academicJournals

Vol. 13(21), pp. 2127-2136, 21 May, 2014
DOI: 10.5897/AJB2013.13331
Article Number: A0B539344751
ISSN 1684-5315

http://www.academicjournals.org/AJB

African Journal of Biotechnology

Full Length Research Paper

Verification of genetic identity of introduced cacao germplasm in Ghana using single nucleotide polymorphism (SNP) markers

Jemmy Takrama[1], Ji Kun[2,3], Lyndel Meinhardt[2], Sue Mischke[2], Stephen Y. Opoku[1], Francis K. Padi[1] and Dapeng Zhang[2*]

[1]The Cocoa Research Institute of Ghana (CRIG), PO Box 8, New Tafo-Akim, Ghana.
[2]USDA/ARS, Beltsville Agricultural Research Center, SPCL, 10300 Baltimore Avenue, Bldg. 001, Rm. 223, BARC-W, Beltsville, MD 20705, USA.
[3]Southwest University, No.2 Tiansheng Road, Beibei, Chongqing, 400715, P.R. China.

Received 29 September, 2013; Accepted 8 May, 2014

Accurate identification of individual genotypes is important for cacao (*Theobroma cacao* L.) breeding, germplasm conservation and seed propagation. The development of single nucleotide polymorphism (SNP) markers in cacao offers an effective way to use a high-throughput genotyping system for cacao genotype verification. In the present study, high-throughput genotyping with SNP markers was used to fingerprint 160 cacao trees in the germplasm collection at the Cocoa Research Institute of Ghana (CRIG). These accessions had been originally introduced from international germplasm collections. The multilocus SNP profiles, generated by the Sequenom Mass Spectrometry platform, were compared with the SNP profiles of reference trees maintained in the international cacao collections. The comparison unambiguously identified mislabeled trees. For materials introduced as hybrid seeds without an available reference genotype, parentage analysis and model-based assignment were applied to verify their recorded parentage and genetic background. Our study shows that a small set of polymorphic SNP markers can provide a robust and accurate result for cacao genotype identification. This protocol can be applied for large-scale genotyping of cacao as well as for many other crops.

Key words: Cacao, conservation, chocolate, DNA fingerprint, molecular marker, tropical plant, off-type, true-to-type, West Africa.

INTRODUCTION

Cacao (*Theobroma cacao* L.) is an important tropical tree crop that provides raw ingredients for the chocolate confectionery industries. This global commodity has an annual production that exceeded 4 million tons in 2010,

*Corresponding author. E-mail: Dapeng.Zhang@ars.usda.gov. Tel: 1 301 504 7477. Fax: 1 301 504 1998.

Abbreviations: SNP, Single nucleotide polymorphism; **CRIG,** Cocoa Research Institute of Ghana; **EST,** express sequence tag; **PIC,** polymorphic information index.

of which 75% was produced in West Africa. Ghana alone produced 850,000 tons of cacao, accounting for 21% of the world's total output in 2010 (FAOSTAT, http://faostat3.fao.org/home/index.html). Cacao originated in the Amazon rainforest in South America and was domesticated by the Maya and Olmec peoples at least 3000 years ago (Cuatrecasas, 1964; Wood and Lass, 1985; Bartley, 2005; Powis et al., 2011). Beginning in the late 1800's and continuing into recent times, cacao has been repeatedly introduced into Ghana. Germplasm was ultimately deposited in an *in situ* germplasm bank at the Cocoa Research Institute of Ghana (CRIG) in Tafo, which currently houses over 1200 clones of various genetic origins (Edwin and Masters, 2005; Adu-Ampomah et al., 2006). Cacao is an outcrossing species (Wood and Lass, 1985) and germplasm is conserved as clonally propagated trees in field genebanks. Cacao germplasm collections have been shown to contain a variety of mislabeled individuals, and mislabeling is estimated at 15 to 44% in global cacao collections (Motilal and Butler, 2003; Motilal, 2004; Sounigo et al., 2006; Takrama et al., 2005). Misidentifications can be attributed to multiplicity of introductions and transfers of plants from point-of-collection to establishment in early holding sites, and to subsequent recollection of budwood and repropagation of material for establishment. The potential for human error during plot demarcations and planting also contributes to this problem. Molecular markers have been used to characterize cacao germplasm since the 1980s (Guiltinan et al., 2008). Mislabeled accessions were identified by using dominant markers (Figueira et al., 1994; Whitkus et al., 1998; Sounigo et al., 2005) as well as codominant DNA markers such as restriction fragment length polymerphisms (Lerceteau et al., 1997; N'Goran et al., 2000). The development of microsatellite markers (Lanaud et al., 1999) greatly increased the efficiency and capacity for cacao fingerprinting and resulted in a wide application of cacao genotype identification (Aikpokpodion et al., 2005; Motilal and Butler, 2003; Efombagan et al., 2008; Motilal et al., 2010).

