacao germplasm collections (Saunders et al. 2001, 2004). This set of SSR primers has been used for cacao genotyping in several germplasm collections (Boccara and Zhang 2006; Zhang et al. 2006a, b). Primers were synthesized by Proligo (Boulder, CO, USA) and forward primers were 5′-labelled using WellRED fluorescent dyes (Beckman Coulter Inc., Fullerton, CA, USA). 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/ml of hot-start *Taq* DNA polymerase (Invitrogen Platinum *Taq*, Carlsbad, CA, USA; or Eppendorf HotMaster *Taq*, Brinkman, Westbury, NY, USA).

The amplified microsatellite loci were separated by capillary electrophoresis as previously described (Saunders et al. 2004; Zhang et al. 2006b). Data analysis was performed using the CEQ 8000 Fragment Analysis software Version 7.0.55 according to manufacturers' 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.).

Data analysis

For the verification of the genetic identity of each accession in the Refractario germplasm, we used a Bayesian test to assign each individual to its corresponding half-sib family. The program Structure Version 2.0 (Pritchard et al. 2000) was used for computation. The assumptions of k varies from 2 (e.g. in Hacienda Amalia

and San Juan, where fruits were taken from two trees) to 20 (e.g. in Hacienda Balao, where fruits were taken from 20 trees). All Structure runs used 10,000 iterations after a burn-in of length 10,000. The assignment probabilities then were computed for each individual, showing the degree to which its genome was classified into each cluster. The allocation of the individual to a particular cluster was set at not <90% probability. The individuals that were not assigned to the ''home family'' were considered as putatively mislabelled and were excluded from the subsequent analysis of intra- and inter-population variation.

The summary statistics for each marker locus, including allele number (Nei 1987), observed heterozygosity (H_o) and gene diversity were computed using PowerMarker Version 3.0 (Liu and Muse 2005). The Exact HW test (Guo and Thompson 1992) was used to test the deviation from HW equilibrium and was performed by the same program.

The within-population inbreeding coefficient (F_{IS}) was calculated and tested for significance by FSTAT (Version 2.9.3, Goudet 2001).

Genetic structures in the Refractario group were analysed by a hierarchical analysis of molecular variance (AMOVA, Excoffier et al. 1992), implemented in the software of Arlequin 3.0 (Excoffier et al. 2005). The total molecular variance was partitioned as among farms, among-families/within farm and among individuals/within family. The significance of Φ statistics was tested by permutation, with the probability of non-differentiation, for 1,000 randomizations. Genetic distance was calculated among all possible pairs of farms, families and individuals using the program PowerMarker Version 3.0 (Liu and Muse 2005). The pair-wise distances between farms followed the definition of Nei et al. (1983) and the distances among all pairs of farms were presented in a dendrogram using the UPGMA algorithm implemented in PowerMarker Version 3.0. Two reference populations, the Parinari population from Peru and the ICS population from Trinidad were included in the cluster analysis. The pair-wise distances among all families were calculated as Euclidian distance. The among-family distances were presented in a two dimensional scaling plot using the Multidimensional Scaling (MDS) procedure of SAS (SAS 1999). The pair-wise distances among all individuals were computed using the program GenAlEx (Peakall and Smouse 2006) and presented using a Principal Coordinates Analysis implemented in the same program. Thirteen international clones and four Nacional accessions were included as references in the Principal Coordinates Analysis.

Results

Identification of putative mislabelling using the Bayesian assignment test

With the prior knowledge of the family membership in each farm, a total of 348 accessions were correctly assigned to the 48 known half-sib families at a 90% threshold value. There were 134 accessions that failed to meet this threshold assignment probability, and were thus categorized as mislabelled (Table 1). Mislabelling ranged from 11.1 (Santa Lucia Farm) to 35.7% (L Farm) with an average rate of misidentification of 27.3% within the Refractario group. Nevertheless, the result showed that even with a high threshold probability (0.90), the majority of accessions (72.7%) could be assigned to home families that correctly corresponded to their membership in known families from different farms. Within each farm, ambiguously classified members were not used in subsequent diversity analysis. An example of the results of the assignment test for the accessions from the Amalia Farm is presented in Table 2. In this case, only two families were involved. Forty-six trees were assigned to the two families in Amalia Farm. For the purpose of the present study, the remaining 25 trees were considered to be mislabelled.

