© NIAB 2012
ISSN 1479-2621

Plant Genetic Resources: Characterization and Utilization (2012) **10**(3); 232–241
doi:10.1017/S1479262112000305

Elucidation of genetic identity and population structure of cacao germplasm within an international cacao genebank

Lambert A. Motilal[1]*, Dapeng Zhang[2], Pathmanathan Umaharan[1], Michel Boccara[1,3], Sue Mischke[2], Antoinette Sankar[1] and Lyndel W. Meinhardt[2]

[1] *Cocoa Research Unit, The University of the West Indies, St. Augustine, Trinidad, Republic of Trinidad and Tobago, West Indies,* [2] *USDA/ARS, Beltsville Agricultural Research Center, PSI, SPCL, 10300 Baltimore Avenue, Bldg. 001, Rm. 223, BARC-W, Beltsville, MD 20705, USA and* [3] *Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Biological Systems Department, Unité Mixte de Recherche Développement et Amélioration des Plantes (UMR DAP) TA A 96/03-34,398, Montpellier, France*

Received 30 May 2012; Accepted 30 August 2012 – First published online 23 October 2012

Abstract

Theobroma cacao L., or cacao, is the source of cocoa products used in the making of chocolate. These tropical trees are conserved in living genebanks. The International Cocoa Genebank, Trinidad is one of the largest *ex situ* collections in the public domain. Mislabelling is a critical problem and the correction of this problem is vital to improve the accuracy and efficiency of genebank management. Using microsatellite DNA markers, we assessed the level of mislabelling in a group of Refractario cacao that originated from Ecuador and determined their population memberships through Bayesian clustering analysis. The microsatellite data revealed a synonymous rate of 7.5% and an error rate of 39.4% in this germplasm subset. The analysis of the population structure grouped the Refractario accessions into four subclusters, indicating intra-population heterogeneity in this germplasm group. Based on the results, we recommend that when the assignment test is used for cacao genotype identification, it should (a) use duplicated samples as internal checks, (b) choose suitable reference accessions, including a known homogeneous group and (c) employ subclustering checks to obtain reliable results. The information framework generated is discussed in relation to the management of the collection, population enhancement and future research of the collection.

Keywords: cocoa; conservation; field genebank; mislabelling; population structure; tree identity

Introduction

Seeds of *Theobroma cacao* L. (cacao) provide cocoa mass and cocoa butterfat, the raw materials of the multi-billion-dollar confectionery industry. Cacao germplasm is conserved live *in situ* or *ex situ* since this outcrossing tropical tree crop has recalcitrant seeds (Toxopeus, 1985). Of the 50 *ex situ* field genebanks (Motilal and Butler, 2003), only two are universal collections: Centro Agronomico Tropical de Universal Investigacion y Enseñanza, Turrialba (CATIE) in Costa Rica and the International Cocoa Genebank, Trinidad (ICG,T) in Trinidad. The latter, managed by the Cocoa Research Unit (CRU) of the University of the West Indies, is the largest and

* Corresponding author. E-mail: lamotilal@yahoo.com

most diverse public domain collection. The ICG,T contains germplasm from multiple expeditions, beginning in 1930, from Amazonian South America, Central America and the West Indies (Kennedy and Mooleedhar, 1993). Details of the ICG,T have been documented in Kennedy and Mooleedhar (1993), Bekele and Bekele (1996), Iwaro *et al.* (2003), Motilal and Butler (2003), Sounigo *et al.* (2005), Motilal *et al.* (2011) and at http://sta.uwi.edu/cru. The ICG,T contains grafted and rooted cuttings representing approximately 0.18% Criollo, 30.2% Forastero, 37.0% Refractario, 15.8% Trinitario and 16.8% unknown accessions.

Mislabelling errors within field germplasm collections have been recognized (http://cropgenebank.sgrp.cgiar.org/index.php?option=com_content&view=article&id=549&Itemid=744). Mislabelling is a major hindrance in the conservation, dissemination and efficient use of crop germplasm (Hurka *et al.*, 2004) including grape (Leão *et al.*, 2009), lettuce (van Treuren *et al.*, 2010) and cacao (Motilal and Butler, 2003; Irish *et al.*, 2010; Motilal *et al.*, 2011). Mislabelled plants that are phenotypically similar but genetically dissimilar will inflate genetic variance instead of phenotypic variance. The ICG,T has a high safety duplication level, with 16 clones (maximum) of an accession in each plot. Mislabelling can primarily exist as (a) an admixture of trees from various accessions, which may or may not include the expected accession and (b) a uniform plot but all trees are of another accession. Mislabelling occurs as a result of (a) inadvertent budwood collection, (b) clerical errors in the transcription of plant tags and map records, (c) incorrect replacement of labels on field trees and (d) inadvertent planting. Overtopping of scion by rootstock material will lead to mislabelling in grafted plants, if the scion insert is weak, broken off or dies back.

Plot admixture in the ICG,T has previously been addressed (Motilal *et al.*, 2011) but the assignation of a tree to a given accession nomenclature is pending. Cacao accessions can be identified from phenotypic examination (Engels *et al.*, 1980; Bekele and Bekele, 1996; Bekele and Butler, 2000; Bekele *et al.*, 2006). Johnson *et al.* (2007) advocated the use of field guides in identifying cacao accessions. DNA fingerprinting techniques (e.g. microsatellite or single nucleotide polymorphism markers), however, are efficient, accurate and unambiguous means of plant identification. This study therefore employed microsatellite markers to (a) determine the percentage of incorrectly named trees (homonymous error), (b) determine the level of duplication within the ICG,T (synonymous error), (c) determine the correct population clustering and hence (d) improve the management strategy for the collection based on a subset of the collection.

