Tree Genetics & Genomes (2013) 9:1395–1411
DOI 10.1007/s11295-013-0645-5

ORIGINAL PAPER

Microsatellite-aided detection of genetic redundancy improves management of the International Cocoa Genebank, Trinidad

Lambert A. Motilal · Dapeng Zhang · Sue Mischke · Lyndel W. Meinhardt · Pathmanathan Umaharan

Received: 6 June 2013 / Accepted: 25 June 2013 / Published online: 20 July 2013

Abstract Cacao (*Theobroma cacao* L.), the tree from which cocoa butter and chocolate is derived, is conserved in field genebanks. The largest of these ex situ collections in the public domain is the International Cocoa Genebank, Trinidad (ICG,T). Reduction of genetic redundancy is essential to improve the accuracy and efficiency of genebank management. This study examined the pedigree and genetic diversity in a subset of 387 accessions in this collection. Sibship reconstruction of this subset revealed 56 full-sib families nested within 189 half-sib families. Sixteen centers of interconnectivity were identified, which suggested a high level of genetic redundancy in the collection. Generally, consistent phylogenetic trees were obtained using different genetic distance measures. However, a principal coordinate analysis of the D_{est} differentiation measure elicited the best representation of accession group clustering, and we recommend this approach when probing fine-scale genetic differentiation among cacao accessions. The composite genetic diversity of 414 cacao accessions was contained in a core set of 59 unique accessions. These results have significant implications in the conservation of genetic resources of the ICG,T and other cacao genebanks. The approach developed in this study is recommended as a strategy to curators in guiding conservation management practices of cacao and other similar ex situ genebanks.

Keywords Cocoa · Conservation management · Sibship · Core collection · Field genebank · Redundancy

Communicated by D. Neale

Electronic supplementary material The online version of this article (doi:10.1007/s11295-013-0645-5) contains supplementary material, which is available to authorized users.

L. A. Motilal (✉) · P. Umaharan
Cocoa Research Centre, The University of the West Indies,
St. Augustine, Trinidad, Republic of Trinidad and Tobago
e-mail: lamotilal@yahoo.com

D. Zhang · S. Mischke · L. W. Meinhardt
Beltsville Agricultural Research Center, SPCL, USDA/ARS,
10300 Baltimore Avenue, Bldg. 001, Rm. 223, BARC-W,
Beltsville, MD 20705, USA

Introduction

Theobroma cacao L. (cacao) is an important tree crop indigenous to the upper regions of the Amazon River basin. The seeds of this understory tropical tree are the source of cocoa mass and cocoa butterfat, essential products in the multibillion-dollar candy industry. Due to its allogamous nature and recalcitrant seeds (Toxopeus 1985), cacao must be conserved in situ or ex situ as living trees. The International Cocoa Genebank, Trinidad (ICG,T), managed by the Cocoa Research Centre of The University of the West Indies, is the largest public domain ex situ field genebank (Motilal and Butler 2003). It contains germplasm from Amazonian South America, Central America, and the West Indies (Kennedy and Mooleedhar 1993). Details of the ICG,T can be found at http://sta.uwi.edu/cru and in Motilal et al. (2012). Cacao accessions can be grouped as Criollo, Forastero, Trinitario, and Refractario varieties (Toxopeus 1985) with the ICG,T containing primarily the latter three types. Accession nomenclature follows Turnbull and Hadley (2013), taking an alphanumeric form where the alpha code represents the accession group and the numeric code indicates the tree from which the fruit or budwood was collected. A description of the process of collecting the original germplasm can be found in Zhang et al. (2009a). The accession NA 702 is purportedly distinct from NA 312, but both belong to the same accession group (NA), members of which were collected along the Nanay River in the Amazon. In contrast, the accession group TRD represents an accession group purportedly distinct from NA and whose members (e.g., TRD 1, TRD 117) were collected from the island of Trinidad. The incompatibility system in cacao (Knight and Rogers 1953, 1955; Cope 1958, 1962) allows for cross-compatibility among and within accession groups, with the probability of inbreeding being increased for self-compatible trees. The ICG,T is a field collection of cacao trees existing as clonal replicates (rooted cuttings or grafted plants) from the original germplasm (seeds or budwood) that were collected. The ICG,T trees are therefore clonal copies of selected trees and seedling progenies, but the trees themselves are not matured seedlings. The ICG,T was planned with an internal safety duplication, with each accession being represented with a maximum of 16 copies in a plot.

It is expensive and difficult to maintain genetic resources as living plants in field genebanks. Thus, it is important to elucidate mislabelling and genetic redundancy so that a reduced number of plants need to be managed, under minimal acreage. The minimum number of accessions that can conserve the maximum genetic diversity is a strategic goal of curators of field genebanks. The creation of core collections is well documented (Frankel 1984; Brown 1989; Pessoa-Filho et al. 2010; Reeves et al. 2012; Odong et al. 2013). Although the majority of studies on core collections have been on annuals, the practice is especially amenable to perennials that are maintained as live plants in field genebanks (Escribano et al. 2008; Belaj et al. 2012). Core germplasm collections have been advocated for Chilean common bean (Mario et al. 2010), saffron (Fernández et al. 2011), apricot (Krishnan et al. 2012), and olive (Haouane et al. 2011) and described for cashew (Dhanaraj et al. 2002), cherimoya (Escribano et al. 2008), grape (Le Cunff et al. 2008), olive (Belaj et al. 2012), pear (Miranda et al. 2010), and Sea Island cotton (Mei et al. 2012) among others. However, at present, core collections have not been described for cacao. About 110 cacao accessions were compiled for favorable disease resistance, seed traits, isozyme diversity, and randomly amplified polymorphic DNA diversity by Sounigo et al. (2006) for international distribution. However, the genetic relation of this subset as compared to cacao germplasm collections worldwide or to the ICG,T was not established. Furthermore, since specific trees were not identified, the high degree of mislabelling in the ICG,T (Motilal et al. 2011) has compromised the value of this set of 110 accessions as a core collection.

Motamayor et al. (2008) elucidated ten genetic groups that encompass wild cacao germplasm at the time of the study. The Forastero germplasm formed a large component of their study, and types that are present in the ICG,T, such as Amelonado, Guiana, Nanay, Parinari, Scavina, and Iquitos Mixed Calabacillo represent six of Motamayor's ten proposed genetic group. However, the Refractario cacao was not covered in the study of Motamayor et al. (2008) since these accessions were selected from a screening program in Ecuador searching for plants resistant to witches' broom disease (Pound 1938, 1943; Bartley 2001; Zhang et al. 2009a). This accession group is a significant and valuable component of the ICG,T (Zhang et al. 2008), and the efficient management of the Refractario accessions is necessary.

Molecular markers have been extensively used in DNA fingerprinting of cacao accessions. Multilocus microsatellite profiles have been determined for each reference accession in the ICG,T (Boccara and Zhang 2006) using an internationally accepted set of 15 microsatellite loci (Saunders et al. 2004). Each accession in the ICG,T is represented as multiple, clonally propagated trees, usually growing in a single plot within one of the five existing fields. Recently, DNA fingerprinting revealed a high degree of mislabelled trees in the plots (Motilal and Butler 2003; Motilal et al. 2011). The elimination of mislabelled samples is an essential objective in genetic diversity studies of cacao germplasm collections (Motamayor et al. 2008; Zhang et al. 2008, 2009a, b; Aikpokpodion et al. 2009, 2010; Irish et al. 2010; Motilal et al. 2010) and is also needed for efficient management.

The preceding researchers have examined genetic redundancy classifying mislabelling as homonymies (accessions with the same name but different DNA profiles) or synonymies (accession with different names but the same DNA profiles). However, there is little research on sibship effects on the genetic diversity of cacao collections. The majority of the accessions of the ICG,T acquired prior to the 1970s, were derived from a limited number of fruits collected during various expeditions (Pound 1938, 1943; Bartley 2001, 2005; Sounigo et al. 2005; Zhang et al. 2009a) since only a few wild cacao trees typically have mature fruits at any specific time. Refractario accessions in particular were said to be derived from seedling progenies based on fruits collected from about 80 mother trees (Pound 1938). In addition, transportation difficulties over long distances contributed to the low survivability of collected cacao germplasm. Each fruit typically contains between 30 and 40 seeds with a limited number of alleles that are independently assorted among the full-sibs and half-sibs. These seeds were planted in quarantine facilities prior to introduction into the genebank, and each relocation and propagation of the collected germplasm increased the chances of mislabelling and proliferation of closely related siblings. Hidden sibship is, therefore, one of the main factors contributing to genetic redundancy in the ICG,T and other cacao collections. Multivariate analysis can reveal genetic diversity patterns in the collection but is unable to confirm the pedigree structure from historical records or uncover new pedigrees that can arise as a result of misidentification. Sibship information is an ideal tool for germplasm curators to gauge degrees of relatedness among accessions, thereby indicating over-representation in the collection. Furthermore, breeders can avoid inbreeding depression by using sibship information to plan mating designs. However, the approach of using pedigree construction to assess genetic redundancy in crop germplasm collections for management decisions was not found in the literature and has never been employed in cacao.

Prior analysis on the ICG,T (Motilal et al. 2012) indicated that the pedigree relationship should be explored further. The current work was therefore undertaken to (a) determine the sibship relationships and genetic diversity within a subsample of the ICG,T collection, (b) determine the redundancy in the collection, and (c) identify a core collection of germplasm in order to improve the management strategy of the ICG,T.

Materials and methods

Plant material

Healthy leaves (flush-mature) were collected from 414 genotypically distinct accessions (Online Resource 1). Of these, 387 accessions came from the ICG,T and represented 17 % of the accessions in the genebank. Accessions were drawn from traditionally assigned Criollo (five accessions), Forastero (72 accessions), Guiana (five accessions), Refractario (269 accessions), Trinitario (23 accessions), and unknown (13 accessions) groups. The remaining 27 accessions were a set of diverse samples that acted as reference accessions known to belong to standard assigned accession groups (Motamayor et al. 2008). An accession was usually represented by one tree. However, synonymies (trees of accessions with several or different names but having the same genotype) were collated and represented by one sample profile, whereas homonymies (trees of accessions with the same name but exhibiting different genotypes) were kept as discrete samples. Further description of sample composition, population structure, and subclustering can be found in Motilal et al. (2012). Refractario accessions were previously arrayed as two main clusters (O and B) and further subdivided into four (OB1–4) subclusters (Motilal et al. 2012).