Recent progress in the development of cacao genomic resources has led to the use of single nucleotide polymorphisms (SNPs) as markers for cacao DNA fingerprinting, since SNPs are the most abundant class of polymorphisms in plant genomes (Buckler and Thornsberry, 2002). Compared with SSR markers, the assays of SNPs can be done without requiring separation of DNA by size, and therefore can be automated in an assay-plate format or on microchips. The diallelic nature of SNPs results in a much lower error rate in allele calling, and the genotyping can be multiplexed, allowing quicker completion at a lower cost than with SSRs. In recent years, SNP markers have been developed to assist cacao breeding and germplasm management (Allegre et al., 2012; Kuhn et al., 2012). TaqMan-based SNP assays have been developed for cacao genotyping under field conditions (Livingstone et al., 2012; Takrama et al., 2012). Using a set of SNP markers derived from express sequence tag (EST) databases, Ji et al. (2013) characterized farmer selections of cacao from Nicaragua and Honduras and demonstrated that the SNP markers constitute a cost-effective marker resource suitable for cacao germplasm characterization. Results for genotyping with SNPs can be compared across different genotyping platforms and laboratories, facilitating the integration and interpretation of SNP data across different genebanks in various cacao-producing countries. The objective of the present study was to test the efficacy of using high-throughput SNP genotyping for molecular characterization of cacao and to assess the extent of mislabeling, or off-type, in the CRIG cacao germplasm collection.

MATERIALS AND METHODS

Sample preparation and SNP genotyping

One hundred and sixty (160) trees from the CRIG germplasm collection, representing 39 cacao accessions (each accession included one to five trees), were sampled for this experiment. Samples were collected from eight plots in the germplasm collection: D8 (2), L6 (34), M6 (32), M6 Ext. (5), Q6 (67), Q6 Ext.2 (8), Q6 Ext.4 (9), and V3 (3) (Table 1). Two young leaves were collected from each individual cacao tree and each sampled branch was tagged for potential revisiting. Both accession name and DNA extraction number were used to label each sample. DNA was extracted from the CRIG samples using the CTAB DNA Extraction Protocol (Doyle and Doyle, 1990). In addition, one hundred international clones were used as references. Preparation of DNA samples for the reference international clones was described in Zhang et al. (2009a; b). DNA concentration was determined with a NanoDrop spectrophotometer (Thermo Scientific, Wilmington, DE). Based on the level of polymorphism and on their distribution across the ten chromosomes in cacao, 54 SNP markers were selected from 1560 candidate SNPs that had been developed using cDNA sequences from a wide range of cacao tissues (Argout et al., 2008). SNP genotyping was performed at the Human Genetics Division Genotyping Core facility, Washington University, St. Louis, using MALDI-TOF mass spectrometry (Sequenom, Inc., San Diego, CA). The heterozygosity and polymorphic information index (PIC) of these SNP markers has been reported by Ji et al. (2013).

Data analysis

Key descriptive statistics for measuring the informativeness of the SNP markers were calculated, including observed heterozygosity, expected heterozygosity, minor allele frequency, inbreeding coefficient and probability of identity (Evett and Weir, 1998; Waits et al., 2001). The program GenAlEx 6.2 (Peakall and Smouse, 2006; 2012) was used for computation. For the identification of mislabeling (off-types), SNP profiles of 100 reference trees maintained in the International Cacao Genebank, Trinidad (ICG,T) were used in the analysis. The genetic identity of the 100 reference trees has been characterized by both SNP (D. Zhang, USDA/ARS, Beltsville, personal communication) and SSR fingerprinting (Zhang et al., 2009b; Motilal et al., 2010; Johnson et al., 2009). Pairwise multilocus matching was applied among each pair of individual trees, including the reference trees from the international germplasm collections, using the same program. Accessions with same names as the reference trees, but not matching them, were declared off-

Table 1. List of the 39 cacao accessions (represented by 160 trees), their field plot and tree stand, from Ghana cacao germplasm collection.

Sample code	Accession name	Number of trees	Plot number	Tree stand
1	ALPHAB 36	5	M6 ext	26; 9; 19; 28; 29
2	AMAZ 3-2	3	Q6 ext 2	225; 260; 243
3	CATONGO	1	Q6	1618
4	EQX 3338	3	Q6	1354; 1536; 1355
5	ICS 43	4	Q6 ext 4	10; 340; 12; 11
6	ICS 95	5	Q6	56; 143; 368; 429; 182
7	IMC 67	5	L6	364; 331; 250; 215; 331
8	IMC 76	5	Q6	738; 737; 729; 727; 724
9	MAN 15-60	5	Q6 ext 2	7; 13; 46; 85; 83
10	MOCORONGO	5	Q6	366; 367; 305; 428; 427
11	NA 33	5	M6	40; 22; 20; 19; 18
12	NA 79	2	Q6	959; 957
13	NA 79	3	Q6	960; 952; 958
14	P 30	5	Q6	906; 904; 901; 900; 898
15	PA 121	4	L6	102; 1396; 1282; 1453
16	PA 150	5	Q6	1019; 1013; 1012; 1010; 1009
17	PA 151	4	L6	687; 850; 686; 689
18	PA 300	5	L6	703; 737; 738; 702; 1246
19	PA 303	4	Q6	777; 1315; 1493; 892
20	PA 7	4	Q6	838; 839; 840; 844
21	PA 70	5	Q6	142; 205; 268; 403; 463
22	PA 88	4	Q6	299; 236; 360; 532
23	POUND 10	5	L6	1146; 843; 996; 995; 994
24	POUND 15	5	L6	1152; 1259; 780; 746; 1553
25	POUND 7	5	Q6	670; 672; 521; 673; 669
26	SCA 12	5	L6	254; 327; 253; 252; 211
27	SCA 6	5	Q6 ext 4	279; 278; 277; 276; 281
28	SCA 9	3	L6	194; 155; 154
29	SUL7	1	Q6	583
30	T16/613	2	D8	130; 128
31	T16/613	3	M6	131; 150; 151
32	T17/524	1	V3 1st planting	515
33	T60/877	5	M6	470; 473; 452; 450; 449
34	T63/967	5	M6	14; 15; 35; 17; 16
35	T63/971	5	M6	8; 9; 12; 10; 11
36	T65/238	5	Q6	1295; 1299; 1301; 1306; 1307
37	T65/326	2	V3 1st planting	745; 180
38	T79/501	4	M6	162; 143; 118; 121
39	T85/799	5	M6	250; 252; 231; 230; 229

types. For the multilocus matching, the option to ignore missing data was used. Discriminating power of the SNP loci was computed using the probability of identity (PID) (Waits et al., 2001) option implemented in the same computer program.

