Published online September 8, 2006

Reproduced from Crop Science. Published by Crop Science Society of America. All copyrights reserved.

Accuracy and Reliability of High-Throughput Microsatellite Genotyping for Cacao Clone Identification

Dapeng Zhang,* Sue Mischke, Ricardo Goenaga, Alaa A. Hemeida, and James A. Saunders

ABSTRACT

Microsatellite-based DNA fingerprinting has been increasingly applied to cacao (*Theobroma cacao* L.) genotype identification. However, the accuracy and reliability of using high throughput microsatellite analysis for cacao clone identification have not yet been rigorously assessed. Despite the use of highly robust fingerprinting protocols, cacao genotype identification has been affected by genotyping errors, which potentially mislead the result of clone identification. In this paper, we calculated the probability of identity for 15 selected microsatellite loci. We then quantified the genotyping error rate through repeated genotyping and simulated the impact of the genotyping error on cacao clone identification. Allelic dropout (ADO), or failure to amplify one allele for a heterozygous locus, and false allele (FA), or an amplicon size error by the polymerase, accounted for 48 and 52% of the genotyping inconsistencies, respectively. The result of simulation showed that 99% of the consensus genotype can be generated for the ambiguous loci through a minimum of three polymerase chain reaction (PCR) repetitions. On the basis of the error rate and probability of identity (PID), we designed a genotyping scheme and applied it to the cacao germplasm held in the USDA cacao collection at Mayaguez, Puerto Rico. Out of the 141 samples, we unambiguously identified nine duplicated groups consisting of 34 cacao accessions. This genotyping scheme is being implemented in large scale fingerprinting of cacao germplasm.

Cacao is an important tropical crop, native to the South American rainforest (Cuatrecasas, 1964; Coe and Coe, 1996; Young, 1994). The species comprises a large number of highly morphologically variable and mutually interfertile populations. A cacao pod may contain numerous seeds, but the seeds do not remain viable for much longer than a week once the pod has been harvested (Coe and Coe, 1996). As a result, cacao germplasm collections must be maintained as clonally propagated, living trees. A dozen major cacao germplasm collections in tropical regions of the world serve as germplasm repositories. The genetic diversity is not fully characterized. A number of molecular tools have been used to examine cacao populations by DNA fingerprinting procedures and a large number of cacao accessions in these collections have been reported to be misidentified, duplicated, or to have no identification at all (Kennedy and Mooleedhar, 1993; Lockwood and End, 1993). The incorrect labeling of accessions is a major limitation to efficient conservation of cacao germplasm. Further, this constraint impedes the progress in genetic improvement of cacao throughout the world.

The development of simple sequence repeats (SSR) markers in cacao (Lanaud et al., 1999) has significantly enhanced the capacity of molecular characterization of cacao germplasm. This technique has been applied for cacao clone identification (Cryer et al., 2006; Saunders et al., 2004), parentage analysis (Schnell et al., 2005), diversity assessment (Lanaud et al., 1999, 2001), and investigation of the origin and dispersal of cacao (Motamayor et al., 2002, 2003). The USDA, together with its collaborating institutions, has undertaken a program to identify cacao genotypes and describe the genetic diversity of living plant germplasm collections that are maintained in 10 to 12 national and international collections located within tropical cacao growing countries of Central and South America. During international forums held in England and in France in 2001, a consortium of scientists and representatives from the cacao industry, academic centers involved in cacao research, and representatives from multiple international, government-sponsored laboratories reached an agreement that a set of standardized SSR primers would be used to characterize all *T. cacao* germplasm collections (Saunders et al., 2001, 2004). The strategy for the identification of mislabeling is to develop a reference SSR profile for each original accession. Then the reference SSR profiles will be compared with the putatively mislabeled accessions. Corrections will then be made as necessary (Turnbull et al., 2004).

However, the accuracy and reliability of using high throughput microsatellite analysis for cacao clone identification have not yet been rigorously assessed. Several questions must be addressed if this tool is to be applied for large-scale genotyping of cacao germplasm. First, we need to know if the set of proposed SSR loci have enough discriminating power to establish unique multilocus profiles. If too few loci are examined, multilocus genotypes in a population may not be unique; thus, closely related clones will be indistinguishable. The problem of finding different clones with the same SSR profile can be solved by increasing the number of loci examined, so that the probability that two different clones have the same multilocus genotype is small. This probability, often called the "probability of identity" (PID) in forensic science, can be estimated from the allele frequencies in a population (Waits et al., 2001).

Dapeng Zhang and Sue Mischke, USDA-ARS, BARC, PSI, SPCL, 10300 Baltimore Ave. Bldg. 50 BARC-W, Beltsville, MD 20705, USA; Ricardo Goenaga, USDA-ARS, Tropical Agric. Research Station, P.O. Box 70, Mayaguez, PR 00681, Puerto Rico; Alaa A. Hemeida, Genetic Engineering & Biotechnology Research Institute (GEBRI), Sadat City, Minufiya Univ., Egypt; James A. Saunders, Molecular Biology, Biochemistry and Bioinformatics (MB3), 360 Smith Hall, 8000 York Road, Towson Univ., Towson, MD 21252, USA. Received 2 Jan. 2006. *Corresponding author (ZhangD@ba.ars.usda.gov).