Microsatellite amplification in cacao DNA

Total leaf genomic DNA was extracted as in Motilal et al. (2010) and amplified at 24 validated, reliable, and discriminatory microsatellite loci (Y16883, Y16980, Y16981, Y16982, Y16984, Y16985, Y16986, Y16988, Y16991, Y16995, Y16996, Y16998, AJ271822, AJ271826, AJ271942, AJ271943, AJ271944, AJ271945, AJ271956, AJ271958, AJ566564, AJ566565, AJ566593, AY389503) as in Motilal et al. (2009). These 24 loci covered all ten chromosomes in cacao (modal value of two loci per chromosome) and detailed characteristics of the microsatellite loci and alleles can be found online at www.ebi.ac.uk and in Lanaud et al. (1999), Pugh et al. (2004), and Saunders et al. (2004). The *Taq* polymerases employed were Eppendorf HotMasterMix (5 Prime Inc., Gaithersburg, MD, USA) or AmpliTaq Gold DNA polymerase (Applied Biosystems, CA, USA). Fragment lengths of amplified loci were sized on a CEQ 8000 or 8800 capillary electrophoresis system (Beckman Coulter Inc., CA, USA) using an internal 400-base-pair DNA size standard according to the manufacturer's instructions (Beckman Coulter Inc., CA, USA). A standard set of seven DNA samples was repeated as controls for each run. Fragment selection and binning were performed as described earlier (Motilal et al. 2009).

Sibship reconstruction

Sibships were reconstructed using 22 microsatellite loci with the most complete allelic data on the 414 accessions. Reliable sibship reconstruction requires determination and implementation of the locus genotyping error rate, attributed to allele dropout (allele present but failed to be amplified during polymerase chain reaction (PCR)) and to false alleles (allele is placed in wrongly sized bin due to mistakes in PCR, electrophoresis, or human error). The error rates were estimated by a maximum likelihood function with PEDANT v1.0 (Johnson and Haydon 2007) on 34 samples run as independent duplicates using the expected heterozygosity (H_e) values calculated with GenAlEx v6.1 (Peakall and Smouse 2006). The mean allele dropout rate was calculated across the loci with non-zero values. The allele dropout rate was set to that obtained from the genotyping error estimate, except for loci with zero values. In these cases, the allele dropout rate was set as the mean allele dropout rate. The false allele rate was zero for 19 loci and was thus set as half of the mean allele dropout rate across the 22 loci.

Simulations were run in Colony (Wang 2004; Wang and Santure 2009) using the cited estimates of typing error under an outbreeding scenario. Three independent medium runs (medium precision, full likelihood analysis, unknown polygamous male and female parents) were conducted. Consistent accessions across runs were identified. An exclusion criterion of 80 % probability was further employed to retain final sibship results. Sibship relationships were diagrammed for linked samples with consensual half-sibs.

Genetic relationships among individual accessions

Pairwise estimates of Nei's genetic distance (Nei et al. 1983) among the individual accessions with 999 bootstrap resamplings across the 24 microsatellite loci were calculated with MSA v. 4.05 (Microsatellite Analyser; Dieringer and Schlötterer 2002). Phylogenetic trees were generated from the resulting distance matrix using the neighbor-joining algorithm (Saitou and Nei 1987) available in PHYLIP (Felsenstein 1989). A consensus tree (50 % majority rule) based on the bootstrapped trees was created using the

CONSENSE procedure implemented in PHYLIP (Felsenstein 1989). The dendrogram was visualized with FigTree Version 1.3.1 (Rambaut 2006–2009).

Genetic relationships among germplasm groups

Homogenous population groups (sample size of at least nine individuals) without admixed individuals were compiled (Motilal et al. 2012) from STRUCTURE v2.3 runs using independent and correlated models (Pritchard et al. 2000; Falush et al. 2003). Sibship clusters with at least ten individuals were identified from the COLONY runs (Wang 2004; Wang and Santure 2009) and prepared as a separate data file. Population diversity of these homogenous groups and sibship clusters was assessed by calculating R_{ST} (Slatkin 1995) in GenAlEx v6.1 (Peakall and Smouse 2006). Two other population measures, the estimator of actual differentiation (D_{est}; Jost 2008) and the G'$_{ST}$ measure of Hedrick (2005), were also

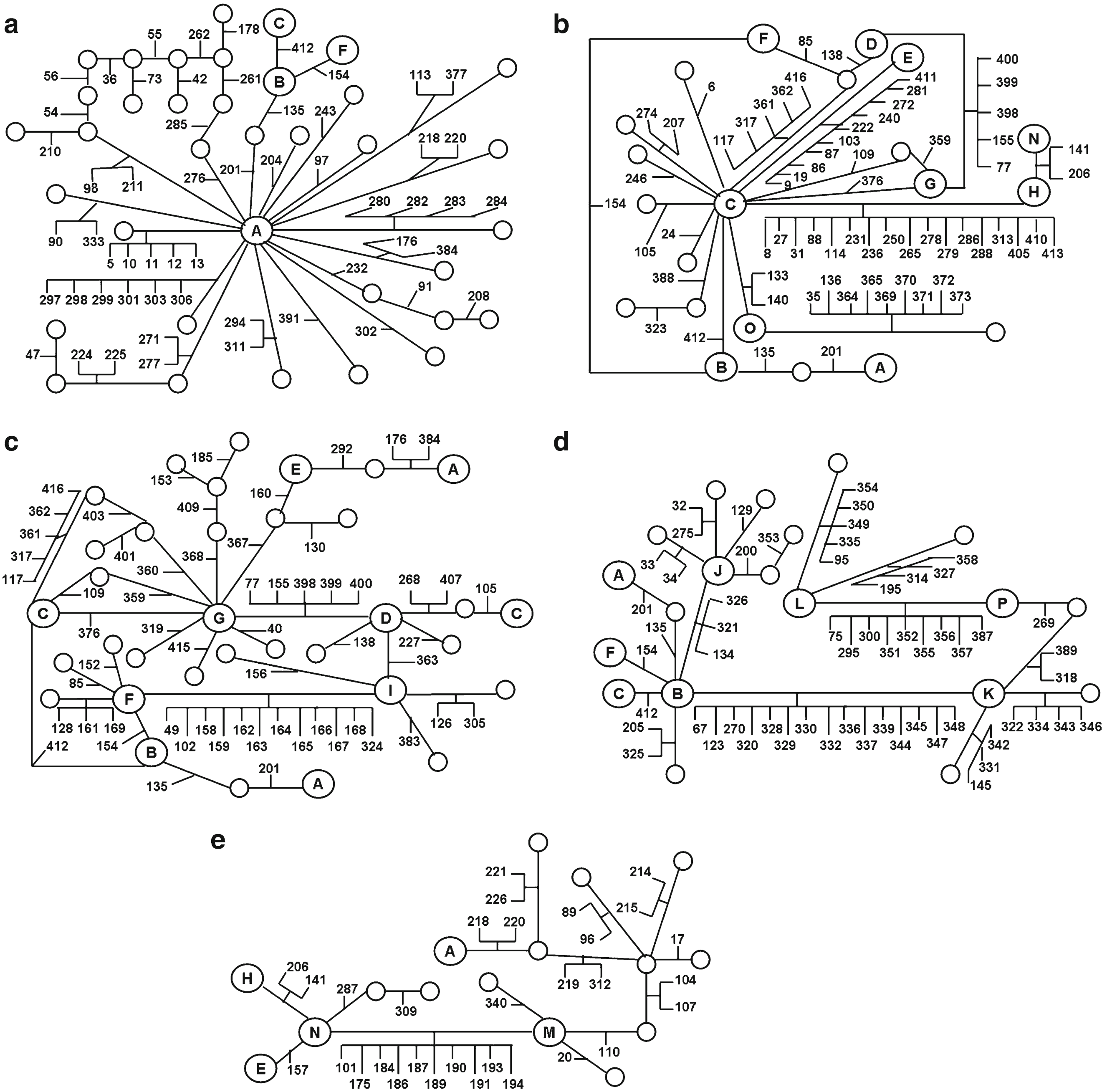

Fig. 1 Consistent sibship linkages (clusters **a**–**e**) detected in 414 cacao accessions. Three independent colony simulations (Wang 2004; Wang and Santure 2009) and a minimum threshold of 0.80 were employed. Putative parents with high connectivity are *encircled* A–P. *Sample numbers* correspond to individual accessions in Online Resource 1

determined with the online program SMOGD (Crawford 2010). The D_{est} measure is reportedly the most accurate measure of population differentiation (Jost 2008, 2009). The matrices of R_{ST} and D_{est} for the homogenous groups and sibship clusters were taken into GenAlEx v6.1 (Peakall and Smouse 2006), and principal coordinate analysis (PCoA) plots were constructed with standardized distances.

Gene diversity, heterozygosity, polymorphism information content (PIC; Botstein et al. 1980), and the inbreeding coefficient with 999 bootstraps via a moment estimator across loci were determined using PowerMarker v3.25 (Liu and Muse 2005) for the homogenous cacao groups. Pairwise distances among all possible pairs of groups were also determined for five measures of genetic differentiation in the same program. Geometric measures were the chord distance (D_{C}) of Cavalli-Sforza and Edwards (1967) and the D_{A} distance of Nei et al. (1983). Evolutionary models employed the θ_{W} distance of Reynolds et al. (1983), the $(\delta\mu)^2$ distance from the stepwise mutation model for microsatellites (Goldstein et al. 1995), and the shared allele distance (D_{SA}) from an infinite allele model for microsatellites (Jin and Chakraborty 1993). Tree files from 999 bootstrappings in the neighbor-joining algorithm (Saitou and Nei 1987) for each distance measure were created in PowerMarker v3.25 (Liu and Muse 2005). The consensus tree