For accessions without a reference tree but with known pedigree record (for example, breeding lines selected in Ghana's breeding program), the genetic identities were verified using parentage analysis and/or model-based assignment test. An example is the T clones (Table 1) that were hybrid families introduced into West Africa in 1944. Since these were the products of hybridization in Trinidad in the early 1940s, and the seed families were evaluated and selected in Ghana (Posnette, 1986), there are no existing reference trees available from the international cacao collections. Nonetheless, because pedigree records for these selections are available, the T clones were used as "offspring" and their parental clones in ICG,T were verified according to the recorded pedigree (Lockwood and Gyamfi, 1979). 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 parent-offspring pair, the natural logarithm of the likelihood ratio (LOD

score) was calculated.

Critical LOD scores were determined for the assignment of parentage to a group of individuals without knowing the maternity or paternity. Simulations were run for 10000 cycles with the assumption that 80% of candidate parents were sampled and a total of 80% of loci were typed, with a typing error rate of 0.5%. The most probable single mother (or father) for each offspring was identified on the basis of the critical difference in LOD scores (Δ) between the most likely and next most likely candidate parent at greater than 95% or 80% confidence (Marshall et al., 1998; Kalinowski et al., 2007).

For accessions lacking a reference tree, assignment test was applied to infer their hidden membership to a known population or germplasm group, using a model-based clustering analysis implemented in the STRUCTURE software program (Pritchard et al., 2000). SNP profiles of 100 reference accessions were included in the analysis. These 100 accessions were taken from six known Forastero germplasm groups, including Amelonado, Scavina (SCA) and Ucayali, Iquitos Mixed Calabacillo (IMC), Morona (MO), Nanay (NA) and Parinari (PA). Classification of these accessions have been reported by (Motamayor et al., 2008; Zhang et al., 2009b). The number of clusters (K-value, which indicated the number of sub-populations of the program attempted to find) was set from two to ten, and the analysis was carried out without assuming any prior information about the genetic group or geographic origin of the samples. Ten independent runs were assessed for each fixed number of clusters (K). The ΔK value was computed to detect the most probable number of clusters (Evanno et al., 2005). Of the 10 independent runs, the one with the highest Ln Pr (X|K) value (log probability or log likelihood) was chosen and represented as a bar plot.

RESULTS

Descriptive statistics of the SNP markers

In total, 53 SNP markers were reliably scored, as assessed by markers producing less than 10% missing genotypic data. Marker TcSNP 174 failed to generate SNP data thus was excluded in subsequent data analysis. The descriptive statistics of the remaining 53 SNP loci are presented in Table 2. The 53 SNP markers were polymorphic across the 39 cacao accessions. The mean expected heterozygosity was 0.343 and the observed heterozygosity was 0.274. An inbreeding coefficient with an average of 0.218 was observed.

Multilocus matching

Comparison of the multilocus SNP profiles with the reference accessions identified seven intraclonal mislabelings in accessions NA 79, PA 150 and IMC 76 (Figure 1). The multilocus matching also found that AMAZ 3-2 and PA 303 were mislabeled. These trees were defined as off-type or homonymous mislabeling because they shared the same name with the reference tree but differed in multilocus SNP profiles. In this experiment the mismatched accessions differed at a minimum of five loci. With all 53 loci considered, the combined probability of identity was in the order of 10^{-9} (Table 2). Overall, the procedure of multilocus matching with known reference trees led to the identification of 149 true-to-type trees out of 160 tested samples. Based on the verified result, 39 samples (a single sample from each accession) were used in the subsequent analyses of population structure and genealogical relationships. Among these 39 samples, the status of the nine T clones could not be decided solely based on multilocus matching, because they were selections made in Ghana and no reference trees were available. For these trees, assignment test and parentage analysis were applied to verify their genetic identity.

Assignment test

Based on the value of delta K, the model-based approach of STRUCTURE indicated K=5 as the most probable number of genetic clusters. The 39 tested cacao accessions from the Ghana cacao collection, as well as the 100 reference accessions, were stratified as germplasm groups of Amelonado, IMC, SCA/Ucayali, Morona, Nanay and Parinari, respectively (Figure 2). The assignment result largely agreed with the previously classified germplasm groups (Figure 2; Zhang et al., 2009b; Motamayor et al., 2008) except that the germplasm groups of SCA/Ucayali and Morona were not separated. The assigned memberships for all the tested trees from Ghana were compatible with their known parentage germplasm groups (Figure 2). The assignment test of the T clones confirmed their recorded parental germplasm groups, as shown in Figure 2. The parental groups of PA and IMC were clearly reflected in the admixed ancestry profiles of T60, T63, T65 and T79. A full genetic background of IMC was revealed for accession T85/799, supporting its recorded parentage of IMC 60 and NA 34 (a member of the IMC germplasm group; Motamayor et al., 2008). In addition, admixed ancestry of IMC and Amelonado was revealed for T16/613 family, which not only supported the recorded parentage of IMC 24, but also detected that the other parent came from the Amelonado group.