Published in Crop Sci. 46:2084–2092 (2006).
Plant Genetic Resources
doi:10.2135/cropsci2006.01.0004
© Crop Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: ADO, Allele dropout; FA, false allele; PID, probability of identity; SSR, single sequence repeats.

Reproduced from Crop Science. Published by Crop Science Society of America. All copyrights reserved.

Probability of identity can also be conditioned on a given relationship (unrelated, parent-offspring, or sibling) between the two individuals (Waits et al., 2001).

Second, microsatellite-based DNA fingerprinting is not an error-free technology. Despite the use of highly robust fingerprinting protocols and the preparation of high quality DNA samples, genotyping errors still occur, which causes duplicate samples of a clone to appear to have different microsatellite profiles. Such error, even at modest rates, can distort the result of cultivar identification. Repeated genotyping is widely recommended as the most reliable approach to tackle the problem of genotyping error (Taberlet and Luikart, 1999; Bonin et al., 2004). With this approach, several PCRs are performed for a given locus and a consensus genotype is developed on the basis of the results of multiple PCRs. It is unlikely that a given allele will drop out in all multiple repeated PCRs. This approach has been applied in projects dealing with low quality DNA extracted from noninvasive samples, such as animal hairs and fecal samples (Taberlet et al., 1996; Taberlet and Luikart, 1999; Paetkau, 2003, 2004). Although this approach is reliable, accomplishment of a consensus profile requires large numbers of amplifications, adding significant extra experimental cost. Therefore, it is important to understand the magnitude of genotyping error so that the minimum number of essential repetitions can be applied.

There are a number of sources of errors encountered in automated microsatellite genotyping. These include human error, sample contamination, and errors occurring during amplification and electrophoresis (Hoffman and Amos, 2005). Whereas human errors and contamination can be avoided by cautious laboratory work or can be detected by controls, errors in amplification during PCR may not be avoided through laboratory procedures. Allelic dropouts (ADO: one allele of a heterozygous clone is not amplified during PCR), and false alleles (FA: PCR-generated allele results from a slippage artifact during the early cycles of the reaction) are the two source of errors that cannot be easily monitored and thus need to be quantified (Taberlet et al., 1996, Taberlet and Luikart, 1999; Broquet and Petit, 2004).

The present study aimed to examine the accuracy and reliability of the 15 SSR loci currently used to identify cacao accessions and to use this information for development of an optimum genotyping scheme. Using the USDA cacao collection at Mayaguez, Puerto Rico, as a test case, we quantified the genotyping error rate through repeated genotyping, and simulated the impact of the genotyping error on cacao clone identification. We then estimated the PID for the 15 standard microsatellite loci, conditioned on unrelated accessions, as well as on full siblings. On the basis of the results, we propose a genotyping protocol for the identification of mislabeling in cacao germplasm collections.

This study is part of an international collaborative project on DNA fingerprinting of cacao germplasm in the Americas. The resultant information will be used to design an optimum scheme for genotyping cacao germplasm and improve our understanding about the extent of mislabeling in the genebanks, thus facilitating the efficient conservation and use of cacao germplasm.

MATERIALS AND METHODS

The cacao samples used for DNA fingerprinting profiles include leaves of various ages collected from individual cacao accessions held at the Tropical Agricultural Research Station (TARS) at Mayaguez, Puerto Rico. This collection represents a unique and diverse cacao resource, with a diverse germplasm base represented in a relatively limited number of accessions. The collection initially began with random cacao accessions of interest to specific researchers, but it is being developed as a core collection of cacao germplasm, representing diverse types available for commercial production. The Mayaguez collection was initially established in 1960 with about 220 cacao clones, many of which arrived from the quarantine facility at Miami, FL, now known as the Subtropical Horticulture Research Station. The cacao germplasm collection was established in Puerto Rico because the island (i) had no commercial cacao plantings, (ii) was free from serious cacao diseases and pests that were present in most commercial producing areas, (iii) had an environment suitable for normal growth and development of cacao, and (iv) was centrally located, with transportation to facilitate germplasm exchange with the Caribbean, Mexico, and Central and South America. Because of their suitability as indicators of cacao viruses, the "Amelonado" clones, and to a smaller extent "EET400" clones, were used as rootstock for the cacao accessions, employing budwood grafting procedures common within the industry. As with most other cacao germplasm collections, records documenting each genotype that had been incorporated into the collection were incomplete. It is noteworthy that several of the primary and secondary contributors of germplasm were unable to guarantee the authenticity of the material supplied. This is considered a common cause of the introduction of mislabeled accessions into cacao collections.

Cacao leaves have very high levels of endogenous phenolics, which can interfere with DNA isolation procedures (Griffiths, 1958; John, 1992; Katterman and Shattuck, 1983). Initial investigations of various DNA isolation protocols identified two methods that worked well for cacao microsatellite DNA analysis, and these were used interchangeably to yield consistent results. DNA was isolated from 50-mg samples of *T. cacao* leaf material using either the DNA Xtract Plus kit (D^2 BioTechnologies Inc., Atlanta, GA) or the DNeasy Plant System (Qiagen, Valencia, CA). For either method, the air-dried and frozen leaf samples were cut into small pieces and placed into a 2-mL tube, sandwiched between ceramic spheres, with garnet matrix (Qbiogene, Inc., Carlsbad, CA). Lysis solution was added following the manufacturers' recommendations, except 10 mg/mL of polyvinylpolypyrrolidone (Sigma, St. Louis, MO) had been added to the Qiagen buffer AP1. Samples were homogenized in a Fast Prep instrument (Bio101-Qbiogene, Inc., Carlsbad, CA) as described previously (Saunders et al., 2001).