Materials and methods

Plant material

Healthy leaves (flush–mature) from 484 samples of 387 cacao accessions (17% of ICG,T accessions) were opportunistically harvested to facilitate the identification of at least one tree of an accession. The samples were constituted as (a) a single tree from 301 accessions, (b) three accessions with two trunks sampled from the same tree number, (c) 76 accessions with two sampled trees and (d) nine accessions with three sampled trees. Sets (b) and (c) contained a common accession. In addition, leaves from a reference set of 26 accessions were retained to act as a pool of distinctive alleles. These reference accessions were composed of two Upper Amazon Forastero accessions from Peru; 15 Criollo accessions from Belize (10), CATIE (2) and Honduras (3); six Lower Amazon Forastero accessions from Brazil (5) and the USDA Tropical Agriculture Research Station cacao germplasm collection in Puerto Rico (1); two Trinitario clones (ICS 97 and MXC 67) from Trinidad and a reference IMC 67 tree from La Reunion Estate of the Ministry of Food Production, Land and Marine Affairs of Trinidad and Tobago. The complete list of accessions can be found in Supplementary Table S1 (available online only at http://journals.cambridge.org).

DNA extraction and quantification

Total leaf genomic DNA was extracted similarly to that described in Motilal *et al.* (2010). Maceration was performed with a 120 V FastPrep instrument (Qbiogene, Inc., Carlsbad, CA, USA) using lysing matrix A. DNA was maintained in sterile deionized water or Tris–EDTA buffer and stored at −20°C. Stock DNA solutions were assayed with either (a) PicoGreen® (Molecular Probes, Eugene, OR, USA) in a Fluroskan Ascent system (Labsystems, Helsinki, Finland), (b) Hoechst dye in a TKO fluorometer or (c) a NanoDrop 8000 spectrophotometer, according to the manufacturer's recommendations. Working solutions were prepared at ~0.1 ng/μl of total DNA.

PCR amplification

Twenty-six microsatellite primer pairs (Supplementary Table S2, available online only at http://journals.cambridge.org) were used to generate independent DNA polymorphisms. Characteristics of these primers can be found online at www.ebi.ac.uk and in Lanaud *et al.* (1999), Pugh *et al.* (2004) and Saunders *et al.* (2004). Microsatellite amplification was as described in

Motilal *et al.* (2009). The *Taq* polymerase employed was Eppendorf HotMasterMix (Brinkmann Instruments Inc., Westbury, NY, USA) or AmpliTaq Gold DNA polymerase (Applied Biosystems, Foster City, CA, USA).

Electrophoresis

Fragment lengths of amplified loci were sized on an 8000 or 8800 capillary electrophoresis system (Beckman Coulter, Inc., Brea, CA, USA) using an internal 400 bp DNA Size Standard Kit as a reference, according to the manufacturer's instructions (Beckman Coulter, Inc.). Binning was performed as described earlier (Motilal *et al.*, 2009).

Multilocus matching

The allelic dataset (4% missing data; dataset I) was checked for binning errors with the Excel Microsatellite Toolkit v.3.1.1 add-in (Park, 2001). The multilocus microsatellite profiles were subjected to all possible pairwise matching, and a mismatch flexibility of three loci with a minimum of 20 matching loci in CERVUS v.3.0.3 (Kalinowski *et al.*, 2007) was implemented. Trees with the same accession name but different multilocus profiles were deemed homonyms. Trees with different accession names but equivalent multilocus profiles were deemed synonyms. Synonymous accessions were replaced with their appropriate single consensus entry. Homonymous accessions were recoded and kept as separate entries. Multilocus profiles in the new dataset (dataset II; 415 individuals, 26 loci, 2.6% missing data; three samples duplicated as internal checks) were then matched manually against the reference tree microsatellite profiles that had been compiled in the CRU/USDA fingerprinting project. An adjusted dataset to align allele bins in the aforementioned project and the present study was created and matching accessions were determined for a mismatch flexibility of two loci with a minimum of 13 matching loci in CERVUS v.3.0.3 (Kalinowski *et al.*, 2007).

Microsatellite loci and dataset II

Probabilities of identity (Waits *et al.*, 2001) of the 26 loci were calculated using the software GIMLET (Valière, 2002). Descriptive statistics for these loci were determined on dataset II (415 samples) with GenAlEx v.6.1 (Peakall and Smouse, 2006). Pairwise genetic distances among all individuals were calculated and the standardized distance (Nei, 1972, 1978) was then used in a principal coordinate analysis with this software. The eigenvectors were graphed with SigmaPlot 2002 v.8.0 (SPSS, Inc., 1986–2001).

Population assignment

Population assignment analysis was conducted on dataset II (containing three known duplicated samples) with STRUCTURE v.2.3 (Pritchard *et al.*, 2000). A burn-in period of 200,000 runs followed by 500,000 Markov Chain Monte Carlo (MCMC) runs was employed under an admixture model with independent allele frequencies. Alpha was inferred in the model. Population groups from $K = 2$ to $K = 12$ were assessed with 50 independent replicates each. The results of the STRUCTURE output were taken into Structure Harvester v.0.6.8 (Earl and von Holdt, 2011) to obtain (a) the minimum number of populations as determined from the method of Evanno *et al.* (2005) and (b) formatted files for alignment in CLUMPP (Jakobsson and Rosenberg, 2007). Alignment of the *Q*-matrices was matched by permutation with the Large K Greedy algorithm under a random matrix option in CLUMPP (Jakobsson and Rosenberg, 2007).

The *ln* Pr results from the original STRUCTURE runs were tabulated and sorted, and a trimmed mean calculated after removing the highest and lowest values. The best-fit number of populations was assessed using the turning points from plots of change in *ln* Pr versus change in *K*. The lowest *K* value that best fitted the data was chosen as the number of effective populations. At this *K* value, the least negative *ln* Pr was chosen to represent membership plots and group contributions (*Q* values).