Parentage analysis

Of the eight candidate parent-offspring relationships, the results of parentage inference confirmed six pairs at the 95% confidence level and one pair (NA 34 - T85/799) at the 80% confidence level (Table 3). For offspring T16/613, only one parent (Amelonado 22) was identified at the >80% confidence level because the reference genotype of maternal parent IMC 24 was not available. The result of parent-offspring assignment supported the outcome of model-based clustering analysis by the STRUCTURE program (Figure 2).

DISCUSSION

Multilocus matching

Over 50 cacao germplasm collections are present worldwide

Table 2. Observed and expected heterozygosities, inbreeding coefficient, minor allele frequency and probability of identity of the 53 SNP loci scored on 39 cacao accessions from the Ghana Cacao germplasm collection.

Locus	Ho	He	Inbreeding coefficient	Minor allele frequency	PID-sib
TcSNP75	0.091	0.127	0.285	0.068	0.879
TcSNP90	0.091	0.127	0.285	0.068	0.879
TcSNP139	0.364	0.483	0.248	0.409	0.604
TcSNP144	0.523	0.500	-0.046	0.489	0.594
TcSNP150	0.310	0.436	0.290	0.321	0.635
TcSNP151	0.273	0.416	0.345	0.295	0.649
TcSNP189	0.295	0.425	0.305	0.307	0.642
TcSNP193	0.364	0.416	0.127	0.295	0.649
TcSNP226	0.318	0.397	0.198	0.273	0.662
TcSNP230	0.409	0.499	0.180	0.477	0.594
TcSNP242	0.488	0.447	-0.093	0.337	0.628
TcSNP309	0.500	0.375	-0.333	0.250	0.678
TcSNP329	0.295	0.312	0.052	0.193	0.725
TcSNP364	0.045	0.087	0.476	0.045	0.916
TcSNP372	0.182	0.201	0.097	0.114	0.814
TcSNP448	0.025	0.117	0.787	0.063	0.888
TcSNP469	0.227	0.298	0.236	0.182	0.736
TcSNP529	0.364	0.474	0.233	0.386	0.610
TcSNP534	0.477	0.487	0.021	0.420	0.602
TcSNP560	0.385	0.479	0.197	0.397	0.607
TcSNP577	0.366	0.442	0.172	0.329	0.631
TcSNP591	0.341	0.487	0.300	0.420	0.602
TcSNP602	0.341	0.456	0.253	0.352	0.622
TcSNP619	0.295	0.487	0.394	0.420	0.602
TcSNP702	0.295	0.469	0.370	0.375	0.614
TcSNP723	0.114	0.146	0.224	0.080	0.862
TcSNP731	0.278	0.424	0.345	0.306	0.643
TcSNP799	0.045	0.044	-0.023	0.023	0.956
TcSNP823	0.318	0.491	0.352	0.432	0.600
TcSNP872	0.317	0.495	0.360	0.451	0.597
TcSNP878	0.114	0.184	0.381	0.102	0.829
TcSNP886	0.227	0.416	0.453	0.295	0.649
TcSNP891	0.000	0.480	1.000	0.400	0.606
TcSNP899	1.000	0.500	-1.000	0.500	0.594
TcSNP928	0.024	0.024	-0.012	0.012	0.977
TcSNP998	0.231	0.416	0.445	0.295	0.649
TcSNP999	0.000	0.176	1.000	0.098	0.836
TcSNP1038	0.045	0.165	0.725	0.091	0.845
TcSNP1060	0.386	0.339	-0.141	0.216	0.704
TcSNP1063	0.091	0.127	0.285	0.068	0.879
TcSNP1111	0.114	0.146	0.224	0.080	0.862
TcSNP1126	0.045	0.087	0.476	0.045	0.916
TcSNP1159	0.068	0.066	-0.035	0.034	0.936
TcSNP1253	0.279	0.357	0.218	0.233	0.691
TcSNP1280	0.364	0.496	0.267	0.455	0.596
TcSNP1309	0.465	0.422	-0.103	0.302	0.645
TcSNP1331	0.326	0.487	0.331	0.419	0.602
TcSNP1378	0.455	0.351	-0.294	0.227	0.695
TcSNP1439	0.250	0.363	0.312	0.239	0.686
TcSNP1442	0.295	0.442	0.331	0.330	0.631

Table 2. Contd.

TcSNP1453	0.116	0.110	-0.062	0.058	0.895
TcSNP1458	0.386	0.479	0.194	0.398	0.607
TcSNP1484	0.523	0.479	-0.091	0.398	0.607
Mean	0.274	0.343	0.218	0.262	$6.5x10^{-9*}$

* Accumulated PID_sibs for 53 SNP locus combinations.