The DNA Xtract Plus procedure was, in brief, lysis, clarification by centrifugation, and solvent phasing followed by precipitation on ice. DNA was collected by centrifugation, washed in 70% (v/v) ethanol, centrifuged, dried, and resuspended in sterile water or buffer. The Dneasy Plant System isolation procedure included tissue lysis and RNase A treatment at 65°, followed by centrifugation and precipitation of detergent, proteins, and polysaccharides. Cell debris and precipitates were removed by centrifugation of the slurry through a QIAshredder spin column (Qiagen) assembly. Ethanol was used to precipitate the DNA in the cleared filtrate, which was loaded onto the DNeasy column. DNA was bound to the col-

Reproduced from Crop Science. Published by Crop Science Society of America. All copyrights reserved.

umn's silica membrane by centrifugation, washed with 70% ethanol, and finally eluted from the membrane with preheated buffer. The presence of double-stranded DNA was verified by measuring DNA quantity with PicoGreen (Molecular Probes, Inc., Eugene OR) using a Fluoroskan Ascent microplate reader equipped with 485/538 excitation/emission filters (Labsystems, Helsinki, Finland).

DNA amplification used primer sets with sequences previously described (Lanaud et al., 1999; Saunders et al., 2004). 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 supermixes that had been fortified with an additional 30 units of the respective hot-start *Taq* DNA polymerase (Invitrogen Platinum *Taq*, Carlsbad CA; Eppendorf HotMaster *Taq*, Westbury NY) added to each milliliter of the supermix.

The amplified PCR products were separated by capillary electrophoresis as previously described (Saunders et al., 2004) using a CEQ 8000 genetic analysis system (Beckman Coulter Inc.). 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 by the CEQ 8000 Genetic Analysis System. Allele determination was performed using the binning wizard software, which automatically generated locus tags in the CEQ 8000 Genetic Analysis System.

Key parameters for measuring informativeness of these 15 loci were calculated using the program Powermarker (Liu and Muse, 2005). These included mean number of alleles per locus, observed heterozygosity, gene diversity (expected heterozygosity), and polymorphic information content (PIC). The same software was also used to measure the two-locus linkage disequilibrium among the 15 loci. For the estimation of probability of identity, we used a conservative estimation of PID to assess the differentiating power of the SSR loci, assuming that the Mayaguez collection may have closely related clones.

We computed the probability of identity among full siblings (PID-sib), which was defined as the probability that two full-sib individuals drawn at random from a natural population have the same multilocus genotype. The overall PID-sib is the upper limit of the possible ranges of PID in a natural population, thus providing the most conservative number of loci required to resolve all individuals, including relatives (Waits et al., 2001). This can be computed using the following equation:

$$P(\mathrm{ID})\mathrm{sib} = 0.25 + (0.5\,\Sigma pi^2) + [0.5(\Sigma pi^2)^2] - (0.25\,\Sigma pi^4)$$

where pi is the frequency of the ith allele (Evett and Weir, 1998; Taberlet and Luikart, 1999). The program GIMLET V.1.3.2 (Valière, 2002) was used for the computation of PID-sib.

To quantify the genotyping error rate, we performed an independent experiment, using the multiple tube approach (Taberlet and Luikart, 1999). This approach involves dividing the sample among multiple tubes (or wells), then amplifying and typing the products in each tube separately. The results are analyzed by a statistical procedure that determines whether a genotype can be conclusively assigned to the DNA sample. Thirty DNA samples were chosen from the Mayaguez collection and genotyped independently by three different persons who performed amplification and capillary electrophoresis at different times, using the established protocols in our laboratory (Saunders et al., 2004). In addition to the three repetitions, data from an experiment performed 2 yr previously, when this subset of samples was genotyped by a fourth person, were also considered. Hence, a total of four independent PCRs were performed on the 30 DNA samples. The genotypes obtained were analyzed for locus-specific error rate following the method of Broquet and Petit (2004), where the ADO rate for a locus was estimated by the number of heterozygous individuals genotyped as homozygous individuals, divided by the total number of genotyped heterozygous samples. The FA rate for a locus is estimated as the number of amplifications leading to the creation of one or more false alleles at the given locus, divided by the total number of genotyped samples. Since the procedure used here does not take into account any possible source of error before the DNA sample (i.e., from collection of the leaf sample in the field through extraction of the DNA), the estimates of error rate are underestimates of the true error rate, and the extent of the underestimation is unknown.

