Tree Genetics & Genomes (2009) 5:595–607
DOI 10.1007/s11295-009-0212-2

ORIGINAL PAPER

Molecular characterization of an earliest cacao (*Theobroma cacao* L.) collection from Upper Amazon using microsatellite DNA markers

Dapeng Zhang · Michel Boccara · Lambert Motilal · Sue Mischke · Elizabeth S. Johnson · David R. Butler · Bryan Bailey · Lyndel Meinhardt

Received: 30 September 2008 / Revised: 17 January 2009 / Accepted: 13 April 2009 / Published online: 23 May 2009

Abstract Cacao (*Theobroma cacao* L.) is indigenous to the Amazon region of South America. The river basins in the Upper Amazon harbor a large number of diverse cacao populations. Since the 1930s, several numbers of populations have been collected from the present-day Peruvian Amazon and maintained as ex situ germplasm repositories in various countries, with the largest held in the International Cacao Genebank in Trinidad. The lack of information on population structure and pedigree relationship and the incorrect labeling of accessions are major concerns for efficient conservation and use of cacao germplasm. In the present study, we assessed the individual identity, sibship, and population structure in cacao populations collected from the present-day Loreto Region, Peru in the 1930–1940s. Using a capillary electrophoresis genotyping system, we analyzed the simple sequence repeat variation of 612 cacao accessions collected from the Marañon, Nanay, and Ucayali river systems. A total of 180 cases of mislabeling were identified using a Bayesian clustering method for admixture detection. Using maximum likelihood-based methods, we reconstructed 78 full-sib families nested in 48 half-sib families, indicating that the pods collected in the 1930s were from 48 mother trees, maximum. Likelihood simulation also identified eight probable parents that are responsible for 117 pairs of mother–offspring relationships in this collection. Principal coordinate analysis (PCoA) and the Bayesian clustering method cohesively demonstrated a pronounced structure of genetic diversity, stratified by the river systems of the Peruvian Amazon. Our results also show that, in spite of the high level of allelic diversity in this collection, it was composed of a large number of related family members collected from a relatively small area, including a couple of sites in the Ucayali and Nanay rivers, as well as the lower Marañon river near Iquitos. The vast majority of the Peruvian Amazon, especially the upper Marañon River and its tributaries, have not been sampled by collecting expeditions. The improved understanding of the individual identities, genealogical relationships, and geographical origin of cacao germplasm in this collection will contribute to more efficient conservation and utilization of these germplasm. Additionally, this study also provides more baseline information to help guide future collecting expeditions in the Peruvian Amazon.

Keywords Amazon · *Theobroma cacao* · Conservation · Germplasm · Genetic diversity · Population structure · Peru · Tropical tree

Communicated by E. Dirlewanger

D. Zhang (✉) · S. Mischke · E. S. Johnson · B. Bailey · L. Meinhardt
USDA ARS PSI SPCL,
10300 Baltimore Avenue, Bldg. 001, Rm. 223, BARC-W,
Beltsville, MD 20705, USA
e-mail: Dapeng.Zhang@ars.usda.gov

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

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

Introduction

Cacao (*Theobroma cacao* L.) is native to the South American rainforest, although it is thought to have been domesticated in Southern Mexico and Central America (Cuatrecasas 1964; Gómez-Pompa et al. 1990; Henderson et al. 2007; Hunter 1990). The hypothesized center of genetic diversity is located in the upper Amazonian regions of Peru, Ecuador, Colombia, and Brazil, which has a series of major river systems flowing into the Marañón and the Amazon (Cheesman 1944; Dias 2001; Bartley 2005). During the past several decades, several expeditions have been made, and a substantial amount of germplasm, as both wild populations and cultivated accessions, has been collected from this region and maintained as ex situ collections in various countries. The majority of the collected germplasm are maintained in the International Cacao Genebank, Trinidad (ICG, T), under the management of the Cacao Research Unit (CRU) of the University of the West Indies (Pound 1938, 1945; Lockwood and End 1993; Kennedy and Mooleedhar 1993).

The first organized cacao germplasm collecting expeditions in the present-day Peruvian Amazon started in 1937–1938 (Pound 1938, 1943, 1945), and the collecting sites, according to the report, included Rio Nanay, Rio Morona, Rio Marañón, and Rio Ucayali (Fig. 1). These expeditions led to the establishment of the germplasm collection in Iquitos, Peru, known as the "Pound collection," named after the collector F. J. Pound. This collection was comprised of five natural populations (or accession groups): "Iquitos Mixed Calabacillo" (IMC), "Morona" (MO), "Nanay" (NA), "Parinari" (PA), and "Scavina" (SCA) (Pound 1938, 1943, 1945). An unknown number of pods (fruits) were collected from trees without any symptoms of witches' broom disease. The seeds were then bulked and sent to Barbados where the seedlings were raised (Pound 1938, 1943, 1945). After the seedlings had developed sufficiently and were declared healthy at quarantine in Barbados, these germplasms were transferred to Trinidad in the form of bud wood and were planted at Marper Estate in Plum Mitan, Manzanilla, Trinidad. In addition to the five accessions groups, this collection also includes a group of clones, which Pound collected in 1943 when he revisited the same sites where the NA, IMC, and SCA were collected. These accessions were collected as bud woods and were referred as "Pound clones" or "P clones". Some of these P clones were believed to be the parental trees from which the pods of NA, IMC, and SCA were taken in the 1937–1938 collecting expedition (Pound 1945; Bartley 2005).

Among the 80 or so different germplasm groups held in the ICG, T, those in Pound collection are among the most widely distributed germplasm, due to their valuable agronomic traits and their potential for resistance to witches' broom disease (Lockwood and End 1993; International Cacao Germplasm Database, http://www.icgd.reading.ac.uk). These germplasm have also been reported to have a high percentage of resistant genotypes to Black Pod disease (caused by *Phytophthora* spp.), indicating that this is a particularly rich source of resistance genes (Iwaro et al. 2003; Wadsworth et al. 1997). In many cacao-producing regions around the world, the Pound selections of Upper Amazonian cacao are either adopted directly as clones or used as parents for the production of seed families. This collection is by far the most widely used germplasm for cacao breeding in the world (Bartley 1994, 2005; International Cacao Germplasm Database, http://www.icgd.reading.ac.uk; Posnette, 1986).

Despite their importance in cacao production and improvement, little detail is known about the Pound's collection regarding its population structure and passport information. It is believed, but not well documented, that these accessions originated from a small number of trees in a few geographical sites. Moreover, errors of documentation often occur during the transportation, propagation, or maintenance of material, resulting in large numbers of trees with unconfirmed identities. The ambiguity in genetic identity, population structure, and family relationships have been a serious concern for the effective use of these germplasms.