For each *K*, the 50 independent runs were examined for individuals with at least 5% Criollo ancestry. An ANOVA was carried out with the Group Differences Program v.3.0 (Chang, 2001). Duncan's multiple range test as implemented in DSAASTAT v.1.1 (Onofri, 2007) was used to distinguish the *K* groups from each other.

Mislabelling from population assignment

A threshold value of $Q \geq 0.85$ was employed as the group membership inclusion criterion. Individuals with $Q < 0.85$ were considered as ambiguous individuals and treated as mislabelled samples. Mislabelled accessions were also identified by running STRUCTURE v.2.3 (Pritchard *et al.*, 2000) on independent datasets of each accession group. Individuals were also partitioned into their appropriate populations based on the results of the 50 replicate analyses. Substructures within these groups without admixed individuals were run as required. Model parameters were similar to those used before except that *K* was set from 1 to *n*, where $n = 5$ or higher as the dataset required (maximum = 12), and 30 iterations were made for each *K* value. A correlation model (Falush *et al.*, 2003) under these parameters was further employed for datasets of Amelonado and Refractario individuals. The best-fit *K* value was chosen as before. Comparisons of membership assignment from these three approaches were then reviewed, and a reduced dataset was obtained with each subpopulation containing only true members as identified by the inclusion criteria.

The population data of Motamayor *et al.* (2008) were reduced to a dataset with individuals with high coefficients of membership for the pure Amelonado, IMC, PA and NA populations involved in the present study. Seventeen loci were common to the present study, and these loci were retained for the Motamayor *et al.* (2008) reference dataset. Individuals with more than ten missing data points were removed. Allele sizes were aligned to those of the present study. One locus was removed due to difficulty in alignment. Mislabelled accessions in the present study, which fell as pure samples into the aforementioned population groups, were assessed for match declaration with CERVUS v.3.0.3 (Kalinowski *et al.*, 2007). Match declarations were guided at a minimum of 13 matching loci and a mismatch of two loci.

Results

Homonymous error and synonymous redundancy

Tree mislabelling as homonyms was present in 17 of the 88 accessions with replicate samples (Supplementary

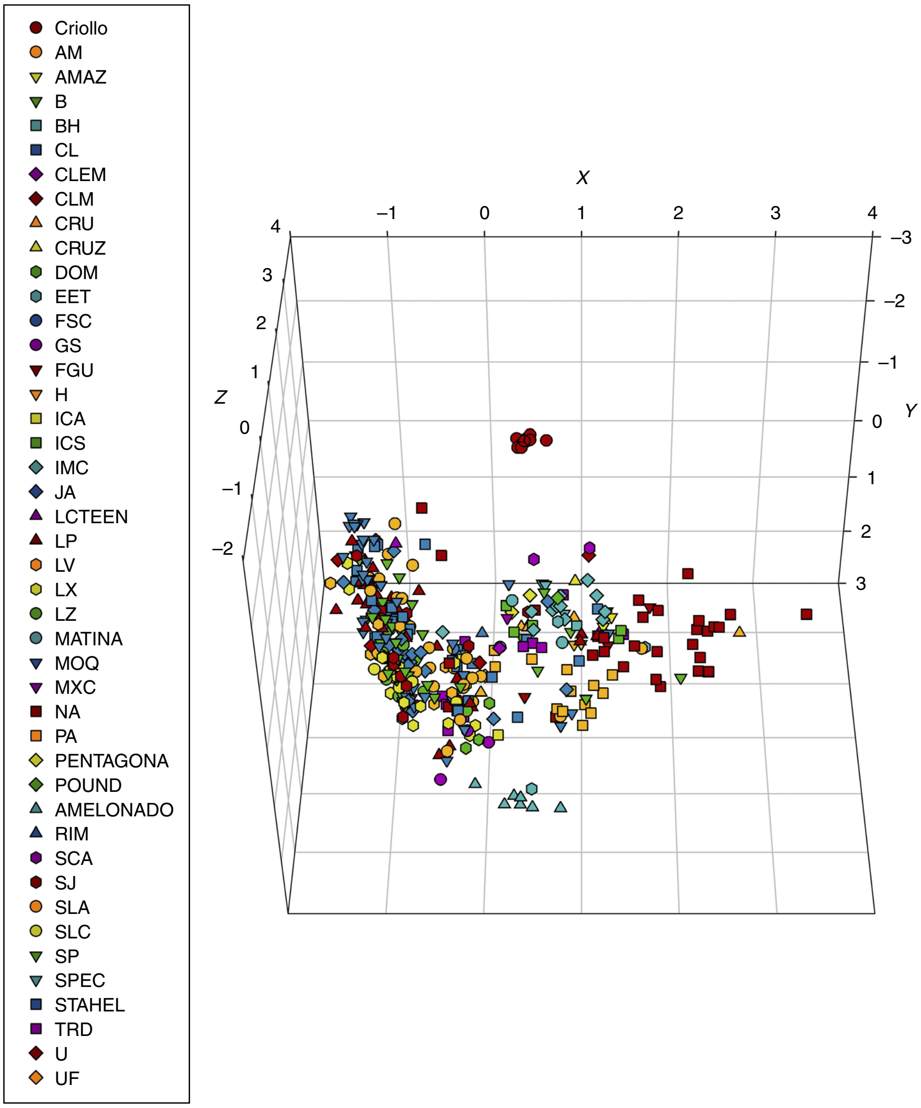

Fig. 1. Principal coordinate analysis on 415 cacao samples with 26 microsatellite loci. The three axes explained 83.22% of the total variation.

Table S3, available online only at http://journals.cambridge.org). Synonymous cases were present for 29 distinct pairs of matched accessions from 388 accessions managed by the CRU (Supplementary Table S3, available online only at http://journals.cambridge.org). This represented error rates of 19.3% homonymy and 7.5% synonymy at the accession level. Of the 208 accessions with accepted true-type reference trees, 82.7% were matched in the current dataset, yielding a 17.3% mislabelling error.