The impact of genotyping error on cacao clone identification was assessed in two ways. First, we evaluated the effectiveness of using a single genotyping for cacao identification. The single genotyping approach assumes that, if the genotyping error is reasonably small, most errors will cause samples from the same individual to differ (mismatch) by a small number of loci, and most frequently only a single locus. If enough loci are scored, each clone in the population will have multilocus genotypes (or a DNA profile) with mismatches at more than one locus. Therefore, if both of these criteria are met, most pairs of samples that mismatch by only a single locus will differ because of genotyping error (Kalinowski et al., 2006; Paetkau, 2004). This shows that if genotyping error is a potential problem, enough loci should be examined so that different clones sampled from the population are likely to have genotypes differing by more than one locus (Kalinowski et al., 2006). To test if this is the case for cacao clone identification in our system, we calculated the probability of mismatch for the 141 cacao accessions in the Mayaguez collection, on the basis of the 15 recommended SSR loci, using the computer program MM-DIST (Kalinowski et al., 2006).

Second, we examined the effectiveness of the multitube approach by simulating how many PCR repetitions would be needed to generate a "consensus genotype." We defined the minimum acceptable standard of a consensus genotype as one that appears at least twice in the repeated genotyping (Valière et al., 2002). The simulation used the allele frequency obtained from the 141 cacao accessions in the Mayaguez collection and the observed error rate from the experiment of repeated genotyping. The simulation was implemented using the program of GEMINI (Valière et al., 2002).

To identify duplicates and mislabeling in the Mayaguez cacao collection, a genotyping protocol that combined the multitube approach with the "mismatch tolerant" approach was applied. Pairwise comparison of all the 141 cacao accessions was performed to identify duplicates on the basis of their multilocus profiles (i.e., 15 SSR loci) generated through a single genotyping. Accessions with different names but fully matched at 15 loci were declared duplicates or synonymously mislabeled accessions. Accessions that differed by one, two, or three loci were repeatedly genotyped three times at the disputed loci to generate the consensus genotypes, followed by pairwise matching.

RESULTS

All microsatellite loci were polymorphic and met the assumptions of independent segregation. An example of the capillary electropherogram is presented in Fig. 1.

Reproduced from Crop Science. Published by Crop Science Society of America. All copyrights reserved.