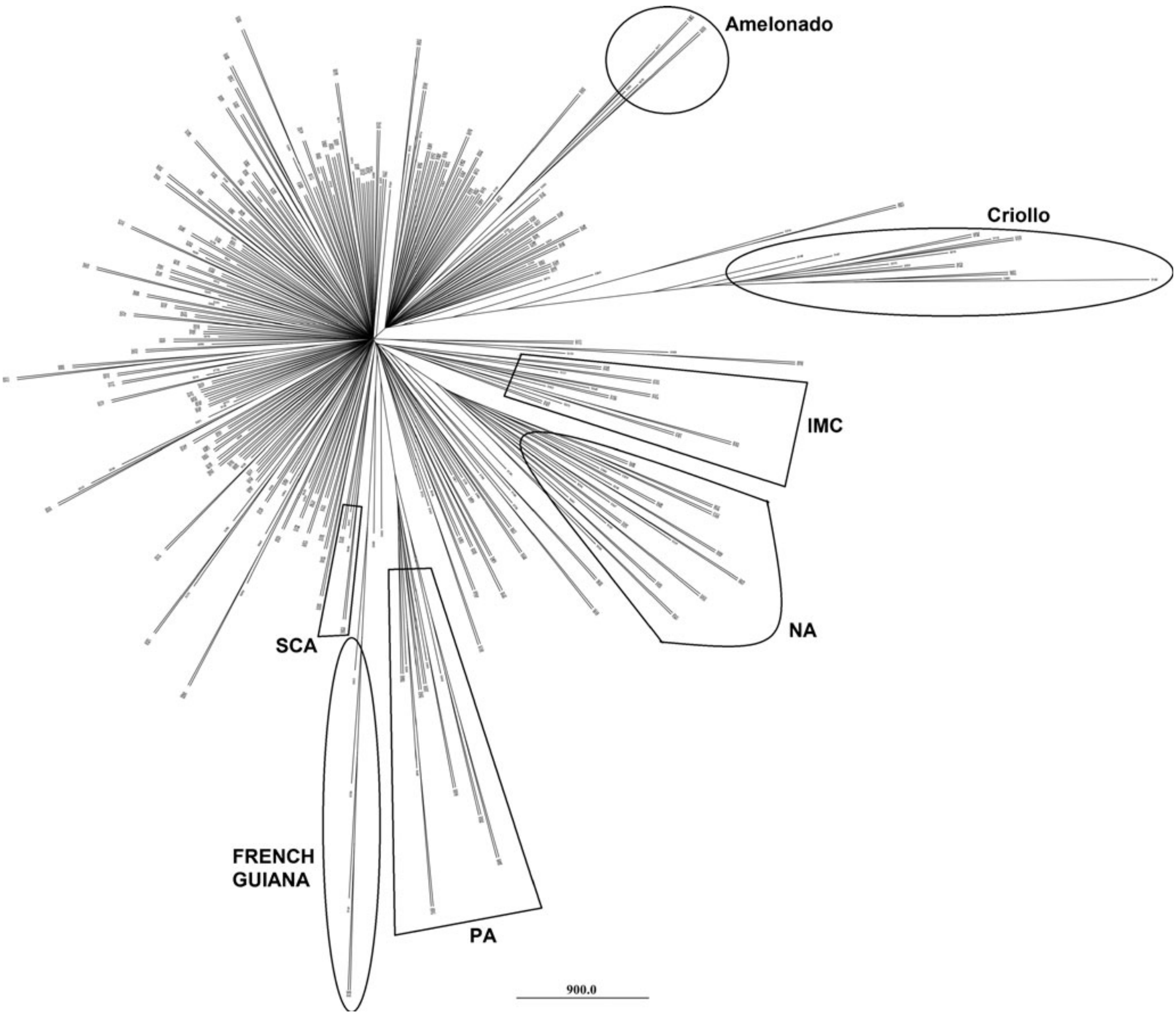

Fig. 2 Consensus radial tree diagram of 414 cacao accessions. Refractario accessions were primarily in the *left semicircle*. Pairwise estimates of Nei's genetic distance (Nei et al. 1983) with 999 bootstraps in MSA v. 4.05 (Dieringer and Schlötterer 2002) were obtained from 24 microsatellite loci. Trees were generated with the neighbor-joining algorithm (Saitou and Nei 1987), and the consensus tree (50 % majority rule) was created with the CONSENSE procedure in PHYLIP (Felsenstein 1989). The figure was visualized with FigTree Version 1.3.1 (Rambaut 2006–2009)

was created as before, and radial topologies were manipulated in TreeDyn (Chevenet et al. 2006).

Core germplasm identification

Heuristic and non-heuristic algorithms were independently employed in PowerCore v1.0 (National Institute of Agricultural Biotechnology 2006) to determine a core set of accessions based on genetic variation as detected by the 24 microsatellite loci. Summary statistics as number of effective alleles, Shannon's Information Index (Lewontin 1972), observed heterozygosity, expected heterozygosity, and PIC (Botstein et al. 1980) of the core and the full datasets were calculated in GenAlEx v6.1 (Peakall and Smouse 2006) and CERVUS v3.0.3 (Kalinowski et al. 2007). Summary values were subjected to analysis of variance in the group differences programs v3.0 of Chang (2001).

Results

Sibship

Nine isolated families were identified and comprised seven Refractario groups, the Criollo, and the French Guiana accessions. In addition, 56 full-sib families were nested in 189 half-sib families. Each of the full-sib families contained two to 17 individuals. The interlinked families with consistent members are provided in Fig. 1, and each sample is coded with a number provided in Online Resource 1. Refractario accessions were predominantly grouped in cluster A (AM, B, CL, CLM, JA, LP, LV, LX, LZ, MOQ, SLA, and SJ accessions) and cluster E (AM, CL, JA, and LP accessions). Amelonado, MXC, and PENTAGONA accessions were in cluster B, IMC and SCA accessions were in cluster C, and cluster D contained AMAZ, NA, and PA accessions. Thirty-

Table 1 Estimated population diversity parameters of G'_{ST_est} (Hedrick 2005) and D_{est} (Jost 2008) on 16 selected accession clusters and ten homogenous groups of *T. cacao* L. Diversity parameters estimated in SMOGD (Crawford 2010)

Microsatellite locus name		Sibship cluster				Homogenous group			
GenBank/EMBL	Locus ID	G'_{ST_est}	G'_{ST_est} 95 % CI	D_{est}	D_{est} 95 % CI	G'_{ST_est}	G'_{ST_est} 95 % CI	D_{est}	D_{est} 95 % CI
Y16883	mTcCIR001	0.346	0.295–0.462	0.144	0.118–0.198	0.617	0.586–0.668	0.248	0.233–0.272
Y16980	mTcCIR006	0.711	0.696–0.761	0.570	0.549–0.618	0.826	0.806–0.852	0.675	0.653–0.705
Y16981	mTcCIR007	0.617	0.602–0.675	0.421	0.406–0.462	0.781	0.759–0.813	0.583	0.563–0.609
Y16982	mTcCIR008	0.594	0.561–0.672	0.424	0.395–0.484	0.780	0.745–0.822	0.612	0.577–0.653
Y16984	mTcCIR010	0.670	0.655–0.721	0.522	0.509–0.566	0.795	0.776–0.823	0.618	0.600–0.643
Y16985	mTcCIR011	0.632	0.598–0.704	0.481	0.448–0.548	0.777	0.749–0.815	0.623	0.592–0.662
Y16986	mTcCIR012	0.653	0.631–0.714	0.509	0.486–0.562	0.825	0.809–0.856	0.688	0.667–0.722
Y16988	mTcCIR015	0.794	0.779–0.833	0.704	0.688–0.742	0.830	0.809–0.859	0.717	0.695–0.745
Y16991	mTcCIR018	0.754	0.743–0.792	0.595	0.582–0.629	0.837	0.823–0.857	0.658	0.641–0.681
Y16995	mTcCIR022	0.430	0.399–0.505	0.181	0.158–0.225	0.760	0.733–0.794	0.457	0.431–0.486
Y16996	mTcCIR024	0.295	0.279–0.378	0.138	0.123–0.189	0.654	0.633–0.691	0.413	0.389–0.446
Y16998	mTcCIR026	0.451	0.436–0.551	0.320	0.303–0.397	0.668	0.640–0.718	0.499	0.472–0.542
AJ271822	mTcCIR029	0.664	0.652–0.718	0.511	0.497–0.556	0.800	0.775–0.841	0.651	0.625–0.690
AJ271826	mTcCIR033	0.695	0.684–0.752	0.592	0.577–0.648	0.786	0.767–0.823	0.672	0.651–0.710
AJ271942	mTcCIR037	0.689	0.677–0.750	0.590	0.575–0.648	0.735	0.719–0.777	0.609	0.592–0.649
AJ271943	mTcCIR040	0.643	0.623–0.700	0.470	0.446–0.520	0.795	0.775–0.825	0.585	0.564–0.614
AJ271944	mTcCIR042	0.677	0.668–0.737	0.575	0.563–0.629	0.747	0.734–0.786	0.632	0.616–0.671
AJ271945	mTcCIR043	0.792	0.777–0.831	0.662	0.642–0.701	0.829	0.801–0.865	0.662	0.629–0.699
AJ271956	mTcCIR057	0.687	0.661–0.736	0.499	0.473–0.543	0.839	0.818–0.865	0.653	0.631–0.678
AJ271958	mTcCIR060	0.694	0.689–0.744	0.571	0.561–0.619	0.706	0.689–0.747	0.524	0.506–0.561
AJ566512	mTcCIR184	0.717	0.697–0.766	0.549	0.528–0.592	nd	nd	nd	nd
AJ566564	mTcCIR243	0.823	0.805–0.854	0.635	0.620–0.661	0.829	0.795–0.866	0.616	0.583–0.651
AJ566565	mTcCIR244	0.787	0.774–0.824	0.659	0.643–0.696	0.839	0.820–0.866	0.716	0.693–0.747
AJ566593	mTcCIR274	0.493	0.464–0.584	0.348	0.318–0.425	0.733	0.700–0.783	0.554	0.512–0.610
AY389503	SHRSTc016	0.633	0.598–0.710	0.407	0.372–0.471	0.671	0.613–0.753	0.331	0.278–0.402
Harmonic mean ($\tilde{N}$) of loci		0.598	0.572–0.674	0.437	0.366–0.464	0.763	0.738–0.803	0.560	0.519–0.587

CI confidence interval, *nd* not determined (removed due to missing data) nine samples, predominantly from the Refractario grouping, lacked sibship linkages.

Genetic relationships among individual accessions

The Refractario accessions were grouped together as a single cluster that was separate from the reference sets of Amelonado, Criollo, French Guiana, IMC, NA, PA, and SCA accessions (Fig. 2). The NA and IMC accessions were in close proximity to each other. The AMAZ, BH, CRU, FSC, H, and SPEC accessions were found between the NA and PA clusters. Accessions falling out of their purported groups were mainly aligned with the Amelonado accessions. The Amelonado-type accessions were adjacent to the Criollo accessions and appeared to be partitioned into three subclusters.