Genetic variation within the Refractario germplasm group

The total number of alleles discovered in the Refractario group was 63, with a range of two to seven alleles per locus and a mean of 4.2 alleles per locus (Table 3). The Refractario accessions appeared highly heterozygous. The mean expected heterozygosity ranged from 0.238 to 0.698, with an average of 0.561, whereas the mean observed heterozygosity ranged from 0.236 to 0.691, with an average of 0.554. Out of the 15 loci, ten significantly deviated from HWE (Table 3). Among the nine farms, the variation of allelic richness was small, ranging from 2.27 alleles per locus in L Farm to 3.60 alleles in Javilla Farm and Moquique Farm (Table 4). No private allele was observed in any of the nine farms. The mean inbreeding coefficient (F_{IS}) was not low and not significant for the 15 loci (F_{IS} = 0.009; Table 3) as well as for all the nine farms (F_{IS} = -0.15 to 0.06; Table 4), showing that there was neither deficiency nor excess of heterozygotes in the Refractario cacao.

AMOVA showed that majority of the molecular variance (76%) was contributed by the within-family variation (Table 5). The inter-family and inter-farm difference accounted for 15 and 9% of the total variance, respectively,

Table 2 List of 71 ''Refractario'' cacao accessions from the Amalia farm, Ecuador and their assigned population membership using Bayesian clustering analysis

Accession	Cluster	Probability of assignment	Accession	Cluster	Probability of assignment
AM 1/1(FP1278)	1	0.993	AM 2/1(FP1288)	2	0.035*
AM 1/10(FP1969)	1	0.995	AM 2/12(FP1286)	2	0.936
AM 1/107(FP45)	1	0.992	AM 2/13(FP1358)	2	0.070*
AM 1/109(FP799)	1	0.984	AM 2/17(FP2022)	2	0.989
AM 1/11(FP1972)	1	0.883*	AM 2/18(FP1282)	2	0.985
AM 1/12(FP1968)	1	0.982	AM 2/18(FP1965)	2	0.985
AM 1/19(FP2145)	1	0.348*	AM 2/20(FP665)	2	0.951
AM 1/21(FP696)	1	0.996	AM 2/21(FP1314)	2	0.016*
AM 1/28(FP702)	1	0.004*	AM 2/3(FP1439)	2	0.988
AM 1/29(FP1716)	1	0.996	AM 2/32(FP2446)	2	0.974
AM 1/3(FP1306)	1	0.981	AM 2/36(FP562)	2	0.810*
AM 1/33(FP1317)	1	0.989	AM 2/38(FP1281)	2	0.009*
AM 1/39(FP714)	1	0.986	AM 2/38(FP1284)	2	0.007*
AM 1/40(FP1593)	1	0.997	AM 2/39(FP1966)	2	0.992
AM 1/42(FP1334)	1	0.993	AM 2/4(FP1315)	2	0.992
AM 1/42(FP708)	1	0.994	AM 2/42(FP2010)	2	0.993
AM 1/48(FP1269)	1	0.995	AM 2/43(FP2034)	2	0.065*
AM 1/49(FP1285)	1	0.990	AM 2/45(FP2297)	2	0.030*
AM 1/5(FP1564)	1	0.985	AM 2/46(FP2288)	2	0.425*
AM 1/53(FP2237)	1	0.984	AM 2/5(FP676)	2	0.908
AM 1/55(FP1337)	1	0.739*	AM 2/50(FP2447)	2	0.989
AM 1/56(FP1967)	1	0.997	AM 2/53(FP1316)	2	0.121*
AM 1/60(FP1313)	1	0.988	AM 2/6(FP1289)	2	0.978
AM 1/63(FP1283)	1	0.993	AM 2/6(FP2298)	2	0.964
AM 1/68(FP1335)	1	0.990	AM 2/60(FP1559)	2	0.986
AM 1/7(FP710)	1	0.989	AM 2/61(FP1336)	2	0.004*
AM 1/70(FP678)	1	0.991	AM 2/63(FP1565)	2	0.004*
AM 1/72(FP703)	1	0.993	AM 2/64(FP264)	2	0.005*
AM 1/73(FP2147)	1	0.004*	AM 2/65(FP1277)	2	0.029*
AM 1/8(FP2000)	1	0.997	AM 2/68(FP1275)	2	0.992
AM 1/85(FP1116)	1	0.980	AM 2/70(FP1338)	2	0.969
AM 1/88(FP62)	1	0.994	AM 2/83(FP2142)	2	0.143*
AM 1/95(FP370)	1	0.980	AM 2/88(FP1195)	2	0.013*
AM 1/96(1)(FP412)	1	0.005*	AM 2/90(FP2143)	2	0.004*
			AM 2/91(FP2144)	2	0.160*
			AM 2/92(FP1606)	2	0.348*
			AM 2/94(FP2146)	2	0.964