A comprehensive assessment of the genetic identity of individual accessions, and the genealogical relationships among them, is essential to reduce redundancy in germplasm collections. This information can lead to a greater understanding of the ecogeographic representation of existing collections and improve the accuracy and effectiveness of germplasm utilization. The development of simple sequence repeat (SSR) markers in cacao (Lanaud et al. 1999) has significantly increased the capacity of molecular characterization of cacao germplasm. SSR-based DNA fingerprinting has been increasingly applied in cacao germplasm characterization (Aikpokpodion et al. 2005; Cryer et al. 2006; Efombagan et al. 2008; Johnson et al. 2007; Lanaud et al. 2001; Motamayor et al. 2002, 2008; Schnell et al. 2005; Sereno et al. 2006; Zhang et al. 2008, 2009). These studies generate important information relating to the genetic diversity and the origin of the crop. To date, however, application of SSR markers to assess individual identity and genealogical relationships among cacao individuals has been very limited.

In this paper, we report a study in which 15 SSR loci were used to characterize the Upper Amazonian cacao germplasm collected in the 1930s–1940s by Dr. F.J. Pound. Our objectives were to (1) identify mislabeling and duplicates in this group of cacao germplasm, (2) analyze the family structure and reconstruct sibships for each accession group, and (3) assess genetic diversity and population structure in this set of Upper Amazon cacao.

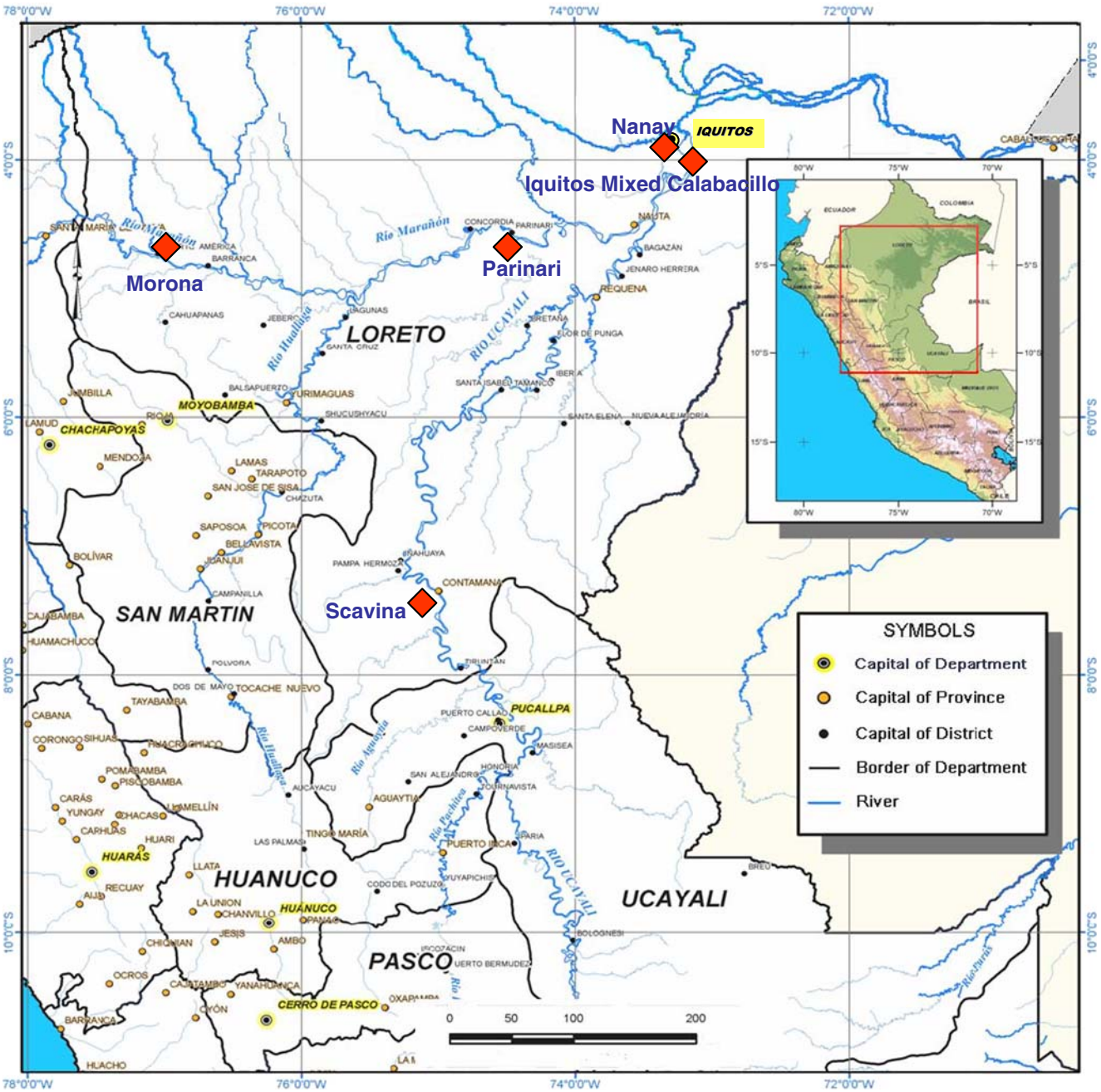

Fig. 1 The putative geographical sites where Pound's collection was taken in present-day Loreto Region, Peru. The exact collecting sites were unknown

Materials and methods

Plant materials

A total of 612 cacao trees including most of the existing accessions of the Pound collection were taken, most of which came from Marper Farm where the original trees are still maintained. In several cases, samples from UCRS were taken in order to confirm if the correct trees were transferred (Table 1).

The cacao samples used for these DNA fingerprinting profiles include leaf samples of variable ages. Each sampled branch was tagged for potential revisiting. In several cases, purported duplicate trees from a different field and plot were sampled, which resulted in two or four samples for these accessions. Therefore, each sample was labeled with both accession name and DNA extraction number. DNA was extracted at CRU following the protocol of Kobayashi et al. (1998) and quantified with ethidium staining in 1% agarose gels. Aliquots of 50 µl were

Table 1 Summary of mislabeled and duplicated genotypes identified in the "Pound collection"

Accession name	No. of analyzed samples	No with mislabeling[a]	Percent mislabeling (%)	Number of eliminated duplicates[b]	Number of accessions used for diversity analysis
IMC	82	12	14.0	18	52
MO	28	11	39.3	2	15
NA	279	118	42.3	34	127
PA	173	27	15.6	46	100
SCA	20	8	40.0	3	9
Pound	30	4	13.3	13	13
Total	612	180		116	316

[a] Mislabeled individuals were identified using the Bayesian clustering analysis (Pritchard et al. 2000)

[b] Duplicated or cloned genotypes were identified using multi-locus genotype matching (Waits et al. 2001). The eliminated duplicates are mainly trees with same names but planted in different farms, fields, and plots

prepared and shipped to the USDA Beltsville Agricultural Research Center.

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

- Thirty-three Ucayali accessions from the National Agricultural University at Tingo Maria, Peru
- Ten Lower Amazon Forastero clones from the collections at Centro Agronómico Tropical de Investigación y Enseñanza (CATIE, Costa Rica): Amelonado-15, BE-3, Catongo Blanco, SIAL 169, SIAL 325, SIC 806, SIC 813, SIC 1, SIC 7, and SIC 256.