Name	Field stand	Genotype	139	144	150	151	189	193	230	242	309	529	534	591	602	619	702	886	1060	1253	1280	1378	1484
PA 150	1019Q6	√	GG	AC	GG	CC	GG	AA	GG	CC	TT	AC	CT	AC	CT	TT	CC	CT	CC	TT	AA	TT	AG
PA 150	1013Q6	√	GG	AC	GG	CC	GG	AA	GG	CC	TT	AC	CT	AC	CT	TT	CC	CT	CC	TT	AA	TT	AG
PA 150	1012Q6	√	GG	AC	GG	CC	GG	AA	GG	CC	TT	AC	CT	AC	CT	TT	CC	CT	CC	TT	AA	TT	AG
PA 150	1010Q6	√	GG	AC	GG	CC	GG	AA	GG	CC	TT	AC	CT	AC	CT	TT	CC	CT	CC	TT	AA	TT	AG
PA 150	1009Q6	Off-type	TT	AC	GG	CC	GG	AA	AA	CT	TT	CC	CT	AA	CC	CT	TT	CC	CT	TT	GG	CT	AA
PA 150	Field D 679 Marper Farm Trinidad	Reference	GG	AC	GG	CC	GG	AA	GG	CC	TT	AC	CT	AC	CT	TT	CC	CT	CC	TT	AA	TT	AG
IMC 76	738Q6	Off-type	TT	AC	TT	CC	AG	AC	AG	CT	CT	CC	CT	AA	CT	CC	TT	CC	CT	GG	AG	CT	AG
IMC 76	737Q6	Off-type	TT	AC	TT	CC	AG	AC	AG	CT	CT	CC	CT	AA	CT	CC	TT	CC	CT	GG	AG	CT	AG
IMC 76	729Q6	√	TT	CC	GT	CC	GG	AC	AG	CT	TT	AC	CC	AC	CC	CC	TT	CC	CC	GT	AG	CT	AG
IMC 76	727Q6	√	TT	CC	GT	CC	GG	AC	AG	CT	TT	AC	CC	AC	CC	CC	TT	CC	CC	GT	AG	CT	AG
IMC 76	724Q6	Off-type	GG	CC	TT	CC	GG	AA	AA	CC	CC	CC	TT	AA	TT	TT	TT	TT	CC	TT	GG	TT	AA
IMC 76	Field D 144 Marper Farm Trinidad	Reference	TT	CC	GT	CC	GG	AC	AG	CT	TT	AC	CC	AC	CC	CC	TT	CC	CC	GT	AG	CT	AG
MAN 15/60	7Q6 e	Off-type	TT	CC	GT	CT	AG	AC	AG	TT	CT	AC	CC	AC	CT	CC	CC	CT	CT	TT	AG	TT	AA
MAN 15/60	85Q6 e	Off-type	TT	AA	GG	CT	AA	AA	AG	TT	TT	CC	CC	AC	CC	CT	TT	CT	TT	TT	AG	CT	GG
MAN 15/60	83Q6 e	√	GT	CC	GT	CC	AG	AC	AA	TT	CT	AC	CC	AC	CT	CC	CC	CT	CT	TT	AG	TT	AA
MAN 15/60	13Q6 e	√	GT	CC	GT	CC	AG	AC	AA	TT	CT	AC	CC	AC	CT	CC	CC	CT	CT	TT	AG	TT	AA
MAN 15/60	46Q6 e	√	GT	CC	GT	CC	AG	AC	AA	TT	CT	AC	CC	AC	CT	CC	CC	CT	CT	TT	AG	TT	AA
MAN 15/60	Field D237 T5 Trinidad	Reference	GT	CC	GT	CC	AG	AC	AA	TT	CT	AC	CC	AC	CT	CC	CC	CT	CT	TT	AG	TT	AA
NA 79	952Q6	√	GT	AC	GG	CC	AA	AA	AG	CT	TT	CC	CT	AA	CT	CC	TT	CC	CC	GT	GG	TT	AA
NA 79	958Q6	Off-type	GG	AC	GG	CC	GG	AA	GG	CC	TT	AC	CT	AC	CT	TT	CC	CT	CC	TT	AA	TT	AG
NA 79	Field D612 Marper farm Trinidad	Reference	GT	AC	GG	CC	AA	AA	AG	CT	TT	CC	CT	AA	CT	CC	TT	CC	CC	GT	GG	TT	AA

Figure 1. Intraclonal mislabeling (off-type) identified in 160 cacao trees from Ghana cacao collections based on 53 SNP markers (of which only 21 loci were presented). The true-to-type clones were marked as "√". The SNP profiles of the reference clones were generated using original trees from International Cacao Genebank, Trinidad.

and of these, two are universal collections (representing nearly all of the known genetic diversity): CATIE (Centro Agronómico Tropical de Investigación y Enseñanza) in Costa Rica and ICG,T in Trinidad and Tobago (Motilal et al, 2013; Wadsworth and Harwood, 2000). Mislabeled plants have been identified as a serious problem in germplasm collections (Hurka et al., 2004). Significant efforts have been made to solve the problem in some international cacao collections (Motilal et al., 2013; Zhang et al., 2009a,b); however, the mislabeling problem in most of the various national collections has not been systematically addressed. Until recently, tools have not been available to clearly identify mislabeled germplasm accessions. Molecular markers such as AFLP (amplified fragment length polymorphism) have sufficient discriminatory power to distinguish cacao accessions; however, these tools often failed to reach clear conclusions, with convincing statistical rigor, that two genotypes are identical (Christopher et al., 1999; Perry et al., 1998; Sounigo et al., 2001).