*Twenty-five accessions failed to be assigned to their home family (threshold $P = 0.90$)

and both were highly significant ($P < 0.01$). Significant differentiation was detected between all pairs of farms by the AMOVA's permutation test of Φ statistics ($P < 0.05$; Table 5). Of the 435 pair-wise Φ statistics for the 30 families, 89% (385 pair-wise Φ-value) were found significant by permutation test.

The genetic relationships among the nine farms, as well as their relationships with the upper Amazon Forastero and Trinitario groups are illustrated by the dendrogram in Fig. 1. Within the Refractario group, the nine farms were divided into two subsets. The first subset included Amalia Farm, Clementina Farm (including Clementina Mixed), La Paz Farm and Moquique Farm whereas the second subset included Balao Farm, Javilla Farm, L Farm, Santa Lucia Farm and San Juan Farm. The dendrogram also showed that the Refractario was clearly separated from the Parinari and the ICS cacao groups.

Table 3 Summary statistics of 15 microsatellite loci in Refractario cacao germplasm collected from nine farms from the coast valley of Ecuador

Locus	N	N_a	H_o	H_e	F_{IS}	HWE test
Y16981	348	4.0	0.675	0.650	–0.039	0.008
Y16980	348	4.0	0.537	0.506	–0.062	0.000
Y16995	348	3.0	0.483	0.466	–0.036	0.000
Y16996	346	5.0	0.691	0.627	–0.101	0.000
Y16982	347	4.0	0.646	0.653	0.011	0.280^{NS}
Y16883	348	2.0	0.236	0.238	0.009	0.864^{NS}
Y16985	347	7.0	0.530	0.553	0.042	0.000
Y16986	347	4.0	0.628	0.650	0.033	0.003
Y16988	348	3.0	0.583	0.551	–0.059	0.062^{NS}
AJ271942	348	6.0	0.529	0.599	0.117	0.000
AJ271826	348	4.0	0.592	0.637	0.070	0.274^{NS}
Y16991	345	4.0	0.490	0.474	–0.033	0.466^{NS}
Y16998	347	5.0	0.625	0.698	0.105	0.000
AJ271943	347	4.0	0.522	0.605	0.137	0.000
AJ271958	348	4.0	0.540	0.508	–0.064	0.000
Mean	347.3	4.2	0.554	0.561	0.009	

Sample size (*N*), Number of alleles (N_a), Observed heterozygosity (H_o), Expected heterozygosity (gene diversity; H_e), Inbreeding coefficient (F_{IS}) and Exact test for deviation from HW equilibrium (*HWE*; Guo and Thompson 1992)

Values marked ^{NS} are not significant

Table 4 Diversity parameters for the Refractario cacao from nine farms in the coast valley of Ecuador

Farm (Hacienda)	N	K	H_e	H_o	F_{IS}
Amalia	46	2.87	0.54	0.62	–0.15
Balao	39	3.33	0.53	0.52	0.03
Clementina	58	2.92	0.51	0.53	–0.04
Javilla	58	3.60	0.55	0.61	–0.10
La Paz	39	3.20	0.50	0.57	–0.12
L	9	2.27	0.47	0.56	–0.13
Moquique	53	3.60	0.50	0.47	0.06
San Juan	22	2.73	0.52	0.54	–0.01
Santa Lucia	24	2.80	0.49	0.56	–0.11
Parinari	26	3.60	0.46	0.42	0.10

Number of accessions that passed the assignment test and were used for computation (*N*), Average number of alleles per locus (*K*), Expected heterozygosity (H_e), Observed heterozygosity (H_o) and Within-population inbreeding coefficient (F_{IS}). A sample of a natural population from Peruvian Amazon—the Parinari group, was included as a reference group for the purpose of comparison