Genetic relationships among germplasm groups

Sixteen clusters (18–86 individuals) with the highest connectivity from the sibship determination were selected (parents A through P in Fig. 1). The overall harmonic means for the measures of Hedrick (2005; G'_{ST}) and Jost (2008; D_{est}) were 0.598 and 0.437, respectively, indicating moderate differentiation among the clusters (Table 1). Analysis of molecular variance showed that the clusters accounted for 6 % of the total molecular variance, while 82 % of the variation existed within the clusters (Online Resource 2). Significant differences ($P<0.01$) were detected at all three hierarchical levels. However, a low ($R_{ST}=0.056$) level of differentiation was present among clusters, with most of the genetic differentiation attributed to the individual accessions. Pairwise permutations of R_{ST} over clusters revealed that 40.0 % of the pairs were non-significant ($P>0.05$; Table 2). The PCoA plots from R_{ST} and D_{est} of sibship clusters were dissimilar from each other (Fig. 3). The plot from R_{ST} pairwise matrix partitioned 76.7 % of the variation on the first two axes, whereas that from D_{est} had 83.5 % of the variation explained by the first two axes.

Ten homogenous groups (Amelonado1, Amelonado2, Criollo, IMC, OB1, OB2, OB3, OB4, NA, and PA) were identified. Positive inbreeding coefficients across all 24 loci were obtained when the 414 accessions were considered (Table 3). However, positive inbreeding coefficients were only present in the Amelonado1, Criollo, NA, and OB4 groups (Table 3). The lowest PIC values were observed in Y16883 and AY389503 (~0.3), whereas the highest PIC value (0.76) was present in Y16988 (Table 3). Analysis of molecular variance demonstrated that the molecular variation was distributed primarily among the individuals (88 %) rather than among the groups (4 %) (Online Resource 3). Even though the population groups were generally not well differentiated from each other ($R_{ST}=0.043$), significant differences ($P<0.01$) were detected at all three hierarchical levels. Pairwise permutations over groups revealed that 35.6 % of the comparisons were non-significant ($P>0.05$; Table 4). The Criollo group was significantly ($P=0.001$)

Table 2 Genetic differentiation coefficient (R_{ST}; Slatkin (1995)) values (below diagonal) for selected clusters in cacao accessions of the International Cocoa Genebank, Trinidad. R_{ST} values are from 999 pairwise permutations in GenAlEx v6.1 (Peakall and Smouse 2006). Probability values are indicated above diagonal. Codes represent clusters of individuals found one link away from putative parents (A–P) as shown in Fig. 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	–	0.001	0.001	0.331	0.001	0.047	0.035	0.001	0.391	0.153	0.001	0.001	0.437	0.034	0.001	0.061
B	0.093	–	0.001	0.053	0.001	0.012	0.002	0.001	0.041	0.346	0.440	0.001	0.002	0.001	0.001	0.013
C	0.115	0.154	–	0.009	0.385	0.001	0.001	0.406	0.001	0.004	0.001	0.001	0.031	0.383	0.011	0.003
D	0.003	0.056	0.075	–	0.001	0.409	0.456	0.001	0.447	0.042	0.288	0.417	0.243	0.181	0.001	0.309
E	0.114	0.16	0.000	0.193	–	0.004	0.001	0.454	0.004	0.001	0.001	0.005	0.135	0.455	0.018	0.001
F	0.020	0.042	0.108	0.000	0.053	–	0.332	0.001	0.417	0.436	0.051	0.004	0.469	0.454	0.001	0.219
G	0.046	0.099	0.18	0.014	0.161	0.007	–	0.001	0.480	0.135	0.053	0.011	0.084	0.002	0.001	0.130
H	0.127	0.17	0.000	0.099	0.000	0.077	0.159	–	0.001	0.001	0.001	0.006	0.049	0.428	0.236	0.002
I	0.001	0.031	0.090	0.000	0.044	0.000	0.000	0.068	–	0.459	0.137	0.013	0.476	0.492	0.001	0.469
J	0.022	0.000	0.110	0.063	0.148	0.000	0.047	0.117	0.000	–	0.373	0.314	0.269	0.319	0.001	0.203
K	0.074	0.000	0.132	0.001	0.097	0.021	0.032	0.120	0.0104	0.000	–	0.001	0.018	0.001	0.001	0.202
L	0.064	0.063	0.070	0.000	0.058	0.044	0.057	0.046	0.030	0.008	0.000	–	0.047	0.055	0.001	0.448
M	0.000	0.087	0.034	0.022	0.030	0.000	0.040	0.032	0.000	0.013	0.043	0.037	–	0.397	0.001	0.040
N	0.037	0.085	0.000	0.023	0.000	0.000	0.072	0.000	0.000	0.008	0.051	0.028	0.000	–	0.200	0.047
O	0.211	0.217	0.066	0.206	0.061	0.082	0.143	0.013	0.085	0.158	0.124	0.088	0.092	0.018	–	0.001
P	0.035	0.060	0.102	0.026	0.142	0.015	0.043	0.104	0.000	0.029	0.014	0.000	0.054	0.039	0.164	–

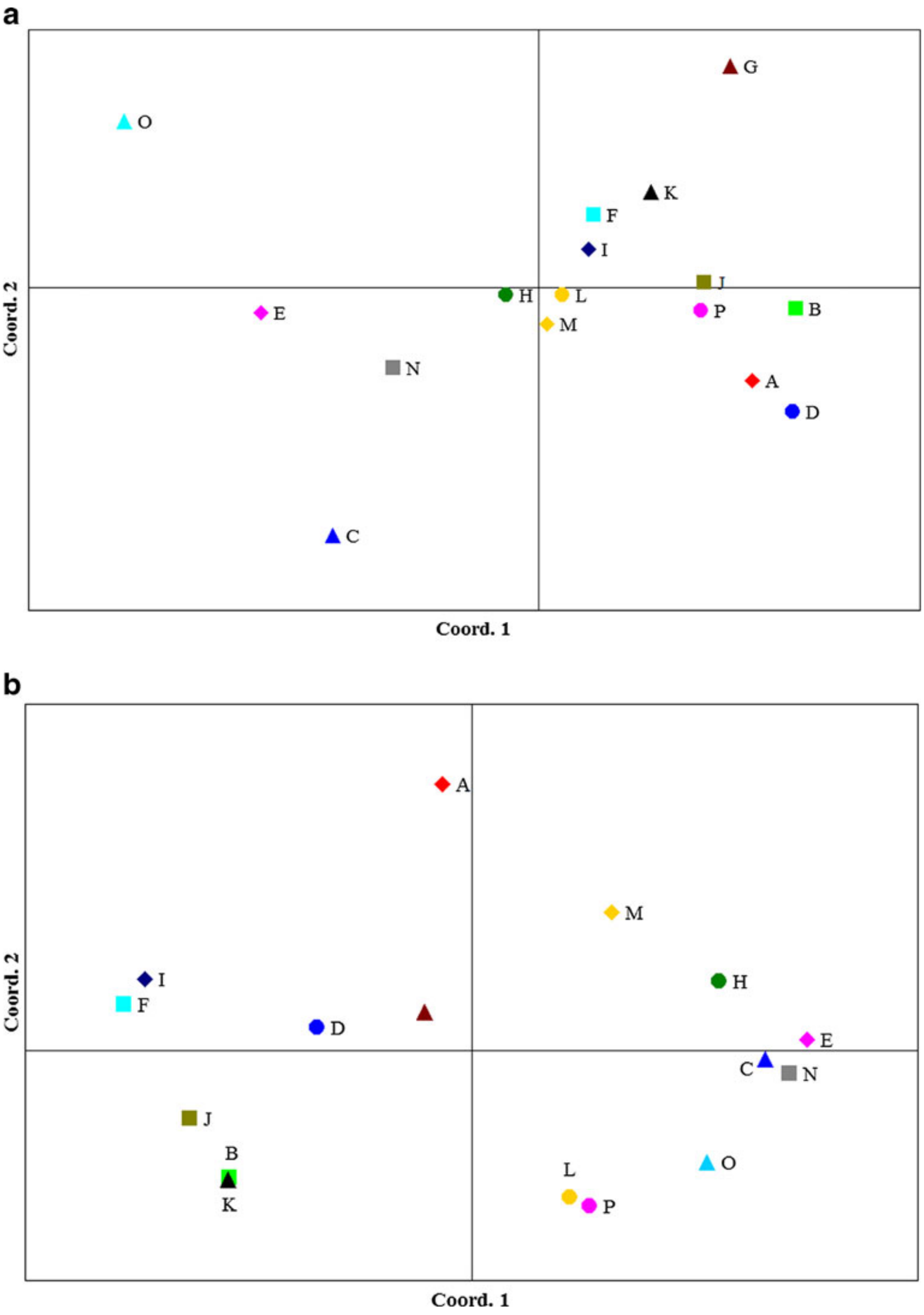

Fig. 3 Principal coordinate analyses from **a** R_{ST} (Slatkin 1995) and **b** harmonic mean of D_{est} (Jost 2008) pairwise matrices for sibship clusters. R_{ST} and D_{est} were determined in GenAlEx v6.1 (Peakall and Smouse 2006) and SMOGD (Crawford 2010). The first two axes in each plot explained 76.4 and 83.6 % of the variation, respectively. Clusters *A–P* are composed of individuals found one link away from the encircled putative parents A–P in Fig. 1

different from the other groups but was genetically closer to the OB1 (R_{ST}=0.064) and OB4 (R_{ST}=0.064) Refractario subgroups. Among the ten groups, the genetic differentiation over loci according to the measures of Hedrick (2005) or Jost (2008) ranged from 0.617 (Y16883) to 0.839 (AJ271956, AJ566565) or 0.248 (Y16883) to 0.717 (Y16988), respectively (Table 1). The overall harmonic mean of D_{est} (0.56) indicated moderate differentiation among the genetic groups. Multivariate analysis of homogenous groups partitioned 96.4 % (R_{ST} matrix; Fig. 4a) or 72.7 % (harmonic mean D_{est}; Fig. 4b) of the molecular variation between the first two axes. Much like the sibship clusters, the PCoA plots from R_{ST} and harmonic mean of D_{est} were dissimilar from each other.

Tree diagrams based on these 24 loci (Fig. 5) consistently located the Amelonado and Criollo groups between the Refractario and Upper Amazon Forastero groups. The NA and IMC groups were consistently closer to each other than PA except in the $(\delta\mu)^2$ distance of Goldstein et al. (1995).