In the past few years, microsatellite markers have been widely used in cacao genotyping and individual identification, enabling systematic assessment of genetic identity in national and international cacao genebanks (Zhang

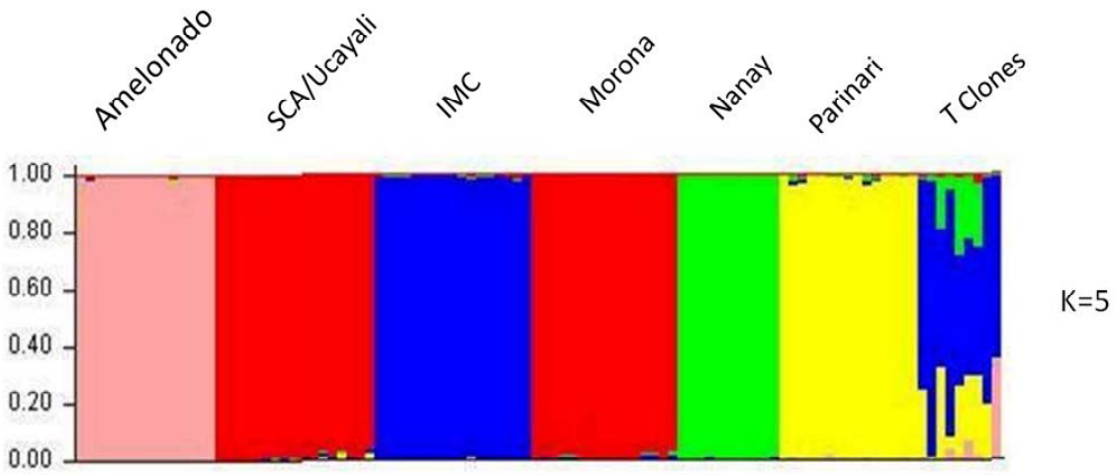

Figure 2. Verification of genetic membership for ten T clones of cacao in Ghana cacao germplasm using assignment test. The computer program STRUCTURE was used, where K is the potential number of genetic clusters that may exist in the overall sample of individuals. Each vertical line represents one individual 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.

Table 3. Parentage verification for cacao selections with known breeding pedigree, based on 53 SNP markers with LOD scores at 80 and 95% probability. The SNP profiles of the parental clones were generated using original trees from International Cacao Genebank, Trinidad.

Offspring ID	Recorded Pedigree	Tested candidate mother/father	LOD score*
T16/613	IMC 24 OP	IMC 24 (N/A)	N/A
T16/613	IMC 24 OP	Amelonado 22	8.83
T60/877	PA 7 x NA 32	PA 7	7.97
T60/877	PA 7 x NA 32	NA 32	7.01
T63/967	NA 32 x IMC 67	NA 32	7.13
T63/967	NA 32 x IMC 67	IMC 67 (N/A)	N/A
T63/971	NA 32 x IMC 67	NA 32	6.41
T63/971	NA 32 x IMC 67	IMC 67 (N/A)	N/A
T65/238	PA 7 x IMC 47	PA 7	2.34
T65/238	PA 7 x IMC 47	IMC 47	3.89
T65/326	PA 7 x IMC 47	PA 7	5.17
T65/326	PA 7 x IMC 47	IMC 47	6.08
T79/501(a)	NA 32 x PA 7	PA 7	0.04
T79/501(a)	NA 32 x PA 7	NA 32	0.45
T79/501(b)	NA 32 x PA 7	PA 7	6.65
T79/501(b)	NA 32 x PA 7	NA 32	7.32
T85/799	IMC 60 x NA 34	IMC 60 (N/A)	N/A
T85/799	IMC 60 x NA 34	NA 34	3.04

*Critical LOD (the natural logarithm of the likelihood) ratio for assignment of maternity and paternity are 5.70 at >95% confidence and 2.75 at >80% confidence.

et al., 2009a; Motilal et al., 2009,2010). In contrast to dominant markers, identical genotypes can have a 100% match in the multilocus SSR profiles without ambiguity, thus accuracy of identification is significantly improved. Reference SSR profiles of cacao clones have been deposited in the International Cacao Germplasm Database at the University of Reading, UK (http://www.icgd.rdg.ac.uk/index.php). However, comparison of genotyping results from different laboratories has not been straight forward. The effectiveness of clone identification via SSR fingerprints depends on the number of loci used for genotyping, as well as the rate of geno-

typing error. For example, it may require multiple repeated genotyping runs to reach the “consensus genotype”. Moreover, data generated from different genotyping platforms can be difficult to compare with one another because the same allele may be binned differently, leading to false conclusions.

The present study demonstrated that using the SNP-based multilocus fingerprints significantly improved the efficiency of genotype identification. Off-type identification, through the comparison with reference SNP profiles, is straightforward when reference trees are available. The reference trees used in the present study were sampled from the original collections maintained at Marper Farm and San Juan Estate in Trinidad, and Cabiria Farm, CATIE, in Costa Rica. These reference trees have been genotyped by SSR markers and passed through rigorous statistical population genetics tests (Motamayor et al., 2008; Zhang et al., 2009a,b; Johnson et al., 2009).

Parentage verification and assignment test

Many national cacao germplasm collections also maintain local varieties and breeding lines, which do not have a reference tree in international germplasm collections. In this situation, indirect verification such as Bayesian assignment test, parentage analysis, and sibship reconstruction need to be applied. The present study demonstrated how parentage analysis and Bayesian assignment test can be used to verify the genetic identity and pedigree information. Of the eight tested accessions, six were confirmed to have the correct maternal or paternal parent matching with the breeding record. Among them, T63/967 and T63/971 were supposed to be siblings and their verified parentage supported each other. T16/63 was recorded as the open pollinated progeny of IMC 24. Parentage analysis identified Amelonado 22, at a 95% confidence level, as the hidden pollen parent. For candidate parents that did not reach the 80% confidence level, the failure indicates mislabeling (off-type). Another possibility is possible conta-mination due to unwanted pollen or self-compatibility.