SSR analysis

Amplifications of microsatellite loci were achieved using 15 primers with sequences previously described (Lanaud et al. 1999; Risterucci et al. 2000; Saunders et al. 2004). These 15 loci have been agreed upon, by multiple international and government-sponsored laboratories in the cacao research community, as standardized SSR primers to characterize all *T. cacao* germplasm collections (Saunders et al. 2001, 2004). These standard loci have been used for cacao genotyping in several germplasm collections (Boccara and Zhang 2006; Zhang et al. 2006a,b). Primers were synthesized by Proligo (Boulder, CO), and forward primers were 5′-labeled using WellRED fluorescent dyes (Beckman Coulter, Inc., Fullerton, CA). PCR was performed as described in Saunders et al. (2004), using commercial hot-start PCR SuperMix that had been fortified with an additional 30 U/ml of hot-start *Taq* DNA polymerase (Invitrogen Platinum *Taq*, Carlsbad, CA; or Eppendorf HotMaster *Taq*, Brinkman, Westbury, NY).

The amplified microsatellite loci were separated by capillary electrophoresis as previously described (Saunders et al. 2004; Zhang et al. 2006b). Data analysis was performed using the CEQ 8000 Fragment Analysis software version 7.0.55 according to manufacturers' recommendations (Beckman Coulter, Inc.). SSR fragment sizes were automatically calculated to two decimal places by the CEQ^{TM} 8000 Genetic Analysis System. Allele calling was performed using the CEQ 8000 binning wizard software (CEQ^{TM} 8000 software version 7.0.55, Beckman Coulter, Inc.), and edited based on the bin list using a SAS program (SAS 1999).

Data analysis

Identification of mislabeled accessions Verification of the genetic identity of each accession was carried out using a Bayesian clustering analysis, which assigns each individual to its corresponding "home population." The program Structure v2.1 (Pritchard et al. 2000) was used for computation. For each population (or accession group), k=2 (population member vs. immigrant) was assumed. All Structure runs used 100,000 iterations after a burn-in of length 200,000. The allocation of the individual to a particular cluster was set at not less than 75% assignment probability (Q value). The individuals that were not assigned to the "home population" were considered as mislabeled and were excluded from the subsequent analysis of sibship reconstruction, parentage analysis, and diversity assessment.

Identification of duplicated genotypes After the elimination of mislabeled individuals, the accessions with confirmed population membership were examined to identify duplicated genotypes, including trees having the same name but planted in different farms, fields, or plots. Pairwise comparisons were made among all individuals based on their multilocus SSR profile. Accessions with different names that were fully matched at 15 loci were declared synonymous accessions. Statistical rigor was assessed for match declaration to determine whether two individuals

Table 2 Examples of identified mislabeling using the assignment test

Accession	*Q*	Accession	*Q*	Accession	*Q*
IMC10(fp612)	0.998	IMC42(fp1127)	0.998	IMC63(fp1077)	0.998
IMC102(fp121)	0.058*	IMC44(fp59)	0.998	IMC65(fp1056)	0.998
IMC104(fp1384)	0.982	IMC45(fp143)	0.226*	IMC65(fp1161)	0.994
IMC105(fp275)	0.998	IMC47(fp1061)	0.994	IMC66(fp1079)	0.998
IMC107(fp1065)	0.998	IMC47(fp1628)	0.998	IMC66(fp1115)	0.995
IMC107(fp203)	0.998	IMC47(fp624)	0.068*	IMC67(fp81)	0.007*
IMC107(fp2714)	0.998	IMC47(fp807)	0.996	IMC68(fp1064)	0.998
IMC11(fp1074)	0.998	IMC50(fp19)	0.998	IMC68(fp57)	0.998
IMC12(fp1054)	0.998	IMC51(fp1386)	0.997	IMC71(fp2675)	0.998
IMC12(fp592)	0.998	IMC53(fp1059)	0.998	IMC73(fp1078)	0.997
IMC13(fp607)	0.998	IMC53(fp1119)	0.998	IMC76(fp1123)	0.998
IMC16(fp1057)	0.997	IMC55(fp1125)	0.998	IMC76(fp1124)	0.998
IMC16(fp302)	0.443*	IMC57(fp1063)	0.996	IMC76(fp622)	0.998
IMC18(fp760)	0.998	IMC57(fp1357)	0.01*	IMC77(fp1052)	0.998
IMC2(fp1385)	0.998	IMC57(fp148)	0.015*	IMC77(fp232)	0.998
IMC20(fp634)	0.998	IMC57(fp361)	0.316*	IMC78(fp1070)	0.998
IMC22(fp617)	0.08*	IMC58(fp1011)	0.997	IMC78(fp157)	0.998
IMC23(fp574)	0.998	IMC58(fp1058)	0.998	IMC81(fp1635)	0.012*
IMC27(fp614)	0.998	IMC58(fp1075)	0.998	IMC85(fp272)	0.998
IMC3(fp1067)	0.998	IMC58(fp276)	0.998	IMC9(fp606)	0.996
IMC30(fp586)	0.997	IMC59(fp314)	0.998	IMC94(fp1080)	0.997
IMC31(fp12)	0.998	IMC59(fp1081)	0.998	IMC94(fp17)	0.998
IMC31(fp561)	0.998	IMC6(fp101)	0.998	IMC94(fp640)	0.997
IMC33(fp1076)	0.998	IMC6(fp1060)	0.997	IMC96(fp1051)	0.998
IMC33(fp557)	0.998	IMC6(fp1068)	0.998	IMC96(fp140)	0.998
IMC38(fp1055)	0.998	IMC60(fp1073)	0.998	IMC97(fp1387)	0.716*
IMC38(fp127)	0.997	IMC61(fp1053)	0.998	IMC98(fp1388)	0.998
IMC41(fp1069)	0.016*				

The list is 82 Iquitos Mixed Calabacillo (IMC) samples

Accessions with membership assigned to the IMC population based on the Bayesian clustering analysis (Pritchard et al. 2000). Only those with assignment probability above the criterion of 0.75 were accepted as IMC accessions. Fourteen accessions (marked with asterisk) were excluded in subsequent analysis for genetic diversity and for sibship reconstruction

may share the same multilocus genotype by chance (Waits et al. 2001). The computer program GenAlEx 6 (Peakall and Smouse 2006) was used for genotype matching and computation of probability of identity (PID) between siblings (PID-sib). In the subsequent analysis of genetic diversity and population structure, only one genotype out of each of the identified synonymous sets was used.