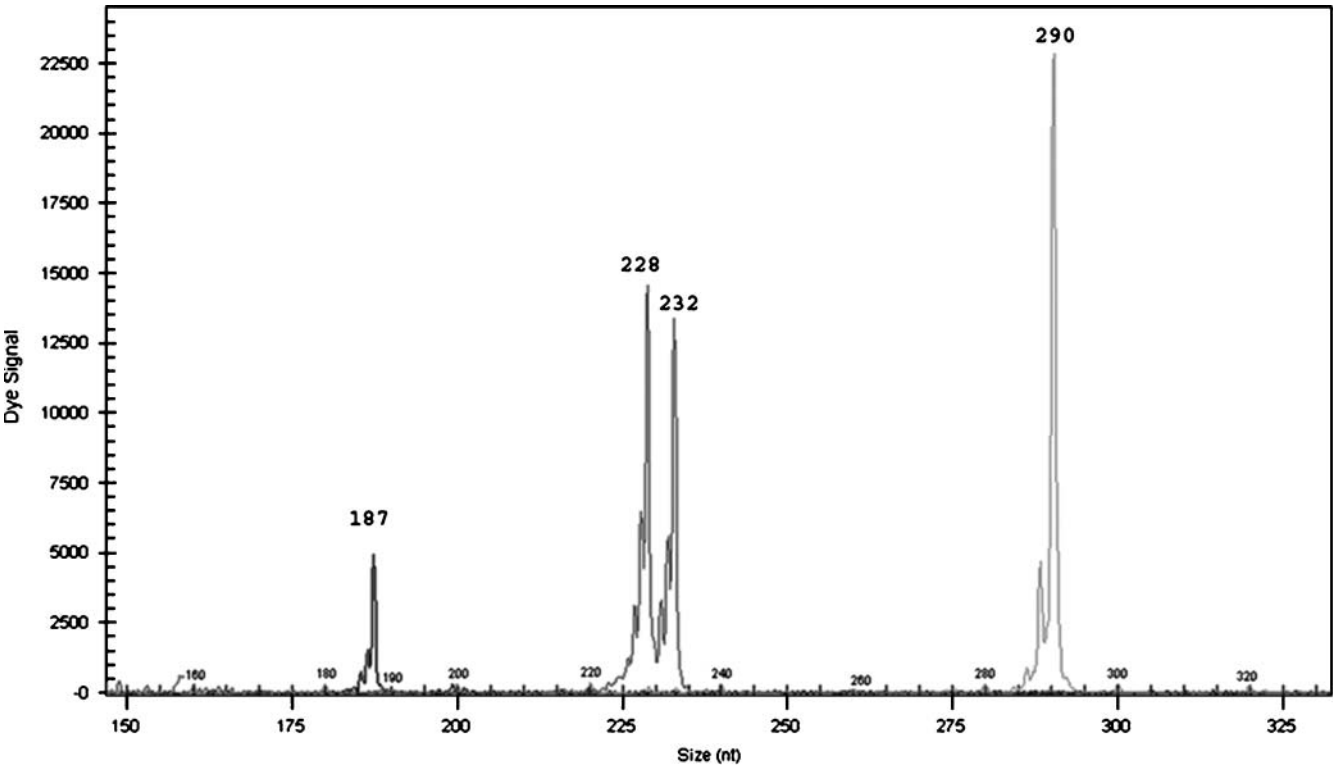

Fig. 1. This electropherogram shows the DNA fragment profiles of SSR primers for three different loci, amplified separately using three different dyes, multiplexed and separated by capillary electrophoresis. The sample was the *T. cacao* accession PA44, an Upper Amazon Forastero type that is part of the USDA germplasm collection maintained in Puerto Rico. Alleles from three loci were amplified as shown in the electropherogram sample: The heterozygous locus Y16980 (mTcCIR6) is located on chromosome 6 and shows alleles sizes of 228 and 232 base pairs. The two homozygous loci, Y16996 (mTcCIR24) located on chromosome 9 and the Y16995 (mTcCIR22) locus on chromosome 1, each show a single allele at 187 and 290 bp, respectively. Internal standards, labeled with red, are run concurrently with all samples for base pair determination during the capillary electrophoresis DNA fragment separations.

The number of alleles ranged from 5 to 17 across the 15 loci, with the average being 10.8 (Table 1). The 141 cacao accessions contained a total of 162 alleles over the 15 loci. High values of observed heterozygosity (Ho = 0.701), gene diversity (He = 0.651), and PIC value (PIC = 0.604) were observed across the 15 loci (Table 1), in comparison with previous published data in cacao (Lanaud et al., 2001).

With all 15 loci considered, the combined probability of identity of siblings (PID-sib) was on the order of 10^{-6} in the 141 germplasm accessions (Table 1). This clearly shows that the 15 microsatellite loci are capable of dis-

Table 1. Informativeness and probability of identity (PID) of the 15 microsatellite loci, estimated from 141 accessions in the USDA Mayaguez collection (K = number of observed alleles; Ho = Observed heterozygosity; He = Expected heterozygosity; PIC = Polymorphism information content).

		Mayaguez collection (n = 141)					
Marker name	EMBL† accession number	K	Ho	He	PIC‡	PID-sib§	PID-sib (cumulative)¶
mTcCIR7	Y16981	9	0.727	0.663	0.611	0.449	4.490e-01
mTcCIR6	Y16980	11	0.745	0.630	0.583	0.476	2.137e-01
mTcCIR22	Y16995	7	0.652	0.569	0.515	0.522	1.115e-01
mTcCIR24	Y16996	8	0.553	0.469	0.415	0.606	6.752e-02
mTcCIR8	Y16982	5	0.671	0.578	0.512	0.517	3.491e-02
mTcCIR1	Y16883	6	0.571	0.477	0.405	0.510	1.782e-02
mTcCIR11	Y16985	15	0.745	0.670	0.627	0.431	7.689e-03
mTcCIR12	Y16986	15	0.770	0.783	0.757	0.367	2.821e-03
mTcCIR15	Y16988	13	0.789	0.796	0.769	0.371	1.048e-03
mTcCIR37	AJ271942	17	0.776	0.719	0.688	0.423	4.429e-04
mTcCIR33	AJ271826	17	0.665	0.682	0.641	0.383	1.697e-04
mTcCIR18	Y16991	10	0.665	0.682	0.641	0.432	7.328e-05
mTcCIR26	Y16998	10	0.714	0.646	0.597	0.454	3.332e-05
mTcCIR40	AJ271943	11	0.789	0.779	0.745	0.378	1.260e-05
mTcCIR60	AJ271958	8	0.683	0.627	0.556	0.477	6.011e-06
	Mean	10.8	0.701	0.651	0.604		

† EMBL is the European Molecular Biology Laboratory Nucleotide Sequence Database.
‡ PIC (Polymorphism Information Content) follows the definition of Powell et al. (1996).
§ PID-sib (Probability of identity among siblings) follows the definition of Evett and Weir (1998).
¶ Cumulated 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.

Reproduced from Crop Science. Published by Crop Science Society of America. All copyrights reserved.

criminating closely related clones. In reality, a smaller number of loci would be sufficient to provide enough differentiation power for cacao clone identification. PID-sib became close to zero after the seven SSR loci with highest expected heterozygosity were applied (Fig. 2). The overall PID-sib is the upper limit of the possible ranges of PID in a population and thus provides the most conservative number of loci required to resolve all cacao genotypes.

The result of the multitube test showed that the genotyping error rate was low. Out of the 3600 alleles (i.e., 30 DNA samples typed for 15 loci with four PCR repetitions), a total of 48 alleles, involving 12 loci, had been misidentified (Table 2). Of the 48 identified genotyping errors, there were 23 allelic dropouts and 25 false alleles. The average error rate over the 15 loci, following the definition of Broquet and Petit (2004) was 0.014 for allele drop out and 0.019 for false allele. The error rate appeared to differ among the 15 loci. The four loci Y16986, Y16996, AJ271826, and AJ271942 accounted for 40% of the total misclassification, whereas no error was found for loci Y16995, Y16998, and AJ271943. There were a few cases where more than one genotyping error occurred in a single locus (i.e., both alleles in that locus were false alleles). However there was no ambiguity in identifying the erroneous genotype because the true genotype (consensus genotype) for that accession was known through repeated genotyping. When the number of errors was calculated on the basis of each multilocus genotype, there were 29 multilocus genotypes that contained only one error (one mistyped locus) among the 120 genotyped samples (30 accessions × four repetitions). Four multilocus genotypes contained two errors (two mismatched loci). No multilocus genotype contained more than two mismatched loci. In summary, allelic dropout accounted for 48% of the inconsistencies and false alleles for 52%.