The SCA/Ucayali and Morona accessions represent two distinct geographical regions and were clustered as two different genetic groups when SSR markers were used (Zhang et al., 2009b; Motamayor et al., 2008). However, in the present study, the Bayesian clustering analysis based on 53 SNP markers did not significantly differentiate these two germplasm groups (Figure 2). Differences in genetic distances quantified by SNP and SSR markers have been reported in other crops. Yang et al. (2011) reported a correlation between kinship coefficient estimated by SSR and SNP of 0.69 in maize. Murray et al. (2009) found that some sorghum individuals shifted groups, depending upon whether SSR or SNP data was used in the STRUCTURE program. The discrepancy in stratification based on the two marker systems could also be due to the relatively small number of SNP markers used in the present study. Yu et al. (2009) showed that kinship estimated using 1,000 SNPs was consistent with that estimated with 100 SSRs in maize. Van Inghelandt et al. (2010) proposed that 7 to 11 times more SNPs than SSR markers should be used for analyzing population structure and genetic diversity in maize germplasm. Given that our previous stratification was based on 15 SSR markers, it would require more than 100 SNP markers to reach the same precision level. Additional SNP markers need to be evaluated for cacao and the correlation between SNP markers and SSR markers needs to be systematically assessed.

In addition to the limitation due to a limited number of SNP markers, the discrepancy between the two marker systems might also be partially explained by the derivation of the SNP markers used in the present study from the EST data. A set of unequivocally neutral SNP markers would be ideal. Despite the lack of differentiation between the SCA/Ucayali and Morona populations, the assignment test correctly excluded both groups in terms of parentage contribution to the tested T clones. The assignment of the T clones is fully consistent with the outcome of parentage analysis and is consistent with the recorded pedigree (Lockwood and Gyamfi, 1979). The high repeatability of the genotyping result, as demonstrated by the multiple trees for some cacao germplasm maintained in the Ghana collection, as well as the consistency in pedigree records and parentage analysis, demonstrated that these SNP markers provide a reliable and efficient solution for cacao genotype identification. This modest set of SNP markers thus constitutes a cost-effective marker resource, suitable for backstopping large-scale clone propagation in cacao. Nonetheless, the study also showed that a larger number of SNP markers would be needed for comprehensive diversity analysis.

Conflict of Interests

The author(s) have not declared any conflict of interests.

ACKNOLEDGEMENTS

The authors would like to thank Michel Bocarra, Xavier Argout, Claire Lanaud and Mathilde Allegre of CIRAD, France for providing the SNP sequences; Shenghui Duan and Cindy Helms of the Human Genetics Division Genotyping Core, Washington University School of Medicine for SNP genotyping; Stephen Pinney and Yan Mei Li of SPCL, USDA-ARS for assistance in DNA sample preparation.

REFERENCES

Adu-Ampomah Y, Adomako B, Opoku IY (2006). Cocoa population breeding approaches in Ghana. 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, The Netherlands/ICCO, London, UK/IPGRI, Rome, Italy, pp 41-46.

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: Denamany G, Lamin K, Ling A, Maisin N, Ahmad, AC, Saripah B, Nuraziawati MY (eds) Sustainable cocoa economy through increase in productivity, efficiency and quality: Proceedings of 4th Malaysian International Cocoa Conference, Kuala Lumpur, Malaysia, 18th-19th July 2005, Malaysian Cocoa Board, Kota Kinubalu, pp 83-86.

Allegre M, Argout X, Boccara M, Fouet O, Roguet Y, Bérard A, Thévenin JM, Chauveau A, Rivallan R, Clement D, Courtois B, Gramacho K, Boland-Augé A, Tahi M, Umaharan P, Brunel D, Lanaud C (2012). Discovery and mapping of a new expressed sequence tag-single nucleotide polymorphism and simple sequence repeat panel for large-scale genetic studies and breeding of *Theobroma cacao* L. DNA Res. 19:23-35.

Argout X, Fouet O, Wincker P et al. (2008). Towards the understanding of the cocoa transcriptome: production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* generated from various tissues and under various conditions. BMC Genomics 9:512.

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

Buckler ES, Thornsberry J (2002). Plant molecular diversity and applications to genomics. Curr. Opin. Plant Biol. 5:107-111.

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

Cuatrecasas J (1964). Cacao and its allies: A taxonomic revision of the genus *Theobroma*. Contributions from the United States National Herbarium Volume 35, part 6, Washington DC, Smithsonian Institution Press, Washington DC. pp. 375-614.

Doyle JJ, Doyle JL (1990), Isolation of plant DNA from fresh tissue. Focus 12:13-15.

Edwin J, Masters WA (2005). Genetic improvement and cocoa yields in Ghana. Exp. Agric. 41:491-503.

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 Genet. Genomes 4:821-831.

Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14:2611-2620.

Evett IW, Weir BS (1998). Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists. Sinauer, Sunderland, Massachusetts, USA.

FAOSTAT. Food and Agricultural commodities production. FAO statistical databases. Available from: http://faostat3.fao.org/home/index.html.

Figueira A, Janick J, Levy M, Goldsbrough P (1994). Reexamining the classification of *Theobroma cacao* L. using molecular markers. J. Am. Soc. Hortic. Sci. 119:1073-1082.

Guiltinan M, Verica J, Zhang D, Figueira A (2008). Genomics of *Theobroma cacao*, the Food of the Gods. In: Moore P, Ming R (eds) Genomics of Tropical Crop Plants, Springer, New York. pp. 145-170.

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

Ji K, Zhang DP, Motilal L, Boccara M, Lachenaud P, Meinhardt LW (2013). Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers. Genet. Resour. Crop Evol. 60:441-453.