Table 3 Summary statistics of allelic diversity over 24 microsatellite loci in ten homogenous groups in cacao. Estimates were derived in PowerMarker v3.25 (Liu and Muse 2005)

Assessment		Parameters				
EMBL code of locus	Locus	N_a	H_e	H_o	PIC	f
Y16883	mTcCIR001	5	0.3713	0.2471	0.3130	0.3361
Y16980	mTcCIR006	10	0.7396	0.4674	0.7008	0.3696
Y16981	mTcCIR007	5	0.6773	0.4504	0.6251	0.3368
Y16982	mTcCIR008	8	0.7233	0.4695	0.6893	0.3526
Y16984	mTcCIR010	8	0.6959	0.4800	0.6431	0.3120
Y16985	mTcCIR011	9	0.7314	0.4809	0.6865	0.3441
Y16986	mTcCIR012	11	0.7437	0.4943	0.7081	0.3371
Y16988	mTcCIR015	11	0.7895	0.5709	0.7584	0.2787
Y16991	mTcCIR018	9	0.6916	0.4731	0.6467	0.3177
Y16995	mTcCIR022	6	0.5434	0.3027	0.4810	0.4446
Y16996	mTcCIR024	6	0.5695	0.4269	0.5297	0.2522
Y16998	mTcCIR026	9	0.6898	0.5076	0.6394	0.2659
AJ271822	mTcCIR029	6	0.7434	0.4923	0.7070	0.3395
AJ271826	mTcCIR033	10	0.7692	0.5458	0.7369	0.2922
AJ271942	mTcCIR037	14	0.7619	0.4733	0.7249	0.3804
AJ271943	mTcCIR040	9	0.6705	0.4521	0.6183	0.3274
AJ271944	mTcCIR042	12	0.7604	0.5840	0.7257	0.2338
AJ271945	mTcCIR043	9	0.6617	0.3837	0.6148	0.4217
AJ271956	mTcCIR057	6	0.7126	0.4695	0.6638	0.3429
AJ271958	mTcCIR060	9	0.6640	0.4601	0.6141	0.3088
AJ566564	mTcCIR243	6	0.6602	0.3840	0.5956	0.4199
AJ566565	mTcCIR244	11	0.7671	0.5058	0.7419	0.3423
AJ566593	mTcCIR274	10	0.6946	0.4643	0.6615	0.3334
AY389503	SHRSTc016	5	0.3391	0.1216	0.3162	0.6427
	Mean	8.5	0.6738	0.4461	0.6309	0.3396
Traditional group	Group[a]					
LAF (9)	Amelonado1	1.5	0.1433	0.1065	0.1227	0.3109
LAF (61)	Amelonado2	2.9	0.3754	0.3857	0.3074	−0.0191
Criollo (16)	Criollo	1.5	0.0683	0.0240	0.0590	0.6680
UAF (14)	IMC	3.1	0.4802	0.6307	0.4223	−0.2794
UAF (27)	NA	3.9	0.4043	0.3538	0.3610	0.1436
Refractario (25)	OB1	3.0	0.5076	0.5663	0.4333	−0.0951
Refractario (34)	OB2	3.0	0.5210	0.5863	0.4432	−0.1105
Refractario (38)	OB3	2.8	0.4824	0.5886	0.3965	−0.2072
Refractario (22)	OB4	3.4	0.3784	0.3824	0.3428	0.0130
UAF (17)	PA	3.9	0.5130	0.5516	0.4692	−0.0450

Numbers in parentheses are sample sizes

N_a number of alleles, H_e gene diversity, H_o observed heterozygosity, *PIC* polymorphism information content (Botstein et al. 1980), f inbreeding coefficient; *LAF* Lower Amazon Forastero, *UAF* Upper Amazon Forastero, *IMC* Iquitos Mixed Calabacillo, *NA* Nanay, *OB1–4* Refractario subclusters, *PA* Parinari

[a] Homogenous cocoa population groups

Core germplasm

A reduced (85.8 %) congruent subsample of 59 accessions was selected from the 414 accessions in PowerCore v1.0 (National Institute of Agricultural Biotechnology 2006), and the same accessions were selected under either a heuristic or non-heuristic search (Table 5). A reduction of 86.9 % of the 387 accessions from the ICG,T was realized. Eight of the core accessions (AMELONADO 15, SIC 256 CC 7, HON 10, HON 11, ST 4/1, H 1, and U 1) were not part of the ICG,T holdings. Fifteen accessions (25.4 %) of the core set were mislabelled trees, having an ancestral contribution different from that expected from their accession name. Thirteen accessions (22.0 %) of the core set were of Refractario origin. This represented a reduction of 95.4 and 93.7 % of the original and assigned Refractario accessions. Forty-one of the core accessions were drawn from 22 full-sib families. Comparison of the 414 accessions with the core 59 accessions yielded R_{ST} and D_{est} measures of −0.004 (P=1) and 0.071, respectively. The number of effective alleles, observed

Table 4 Genetic differentiation coefficient (R_{ST}; Slatkin (1995)) values (below diagonal) for selected homogenous groups of cacao accessions from the International Cocoa Genebank, Trinidad. R_{ST} values are from 999 pairwise permutations in GenAlEx v6.1 (Peakall and Smouse 2006). Probability values are indicated above diagonal

	PA	OB1	OB3	OB2	OB4	NA	AML1	AML2	CRI	IMC
PA	–	0.326	0.005	0.056	0.001	0.044	0.001	0.008	0.001	0.166
OB1	0.004	–	0.026	0.427	0.158	0.196	0.062	0.004	0.001	0.402
OB3	0.040	0.018	–	0.008	0.063	0.001	0.335	0.002	0.001	0.083
OB2	0.019	0.000	0.019	–	0.165	0.004	0.006	0.001	0.001	0.430
OB4	0.066	0.011	0.018	0.008	–	0.001	0.002	0.001	0.001	0.291
NA	0.025	0.008	0.045	0.029	0.040	–	0.001	0.001	0.001	0.442
AML1	0.237	0.037	0.002	0.084	0.143	0.151	–	0.351	0.001	0.001
AML2	0.036	0.032	0.016	0.059	0.085	0.066	0.000	–	0.001	0.013
CRI	0.201	0.064	0.117	0.103	0.064	0.119	0.430	0.127	–	0.001
IMC	0.015	0.000	0.017	0.000	0.005	0.000	0.117	0.040	0.072	–

AML1 Amelonado 1 (nine samples), *AML2* Amelonado 2 (61 samples), *CRI* Criollo (16 samples), *IMC* Iquitos Mixed Calabacillo (14 samples), *NA* Nanay (27 samples), *OB1–4* Refractario subclusters (25, 34, 38, and 22 samples, respectively), *PA* Parinari (17 samples)

heterozygosity, Shannon's Information Index, and PIC were significantly different between these two sets (Table 6). However, private alleles were absent when the core set was compared to the full set of 414 accessions.

Discussion

The ICG,T subsample in this study contained a high degree of redundancy. The number of full-sib families was high (56), and a substantial 189 half-sib families were detected. A high redundancy was also indicated for sorghum, with highly genetically related accessions appearing after a random size of 90 accessions was exceeded (Cuevas and Prom 2013). A high level of redundancy was also detected in grape germplasm (Le Cunff et al. 2008). In contrast, a study of 37 peach accessions identified only two full-sib families with a total of five accessions (Viji et al. 2012).

The present study confirmed a high degree of relatedness within the ICG,T (Zhang et al. 2008, 2009a), which is not unexpected as many accessions in the ICG,T were derived from seeds of fruits from single trees or multiple trees (Pound 1938, 1943; Bartley 2001, 2005; Sounigo et al. 2005; Zhang et al. 2009a). The other universal cacao collection in Costa Rica (Centro Agronómico Tropical de Investigacíon y Enseñanza) also contained a high redundancy rate (Zhang et al. 2009b).

Elucidation of sibship relationships is therefore recommended as a routine practice to be conducted in field genebanks. The curator may be faced with the dilemma of retaining individuals of several families. Provided that land space is not an issue, we propose that these full-sib and half-sib families be retained at the ICG,T since reliable estimates of heritability can be obtained by treating these families as experimental units and by replicating them over geographic areas with diverse environmental conditions. Furthermore, the sibship identified for cacao herein and elsewhere (Zhang et al. 2008, 2009a) will be useful to researchers desirous of making controlled crosses.

Participatory conservation with farmers, locally, regionally, and internationally, can be a useful strategy for safety duplication of these families while obtaining valuable phenotypic field data.

The high (~80 %) molecular variation due to individual accessions supported the existence of greater variation within a group rather than greater variability among groups as might be expected for cacao in accordance with its outcrossing nature. A similar effect has been found in other studies in cocoa (Whitkus et al. 1998; Sereno et al. 2006; Zhang et al. 2008, 2009b, 2011; Motamayor et al. 2008) and in coffee (Krishnan et al. 2012). The high individual molecular variation suggested that representatives of accession groups, rather than particular populations, should be retained and conserved as a germplasm management strategy for allele diversity. The prioritization of accessions can be guided by parameters other than molecular differentiation, e.g., pathogen reaction, morphological traits, and flavor attributes.

Positive inbreeding coefficients were found, indicating an excess of homozygotes, similar to previous studies (Sereno et al. 2006; Aikpokpodion et al. 2009; Zhang et al. 2008, 2009a, 2011). In three of the reference groups (Amelonado, Criollo, and NA) and in the OB4 Refractario (MOQ) subcluster, there was a lack of heterozygotes. Amelonado and Criollo are known to be self-compatible, while Amazonian populations are generally self-incompatible (Toxopeus 1985). Thus, the results may be taken to support the theory that a high degree of