Microsatellite loci and dataset II

From the 415 samples of dataset II, the microsatellite loci detected 5–15 alleles, with a range of 0.193–0.462 for the fixation index and a range of 0.3643–0.6954 for the PID_{sib} (Supplementary Table S2, available online only at http://journals.cambridge.org). The combined probability from all 26 loci was 4.097×10^{-10} and the probability ranged between 4.73×10^{-7} and 5.31×10^{-14} for matching an individual. The four most informative loci (unordered) were Y16996, Y16988, AJ271942 and AJ566565 from the PID_{sib} and Shannon's information index, respectively. The fifth most informative locus was Y16995 or AJ271944 for these two respective measures.

Multidimensional scaling revealed a clustering of individuals, with a clear separation of the Criollo and Amelonado samples from all other accessions (Fig. 1). The three axes explained 83.22% of the total variation, with the first two axes explaining 59.35% of the variation. The NA, PA and IMC accessions tended to cluster together. The reference accession U 1 was in close proximity to the SCA accessions (Fig. 1).

Population structure

The 415 individual representative samples (inclusive of the reference accessions) could be fitted into three groups (Criollo–Amelonado–Trinitario, Forastero and Refractario) by the method of Evanno *et al.* (2005). With the alternative graphing method described here, the dataset could be assigned to four populations, and with subclustering into eight or ten populations (Supplementary Fig. S1, available online only at http://journals.cambridge.org). As the dataset was partitioned, several events were noticed. First, the duplicated internal checks were consistently assigned across *K* assignations. Second, the Forastero group began to be partitioned at $K = 4$ (individual plots) or 5 (CLUMPP alignment), as the French Guiana and PA accessions were separated. The French Guiana group (ELP and GU accessions) and the PA clustering were separated from each other at $K = 10$ when representative individual plots were examined, but remained clustered according to the CLUMPP alignment. Third, at $K = 4$, the reference Amelonado accessions, together with the Trinitario accessions, separated from the Criollo group. The Amelonado and Trinitario accessions remained clustered together at all *K* groups assessed. Fourth, the SCA group separated out at $K = 7$ from the other Forastero accessions. Fifth, as $K = 8$ moved to $K = 10$, three samples (NA 471 Field 6A B86 T9 = Field 4A D412 T1; EET 400 Field 6B F455 T6 and CRUZ 7/8 Field 6B B83 T1 = T9) were further subdivided. Lastly, the number of accessions with Criollo ancestry became progressively less and was significantly different ($P < 0.05$) up to $K = 5$ but was relatively the same thereafter (Fig. 2). Criollo individuals appeared admixed at $K = 2$, 4 and 5 according to the CLUMPP permuted matrix.

Generally, individuals were either admixed (96 samples, 23.1%) or they fell into one of eight main groups: Amelonado (75 samples), Criollo (16 samples), French Guiana (four samples), IMC (14 samples), NA (24 samples), PA (18 samples), Refractario (165 samples) and SCA (three samples). The last was composed of the two SCA samples (SCA 3 and SCA 6) and the U 1 reference accession.

The material with Amelonado ancestry could be partitioned into two or three main clusters under the independent or the correlated allele model, respectively. However, the increased partitioning under the correlated model did not coincide with any biological clustering and resulted in several admixed individuals. The material with Amelonado ancestry was therefore separated into two subclusters, consisting of the reference Amelonado accessions in one group and all other accessions with Amelonado ancestry in the other group.

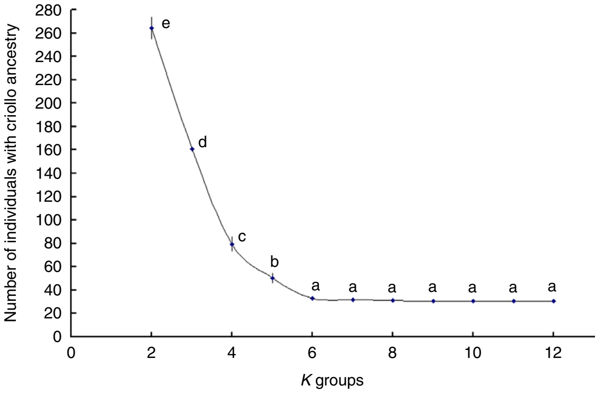

Fig. 2. Influence of population clustering on the number of individuals with Criollo ancestry. *K* points (number of population groups) with different letters are significantly different ($P < 0.05$) from each other. Duncan's multiple range test was conducted with DSAASTAT v.1.1. (Onofri, 2007) after ANOVA with the Group Differences Program v.3.0 (Chang, 2001).

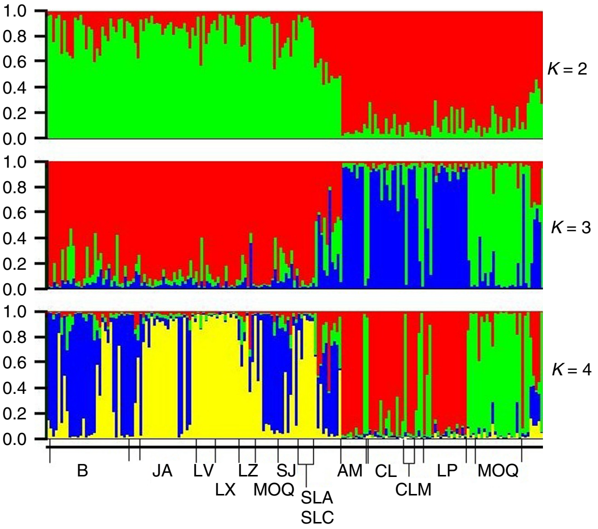

Fig. 3. Subclustering within the Refractario accessions. *K* is the number of population groups.