Johnson ES, Bekele FL, Brown SJ, Song Q, Zhang D, Meinhardt LW, Schnell RJ (2009). Population structure and genetic diversity of the Trinitario cacao (*Theobroma cacao* L.) from Trinidad and Tobago. Crop Sci. 49:564-572.

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.

Kuhn DN, Livingstone D, Main D, Zheng P, Saski C, Feltus FA, Mockaitis K, Farmer AD, May GD, Schnell RJ, Motamayor JC (2012). Identification and mapping of conserved ortholog set (COS) II sequences of cacao and their conversion to SNP markers for marker-assisted selection in *Theobroma cacao* and comparative genomics studies. Tree Genet. Genomes 8:97-111.

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.

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

Livingstone DS, Freeman B, Motamayor JC, Schnell RJ, Royaert S, Takrama J, Meerow AW, Kuhn DN (2012). Optimization of a SNP assay for genotyping *Theobroma cacao* under field conditions. Mol. Breed. 30:33-52.

Lockwood G, Gyamfi MMO (1979). The CRIG cocoa germplasm collection with notes on codes used in the breeding programme at Tafo and elsewhere. Tech. Bull. 10, Cocoa Research Institute, Ghana, 62 pp.

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, Lachenaud P, da Silva e Mota JW, Loor G, Kuhn DN, Brown JS, Schnell RJ (2008). Geographic and genetic population differentiation of the Amazonian chocolate tree. PLoS ONE 3:e3311. 10.1371/journal.pone.0003311.

Motilal L, 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. 8:106-110.

Motilal LA (2004). The potential of cacao microsatellites amplification across diverse plant taxa. In: Thangadurai D, Pullaiah T, Balatti PA (eds) Genetic Resources and Biotechnology, Vol. 2, , Regency Publications, New Delhi, India. pp. 24-49.

Motilal LA, Butler D (2003). Verification of identities in global cacao germplasm collections. Genet. Resour. Crop. Ev. 50:799-807.

Motilal LA, Zhang D, Mischke S, Meinhardt LW, Umaharan P (2013). Microsatellite-aided detection of genetic redundancy improves management of the International Cocoa Genebank, Trinidad. Tree. Genet. Genomes 9:1395-1411.

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.

Murray SC, Rooney WL, Hamblin MT, Mitchell SE, Kresovich S (2009). Sweet sorghum genetic diversity and association mapping for brix and height. Plant Genome 2:48-62.

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

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

Peakall R, Smouse PE (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28:2537-2539.

Perry MD, Davey MR, Power JB, Lowe KC, Bligh HFJ, Roach PS, Jones C (1998). DNA isolation and AFLP genetic fingerprinting of *Theobroma cacao* L. Plant Mol. Biol. Rep. 16:49-59.

Posnette AF (1986). Fifty years of cocoa research in Trinidad and Tobago. Cocoa Research Unit, University of the West Indies, St. Augustine, Trinidad, 131 pp.

Powis TG, Cyphers A, Gaikwad NW, Grivetti L, Cheong K (2011). Cacao use and the San Lorenzo Olmec. Proc. Natl. Acad. Sci. USA 108:8595-8600.

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

Sounigo O, Christopher Y, Bekele F, Mooleedhar V, Hosein F (2001). The detection of mislabelled trees in the International Cocoa Genebank, Trinidad (ICG,T). In: Proceedings of the International Workshop on the New Technologies and Cocoa Breeding, 16th–17th October 2000, Kota Kinabalu, Sabah, Malaysia, INGENIC (International Group for Genetic Improvement of Cocoa). pp. 34-39.

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, Risterucci A-M, Clement D, Fouet O, Lanaud C (2006). Identification of off-types of clones used in the International Clone Trial using DNA analyses. 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, The Netherlands/ICCO, London, UK/IPGRI, Rome, Italy, pp. 82-86.

Takrama J, Dadzie AM, Opoku FK, Padi FK, Adomako B, Asu-Ampomah Y, Livingstone DS, Motamayor JC, Schnell RJ, Kuhn RJ (2012). Applying SNP marker technology in the cacao breeding programme in Ghana. Afr. Crop. Sci. J. 20:67-75.

Takrama JF, Cervantes-Martinez C, Phillips-Mora W, Brown JS, Motamayor JC, Schnell RJ (2005). Determination of off-types in a cocoa breeding programme using microsatellites. INGENIC Newsletter 10:2-8.

Van Inghelandt D, Melchinger AE, Lebreton C, Stich B (2010). Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor. Appl. Genet. 120:1289-1299.

Wadsworth RM, Harwood T (2000). International Cocoa Germplasm Database, ICGD 2000 V4.1. London International Financial Futures and Options Exchange and the University of Reading, 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.

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.

Wood GAR, Lass RA (1985). Cocoa. 4th Edn, Essex: Longman Group Ltd, 620 pp.

Yang XH, Xu YB, Shah T, Li HH, Han ZH, Li JS, Yan JB (2011). Comparison of SSRs and SNPs in assessment of genetic relatedness in maize. Genetica 139:1045-1054.

Yu JM, Zhang ZW, Zhu CS, Tabanao DA, Pressoir G, Tuinstra MR, Kresovich S, Todhunter RJ, Buckler ES (2009). Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. Plant Genome 2:63-77.

Zhang D, Boccara M, Motilal L, Mischke S, Johnson ES, Butler D, Bailey BA, Meinhardt, LW (2009b). Molecular characterization of an earliest cacao (*Theobroma cacao* L.) collection from Peruvian Amazon using microsatllite DNA markers. Tree Genet. Genomes 5:595-607.

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