The relations among different families were illustrated by the MDS plot (Fig. 2). The nine farms showed different level of within-farm heterogeneity. In Clementina Farm, Moquique Farm and Santa Lucia Farm, families from the same farm were largely grouped together, suggesting their common parentage among these families. Larger within-farm heterogeneity was observed in Amalia Farm, Balao Farm, Javilla Farm, La Paz Farm and San Juan Farm. In these farms, a few family pairs appeared to have larger difference (i.e. in the JA2–JA5 in Javilla Farm, AM1–AM2 in Amalia Farm, LP1–LP4 in La Paz Farm and SJ1–SJ2 in San Juan Farm).

The relation between the Refractario cacao and a diverse set of cacao germplasm groups was presented by the Principal Coordinates Analysis (Fig. 3). The plane of the first two main PCO axes, which accounted for 66.9% of total variation, showed that all the Refractario germplasm was clearly separated from the international clones used as controls, including lower and upper Amazon Forasteros, Trinitario, Criollo and French Guiana wild cacao. However, all clones which are known hybrids of Nacional cacao, i.e. clone EET 96 [ECU], UF 705, NAL 1, NAL 2, NAL 3 and NAL 4, intermingled with the Refractario group and were different from the true Nacional cacao (Fig. 3).

Table 5 Analysis of molecular variance (AMOVA) for SSR variation among and within nine cacao populations from the coast valley of Ecuador

Source	*df*	SSD[a]	MSD[b]	Variance component	% Total[c]	*P*-value[d]
Among pops	8	399.84	49.98	0.81	9%	0.001
Among families/pop	21	386.58	18.41	1.25	15%	0.001
Within pops	285	1,867.58	6.55	6.55	76%	0.001
Amalia	46	311.74	6.78	–	–	–
Balao	16	113.69	7.11	–	–	–
Clementina	58	442.58	7.63	–	–	–
Javilla	55	398.68	7.25	–	–	–
La Paz	36	228.61	6.35	–	–	–
L	9	51.78	5.75	–	–	–
Moquique	49	384.49	7.85	–	–	–
San Juan	22	167.68	7.62	–	–	–
Santa Lucia	24	154.90	6.45	–	–	–
Total	315	2,653.99	74.94	–	–	–

[a] Sum of squared deviations

[b] Mean squared deviations

[c] Per cent of total molecular variance

[d] Probability of obtaining a larger component estimate. Number of permutations = 1,000

Discussion

Identification of putative mislabelling and verification of family memberships

Mislabelling of germplasm accessions has been acknowledged as a serious problem in national and international cacao collections. However, until recently tools have not been available to clearly identify mislabelled accessions. SSR markers are highly suitable for the application of the assignment test, which determines the population of origin of a single individual through Bayesian method (Pritchard et al. 2000). The method of Pritchard et al. (2000) to assign individuals to populations does not assume any particular mutation model. This method needs a relatively small number of loci to detect a very strong signal of population structure and assign individuals appropriately (Pritchard et al. 2000). In our previous reports, we have demonstrated the effectiveness of using Bayesian assignment tests in cacao germplasm identification (Zhang et al. 2006a).

In the present study, all the Refractario accessions were labelled by the location (farm) names and their tree number. This labelling provides information of the ''home family'' as a priori for the assignment test. With the threshold probability at 0.90, a total of 134 accessions (28%) failed to be assigned to their claimed families (Table 1). Similar overall misidentification rates have been recorded for the ICG,T (Motilal and Butler 2003; Boccara and Zhang 2006) as well as other genebanks (Zhang et al. 2006b). The correctly assigned trees would serve as reference true-type trees for future work on verification of the multiple trees in the ICG,T as well as in other international and national germplasm collections. However, it needs to be pointed out that the decision of how stringent the threshold probability should be is subjective, and depends on the purpose of the assignment test. In the present study, our main purpose was to assess the population structure of the Refractario germplasm group. Our goal was to eliminate any accessions with possible ambiguous membership, so that the confounding factors in subsequent diversity analyses could be minimized. Therefore, we took a highly stringent threshold for the assignment test. In the routine use of the multilocus fingerprints for germplasm identification, we usually take a lower threshold ($P = 0.75$), and also combine assignment test with other methods such as sib-ship reconstruction to verify the putative mislabelling.