Table 2. Genotyping error rate estimated from repeated PCR of 30 cacao accessions, selected from the Mayaguez cacao collection.

Marker name	EMBL Accession number	ADO rate†	FA rate‡
mTcCIR7	Y16981	0.017	0.000
mTcCIR6	Y16980	0.017	0.000
mTcCIR22	Y16995	0.000	0.000
mTcCIR24	Y16996	0.008	0.023
mTcCIR8	Y16982	0.025	0.047
mTcCIR1	Y16883	0.000	0.045
mTcCIR11	Y16985	0.017	0.015
mTcCIR12	Y16986	0.075	0.000
mTcCIR15	Y16988	0.000	0.020
mTcCIR37	AJ271942	0.017	0.034
mTcCIR33	AJ271826	0.033	0.031
mTcCIR18	Y16991	0.000	0.045
mTcCIR26	Y16998	0.000	0.000
mTcCIR40	AJ271943	0.000	0.000
mTcCIR60	AJ271958	0.000	0.025
Mean		0.014	0.019

† Allelic dropout: the number of heterozygous individuals (from the consensus genotype) typed as homozygous individuals, divided by the total number of typed heterozygous samples (Broquet and Petit, 2004).
‡ False allele: the number of amplifications leading to the creation of one or more false alleles at the given locus, divided by the total number of genotyped samples (Broquet and Petit, 2004).

To understand the importance of the observed error rate for cacao clone identification, we estimated the mismatch distribution of the 141 genotype profiles generated in a single genotyping. The expected distribution predicted that if the 15 SSR loci are genotyped, clones are likely to differ by more than five loci if they are unrelated accessions and three loci if the clones are full siblings (Fig. 3). The empirical distribution departed greatly from expected distribution of full siblings and resembled the expected distribution of unrelated accessions, suggesting that most of the 141 clones in the germplasm collection were unrelated.

This result confirmed that when the 15 SSR loci are used for fingerprinting, the true, positive genetic difference between two genotypes would result a mismatch

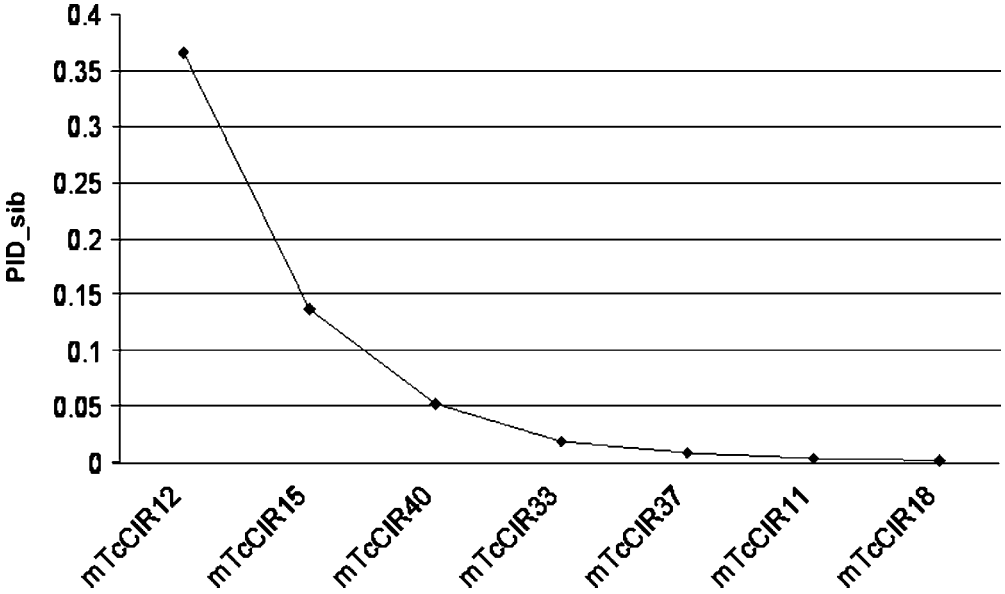

Fig. 2. Sibling probabilities of identity (PID-sib) from 141 cacao accessions in the USDA cacao collection maintained at Mayaguez, Puerto Rico. The probability that two sibling individuals drawn at random from this collection have the same mutilocus genotype became close to zero after seven SSR loci with highest PID-sib were applied.

Reproduced from Crop Science. Published by Crop Science Society of America. All copyrights reserved.