Analysis of genetic diversity and population structure After the elimination of mislabeled and duplicated genotypes, the summary statistics for each marker locus, including allele number, observed heterozygosity (H_O), gene diversity, and inbreeding coefficient were computed using PowerMarker v. 3.0 (Liu and Muse 2005). To assess the relationships among the correctly identified individuals, pairwise Euclidian distances were computed for every pair of the accessions using the genetic distance procedure in GenAlEx 6 (Peakall and Smouse). The pairwise distances were then presented by PCoA using the same program. Both distance and covariance were standardized. Thirty-three samples from a Peruvian cacao population collected from the Ucayali river in the late 1980s (Zhang et al. 2006b) and ten international clones from the CATIE cacao germplasm collection (Zhang et al. 2009) were included as references in the PCoA. The population structure of the Pound collection was determined by the model-based clustering method implemented in the software program Structure (Pritchard et al. 2000). Based on the previous knowledge about the Pound collection, the number of clusters (K value) was set from 3 to 8. The permutation used 200,000 iterations after a burn-in period of 100,000. Ten independent runs were assessed for each fixed number of populations (K). Results of runs with the highest ln Pr ($G|K$) value of the ten runs were chosen and presented as bar plots. The 33 samples from the Ucayali population (Zhang et al. 2006b) and ten Lower Amazon Forastero accessions from the CATIE cacao germplasm collection (Zhang et al. 2009) were included as references in the analysis.

Sibship reconstruction and parentage analysis The program Colony (Wang 2004) was used for sibship reconstruction. Based on the multilocus SSR genotypes, Colony

used a moment estimator for pairwise relatedness between individuals and a maximum likelihood method to assign individuals sampled from a single population into full-sib families nested within half-sib families without parental information. The dropout and error rates for genotyping were both set at 0.02. The calculation was repeated three times using different computation seed numbers. Only those family members consistently assigned in three replications were considered as reconstructed sibships. For populations that have hypothesized parental clones (i.e., SCA, IMC, and NA), the probable mother trees were tested using parentage estimation. A likelihood-based method implemented in the program Cervus 3.0 (Marshall et al. 1998; Kalinowski et al. 2007) was used for computation. For each mother–offspring pair, the natural logarithm of the likelihood ratio [logarithm of the odds (LOD) score] was calculated. This score is the likelihood of maternity of a particular candidate parent, relative to an arbitrary individual. Critical LOD scores were determined for the assignment of maternity to a group of cacao accessions collected in 1937–1938, without knowing the maternity. The most probable single mother for each produced offspring was identified on the basis of the critical difference in LOD scores (Δ) between the most likely and next most likely candidate parent necessary for assignment at greater than 95% confidence.

Results

Individual identification

Out of the 612 examined accessions, a total of 432 accessions were correctly assigned to the six accession groups (home populations) at a 75% threshold value. The criterion for allocation was set such that when an individual's probability of being in one cluster was more than 0.75, it was classified into this cluster. In other words, an individual with more than a three-fourth proportion of genetic background in the cluster should be allocated into the corresponding population, and one with less than a three-fourth background in either of the two clusters should be treated as an ambiguous class member. The 180 ambiguously classified members were not used in subsequent analyses for sibship reconstruction (Table 1).

The rate of mislabeling varies among the six accessions groups (Table 1), ranging from 13.3% (Pound clones) to 42.3% (NA) with an average rate of 28.8%. Nevertheless, the result showed that the majority of accessions (71.2%) correctly corresponded to their membership in home populations (or families). An example of the results of the assignment test for the accessions from the IMC is presented in Table 2. Among the 82 IMC trees, 12 could not be assigned to their home population (or family) and thus were considered to be mislabeled and excluded from the subsequent analysis (Table 3).

The 432 trees that passed the assignment test were then subjected to identification of duplicated genotypes, using the method of multi-locus matching (Waits et al. 2001). The comparison of multi-locus microsatellite profiles identified a total of 45 duplicate groups comprised of 116 individuals, the majority of which are confirmed trees with the same name but planted in different farms, fields, and plots. Individuals labeled with different names in the same group who shared exactly the same alleles at all 15 loci were defined as synonymous sets because they shared exactly the same alleles in all 15 loci but were labeled with different names. From each duplicate set, only one individual was retained, and the rest were excluded from the subsequent analyses. As a result, a total of 316 accessions were retained for the analysis of genealogical relationship and population structure (Table 4).

Sibship reconstruction and parentage analysis

Colony inferred a total of 78 full-sib families nested within 48 half-sib families, consisting of 270 individuals (Table 5). The largest reconstructed families were found in the NA population, with 38 full-sib families nested within 22 half-sib families (indicating the maximum number of mother trees from which the pods were collected by Pound in 1937). The PA population was second in the number of reconstructed families, consisting of 28 full-sib families nested within 20 half-sib families. In contrast to the NA and PA accession groups, only two half-sib families were found in MO, IMC, and SCA, suggesting that the three accession groups were derived from either a two or a single mother tree. Among the half-sib families, the sizes of the reconstructed families vary greatly. An example of the reconstructed sibships is presented in Table 6. The largest half-sib family in the IMC population includes as many as 38 individual accessions.

Of the 13 P clones used as candidate parents, eight of them (P-1, P-2, P-5, P-7, P-10, P15, and P-27) were identified, at 95% confidence level, as probable mothers of 117 accessions in the collection. Among the P clones, seven clones were identified as candidate parents for 83 NA accessions, whereas one (P-27) was found as mother tree for 34 IMC accessions. No parentage was detected for the MO, PA and SCA accessions (Table 6).

Genetic structure and inter-population relationship

The relation among the Peruvian cacao germplasm, as well as the ten reference clones (Lower Amazon Forastero), was shown by PCoA (Fig. 2). The result demonstrates a clear

Table 3 Allelic diversity and probability of identity of the 15 microsatellite loci scored in the Parinari accessions (PA)

Genebank Designation	CIRAD accession name	N	H_O	H_E	Inbreeding coefficient[a]	PID-sib[b]
TCA16981	mTcCIR7	5	0.329	0.389	0.155	5.23E−01
TCA16980	mTcCIR6	13	0.462	0.654	0.293	5.12E−01
TCA16995	mTcCIR22	7	0.212	0.328	0.353	6.77E−01
TCA16996	mTcCIR24	16	0.190	0.312	0.392	9.34E−01
TCA16982	mTcCIR8	11	0.251	0.686	0.634	6.61E−01
TCCTREP	mTcCIR1	8	0.156	0.410	0.621	8.03E−01
TCA16985	mTcCIR11	18	0.584	0.817	0.285	5.17E−01
TCA16986	mTcCIR12	11	0.415	0.568	0.271	6.92E−01
TCA16988	mTcCIR15	14	0.554	0.824	0.328	5.01E−01
TCA271942	mTcCIR37	14	0.573	0.817	0.299	4.60E−01
TCA271826	mTcCIR33	18	0.642	0.805	0.202	5.25E−01
TCA16991	mTcCIR18	8	0.427	0.717	0.405	5.12E−01
TCA16998	mTcCIR26	11	0.376	0.743	0.494	5.33E−01
TCA271943	mTcCIR40	16	0.256	0.518	0.505	8.86E-01
TCA271958	mTcCIR60	10	0.573	0.831	0.311	5.72E-01
Mean over 15 loci		12	0.4	0.628	0.370	1.07E-06[c]

Values for PID-sib are given in scientific notation.