The Refractario accessions were clustered into two main groups (B and O) from the dataset of 415 individuals. The exclusion of non-Refractario accessions revealed that each Refractario cluster was composed of two subpopulations under both the independent and correlated allele models (Fig. 3). Cluster B was composed of OB1 (B and SJ accessions) and OB3 (JA, LV, LX, LZ, SLA, SLC and SJ accessions). Cluster O was composed of OB2 (AM, CL, CLM and LP accessions) and OB4 (MOQ accessions) (Supplementary Table S1, available online only at http://journals.cambridge.org). STRUCTURE analysis of a dataset of only SLA and SLC accessions revealed that these two accessions stayed as one cluster. In contrast, a dataset of CLM and CLEM accessions was clearly separated into these two accessions.

Typing trees

The percentage of true-type trees in the accession groups ranged from 32% (AM) to 100% (CRU) in the 16 groups that were assessed (Fig. 4). The distribution of true-type trees by accession group was non-significant ($\chi^2 = 12.77$; df = 15; P = 0.62). Of the 401 samples from the ICG,T (Fields 4A, 5A, 5B, 6A and 6B), 158 samples were misidentified given an estimated 39.4% error rate. Approximately 34% of the Refractario accessions in these fields were misidentified.

Several mislabelled or non-reference trees were matched to their appropriate nomenclature or ancestry (Table 1). Amelonado ancestry was evident in many mislabelled accessions, particularly AM (16), CL (11) and MOQ (11) as shown in Supplementary Table S1 (available online only at http://journals.cambridge.org). Accessions with primarily Amelonado–Criollo ancestry included MXC 67 UWI Field 12 x3y6, PENTAGONA 1 Field 6B F491 T5, PENTAGONA 2 Field 6B F492 T8, RIM 113 Field 4A T2, RIM 117 Field 4A T1 and TRD 66 Field 4A A50 T1. The SPEC accessions (SPEC 138/11 Field 6B C141 T1, SPEC 184/2 Field 6B D194 T1 and SPEC 194/44 Field 6B D195 T2) were of IMC–SCA ancestry, except for the mislabelled SPEC 194/48 Field 6B D219 T9, which grouped with Amelonado accessions. The accession CLEM /S-62-1 Field 5B I745 T2 had contributions from the SCA, Refractario Cluster B and NA accession groups. Mixed ancestry was also present in FSC 13 Field 4A C321 T1 (IMC–Amelonado), H 1 (IMC–NA), ICS 39 Field 4A C305 T4 (IMC–Amelonado–Criollo), LCT EEN 162 /S-1010 Field 4A A60 T1 (NA–IMC–PA), MATINA 1/7 Field 6B D236 T12 (IMC–Criollo–Amelonado) and MATINA 1/7 Field 6B D236 T15 (French Guiana–NA). Further details on accession composition can be found in Supplementary Table S1 (available online only at http://journals.cambridge.org).

Discussion

The population structure of a subset of the ICG,T was documented in a previous study (Motilal *et al.*, 2011) which estimated that the collection contained on average 25% mixed plots. From the present study, a 39.4% misidentification rate was estimated. The estimate is in agreement with previous studies on this genebank which employed dominant markers (Christopher *et al.*, 1999; Sounigo *et al.*, 2001), or the same marker system but on only Upper Amazon Forastero accessions (Zhang *et al.*, 2009a). Aikpokpodion *et al.* (2010) determined a 46.4% error rate in a Nigerian field genebank. A prior conservative mislabelling estimate of 24.7% across international cacao genebanks (Motilal *et al.*, 2011) can

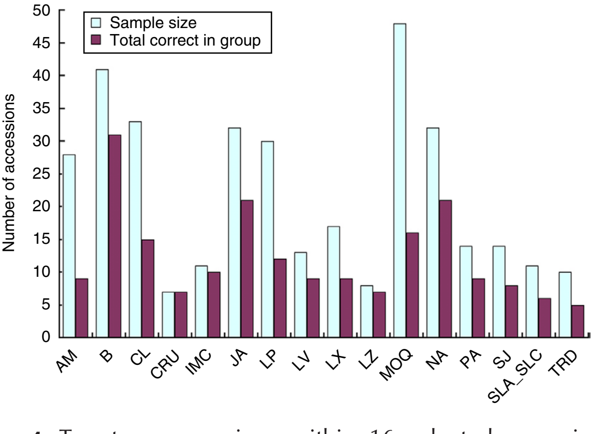

Fig. 4. True-type accessions within 16 selected accession groups in the International Cocoa Genebank, Trinidad. Distribution of true-type trees by accession group was non-significant ($\chi^2 = 12.77$; df = 15; $P = 0.62$).