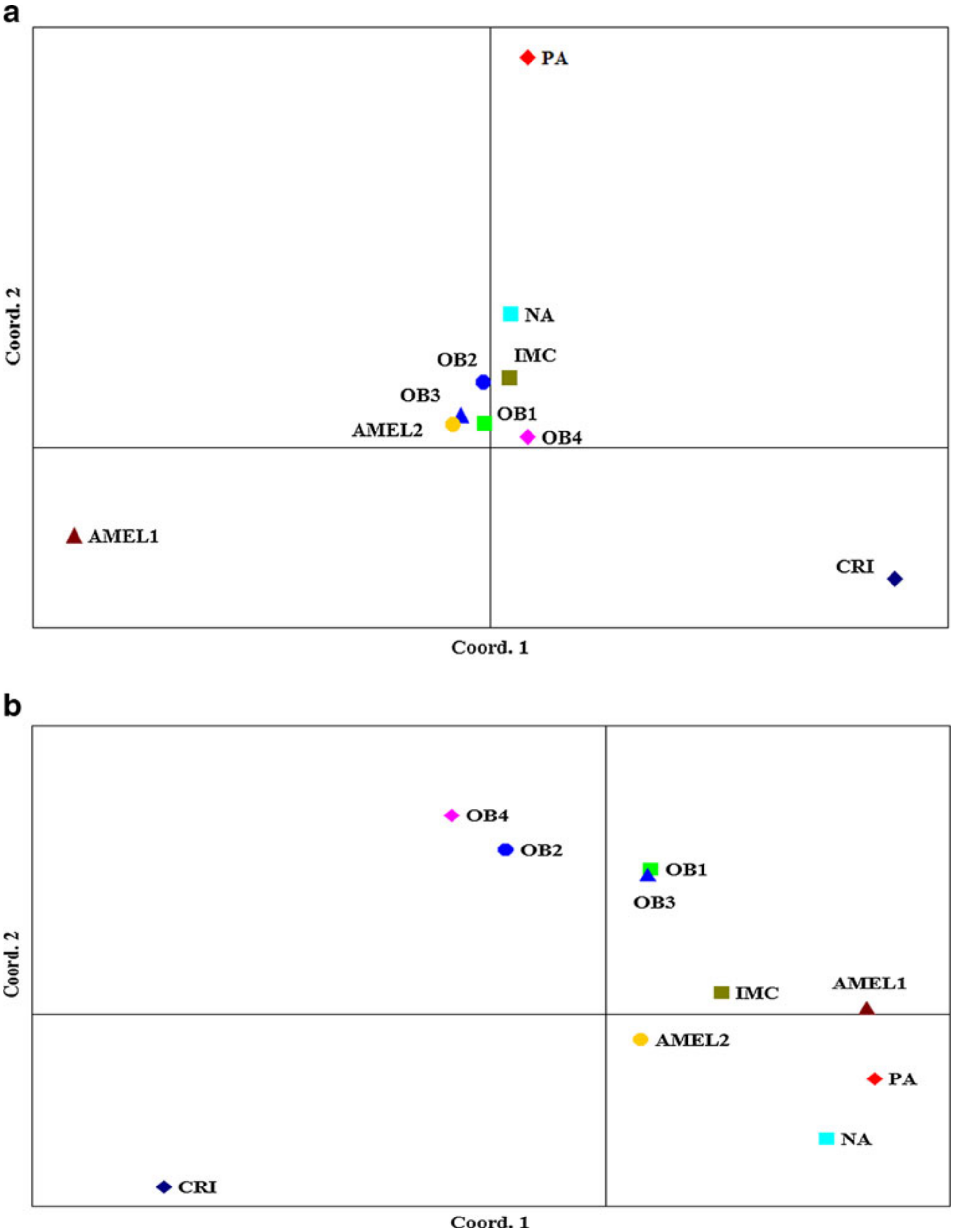

Fig. 4 Principal coordinate analyses for ten homogenous groups of cacao accessions from **a** R_{ST} (Slatkin 1995) and **b** harmonic mean of D_{est} (Jost 2008) pairwise matrices. R_{ST} and D_{est} were determined in GenAlEx v6.1 (Peakall and Smouse 2006) and SMOGD (Crawford 2010). The first two axes in each plot explained 96.4 and 72.7 % of the variation, respectively. Cacao accession groups were *AMEL1* Amelonado 1, *AMEL2* Amelonado 2, *CRI* Criollo, *IMC* Iquitos Mixed Calabacillo, *NA* Nanay, *OB1–4* Refractario subclusters, *PA* Parinari

inbreeding as increased incidence of self-pollination led to the Criollo and Amelonado clusters. Due to their probable self-incompatibility, cross-pollination that resulted in homozygosity is indicated for the NA and MOQ clusters.

Kalinowski (2005) demonstrated that the estimate of genetic distance from highly polymorphic loci required lower sample sizes and suggested a sample size of 20 individuals per population when F_{ST} was 0.05 or higher. Thus, our sample sizes can be presumed to yield reliable calculations of estimated genetic distances. Our data demonstrated low to moderate genetic differentiation between cacao groups or clusters in the traditional Refractario and Forastero groups (Figs. 2 and 5); nonetheless, many of the sibship clusters and population groups were significantly different from each other (Tables 2 and 4). This information will provide guidance to curators in cataloging and assessing the genetic diversity of the collection and will enable breeders to select appropriate accession groups in searching for heterosis.

The distance measures described in this study were consistent with previously observed subclustering of the Refractario

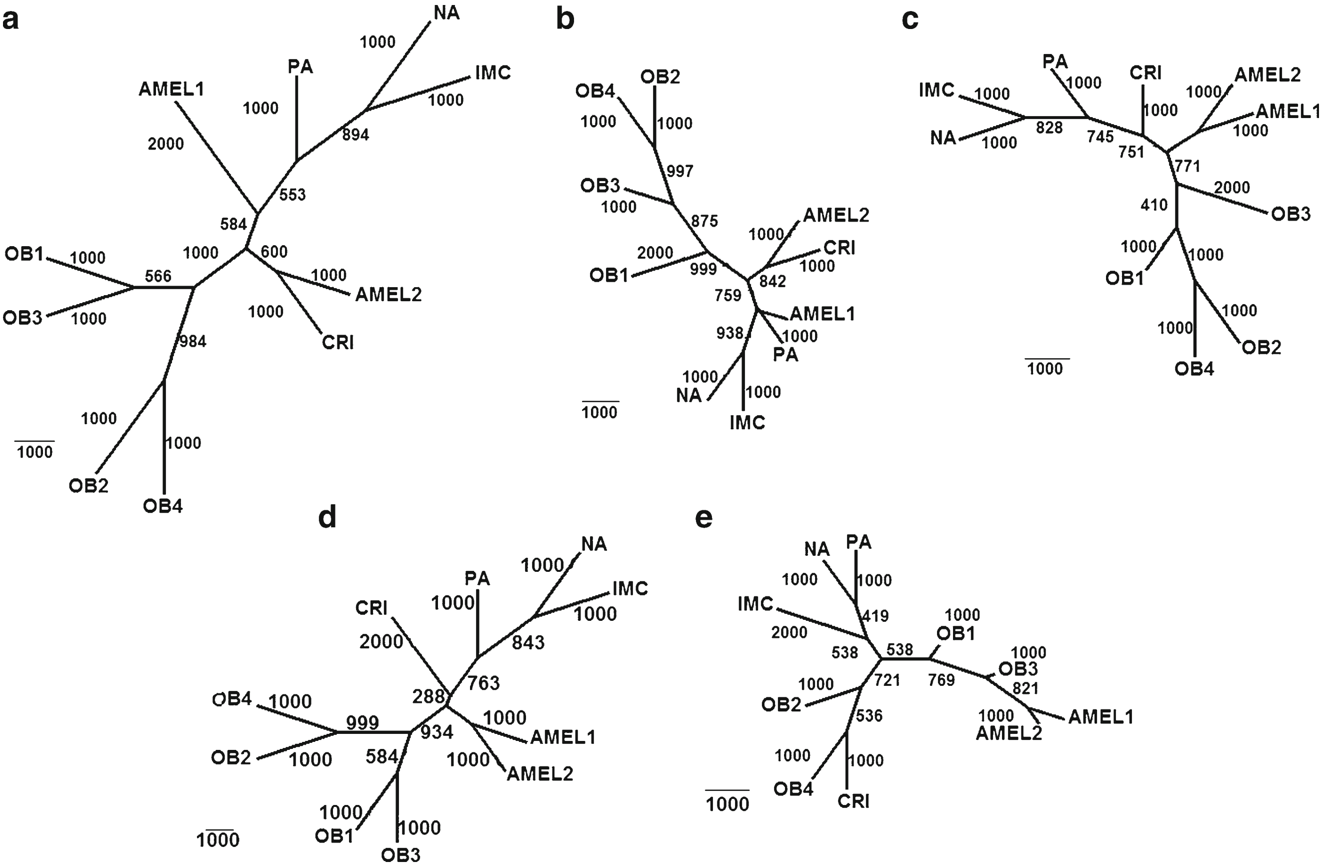

Fig. 5 Consensus radial tree diagrams for ten homogenous groups of cacao accessions from **a** chord distance (D_C) of Cavalli-Sforza and Edwards (1967), **b** Nei et al. (1983) D_A distance, **c** Reynolds et al. (1983) θ_W distance, **d** shared allele distance (D_{SA}) of Jin and Chakraborty (1993), and **e** the $(\delta\mu)^2$ distance of Goldstein et al. (1995). Tree files from 999 bootstrapping in the neighbor-joining algorithm (Saitou and Nei 1987) for each distance measure were created in PowerMarker v3.25 (Liu and Muse 2005). Consensus trees (50 % majority rule) were created with the CONSENSE procedure in PHYLIP (Felsenstein 1989). Radial topologies were manipulated in TreeDyn (Chevenet et al. 2006). Cacao accession groups were *AMEL1* Amelonado 1, *AMEL2* Amelonado 2, *CRI* Criollo, *IMC* Iquitos Mixed Calabacillo, *NA* Nanay, *OB1–4* Refractario subclusters, *PA* Parinari

accessions from STRUCTURE analyses of Zhang et al. (2008) and Motilal et al. (2012). This study demonstrated the superior fit of the D_{est} measure of Jost (2008) compared to the R_{ST} measure of Slatkin (1995) and corroborated previous recommendations for the use of D_{est} (Jost 2008, 2009; Heller and Siegismund 2009). In Fig. 3, PCoA graphs illustrated that the two sibship clusters B and K, both detected by prior STRUCTURE results as belonging to the NA group, could be placed into one group by the D_{est} but not the R_{ST} pairwise matrices. Similarly, clusters L and P (PA accessions) and M (JA accessions) had optimal alignment in the D_{est} PCoA graph. The D_{est} PCoA graph (Fig. 4) also showed Refractario subclustering of OB1 with OB3 and OB2 with OB4, which was in agreement with analyses from STRUCTURE (Motilal et al. 2012) but is in contrast to the Refractario subclustering observed when R_{ST} was used. A similar effect was noted for the Amelonado subgroups. Our use of eigenanalysis to infer population structure and to visualize the relationship that exists among multiple groups demonstrated the value of the approach suggested by Patterson et al. (2006). The approach has also been used successfully for analysis of animal (Gonder et al. 2011) and plant (van Heerwaarden et al. 2011) populations. We therefore recommend that the genetic differentiation measure of D_{est} (Jost 2008, 2009; Heller and Siegismund 2009) be used in conjunction with PCoA in elucidating genetic structure in cacao.