N total number of alleles, H_O observed heterozygosity, H_E expected heterozygosity, *PID-sib* probability of identity of siblings

[a] Definition of Inbreeding Coefficient follows Wright (1965)

[b] PID-sib Probability of identity among siblings follows the definition of Evett and Weir (1998)

[c] Accumulated PID-sib as the loci add up, i.e., the PID-sib value of the second locus is the product of PID-sib of the first two loci

population differentiation, which is stratified by the river systems in the Peruvian Amazon. The plane of the first two main PCoA axes, which accounted for 66.9% of total variation, showed that all the accessions were grouped into four clusters. The first cluster included NA and seven Pound clones, which was clearly separated from the other groups. The second cluster included Parinari and the lower Amazon accessions used as reference. The third cluster included the accessions of Scavina, Morona, and Ucayali accessions, and the three accession groups were partially overlapped. The IMC accessions and three P clones (P-4, P-18, and P-27) formed the fourth cluster, which falls in between the NA and Ucayali accessions (Fig. 3).

The results of Bayesian clustering analysis are compatible with that of PCoA. The population differentiation in the Pound collection changes only slightly at the assumption of different *K* values (Fig. 3). At $K=3–5$, the collection is predominantly composed of four clusters. The first cluster is comprised of the IMC accessions only. The second cluster consists of the NA and most of the Pound clones, which were collected from same site in 1937 and 1943, respectively. The third cluster is composed of MO,

Table 4 Summary of reconstructed half- and full-sib families for the five cacao accession groups in the Pound collection

Accession Group	No. of reconstructed half-sib fam.	No. of Reconstructed full-sib fam.	No. of examined individuals	No. of assigned individuals	Percent assigned (%)
IMC	2	4	52	48	92.3
MO	2	5	15	13	86.7
NA	22	38	127	112	88.2
PA	20	28	100	88	88
SCA	2	3	9	7	77.7
Total	48	78	303	268	

The 13 "P-clones" were collected as budwood in 1943 and thus were not included in the analyses of sibship reconstruction

Table 5 Examples of four reconstructed full-sib families nested within the two half-sib families in the Iquitos Mixed Calabacillo (IMC) accession group using microsatellite fingerprints

Half-sib[a]	Full-sib[b]	Reconstructed family members				
1	1–1	IMC 2	IMC 3	IMC 57	IMC 194	IMC 51
		IMC 105	IMC 66	IMC 73		
2	2–1	IMC 6	IMC 10	IMC 12	IMC 18	IMC 23
		IMC 27	IMC 30	IMC 33	IMC 38	IMC 42
		IMC 44	IMC 47	IMC 53	IMC 53	IMC 58
		IMC 60	IMC 61	IMC 63	IMC 65	IMC 68
		IMC 76	IMC 76	IMC 85	IMC 96	IMC 98
		IMC 107				
	2–2	IMC 11	IMC 13	IMC 55	IMC 58	IMC 59
		IMC 71	IMC 78			
	2–3	IMC 16	IMC 58	IMC 77	IMC 9	IMC 47
Undecided accessions[c]	IMC 20	IMC 104				

[a] Reconstructed half-sib families

[b] Reconstructed full-sib families nested in half-sib families

[c] Accessions that were not able to consistently fall into the same families over three replicated runs.

SCA, and Ucayali. The last cluster consists of the PA population and the accessions used as "controls" to Lower Amazon Forastero. At K=6–8, MO split from SCA and Ucayali and became a distinctive subgroup. At K=7, Parinari separated with Lower Amazon Forastero (Fig. 3). At all different K values (3–8), no genotypes in the IMC, MO, PA, and SCA accessions show significant levels of admixture, whereas in NA and the corresponding P clones, admixture was detected at K=5 or above.

Descriptive statistics

The main descriptive statistics are presented in Table 3. The number of alleles ranged from 5 to 18 across the 15 loci, with the average being 12. The observed heterozygosity (H_O) and gene diversity (H_E) vary greatly across the 15 loci. The locus mTcCIR24 has the second lowest H_O (0.156) and lowest H_E (0.312), whereas the locus mTcCIR33 has the highest H_O (0.642) and a high H_E (0.805; Table 2). This broad range of variation results from the large variation in the number of alleles per locus and allele frequency distribution in this collection. Loci with smaller numbers of alleles, such as mTcCIR24 and mTcCIR1, tend to have lower heterozygosity. The average observed heterozygosity (H_O) and gene diversity were 0.40 and 0.628, respectively. The inbreeding coefficient was positive for all of the 15 loci, ranging from 0.155 at mTcCIR7 to 0.634 at mTcCIR8, with an average of 0.37. Overall, these Peruvian cacao populations have a high level of genetic diversity in terms of allelic richness, which is higher than that in the reported cacao germplasm group from Ecuador (Zhang et al. 2008), Trinidad (Johnson et al. 2007), and Brazil (Sereno et al. 2006). However, the level of heterozygosity in this cacao collection is moderate, partially due to the fact that these populations are composed of a large number of siblings.

Discussion

Identification of putative mislabeled and duplicated genotypes

The cacao trees in the Pound collection were obtained about 70 years ago with limited information on their correct identity. Mislabeling of germplasm accessions has been acknowledged as a problem in the early cacao collections. We show that the assignment tests can verify if a given accession has a membership in the "home family." With the threshold probability at 0.75, a total of 180 individuals (29.4%) failed to be assigned to the home population signified by their accession name (Table 1). It must be pointed out that the decision of how stringent the threshold probability should be is subjective and depends on the purpose of the assignment test. If the threshold probability is set at 0.50, then the total number of identified mislabelings will be reduced to 163. In the present study, our objectives were to verify the identity and assess population structure in the five cacao populations collected by F.J. Pound in the 1930s, so we used a high stringent threshold for the assignment test, in order to exclude accessions with ambiguous assignment status. The correctly assigned trees will serve as reference true-to-type trees for future work on verification of the multiple trees of the same accessions in the ICG, T as well as in other international and national germplasm collections.

Population structure in the pound collection

At all K values (K=3–8), the Bayesian clustering analysis consistently assigned the SCA accessions to the Ucayali population. This result is in agreement with the PCoA plot where all nine SCA accessions fall in the same cluster as the Ucayali accessions (Fig. 2). The Scavina cacao was

Table 6 Likelihood assignment of 117 mother-offspring pairs in the Pound collection, based on 13 candidate mother trees (the "P clones")