Table 1. Matching of selected non-reference and mislabelled accessions

Sample ID[a]	Grouping[b]	Sample ID	Grouping
AMAZ 12 F6B B94 T2	IMC-NA	CRU 72 F6A A50 T1	IMC
AMAZ 15/15 F4A A101 T1	IMC-NA	CRU 94 F5B D278 T2	NA-IMC
AMAZ 32 F6B F433 T7	IMC-NA	CRUZ 7/8 F6B B83 T1,9	SCA-FGU
B 14/13 F5B A68 T5	SCA-AML	EET 48 F6B C126 T8	AML
CL 9/15 F5B A64 T7			
B 14/14 F5B A44 T11	SCA-AML	ICA 70 F4A C290 T2	AML
B 17/10 F5B G557 T1, 4	IMC-AML	ICS 3 F4A C288 T1	IMC-NA
B 18/5 F6B C152 T2	IMC (42)	ICS 53 F4A C279 T1	IMC-NA
B 7/38 F6B F441 T7, 9	NA (678)	ICS 56 F4A C285 T2	IMC-SCA
B 9/10–33 F5B I768 T1	PA (141)	IMC 47 F6B F401 T3	SCA-AML
BH UWI Field 7 x11 y2	IMC-SCA	JA 5/23 F5B G594 T1	IMC-SCA
CL 9/15 F5B A64 T4	IMC (60, 77)	LP 1/45 F5B B96 T2	NA
CL 19/10 F5B A69 T11	AML-SCA	MOQ 2/18 F5B C171 T10	NA-IMC
CRU 100 F5B G582 T2	NA	MOQ 6/109 F5B C209 T1, 2	PA
CRU 128 F5B G569 T1,2	SCA-IMC	MOQ 6/95 F5B C221 T3	PA
CRU 133 F5B D343 T3, 11	IMC-AML	NA 176 F5B E403 T1 & F4A D389 T1	PA (176)
CRU 47 F5B G621 T1 & F6A A44 T1	AML-IMC	NA 312 F5B G614 T1, 4	PA (312)

[a] F4A, F5B, F6A, F6B = Field 4A, 5B, 6A, 6B, respectively. [b] Accession group, AML = AMELONADO, putative accession match is given in parentheses.

be revised upwards to 29.8% mislabelling. The synonymous error rate was estimated here at 7.5% from 388 accessions, which was within the modelled synonymy estimate of 14.4% of 2000 accessions (Motilal *et al.*, 2011). Both true-type and off-type trees should be documented in the field with appropriate labels and CRU should add this information to its database. Off-type trees should be renamed and retained until all the trees in the genebank are fingerprinted. A decision to remove off-type trees can then be considered. Homonymous cases should be retained provided that they remain unique cases. Accessions arising out of homonymous identification and with a safety duplication of less than four trees should be clonally propagated and maintained in the field genebank. Synonymy will inflate the safety duplication level of some accessions while concomitantly decreasing the safety duplication level in other accessions. Removal of extraneous trees should only be undertaken if there is an excess of duplicated accessions. New unique accessions can then be introduced so that a greater number of accessions can be maintained on the same area of land.

The Refractario accessions were grouped into OB1 (B and SJ), OB2 (AM, CL, CLM and LP), OB3 (JA, LV, LX, LZ, SLA, SLC and SJ) and OB4 (MOQ) subclusters. Subclusters OB1 and OB3 formed a larger cluster as did OB2 with the OB4 cluster. The results obtained are in agreement with the Refractarios being derived from multiple closely related parents (Zhang *et al.*, 2008). Moreover, the grouping presented here suggested that the Refractarios had a narrower origin than was traditionally expected (Pound, 1938, 1943; Toxopeus, 1985; Bartley, 2005). Cacao breeders seeking to exploit the variability within Refractario are advised to select parents from different subclusters. The SLA and SLC accession groups were not separated from each other. These accessions were collected from trees A and C from the farm Santa Lucia (Bartley, 2000), which would be consistent with the SLA and SLC nomenclature present in the genebank. Full phenotypic evaluation of these two groups is recommended and if similar, they should be lumped into an SL accession group.

The approach to population clustering indicated that the method of inferring K, described in this paper, can adequately detect the true population structure when compared with that of Evanno *et al.* (2005). A low number (10) of iterations are usually employed (Kaeuffer *et al.*, 2007; Efombagn *et al.*, 2008; Motamayor *et al.*, 2008; Schmidt *et al.*, 2009; Zhang *et al.*, 2009a; Aradhya *et al.*, 2010). A larger number of iterations were employed in Aikpokpodion *et al.* (2010), Motilal *et al.* (2010) and the present study. This may be a better approach to obtaining a normal sample size of iterations but is hindered by the length of time required by the software, especially on larger datasets. Further, submitting all the runs to CLUMPP may result in biologically invalid results as evidenced by the hybrid Criollo nature at $K = 2$, 4 or 5 under the Large K Greedy algorithm. The use of selected consistent representative runs per required K is therefore supported (Zhang *et al.*, 2009a; Motilal *et al.*, 2010; Aikpokpodion *et al.*, 2010). Using separate runs that employed putative clusters to decide on subclustering (Pritchard *et al.*, 2000; Dawson and Belkhir, 2009) was a valuable corroborating tool. A methodological tool employed in the present study was the inclusion of known samples. Here, a known homozygous population (Criollo) was used to track the population structure. In addition, duplicated samples were used as independent unknowns and acted as spiked samples. These two inclusions advocated for consistency and biological interpretation of the population subdivision. The SCA and U accessions partitioned away from other accessions into the same subcluster. This agreed with the Contamana group of Motamayor *et al.* (2008) and their collection history (Bartley, 2005). The inferred population structure in the present study is therefore reliable. The close grouping of PA and French Guiana accessions is consistent with earlier researchers (Sounigo *et al.*, 2005) with the PA and French Guiana groups suggested to be derived from the human selection of Lower Amazon Forastero material (Bartley, 2005). The absence of French Guiana accessions in Zhang *et al.* (2009a) precluded a similar assessment but did indicate a Lower Amazon Forastero profile for the PA group. The results supported the proposition that attention should be paid to sample composition effects when inferring structure relationships (Motilal *et al.*, 2010).

The choice of K will influence the interpretation of the results. Criollo ancestry was highly influenced by K (Fig. 2). At a choice of $K = 3$ (method of Evanno *et al.* (2005)) or $K = 4$, the number of individuals with Criollo ancestry was probably overestimated. The fit of the genetic data to the finalized population structure should therefore be accepted only after probing for substructure in the entire dataset and in putative homogeneous clusters. This study has demonstrated that three accession groups (MXC, PENTAGONA and STAHEL) traditionally assigned to the Criollo group in the ICG,T must be reassigned to the Trinitario (MXC and PENTAGONA) and Forastero (STAHEL) groups. A similar result was found by Motilal *et al.* (2010).