The genetic identity and population structure of the Refractario cacao

Little information is available about the genetic identity of the Refractario cacao from Ecuador. It was presumed that a group of different varieties had contributed to the parentage of the Refractario accessions. This presumption was based on unpublished reports regarding the cacao varieties grown in the coastal valley of Ecuador in the 1920s and their resistance to witches' broom disease (Bartley 2001). The present results, based on microsatellite analysis, substantiated the hypothesis that the Refractario is a group of

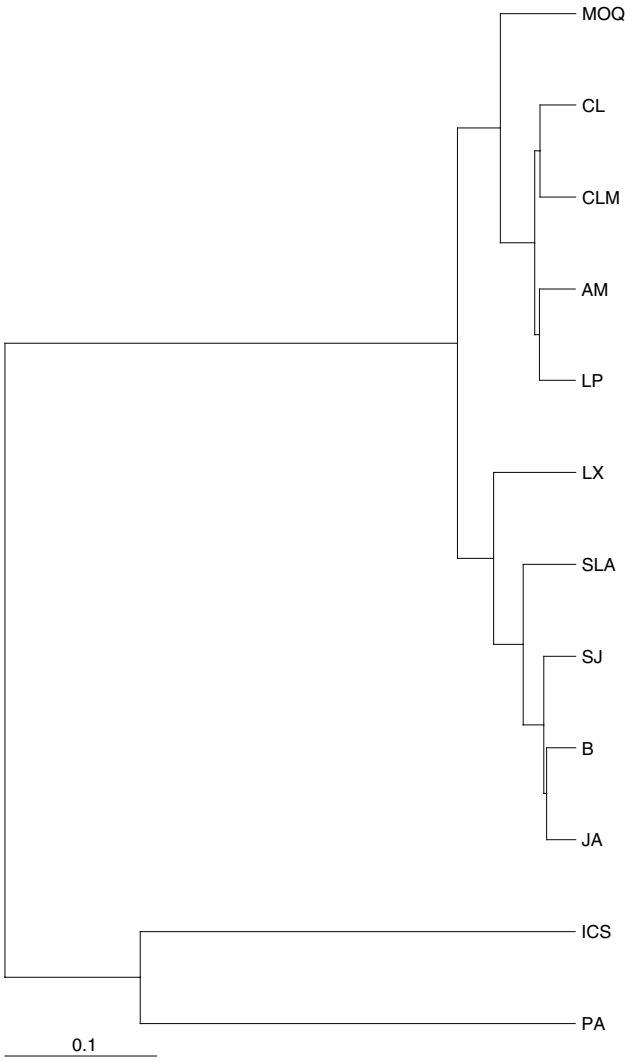

Fig. 1 Dendrogram of nine cacao populations (including 348 accessions) from the coastal valley of Ecuador and their relationship to an Upper Amazon Forastero population from Peru and a Trinitario population from Trinidad. The cluster analysis was based on Nei's distance (Nei et al. 1983)

hybrids derived from multiple parents, but that these parents appear to have close genetic relationships to each other. The level of heterozygosity was high in Refractario and there was no sign of inbreeding.

Our result also showed that the Refractario had a distinctive genetic profile among the cacao germplasm groups. The close genetic relationship between the Refractario and the Nacional hybrids [UF 705, NAL 1, NAL 2, NAL 3, NAL 4 and EET 96 (ECU)] suggested that the Nacional cacao was one of the parents of the Refractario group. However, whether the Nacional cacao contributed pollen to those selected seedling progenies when they were planted in different farms, or whether those selected progenies came directly from Nacional trees is as yet undetermined.

The present study also illustrated a population substructure in the Refractario group. Significant Φ statistics ($P < 0.05$) were observed in all 36 pair-wise comparisons among the nine farms. Cluster analysis grouped the nine farms into two main subsets. Within each subset, the genetic profiles of different farms overlapped to a various extent. The result again suggested that in addition to the putative common parentage from the Nacional trees, other different parental varieties contributed to the formation of Refractario. However, these parental varieties all shared a similar genetic background.