Offspring	FM[a]	LOD	FM	FM	LOD	Offspring	FM	LOD
NA62(fp360)	P 1	7.95	NA191(fp18)	P 10	5.36	NA7/10(fp96)	P 15	5.61
NA144(fp166)	P 1	7.05	NA194(fp111)	P 10	8.16	NA141(fp6)	P 15	7.10
NA179(fp31)	P 1	10.10	NA678(fp547)	P 10	7.61	NA186(fp30)	P 15	9.03
NA181(fp1391)	P 1	10.13	NA708(fp100)	P 10	7.45	NA232(fp14)	P 15	4.94
NA183(fp7)	P 1	11.47	NA756(fp751)	P 10	7.62	NA686(fp92)	P 15	6.00
NA206(fp1392)	P 1	4.88						
NA226(fp1408)	P 1	3.47	NA13(fp608)	P 7	5.17	IMC6(fp1068)	P 27	12.55
NA227(fp651)	P 1	5.10	NA170(fp381)	P 7	7.88	IMC9(fp606)	P 27	4.27
NA229(fp379)	P 1	4.88	NA32(fp2494)	P 7	3.74	IMC10(fp612)	P 27	9.01
NA283(fp20)	P 1	11.14	NA719(fp98)	P 7	4.94	IMC11(fp107)	P 27	8.93
NA286(fp660)	P 1	8.08	NA867(fp186)	P 7	5.56	IMC12(fp105)	P 27	11.48
NA311(fp2)	P 1	8.37	NA888(fp316)	P 7	5.08	IMC13(fp607)	P 27	11.46
NA322(fp1174)	P 1	6.59				IMC16(fp105)	P 27	3.49
NA327(fp24)	P 1	9.30	NA90(fp779)	P 2	5.34	IMC18(fp760)	P 27	9.69
NA342(fp635)	P 1	11.15	NA168(fp788)	P 2	5.67	IMC23(fp574)	P 27	11.45
NA406(fp23)	P 1	7.10	NA184(fp661)	P 2	11.73	IMC27(fp614)	P 27	10.7
NA427(fp158)	P 1	12.68	NA187(fp235)	P 2	8.35	IMC30(fp586)	P 27	6.44
NA48(fp1131)	P 1	3.61	NA189(fp716)	P 2	7.67	IMC33(fp557)	P 27	9.82
NA507(fp8)	P 1	11.83	NA217(fp312)	P 2	11.73	IMC38(fp105)	P 27	10.99
NA528(fp112)	P 1	7.10	NA228(fp1)	P 2	11.73	IMC42(fp112)	P 27	13.56
NA687(fp545)	P 1	7.12	NA235(fp657)	P 2	9.01	IMC44(fp59)	P 27	6.40
NA689(fp544)	P 1	7.12	NA241(fp271)	P 2	10.37	IMC47(fp162)	P 27	13.4
NA697(fp87)	P 1	6.96	NA244(fp16)	P 2	9.70	IMC53(fp111)	P 27	13.24
NA720(fp554)	P 1	4.42	NA254(fp217)	P 2	11.73	IMC55(fp112)	P 27	6.02
NA730(fp332)	P 1	6.59	NA266(fp25)	P 2	11.73	IMC58(fp105)	P 27	8.23
NA733(fp274)	P 1	6.26	NA279(fp139)	P 2	5.24	IMC59(fp108)	P 27	9.58
NA753(fp583)	P 1	5.27	NA280(fp255)	P 2	11.73	IMC60(fp107)	P 27	13.61
NA79(fp385)	P 1	4.28	NA289(fp662)	P 2	11.73	IMC61(fp105)	P 27	11.80
NA824(fp22)	P 1	4.23	NA329(fp741)	P 2	11.73	IMC63(fp107)	P 27	11.80
NA841(fp56)	P 1	3.56	NA331(fp383)	P 2	7.31	IMC65(fp116)	P 27	4.85
NA870(fp736)	P 1	4.40	NA335(fp5)	P 2	8.35	IMC68(fp57)	P 27	10.86
NA916(fp2365)	P 1	7.89	NA337(fp141)	P 2	9.70	IMC71(fp267)	P 27	14.11
NA929(fp188)	P 1	6.89	NA399(fp358)	P 2	6.27	IMC76(fp112)	P 27	10.89
			NA435(fp659)	P 2	11.73	IMC77(fp232)	P 27	8.72
NA326(fp1183)	P 5	9.64	NA46(fp769)	P 2	4.0	IMC78(fp107)	P 27	9.17
NA702(fp819)	P 5	10.00	NA504(fp648)	P 2	7.31	IMC85(fp272)	P 27	8.92
NA705(fp15)	P 5	8.54	NA724(fp215)	P 2	7.34	IMC96(fp105)	P 27	13.67
NA715(fp555)	P 5	9.91	NA734(fp377)	P 2	7.34	IMC98(fp138)	P 27	9.15
NA771(fp27)	P 5	9.23	NA770(fp187)	P 2	6.66	IMC105(fp27)	P 27	8.16
			NA84(fp390)	P 2	9.70	IMC107(fp1065)	P 27	8.75
NA670(fp736)	P8	11.59	NA860(fp1167)	P 2	7.34			
NA672(fp168)	P8	9.18						

Critical LOD (the natural logarithm of the likelihood) ratio for assignment of maternity are 3.46 at >95% confidence and 0.79 at >80% confidence

[a] Female parents, refers to the putative mother trees Pound collected in 1943 when he revisited the same collecting sites of his 1937–1938 expeditions

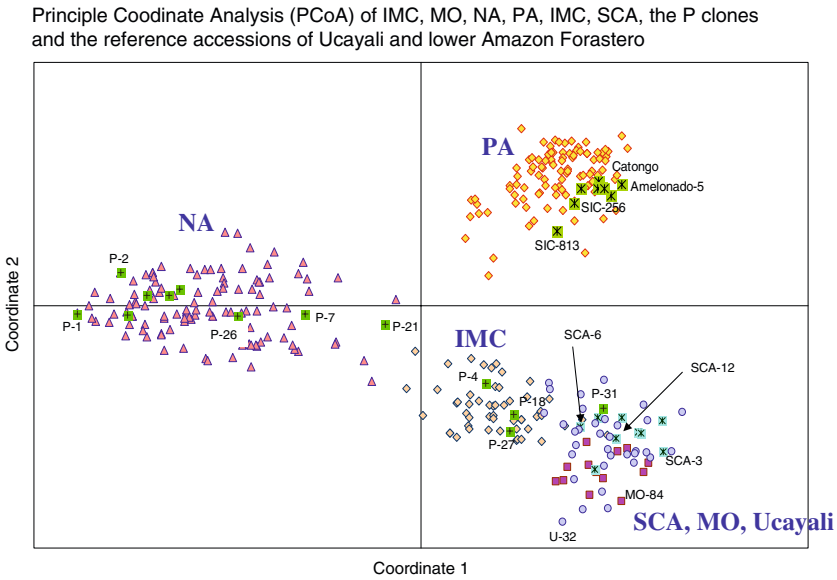

Fig. 2 PCoA plot of 359 cacao accessions, including 316 accessions from the Pound collection and 43 reference accessions. First axis=38.9% of total information and the second=28.3%