In cacao field genebanks, an accession is a clone arising from budwood or seed that may then be vegetatively propagated to exist as a single tree or more than one tree. Users of a collection often assume that the multiple trees of an accession are indeed of the same genetic identity. However, this has been proven otherwise (Zhang *et al.*, 2009b; Irish *et al.*, 2010; Motilal *et al.*, 2011; and references therein). Determination of homonymies and synonymies is therefore useful in determining the proper accession nomenclature or accession group. We recommend that duplicated samples, appropriate reference samples and proper compilation of the STRUCTURE runs be used when elucidating population structure. Only then will the elucidation of identities become reliable to enable the adoption of correct management strategies in field genebanks.

Acknowledgements

Thanks to Ms Alisha Omar-Ali for assisting with DNA extractions and to Mr Stephen Pinney for his assistance with the electrophoresis work. Ms Zainab Ali and Mr Kasey Gordon are thanked for their assistance with data entry. Two anonymous reviewers are thanked for critiquing the manuscript. This research was made possible in part by a grant from the Government of Trinidad and Tobago Research Development Fund.

References

Aikpokpodion PO, Kolesnikova-Allen M, Adetmirin VO, Guiltinan MJ, Eskes AB, Motamayor JC and Schnell RC (2010) Population structure and molecular characterization of Nigerian field genebank collections of cacao, *Theobroma cacao* L. *Silvae Genetica* 59: 273–285.

Aradhya MK, Stover E, Velasco D and Koehmstedt A (2010) Genetic structure and differentiation in cultivated fig (*Ficus carica* L.). *Genetica* 138: 681–694.

Bartley B (2000) The nomenclature of the accessions derived from Dr. F.J. Pound's collection in Ecuador in 1937. *INGENIC Newsletter* 5: 4–6.

Bartley BGD (2005) *The Genetic Diversity of Cacao and its Utilization*. Wallingford: CABI Publishing, 341 pp.

Bekele F and Bekele I (1996) A sampling of the phenetic diversity of cacao in the International Cocoa Gene Bank of Trinidad. *Crop Science* 36: 57–64.

Bekele F, Butler DR, *et al.* (2000) Proposed short list of descriptors for characterization. In: Eskes AB (ed.) *Working Procedures for Cocoa Germplasm Evaluation and Selection: Proceedings of the CFC/ICCO/IPGRI Project Workshop,* 1–6 February 1998, Montpellier, France. Rome: IPGRI.

Bekele F, Bekele I, Butler DR and Bidaisee GG (2006) Patterns of morphological variation in a sample of cacao (*Theobroma cacao* L.) germplasm form the International Cocoa Genebank, Trinidad. *Genetic Resources and Crop Evolution* 53: 933–948.

Chang A (2001) Group differences program v.3.0. Available at http://department.obg.cuhk.edu.hk/researchsupport/download/downloads.asp

Christopher Y, Mooleedhar V, Bekele F and Hosein F (1999) Verification of accessions in the ICG,T using botanical descriptors and RAPD analysis. *Annual Report 1998*. St. Augustine: Cocoa Research Unit, The University of the West Indies, pp. 15–18.

Dawson KJ and Belkhir K (2009) An agglomerative hierarchical approach to visualization in Bayesian clustering problems. *Heredity* 103: 32–45.

Earl DA and von Holdt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method *Conservation Genetics Resources*. doi:10.1007/s12686-011–9548-7. Available at http://taylor0.biology.ucla.edu/structureHarvester/

Efombagn IBM, Motamayor JC, Sounigo O, Eskes AB, Nyassé S, Cilas C, Schnell R, Manzanares-Dauleux MJ and Kolesnikova-Allen M (2008) Genetic diversity and structure of farm and GenBank accessions of cacao (*Theobroma* *cacao* L.) in Cameroon revealed by microsatellite markers. *Tree Genetics & Genomes* 4: 821–823.

Engels JMM, Bartley BGD and Enriquez GA (1980) Cacao descriptors, their states and modus operandi. *Turrialba* (Costa Rica) 30: 209–218.

Evanno G, Regnaut S and Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14: 2611–2620.

Falush D, Stephens M and Pritchard JK (2003) Inference of population structure: extensions to linked loci and correlated allele frequencies. *Genetics* 164: 1567–1587.

Hurka H, Neuffer B and Friesen N (2004) Plant genetic resources in botanical gardens. In: Forkmann G and Michaelis S (eds) *Proceedings of the 21st International Symposium on Breeding Ornamentals, Part II. Acta Horticulturae* 651: 35–44.

Irish BM, Goenaga R, Zhang D, Schnell R, Brown JS and Motamayor JC (2010) Microsatellite fingerprinting of the USDA-ARS Tropical Agriculture Research Station cacao (*Theobroma cacao* L.) germplasm collection. *Crop Science* 50: 656–667. doi:10.2135/cropsci2009.06.0299

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

Jakobsson M and Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23: 1801–1806.

Johnson SE, Mora A and Schnell RJ (2007) Field guide efficacy in the identification of reallocated clonally propagated accessions of cacao. *Genetic Resources and Crop Evolution* 54: 1301–1313.

Kaeuffer R, Réale D, Coltman DW and Pontier D (2007) Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity* 99: 374–380. doi:10.1038/sj.hdy.6801010

Kalinowski ST, Taper ML and Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology* 16: 1099–1106. doi:10.1111/j.1365–294X.2007.03089.x

Kennedy AJ and Mooleedhar V (1993) Conservation of cocoa in field genebanks – the International Cocoa Genebank, Trinidad. pp. 21–23. *Proceedings of International Workshop on Conservation, Characterisation and Utilisation of Cocoa Genetic Resources in the 21st Century,* 13–17 September, 1992, Port of Spain, Trinidad. Port of Spain: Cocoa Research Unit, The University of the West Indies.

Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A and Lagoda PJL (1999) Isolation and characterization of microsatellites in *Theobroma cacao* L. *Molecular Ecology* 8: 2141–2143. doi:10.1046/j.1365–294x.1999.00802.x

Leão PCS, Riaz S, Graziani R, Dangl GS, Motoike SY and Walker MA (2009) Characterization of a Brazilian grape germplasm collection using microsatellite markers. *American Journal of Enology and Viticulture* 60: 517–524.

Motamayor JC, Lachneaud P, da Silva e Mota JW, Loor R, Kuhn DN, Brown JS and Schnell RJ (2008) Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L.). *PLoS ONE* 3: e3311. doi:10.1371/journal.pone.0003311

Motilal L and Butler D (2003) Verification of identities in global cacao germplasm collections. *Genetic Resources and Crop Evolution* 50: 799–807. doi:10.1023/A:1025950902827

Motilal LA, Zhang D, Umaharan P, Mischke S, Boccara M and Pinney S (2009) Increasing accuracy and throughput in large-scale microsatellite fingerprinting of cacao field germplasm collections. *Tropical Plant Biology* 2: 23–27. doi:10.1007/s12042-008-9016-z

Motilal LA, Zhang D, Umaharan P, Mischke S, Mooleedhar V and Meinhardt LW (2010) The relic Criollo cacao in Belize – genetic diversity and relationship with Trinitario and other cacao clones held in the International Cocoa Genebank Trinidad. *Plant Genetic Resources: Characterization and Utilization* 8: 106–115. doi:10.1017/S1479262109990232

Motilal LA, Zhang D, Umaharan P, Mischke S, Pinney S and Meinhardt LW (2011) Microsatellite fingerprinting in the International Cocoa Genebank Trinidad: accession and plot homogeneity information for germplasm management. *Plant Genetic Resources: Characterization and Utilization* 9: 430–438. doi:10.1017/S147926211100058X

Nei M (1972) Genetic distance between populations. *American Naturalist* 106: 283–392.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89: 583–590.

Onofri A (2007) Routine statistical analyses of field experiments by using an Excel extension In: *Proceedings 6th National Conference Italian Biometric Society: "La statistica nelle scienze della vita e dell'ambiente"*, 20–22 June 2007, Pisa, pp. 93–96, version 1.1 (update 18 March 2011). Available at http://www.unipg.it/∼onofri/DSAASTAT/DSAASTAT.htm

Park SDE (2001) Trypanotolerance in West African cattle and the population genetic effects of selection. PhD Thesis, University of Dublin.

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

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

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

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

Pugh T, Fouet O, Risterucci AM, Brottier P, Abouladze M, Deletrez C, Courtois B, Clement D, Larmande P, N'Goran JAK and Lanaud C (2004) A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. *Theoretical and Applied Genetics* 108: 1151–1161. doi:10.1007/s00122-003-1533-4

Saunders JA, Mischke S, Leamy EA and Hemeida AA (2004) Selection of international molecular standard for DNA fingerprinting of *Theobroma cacao*. *Theoretical and Applied Genetics* 110: 41–47. doi:10.1007/s00122-004-1762-1

Schmidt JI, Hundertmark KJ, Bowyer RT and McCraken KG (2009) Population structure and genetic diversity of moose in Alaska. *Journal of Heredity* 100: 170–180. doi:10.1093/jhered/esn076

Sounigo O, Christopher Y, Bekele F, Mooleedhar V and Hosein F (2001) The detection of mislabelled trees in the

International Cocoa Genebank, Trinidad (ICG,T). In: *Proceedings of the Third International Group for Genetic Improvement of Cocoa (INGENIC) International Workshop on the New Technologies and Cocoa Breeding,* 16–17. October 2000. Malaysia: Kota Kinabalu, pp. 34–39.

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

SPSS, Inc. (1986–2001) *SigmaPlot 2002 for Windows, version 8.02*. Chicago, IL: SPSS, Inc.

Toxopeus H (1985) Botany, types and populations. In: Wood GAR and Lass RA (eds) *Cocoa*. 4th edn. London: Longman Group Ltd, pp. 11–37.

Valière N (2002) GIMLET: a computer program for analyzing genetic individual identification data. *Molecular Ecology Notes* 2: 377–379. doi:10.1046/j.1471–8286.2002.00228.x-i2

van Treuren R, de Groot EC, Boukema IW, van de Wiel CCM and van Hintum ThJL (2010) Marker-assisted reduction of redundancy in a genebank collection of cultivated lettuce. *Plant Genetic Resources: Characterization and Utilization* 8: 95–105. doi:10.1017/S1479262109990220

Waits LP, Luikart G and Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology* 10: 249–256. doi:10.1046/j.1365–294X.2001.01185.x

Zhang D, Boccara M, Motilal L, Butler DR, Umaharan P, Mischke S and Meinhardt L (2008) Microsatellite variation and population structure in the "Refractario" cacao of Ecuador. *Conservation Genetics* 9: 327–337. doi:10.1007/s10592-007-9345-8

Zhang D, Boccara M, Motilal L, Mischke S, Johnson ES, Butler DR, Bailey B and Meinhardt L (2009*a*) Molecular characterization of an earliest cacao (*Theobroma cacao* L.) collection from Upper Amazon using microsatellite DNA markers. *Tree Genetics & Genomes* 5: 595–607. doi:10.1007/s11295-009-0212-2

Zhang D, Mischke S, Johnson ES, Phillips-Mora W and Meinhardt L (2009*b*) Molecular characterization of an international cacao collection using microsatellite markers. *Tree Genetics & Genomes* 5: 1–10. doi:10.1007/s11295-008-0163-z