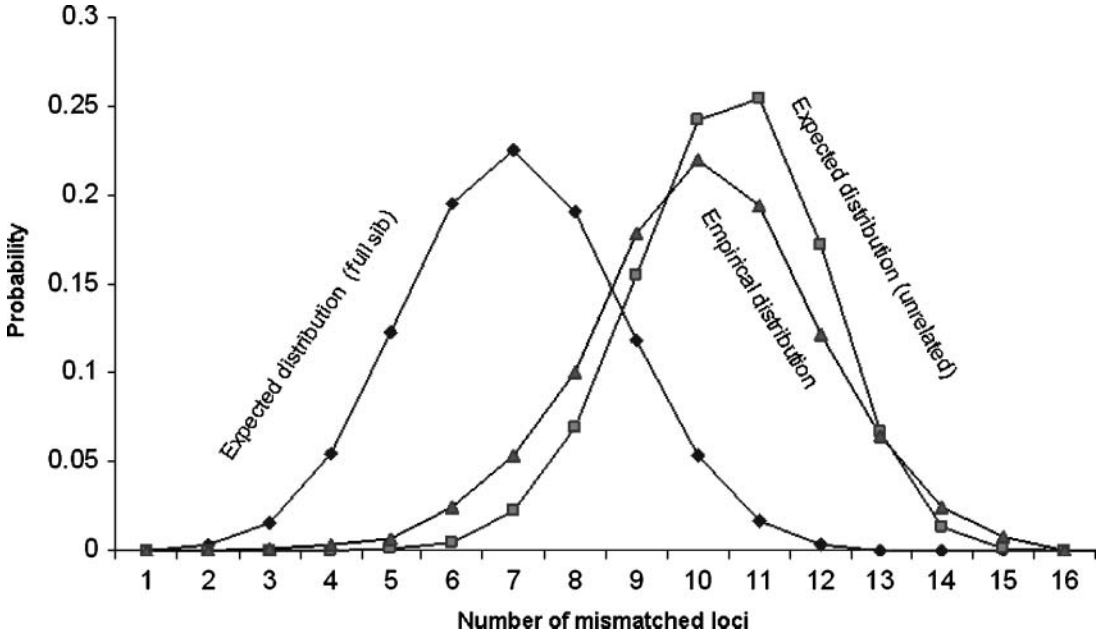

Fig. 3. Mismatch distribution of 141 cacao accessions in the USDA cacao collection maintained at Mayaguez, Puerto Rico. Computer program MM-DIST (Kalinowski et al., 2006) was used for computation of mismatch distribution. The data set from the single genotyping was used to compute the probability that two individuals differ at *k* loci. Both empirical and expected distributions showed that the full siblings will likely differ by at least three out of 15 loci in their multilocus SSR profiles. The unrelated individuals will likely differ by at least five out of 15 loci. The error mismatch distributions are therefore unlikely to overlap with distributions of true genotypic difference (McKelvey and Schwartz, 2004; Kalinowski et al., 2006).

of at least three loci, distinguishing this result from a false positive caused by genotyping error. As found in the repeated genotyping test, 88% of the genotyping errors caused a single locus mismatch between identical samples and 12% caused mismatch of two loci. No mismatch of greater than two loci was caused by genotyping error. In this case, the genotypes were genuinely different.

We then determined how many repeated PCRs would be needed to obtain a consensus genotype at the disputed loci. The minimum acceptable standard of a consensus genotype was defined as a multilocus profile appearing at least twice in the repeated genotyping. The simulation result, on the basis of the observed error rate of the repeated genotyping, shows that 98.7% of the consensus genotype can be identified with a minimum of three PCR repetitions, and 99.7% of the consensus genotypes can be identified with four PCR repetitions. Further increase of PCR repetitions offers little improvement in terms of genotyping accuracy (Fig. 4).

The comparison of the multilocus microsatellite profile led to the identification of nine synonymous groups including 34 accessions (Table 3). Within each group, the accessions were labeled with different names but shared exactly the same alleles at all 15 loci.

DISCUSSION

Incorrectly labeled accessions have been a serious problem in the national and international cacao collections, but until recently tools have not been available to clearly identify mislabeled germplasm accessions. Molecular markers such as RAPD and AFLP have sufficient discriminatory power to distinguish accessions, but these tools often failed to reach clear conclusions in the identification of duplicates. These markers did not identify duplicates by exact match of banding pattern; rather, assessment was by similarity estimation. "Identical" genotypes were declared when the similarity reached certain threshold values (Christopher et al., 1999; Perry et al., 1998; Sounigo et al., 2001).

Progress in developing microsatellite markers in cacao and the availability of high throughput genotyping

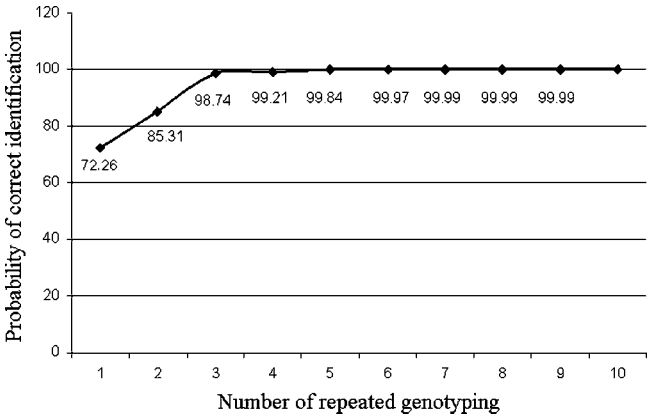

Fig. 4. The probability of obtaining a consensus genotype (at least two correct microsatellite genotypes) for a given number of repeated genotypings (independent PCRs). The simulation was based on the observed error rate in an independently repeated genotyping experiment as described in this paper. The simulation result shows that 98.7% of the consensus genotype can be identified with a minimum of three PCR repetitions, and 99.7% of the consensus genotypes can be identified with four PCR repetitions. Further increase of PCR repetitions offers little further improvement in terms of genotyping accuracy.