The global allelic richness of the Refractario was moderate (4.2 alleles per locus). Using the same set of 15 SSR markers, we identified 8.0 alleles in a group of germplasm from the Ucayali river valley of Peru (Zhang et al. 2006a) and an average of 7.3 alleles from other upper Amazon populations in our unpublished diversity survey in the valleys of Rio Nanay, Rio Morona and Rio Marañón (D. Zhang, M. Michell and D. R. Butler). This suggests that the Refractario cacao was derived from a limited range of genetic diversity.

Implications for cacao germplasm conservation and crop improvement

Three major cacao diseases, witches' broom, frosty pod rot and black pod, constitute a serious threat to the livelihoods of cacao farmers in the Americas. Cacao production in the Americas has dropped by 75% in last 16 years largely due to these three diseases. The challenges posed by these devastating diseases create a need to explore new sources of resistance for the present and future genetic improvement of cacao.

The ICG,T holds several core sets of germplasm from Ecuador, among which the Refractario group does not represent a geographical population as do the other groups. However, this group contains a useful proportion of accessions with low pod index, high bean number, heavy beans and resistance to *Phytophthora palmivora* (Iwaro et al. 2003) and continues to show resistance to witches' broom disease in Trinidad (Thévenin et al. 2005), and therefore represents an important source of breeding material. The present study verified genetic identity and sibship relationships in the Refractario germplasm group. We also showed that the Refractario group has a unique genetic profile among the existing cacao germplasm. Although it is composed of a large number of half-sib families, the group is relatively homogeneous suggesting that they were derived from a small number of parents that share similar genetic background. Their uniqueness among the various germplasm groups from the Americas, in terms of genetic composition, highlights the need to strengthen the evaluation of

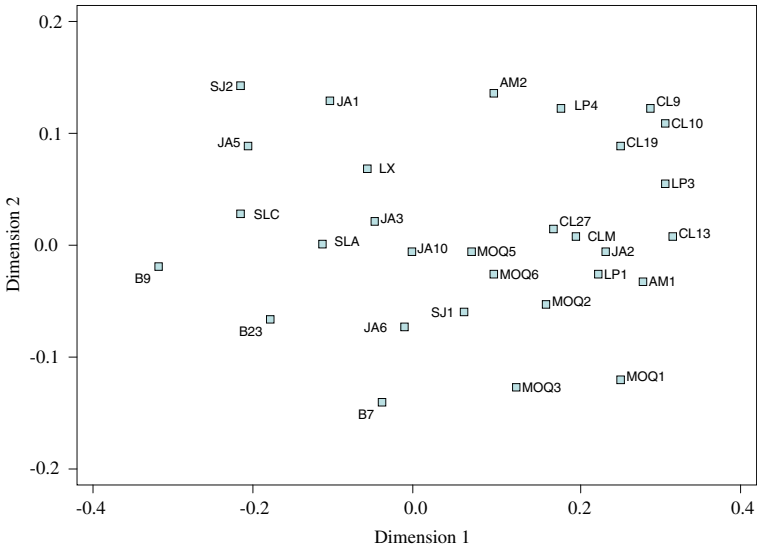

Fig. 2 Inter-family relationship in the Refractario cacao revealed by multidimensional scaling plot. The relationship was based on pair-wise Euclidian distance among 30 families of the Refractario cacao. The family acronym corresponds to sample list in Table 1. Accessions with low assignment probability ($P < 0.90$) assessed by Bayesian's cluster method (Pritchard et al. 2000) and families with progeny size smaller than 5 were not included in this analysis