Based on microsatellite analysis, a core set of 59 unique samples (Table 5) can adequately represent the 414 cacao accessions in the present study, as evidenced by the absence of private alleles and the low R_{ST} and D_{est} measures. This core set should be characterized for all of the phenotypic traits that are under study in the genebank collection. We suggest that compatibility interactions should be included to facilitate breeding purposes. Subsequently, a minicore (10 % of core) can then be identified as priority accessions in the genebank. This study had 27 accessions in common with the 110 accessions proposed by Sounigo et al. (2006), but only five of these (IMC 20, IMC 47, IMC 94,

Table 5 Accessions selected for core germplasm based on microsatellite variation from 414 cacao accessions

Sample[a]	Accn group[b]	Genetic group
AMELONADO 15 [MAY]	AMELONADO	Amelonado
MOQ 5/34 F4A D358 T1, T2	MOQ_mis	Amelonado
SIC 256	SIC	Amelonado
SJ 2/26 [POU] F4A E447 T2	SJ_mis	Amelonado
STAHEL F6B D267 T13	STAHEL	Amelonado
B 8/9 [POU] F6A A21 T1	B	B
B 9/10-23 [POU] F5B A5 T12	B	B
CL 19/31 F5A B105 T8, T9	CL	CL
CLM 101 F5A C161 T15	CLM	CLM
CC 7 [BLZ]	CC	Criollo
HONDURAS 10	HON	Criollo
HONDURAS 11	HON	Criollo
ST 4/1 [BLZ]	ST	Criollo
ELP 1/S6	ELP	French Guiana
GU 300/P F4A B197 T2	GU	French Guiana
GU 310/P F4A B233 T1	GU	French Guiana
YAL 6	YAL	French Guiana–Amelonado
CRU 72 F6A A50 T1	CRU	IMC
IMC 10 F4A F531 T1	IMC	IMC
IMC 20 F6B F422 T1	IMC	IMC
IMC 76 F6B A55 T4	IMC	IMC
IMC 94 F6B A17 T3	IMC	IMC
LP 519 [POU] F6A B95 T2, T8 ≡ MOQ 6/28 F5B H642 T3	LP_mis, MOQ_mis	IMC
MOQ 6/68 F5B H657 T2	MOQ_mis	IMC_Criollo_Refractario Cluster B_Amelonado
AMAZ 12 [CHA] F6B B94 T2	AMAZ	IMC–NA
AMAZ 15/15 [CHA] F4A A101 T1	AMAZ	IMC–NA
AMAZ 3/2 [CHA] F6B F433 T7	AMAZ	IMC–NA
H 1	H	IMC–NA
SPEC 184/2 F6B D194 T1	SPEC	IMC–SCA
LP 4/5 [POU] F6A B79 T10	LP	LP
B 14/8 [POU] F6B F481 T1	B_mis	MOQ
LCT EEN 83/S-8 F4A A174 T1	LCT EEN	MOQ
MOQ 2/22 F6A B81 T8	MOQ	MOQ
MOQ 4/21 F4A D361 T1	MOQ	MOQ
MOQ 5/21 F5B H721 T10	MOQ	MOQ
NA 168 F5B F493 T2	NA	NA
NA 246 F5B E404 T7	NA	NA
NA 58 F4A D374 T2	NA	NA
NA 14 F5B E351 T1	NA	NA
SLA 45 F5A D299 T3, T4≡SLA 20 F5A D261 T8	SLA_mis	NA
MATINA 1/7 F6B D236 T15	MATINA_mis	NA–French Guiana
EET 400 [ECU] F6B F455 T6	EET	NA hybrid
B 9/10-33 [POU] F5B I768 T1	B_mis	PA
MOQ 6/95 F5B C221 T3	MOQ_mis	PA
NA 176 F5B E403 T1, F4A D389 T1	NA_mis	PA
PA 107 [PER] F5A D247 T2	PA	PA
PA 218 [PER] F6B C160 T2	PA	PA
PA 34 [PER] F5B E374 T1, T7	PA	PA
PA 90 [PER] F4A F509 T1 ≡ MOQ 6/5 F4A D344 T4	PA	PA
SJ 2/10 [POU] F5A D277 T2	SJ_mis	Refractario hybrid Cluster B–cluster O
NA 471 F6A B86 T9, F4A D412 T1	NA_mis	Refractario cluster O hybrid
NA 111 F4A E495 T1	NA_mis	Refractario cluster O–SCA
SCA 3 F6B A26 T1	SCA	SCA
SCA 6 F6B A16 T14	SCA	SCA
U 1	U	SCA
IMC 47 F6B F401 T3	IMC_mis	SCA–Amelonado
CRUZ 7/8 F6B B83 T1, T9	CRUZ	SCA–French Guiana
TRD 15 F4A A43 T1	TRD_mis	SCA–IMC
CRU 128 F5B G569 T1	CRU	SCA–IMC

[a] Accession nomenclature after Turnbull and Hadley (2013). Accessions from the International Cocoa Genebank, Trinidad, are provided with field, section, plot, and tree number, e.g., TRD 15 F4A A43 T1 represents TRD 15 found in field 4A, section A, plot43, Tree #1

[b] Accession group; hyphenated suffix indicates mislabelled sample that does not belong to the designated accession group

MATINA 1/7, and MOQ 6/95) were present in the core set identified in this study. These five accessions should, therefore, be considered as an obligate set when compiling core sets for cacao.

The core set of accessions identified herein contained several accessions that are currently absent from the ICG,T. Our demonstration of the need to include accessions in

Table 6 Summary statistics for core and full datasets of cacao accessions

Pop		N_e	I	H_o	H_e	PIC
Full dataset (n=414)	Mean	3.550	1.485	0.494	0.6939	0.6539
	SE	0.191	0.066	0.021	0.0232	0.0242
Core (n=59)	Mean	5.440	1.870	0.472	0.7914	0.7659
	SE	0.404	0.076	0.017	0.0172	0.0197
Significance from analysis of variance	P	0.0001	0.0004	0.4380	0.0016	0.0009

N_e number of effective alleles, I Shannon's information index (Lewontin 1972), H_o observed heterozygosity, H_e expected heterozygosity (gene diversity), *PIC* polymorphism information content of Botstein et al. (1980)

experiments that are not currently present in the ICG,T effectively established the need for this field genebank to renew its efforts to expand their acquisitions. Furthermore, mislabelled trees (trees having an ancestral contribution different from that expected from their accession name) formed 25 % of the core. Curators of collections must, therefore, be cautious in the treatment of mislabelled trees as they can prove valuable in capturing genetic diversity. Curators should seek to have an inclusive conservative management strategy whereby mislabelled material must be properly recoded and recorded, digitally or otherwise, and phenotypically characterized. Since the ICG,T is already safely duplicated with a maximum of 16 trees per plot, mislabelled trees in a plot that present as duplicates of an accession present in another plot elsewhere in the genebank may be tagged for removal only if (a) the accession is already at its maximum safety duplication level, (b) the accession is not in high simultaneous demand by researchers, (c) the accession does not possess a unique phenotype, (d) removal of tree(s) does not create a high stress environment by canopy or shade disruption, and (e) there is an absence of alternative planting sites for an incoming accession. Further, when the core set of accessions is based on molecular markers, as in this study, it is essential to fully characterize the candidate trees to determine the extent of the phenotypic diversity within the core set. This may not necessarily adequately capture the phenotypic diversity of the traits under evaluation in the genebank. Hence, it would be wise to design core collections around thematic areas so that there can be core collections representing the diversity of traits of interest, for example, for disease resistance or fruit morphology. It may be possible though that the same set of accessions that captures the diversity of a particular thematic area may also capture that of another thematic area either without additional accessions or by the inclusion of few additional accessions.

The core collection (Table 5) should be protected from genetic erosion as trees can be lost from the genebank due in part to any one, or a combination over time, of events such as drought, flooding, fire, storm damage, disease, and malicious activity. Additional protection from genetic erosion would necessitate propagation of trees, especially those in single copy. In the case of field genebanks with multiple trees per accession, a DNA fingerprinting activity should be conducted to establish which trees are true clonal replicates of the appropriate accession. Misidentified trees may then be assigned to their correct accession group and accession name. In the case of the ICG,T collection, these activities should be given high priority since several accessions in this list are currently represented by only one tree in the field, e.g., MATINA 1/7, and a high error rate is present in the collection. Subsequent to propagation, the core set should be established in at least one other location. This may include greenhouses, in national conservation trusts, non-governmental organizations interested in conservation, or in participatory conservation. We propose that the core set should be also maintained as a discrete unit in an exclusively assigned field in addition to the original placement of trees spread over various fields in the germplasm collection.

In conclusion, pedigree construction in examining redundancy and the use of PCoA with D_{est} to probe fine structure of differentiation among cacao accessions is recommended to curators of ex situ germplasm collections. The identification of core collections as a management strategy in cacao is also encouraged. For conservation of other species in field genebanks, analyses similar to those used in this study are highly recommended for efficient and cost-effective management of accessions to maximize genetic diversity.

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

Ethical standards The research conducted was in compliance with the laws of the country in which the research was conducted.

Conflict of interest The authors declare that they have no conflict of interest.

Data archiving statement Germplasm materials used in this study are provided in Online Resource 1. Microsatellite data will be uploaded to the International Cocoa Germplasm Database (http://www.icgd.rdg.ac.uk/; Turnbull and Hardy 2013).