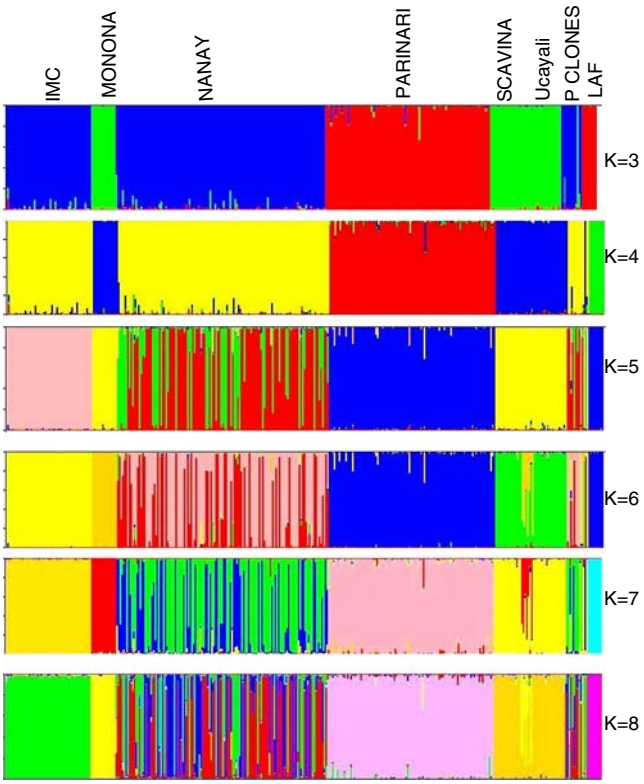

Fig. 3 Inferred clusters in the Pound collection using Structure, where *K* is the potential number of genetic clusters that may exist in the overall sample of individuals. Each *vertical line* represents one individual multilocus genotype. Individuals with multiple colors have admixed genotypes from multiple clusters. Each *color* represents the most likely ancestry of the cluster from which the genotype or partial genotype was derived. Clusters of individuals are represented by colors

reportedly collected near Rio Ucayali (Pound 1938; Bartley 2005). However, the exact location has been controversial. Bartley (1994) considers that it is likely that the material comes from an area between the towns named Contamana and Pucallpa (on the Ucayali River; Fig. 1). In 1943, Pound revisited the same place and collected a clone, which is identified as "P 31" from the same group. Alternatively, it has been suggested that the Scavina originates from a place around the Nanay River, near Iquitos, and was named after the farm of the Escavino family (International Cacao Germplasm Database, http://www.icgd.reading.ac.uk). Our result show that "P-31" is the only clone that is assigned in the same cluster with SCA and Ucayali accessions, which supports that the SCA accessions and clone P-31 both originated from the location near Contamana on the Ucayali River (Fig. 1). Reconstructed sibship for the SCA accessions indicates that the SCA accessions probably came from two trees, and SCA 3, SCA 6, SCA 9, SCA 12, and SCA19 came from a single pod. However, parentage analysis shows that "P-31" is neither the male nor the female parent of SCA (data not shown).

Results of both Bayesian clustering method (Fig. 3) and PCoA plot (Fig. 2) both show that the MO accessions overlapped with the Ucayali clones, indicating that the MO accessions were likely collected from a region near the Ucayali river as well. The collection site should be different from the place where the SCA accessions and other Ucayali clones were originated because MO split from the SCA and Ucayali population at $K=6$ in the Bayesian clustering analyses. Nonetheless, the group's name "Morona" was questionable. In fact, Pound did not mention from which river the pods were collected. Instead, he only mentioned the "region of Rio Morona." This may explain the fact that

the list sent to Trinidad from Pound's expedition did not include any reference of the Morona River (Bartley 2005).

The PA population is another group with ambiguous and controversial passport information. The collecting report shows that the name "Parinari" refers to a region rather than a specific place in the Marañon River (Pound 1945; Bartley 2005), but the exact location remains unknown. Among the Upper Amazon Forastero groups, the PA population was reported as the most distinctive (Bartley 2005; Risterucci et al. 2001) and has a lower level of heterozygosity (Risterucci et al. 2001). This uniqueness was also supported by the results of random amplification of polymorphic DNA analysis (Sounigo et al. 2000). Based on the morphological dissimilarity of the PA population with the other native populations in the Marañon river system, Bartley (2005) suggested that the PA group likely belonged to a cultivated population descended from imported planting material from Brazil. The high level of similarity between the PA population and some Brazilian cacao germplasm was reported by Motamayor et al. (2008). In their report, the germplasm from Ji-Paraná river, Brazil was classified in the same cluster with the PA accessions and was named as "Marañón cluster". The present result confirmed that the PA population has a similar genetic profile as that of the Lower Amazon Forastero from Brazil (Figs. 2 and 3), supporting the hypothesis that the PA accessions were likely cultivated materials originated from Brazil. The result of 27 reconstructed half-sib families shows that the pods were collected from a relatively large number of trees, possibly resulting from introduced seed families.

The IMC, according to Pound (1938), was collected from an island in front of the city of Iquitos. The putative parental trees were collected as budwoods for vegetative propagation by Pound in 1943. The resulting clones were included in the "P" series in the Pound collection. Pods harvested from the P clones were mixed together, and the derived trees were called IMC (Pound 1938; Bartley 2005). However, it was not clear which P clone (or clones) were the parental tree. Based on the similarity of flower and pod characteristics between some accessions of IMC and P clones, Bartley suggested that clones P-18 and P-21 were possible parents for IMC. The present study showed that the IMC accessions formed a distinctive cluster, together with some four P clones (P-4, P-18, P-21, and P-27). The result of sibship reconstruction suggests that the IMC accessions were derived from two half-sib families (Table 6). However, parentage analysis only identified P-27 as the maternal parent, with a total of 34 assigned mother–offspring pairs (Table 6).

The NA group is the largest group in the Pound collection with a total of 708 plants originally planted at the Marper farm (International Cacao Germplasm Database, http://www.icgd.reading.ac.uk). Pound (1938) mentioned that the pods of NA were probably collected from 14 mother trees, which were free of witches' broom disease. Later, in 1943, materials for vegetative propagation were collected from 22 trees in the same area (Pound 1945; Bartley 2005). The PCoA confirmed that clones P-1, P-2, P-5, P-7, P-8, P-10, P-15, P-21, and P-26 all originated from the Nanay river area (Fig. 2). Their membership in the NA population was also supported by the assignment test (Fig. 3). Moreover, out of the nine P clones, seven were identified as maternal parents for 83 NA accessions.

Implications for cacao germplasm conservation and crop improvement

The Pound collection is the first cacao collection established in South America. Much of the plant material in these collections is old, and some labels have been lost. The mislabeling problem restricts the sharing of information and materials among cacao researchers and hampers the use of cacao germplasm in breeding programs. On the other hand, this collection has the much needed genetic variation for cacao breeding, specifically disease resistance. Three major cacao diseases, witches' broom, frosty pod rot, and black pod, constitute a serious threat to the livelihoods of cacao farmers in the Americas. Cacao production in the Americas has dropped by 75%, largely due to these three diseases (Bowers et al. 2001). The challenges posed by these devastating diseases create a need to explore new sources of resistance for the present and future genetic improvement of cacao. Our results show that the multi-locus SSR fingerprinting data, in combination with the model-based statistical method for individual assignment, are powerful tools to address issues in individual and population assignment and sibship reconstruction. This study significantly improves our understanding of the earliest cacao collection from the Upper Amazon by clarifying and adding the passport data, which were largely missing in the records. In addition, true identities of many mislabeled trees were clarified by using multi-locus matching, cluster analyses, and assignment (data not shown). The information on population structure, genealogical relationship, and genotype identity in this germplasm group will be submitted to the International Cocoa Germplasm Database (http://www.icgd.reading.ac.uk/acknowledgements.php) hosted by the University of Reading and CocoaGenDB (http://cocoagendb.cirad.fr/) hosted by Centre de Coopération Internationale en Recherche Agronomique pour le Développement, France. This information will enable improvements in the efficiency and accuracy of cacao germplasm conservation.