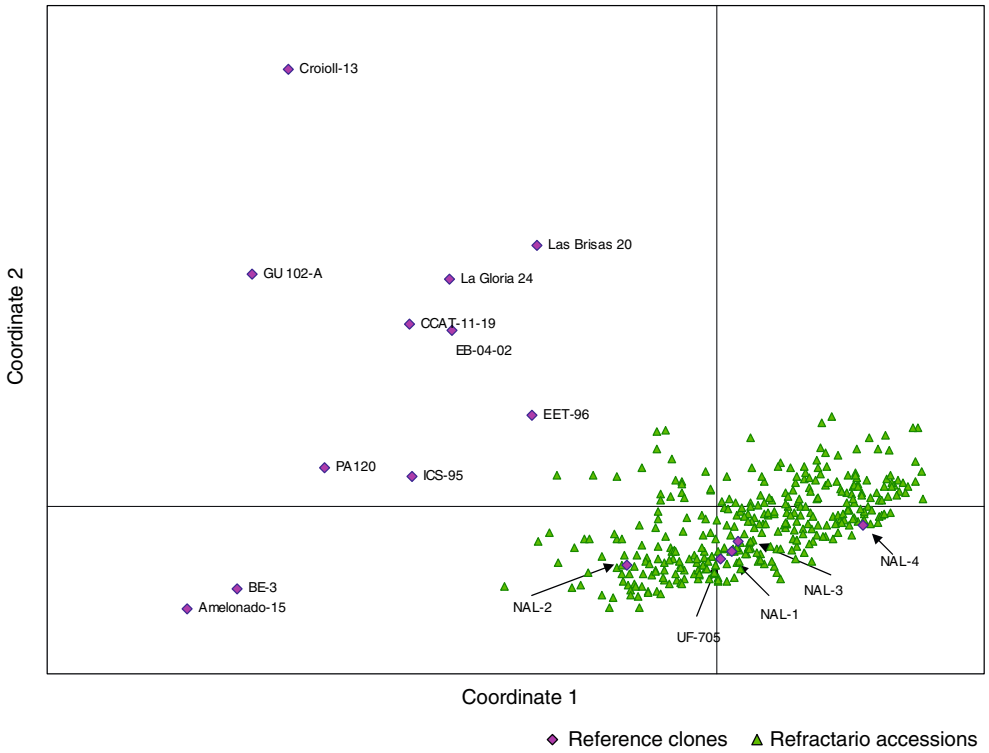

Fig. 3 Principal coordinates analysis of the Refractario cacao and the reference Forastero, Trinitario, Criollo, Nacional hybrids and the Nacional germplasm accessions (first axis = 43.5% of total information and the second = 23.4%)

Refractario accessions and their usage in breeding. The information on population structure in this germplasm group will also allow improvement in the efficiency and accuracy of cacao germplasm conservation.

Acknowledgements We thank Stephen Pinney and Eric Tilson for their contributions to the SSR genotyping; Freddy Amores and Rey Loor for providing the DNA samples of the ''Nacional'' accessions. Antoinette Sankar is thanked for performing the DNA extractions at CRU. Special thanks are due to Drs. Lizz Johnson, Ainong Shi and an anonymous reviewer for their review of the manuscript.

References

Bartley BGD (2000) The nomenclature of the accessions derived from Dr. F. J. Pound's collections in Ecuador in 1937. INGENIC Newsl 5:4–6

Bartley BGD (2001) Refractario—an explanation of the meaning of the term and its relationship to the introductions from Ecuador in 1937. INGENIC Newsl 6:10–15

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

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: Proceedings of the CRU annual report 2005, Cacao Research Unit, The University of the West Indies, St. Augustine, Trinidad and Tobago

Charters YM, Wilkinson MJ (2000) The use of self-pollinated progenies as ''in-groups'' for the genetic characterization of cacao germplasm. Theor Appl Genet 100:160–166

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

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

Coe SD, Coe MD (1996) The true history of chocolate. Thames and Hudson, London

Cuatrecasas J (1964) Cacao and its allies. A taxonomic revision of the genus theobroma. Contrib. US Nat. Herbarium 35(6). Smithsonian Institution, Washington, DC

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

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. Available at http://www.la-press.com/EBO-1-Excoffier(Pr).pdf

Goudet J (2001) FSTAT, a Program to estimate and test gene diversities and fixation indices (V. 2.9.3). Available from http://www.unil.ch/izea/softwares/fstat.html

Guo SW, Thompson EA (1992) Performing the exact test of Hardy–Weinberg proportion for multiple alleles. Biometrics 48:361–372

Iwaro AD, Bekele FL, Butler DR (2003) Evaluation and utilisation of cacao (*Theobroma cacao* L.) germplasm at the international cacao genebank, Trinidad. Euphytica 130:207–221

Kennedy AJ, Mooleedhar V (1993) Conservation of cacao in field genebanks—the international cacao genebank, Trinidad. In: Proceedings of the international workshop on conservation, characterization and utilization of cacao genetic resources in the 21st century. The Cacao Research Unit, Port-of-Spain, Trinidad, pp 21—23, 13–17 September 1992

Kobayashi N, Horikoshi T, Katsuyama H, Handa T, Takayanagi K (1998) A simple and efficient DNA extraction method for plants, especially woody plants. Plant Tissue Cult Biotechnol 4:76–80