References

Aikpokpodion PO, Motamayor JC, Adetimirin VO, Adu-Ampomah Y, Ingelbrecht I, Eskes AB, Schnell RJ, Kolesnikova-Allen M (2009) Genetic diversity assessment of sub-samples of cacao, *Theobroma cacao* L. collections in West Africa using simple sequence repeats marker. Tree Genet Genomes 5:699–711

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

Bartley BGD (2001) Refractario—an explanation of the meaning of the term and its relationship to the introductions from Ecuador in 1937. INGENIC Newsl 5:4–6

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

Belaj A, Dominguez-García MC, Atienza SG, Urdíroz NM, de la Rosa R et al (2012) Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. Tree Genet Genomes 8:365–378

Boccara M, Zhang D (2006) Progress in resolving identity issues among the Parinari accessions held in Trinidad: the contribution of the collaborative USDA/CRU project. In: Annual report 2005. Cocoa Research Unit, The University of the West Indies: St. Augustine, Trinidad, pp. 25–32

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

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

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet 19:233–257

Chang A (2001) Group differences program v3.0. Software available at: http://department.obg.cuhk.edu.hk/researchsupport/download/downloads.asp. Accessed 7 November 2012

Chevenet F, Brun C, Bañuls A-L, Jacq B, Christen R (2006) TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinforma 7:439, 9 pp. Software available at: http://www.treedyn.org/. Accessed 7 November 2012

Cope FW (1958) Incompatibility in *Theobroma cacao*. Nature 181:279

Cope FW (1962) The mechanism of pollen incompatibility in *Theobroma cacao* L. Heredity 17:157–182

Crawford NG (2010) SMOGD: software for the measurement of genetic diversity. Mol Ecol Resour 10:556–557, Online usage at: http://www.ngcrawford.com/django/jost/. Accessed 7 November 2012

Cuevas HE, Prom LK (2013) Assessment of molecular diversity and population structure of the Ethiopian sorghum [*Sorghum bicolor* (L.) Moench] germplasm collection maintained by the USDA-ARS National Plant Germplasm System using SSR markers. Genet Resour Crop Evol. doi:10.1007/s10722-013-9956-5, Accessed 15 February 2013

Dhanaraj AL, Rao EVVB, Swamy KRM, Bhat MG, Prasad DT, Sondur SN (2002) Using RAPDs to assess diversity in Indian cashew (*Anacardium accidentale* L.). J Hort Sci Biotechnol 77:41–47

Dieringer D, Schlötterer C (2002) Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. Mol Ecol Notes 3:167–169

Escribano P, Viruel MA, Hormaza JI (2008) Comparison of different methods to sequence markers. A case study in cherimoya (*Annona cherimola*, Annonaceae), an underutilised subtropical fruit tree species. Ann Appl Biol 153:25–32

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

Felsenstein J (1989) PHYLIP—phylogeny inference package (Version 3.2). Cladistics 5:164–166, Software available at: http://evolution.genetics.washington.edu/phylip/getme.html. Accessed 7 November 2012

Fernández J-A, Santana O, Guardiola J-L, Molina R-V, Heslop-Harrison P, Borbely G et al (2011) The world saffron and *Crocus* collection: strategies for establishment, management, characterisation and utilisation. Genet Resour Crop Evol 58:125–137

Frankel OH (1984) Genetic perspectives of germplasm conservation. In: Arber WK, Llimensee K, Peacock WJ, Starlinger P (eds) Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge, pp 161–170, part III, paper no. 15

Goldstein DB, Luiz LA, Cavalli-Sforza LL, Feldman MM (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92:6723–6727

Gonder MK, Locatelli S, Ghobrial L, Mitchell MW, Kujawski JT, Lankester FJ, Stewart C-B, Tishkoff SA (2011) Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. Proc Natl Acad Sci USA 108:4766–4771, Available online at: www.pnas.org/cgi/doi/10.1073/pnas.1015422108. Accessed 7 November 2012

Haouane H, Bakkali AE, Moukhli A, Tollon C, Santoni S, Oukabli A, Modafar CE, Khadari B (2011) Genetic structure and core collection of the World Olive Germplasm Bank of Marrakech: towards the optimised management and use of Mediterranean olive genetic resources. Genetica 139:1083–1094

Hedrick P (2005) A standardized genetic differentiation measure. Evolution 59:1633–1638

Heller R, Siegismund HR (2009) Relationship between three measures of genetic differentiation G_{ST}, D_{EST} and G'_{ST}: how wrong have we been? Mol Ecol 18:2080–2083

Irish BM, Goenaga R, Zhang D, Schnell R, Brown JS, Motamayor JC (2010) Microsatellite fingerprinting of the USDA-ARS Tropical Agriculture Research Station cacao (*Theobroma cacao* L.) germplasm collection. Crop Sci 50(2):656–667

Jin L, Chakraborty R (1993) Estimation of genetic distance and coefficient of gene diversity from single-probe multilocus DNA fingerprinting data. Mol Biol Evol 11:120–127

Johnson PCD, Haydon DT (2007) Maximum-likelihood estimation of allelic dropout and false allele error rates from microsatellite genotypes in the absence of reference data. Genetics 175:827–842

Jost L (2008) G_{ST} and its relatives do not measure differentiation. Mol Ecol 17:4015–4026

Jost L (2009) D vs. G_{ST}: response to Heller and Siegismund (2009) and Ryman and Leimar (2009). Mol Ecol 18:2088–2091

Kalinowski ST (2005) Do polymorphic loci require large sample sizes to estimate genetic distances? Heredity 94:33–36

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol Ecol 16:1099–1106

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

Knight R, Rogers HH (1953) Sterility in *Theobroma cacao* L. Nature 172:164

Knight R, Rogers HH (1955) Incompatibility in *Theobroma cacao* L. Heredity 9:69–77

Krishnan S, Ranker TA, Davis AP, Rakotomalala JJ (2012) An assessment of the genetic integrity of ex situ germplasm collections of three endangered species of *Coffea* from Madagascar: implications for the management of field germplasm collections. Genet Resour Crop Evol. doi:10.1007/s10722-012-9898-3

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

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

Lewontin RC (1972) The apportionment of human diversity. Evol Biol 6:381–398

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 3.25, distributed by author, available at: http://www.powermarker.net). Accessed 7 November 2012

Mario PC, Viviana BV, Juan TU, Blair MW, Gabriel BB (2010) Selection of a representative core collection from the Chilean common bean germplasm. Chilean J Agric Res 70:3–15

Mei Y, Zhou J, Xu H, Zhu S (2012) Development of Sea Island cotton (*Gossypium barbadense* L.) core collection using genotypic values. Aust J Crop Sci 6:673–680

Miranda C, Urrestarazu J, Santesteban LG, Royo JB, Uribina V (2010) Genetic diversity and structure in a collection of ancient Spanish pear cultivars assessed by microsatellite markers. J Am Soc Hort Sci 135:428–437

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

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

Motilal LA, Zhang D, Umaharan P, Mischke S, Boccara M, Pinney S (2009) Increasing accuracy and throughput in large-scale microsatellite fingerprinting of cacao field germplasm collections. Trop Plant Biol 2:23–27

Motilal LA, Zhang D, Umaharan P, Mischke S, Mooleedhar V, 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 Genet Resour 8:106–115

Motilal LA, Zhang D, Umaharan P, Mischke S, Pinney S, Meinhardt LW (2011) Microsatellite fingerprinting in the International Cocoa Genebank Trinidad: accession and plot homogeneity information for germplasm management. Plant Genet Resour 9:430–438

Motilal LA, Zhang D, Umaharan P, Boccara M, Mischke S, Sankar A, Meinhardt LW (2012) Elucidation of genetic identity and population structure of cacao germplasm within an international cacao genebank. Plant Genet Resour 10:232–241

National Institute of Agricultural Biotechnology (2006) PowerCore (v. 1.0). A program applying the advanced M strategy using heuristic search for establishing core or allele mining sets. Rural Development Administration (RDA), Republic of Korea, software available at: http://www.genebank.go.kr/eng/PowerCore/powercore.jsp. Accessed 7 November 2012

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

Odong TL, Jansen J, van Eeuwijk FA, van Hintum TJL (2013) Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theor Appl Genet 126:289–305

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genetics 2(e190):2074–2093

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

Pessoa-Filho M, Rangel PHN, Ferreira ME (2010) Extracting samples of high diversity from thematic collections of large gene banks using a genetic-distance based approach. BMC Plant Biology 10:127, http://www.biomedcentral.com/1471-2229/10/127

Pound FJ (1938) Cacao and witches' broom disease (*Marasmius perniciosus*) of South America with notes on other species of *Theobroma*. Yuille's Printery, Port-of-Spain

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

Pritchard JK, Stephens M, 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, Lanaud C (2004) A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. Theor Appl Genet 108:1151–1161

Rambaut A (2006–2009) FigTree. Tree Figure Drawing Tool v.1.3.1. Institute of Evolutionary Biology, University of Edinburgh. Software available at: http://tree.bio.ed.ac.uk/. Accessed 7 November 2012

Reeves PA, Panella LW, Richards CM (2012) Retention of agronomically important variation in germplasm core collections: implications for allele mining. Theor Appl Genet 124:1155–1171

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basic for a short-term genetic distance. Genetics 105:767–779

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

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

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

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462

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

Sounigo O, Bekele FL, Iwaro AD, Thévenin J-M, Bidaisee G, Umaharan R et al (2006) Description of cocoa clones proposed for the "CFC/ICCO/IPGRI Project Collection". In: Eskes AB, Efron Y (eds) Global approaches to cocoa germplasm utilization and conservation. Final report of the CFC/ICCO/IPGRI project on "Cocoa germplasm utilization and conservation: a global approach" (1998–2004). CFC, Amsterdam, pp 67–81

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

Turnbull CJ, Hadley P (2013) International Cocoa Germplasm Database (ICGD). [Online database]. CRA Ltd//NYSE Liffe/University of Reading, UK. Available at: http://www.icgd.reading.ac.uk. Accessed 7 February 2013

van Heerwaarden J, Doebley J, Briggs WH, Glaubitz JC, Goodman MM, Gonzalez JJS, Ross-Ibarra J (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc Natl Acad Sci USA 108:1088–1092

Viji S, Zhang D, Dhekney SA, Harris DL, Yadav AK, Okie WR (2012) Cultivar identification, pedigree verification and diversity analysis among peach cultivars based on simple sequence repeat markers. J Amer Soc Hort Sci 137:114–121

Wang J (2004) Sibship reconstruction from genetic data with typing errors. Genetics 166:1963–1979

Wang J, Santure AW (2009) Parentage and sibship inference from multi-locus genotype data under polygamy. Genetics 181:1579–1594

Whitkus R, de la Cruz M, Mota-Bravo L, Gómez-Pompa A (1998) Genetic diversity and relationships of cacao (*Theobroma cacao* L.) in southern Mexico. Theor Appl Genet 96:621–627

Zhang D, Boccara M, Motilal L, Butler DR, Umaharan P, Mischke S, Meinhardt L (2008) Microsatellite variation and population structure in the "Refractario" cacao of Ecuador. Conserv Genet 9:327–337

Zhang D, Boccara M, Motilal L, Mischke S, Johnson ES, Butler DR, Bailey B, Meinhardt L (2009a) Molecular characterization of an earliest cacao (*Theobroma cacao* L.) collection from Upper Amazon using microsatellite DNA markers. Tree Genet Genomes 5:595–607

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

Zhang D, Martínez WJ, Johnson ES, Somarriba E, Phillips-Mora W, Astorga C, Mischke S, Meinhardt LW (2011) Genetic diversity and spatial structure in a new distinct *Theobroma cacao* L. population in Bolivia. Genet Resour Crop Evol. doi:10.1007/s10722-011-9680-y