Inadequate representation of genetic diversity in the germplasm collections is another major constraint to

efficiently conserve the genetic diversity of cacao. The wild germplasm in the existing collections was primarily acquired during a few collecting expeditions, with the majority of them obtained in the 1930s–1940s. As the current study shows, Pound's expedition probably only covered the area around Iquitos and one or two sites in Rio Ucayali. A fraction of this collection is probably cultivated varieties or derived from cultivated varieties. New collecting expeditions into the upper Marañon and its tributaries are necessary to fill diversity gaps in the collections. A recent collecting expedition carried out in the northwestern part of Peruvian Amazon proves that there remains a large amount of untapped genetic variation in Peru, which will significantly expand the genetic foundation for cacao genetic improvement.

Acknowledgements We wish to thank Stephen Pinney, Eric Tillson, Emily Leamy, Elizabeth Gingold, and Sarah Gingold for assisting with the microsatellite genotyping. Enrique Arevalo is thanked for providing leaf samples of the Ucayali clones. Antoinette Sankar is thanked for performing the DNA extractions at CRU, Trinidad. Special thanks are due to Dr. Ainong Shi and two anonymous reviewers for their review of the manuscript and suggestions for the revision.

References

Aikpokpodion PO, Adetimirin VO, Ingelbrecht I, Schnell RJ, Kolesnikova-Allen M (2005) Assessment of Genetic diversity of cacao, *Theobroma cacao* L., Collections in Nigeria using simple sequence repeat markers. In: Malaysian Intl Cocoa Conf Sustainable Cocoa Economy through Increase in Productivity, Efficiency and Quality, pp 83–86

Bartley BGD (1994) A review of cacao improvement: fundamentals, methods and results. Proc Intl Workshop Cocoa Breeding Strategies, pp 3–16. Available at http://ingenic.cas.psu.edu/proceedings.htm

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

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

Bowers JH, Bailey BA, Hebbar PK, Sanogo S, Lumsden RD (2001) The impact of plant diseases on world chocolate production. Plant Health Progress (online). doi:10.1094/PHP-2001-0709-01-RV

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

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

Cuatrecasas J (1964) Cacao and its allies: A taxonomic revision of the genus *Theobroma*. Contributions from the United States National Herbarium, vol. 35. Smithsonian Institution Press, Washington, DC, pp 375–614

Dias LAS (2001) Genetic improvement of cacao. Editora Folha de Vicosa Ltda. Trans. Abreu-Reichart CE, Vicosa, MG, Brazil, supported by FAO (2005). Available at http://ecoport.org/ep?SearchType=earticleView&earticleId=197

Efombagan IB, Motamayor JC, Sounigo O, Eskes AB, Nyasse S, Cilas C, Schnell RJ, Manzanares-Dauleux M, Kolesnikova-Allen M (2008) Genetic diversity and structure of farm and genebank accessions of cacao (*Theobroma cacao* L.) in Cameroon revealed by microsatellite markers. Tree Genetics & Genomics 4:821–831. doi:10.1007/s11295-008-0155-z

Evett IW, Weir BS (1998) Interpreting DNA evidence: statistical genetics for forensic scientists. Sinauer, Sunderland, MA

Gómez-Pompa A, Flores JS, Fernandez MA (1990) The sacred cacao groves of the Maya. Latin Am Antiq 1:247–257. doi:10.2307/972163

Henderson JS, Joyce RA, Hall GR, Hurst WJ, McGovern PE (2007) Chemical and archaeological evidence for the earliest cacao beverages. Proc Natl Acad Sci U S A 104:18937–18940

Hunter RJ (1990) The status of cocoa (*Theobroma cacao*, Sterculiaceae) in the western hemisphere. Econ Bot 44:425–439

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

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

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol Ecol 16:1099–1006

Kennedy AJ, Mooleedhar V (1993) Conservation of cocoa in field genebanks—the International Cocoa Genebank, Trinidad. In: Proceedings of the International Workshop on Conservation, Characterization and Utilization of Cocoa Genetic Resources in the 21st Century. The University of the West Indies, Cocoa Research Unit, Port-of-Spain, Trinidad and Tobago, pp 21–23

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

Lanaud C, Motamayor JC, Risterucci AM (2001) Implications of new insight into the genetic structure of *Theobroma cacao* L. for breeding strategies. In: Proceedings of the International Workshop on New Technologies for Cocoa Breeding. Ingenic Press, Kota Kinabalu, Malaysia, London, pp 89–107

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

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

Lockwood C, End M (1993) History, technique and future needs for cacao collection. In: Proceedings of the Workshop on the Conservation, Characterization and Utilization of Cacao Genetic Resources in the 21st Century, 13–17 September, 1992, Port-of-Spain. Trinidad and Tobago, The University of the West Indies, The Cacao Research Unit, pp 1–14

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

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

Motamayor JC, Lachenaud P, Wallace J, Loor G, Kuhn DN, Brown JS, Schnell Ii RJ (2008) Geographic and genetic population differentiation of the Amazonian chocolate tree. PLoS One 3: e3311. doi:10.1007/s12042-008-9011-4

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

Posnette AF (1986) Fifty years of cocoa research in Trinidad and Tobago. Cocoa Research Unit, University of the West Indies, St. Augustine, Trinidad

Pound FJ (1938) Cacao and witches' broom disease (*Marasmius perniciosus*) of South America. Arch Cacao Res 1:20–72

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

Pound FJ (1945) A note on the cocoa population of South America. In: Report and Proceedings of the 1945 Cocoa Conference. London, pp 131–133

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

Risterucci AM, Grivet L, Ngoran JA, Pieretti I, Flamen MH et al (2000) A high-density linkage map of *Theobroma cacao* L. Theor Appl Genet 101:948–955

Risterucci AM, Eskes B, Fargeas D, Motamayor JC, Lanaud C (2001) Use of microsatellite markers for germplasm identity analysis in cocoa. In: Proc Intl Workshop on New Technologies and Cocoa Breeding, pp 25–33. Available via http://ingenic.cas.psu.edu/proceedings.htm.

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

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

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

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

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

Sounigo O, Christopher Y, Ramdahin S, Umaharan R, Sankar A (2000) Evaluation and use of the genetic diversity present in the International Cacao Genebank (ICG,T), in Trinidad. In: Proceedings of the 3rd INGENIC International Workshop on the New Technologies and Cocoa Breeding, Kota Kinabalu, Sabah, Malaysia.

Wadsworth RM, Ford CS, End MJ, Hadley P (eds) (1997) International cacao germplasm database. The London International Financial Futures and Options Exchange (LIFFE), London, UK.

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

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

Wright S (1965) The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution 19:395–420

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

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

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 Equador. Conserv Genet 9:327–337. doi:10.1007/s10592-007-9345-8

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