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, Motanayor JC, Risterucci AM (2001) Implications of new insight into the genetic structure of *Theobroma cacao* L. for breeding strategies. In: Bekele F, End M, Eskes AB (eds) Proceeding of the international workshop on new technologies and cacao breeding. Kota Kinabalu, Sabah, Malaysia, INGENIC Press, Malaysia, pp 89—107, 16–17 October 2000. Available at http://www.koko.gov.my/CacaoBioTech/ING_Workshop(89–97).html

Laurent V, Risterucci AM, Lanaud C (1993) Genetic diversity in cacao revealed by cDNA probes. Theor Appl Genet 88:193–198

Laurent V, Risterucci AM, Lanaud C (1994) RFLP study of genetic diversity of *Theobroma cacao*. Angew Bot 68:36–39

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

Liu J, Muse S (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinform Appl Note 21:2128–2129. doi:10.1093/bioinformatics/bti282 (Free Program, V 3.23, distributed by author, Available at http://www.powermarker.net)

Lockwood C, End M (1993) History, technique and future needs for cacao collection. In: Proceedings of the workshop on the conservation, characterization and utilization of cacao genetic resources in the 21st century. The Cacao Research Unit, Port-of-Spain, Trinidad, pp 1—14, 13–17 September 1992

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

Motamayor JC, Risterucci AM, Heath M, Lanaud C (2003) Cacao domestication II: progenitor germplasm of the trinitario cacao cultivar. Heredity 91:322–330

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

N'Goran JAK, Laurent V, Risterucci AM, Lanaud C (2000) The genetic structure of cacao populations (*Theobroma cacao* L.) revealed by RFLP analysis. Euphytica 115:83–90

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

Pound FJ (1938) Cacao and witches' broom disease (Marasmius perniciosus) of South America. In: Toxopeus H (eds) Archives cacao research, vol 1. American Cacao Research Institute, Washington DC and International Office of Cacao and Chocolate, Brussels, Belgium, pp 20–72

Pound FJ (1943) Cacao and witches' broom disease (Marasmius perniciosa). Report on a recent visit to the Amazon territory of Peru, September 1942–February 1943. Yuille's Printery, Port of Spain, Trinidad and Tobago

Pound FJ (1945) A note on the cacao population of South America. In: Report and Proceedings of the cacao research conference held at colonial office. The Colonial Office, His Majesty's Stationery Office, London, pp 131—133, May–June 1945

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

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

SAS (1999) SAS Version 8.02: SAS/STAT Software: changes and enhancements through Release 8.02. SAS Institute Inc., Cary, NC

Saunders JA, Hemeida AA, Mischke S (2001) USDA DNA fingerprinting programme for identification of *Theobroma cacao* accessions. In: Bekele F, End M, Eskes AB (eds) Proceeding of the international workshop on new technologies and cacao breeding. Kota Kinabalu, Sabah, Malaysia, INGENIC Press, Malaysia, pp 108—114, 16–17 October 2000

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

Sounigo O, Christopher Y, Umaharan R (1996) Genetic diversity assessment of *Theobroma cacao* L. using isoenzyme and RAPD analyses. In: Cacao Research Unit, Report for 1996, The University of the West Indies, St. Augustine, Trinidad, pp 35–51

Sounigo O, Christopher Y, Bekele F, Mooleedhar V, Hosein F (2001) The detection of mislabelled trees in the international cacao genebank, Trinidad (ICG,T) and options for a global strategy for identification of accessions. In: Bekele F, End M, Eskes AB (eds) Proceeding of the international workshop on new technologies and cacao breeding. Kota Kinabalu, Sabah, Malaysia, INGENIC Press, Malaysia, pp 34—39, 16–17 October 2000

Takrama JF, Cervantes-Martinez C, Philips-Mora W, Brown JS, Motamayor JC, Schnell RJ (2005) Determination of off-types in a cacao breeding program using microsatellites. INGENIC Newsl 10:2–7

Thévenin JM, Umaharan R, Surujdeo-Maharaj S, Latchman B, Cilas C, Butler DR (2005) Relationships between black pod and witches'-broom diseases in *Theobroma cacao*. Phytopathology 95:1301–1307

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

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

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