Reproduced from Crop Science. Published by Crop Science Society of America. All copyrights reserved.

Table 3. List of duplicates identified in the USDA Mayaguez cacao collection. Accessions in same synonymous group shared identical multilocus SSR profiles.

Synonymous group	Accession	Genebank code	Synonymous group	Accession	Genebank code
1	CC10	PR-43	3	GS7	PR-148
1	EET250	PR-90	3	GS29	PR-134
1	EET308	PR-197			
1	EET381	PR-103	4	EET407	PR-133
1	P23	PR-44	4	EET129	PR-169
1	R13	PR-153			
1	SGU59	PR-84	5	UF122	PR-73
1	M-R 8	PR-185	5	UF221	PR-183
2	P22	PR-42	6	CC38	PR-40
2	R10	PR-18	6	MCC-38	PR-184
2	R021-R	PR-162			
2	R34	PR-118	7	CC39	PR-39
2	R41	PR-22	7	CC49	PR-123
2	R52	PR-25			
2	R75	PR-29	8	CAS3	PR-122
2	R105	PR-32	8	IMC47	PR-105
2	SGU91	PR-85	8	SIC1	PR-129
2	SIC72	PR-155			
2	TSAN812	PR-97	9	APA 4	PR-63
2	P8	PR-34	9	APA 5	PR-110

facilities now enable systematic assessment of genetic identity in the national and international cacao gene banks. In contrast to dominant markers such as AFLP, RAPD and ISSR, identical genotypes can have a 100% match in the multilocus SSR profiles without ambiguity; thus, accuracy of identification is significantly improved. Nevertheless, the effectiveness of clone identification via SSR fingerprints depends on the number of loci used for genotyping, as well as on the rate of genotyping error. Paetkau (2004), for example, recommends genotyping samples one time at six loci and using stringent quality control protocols to avoid genotyping error.

The multitube approach has emerged as a widely accepted protocol for eliminating errors from noninvasive samples (Taberlet and Luikart, 1999). Taberlet et al. (1996) recommended genotyping such samples three to seven times.

However, repeated genotyping will increase the costs. Recently, an alternative approach, the mismatch tolerant approach, was proposed to tackle the problem of genotyping error. This approach accepts that genotyping error might not be completely eliminated, but it proposes that accurate clone identification can be achieved by using a relatively large number of loci (McKelvey and Schwartz, 2004; Kalinowski et al., 2006). The mismatch tolerant approach assumes that the genotyping error is reasonably small; thus, these errors will cause samples from the same clone to differ by a very small number of loci. If enough loci are scored, different clones in the population will have multilocus genotypes that mismatch by more than one locus. In this way, the error mismatch distributions are unlikely to overlap with genotyping difference distributions, which would allow the differentiation of mismatch caused by genotyping error, from that caused by real genotype differences (Kalinowski et al., 2006; McKelvey and Schwartz, 2004). When error rate is small, this method was shown to have the potential to reduce the number of PCR amplifications required in the multitube approach (Kalinowski et al., 2006). For example, McKelvey and Schwartz (2004) recommend using as many as 12 to 15 loci so that genotyping error mismatch distributions are unlikely to overlap with genotyping difference distributions.

Our results showed that a combination of the mismatch tolerant approach with the multitube approach was suitable for cacao clone identification using high throughput genotyping. This scheme used the multilocus profile, generated by single genotyping at 15 SSR loci, to identify clones with unique genotypes (i.e., SSR profile mismatched at more than three loci). Then the accessions with mismatch caused by genotyping errors (with mismatch at three loci or fewer) were genotyped three times to verify their true genotype at the disputed loci. In this manner, the need for PCR replication was limited to a small number of loci. As demonstrated in the case of USDA Mayaguez cacao collection, reliable clone identification was obtained. This method is practical and cost effective for assessing genetic identity of a large number of cacao germplasm accessions. Similar strategies can also be applied for cultivar identification in other crops and species.

Among the 15 SSR loci used in the present study, a few of them are less than ideal to serve the purpose of cacao clone identification (i.e., loci Y16986, Y16996, AJ271826, and AJ271942) because of their relatively high error rate. These 15 loci constitute the first set of SSR agreed on by various collaborating institutions for cacao clone identification (Saunders et al., 2004). The selection was made at the time when only limited cacao microsatellite loci were available. Since then, many more SSR sequences have been developed (Pugh et al., 2004). Screening for more informative and less erroneous SSRs is a continuous process in our laboratory. Meanwhile, optimization of the PCR reaction is also ongoing. The genotyping error rate will be reduced with the use of new primers and fully optimized PCR reactions.

Like many other cacao germplasm collections, the USDA Mayaguez cacao germplasm collection suffers from duplicate accessions and other types of mislabeling. In the present study, we only used samples taken from a single tree per accession. We conclude that accessions that have been transferred between cacao germplasm

Reproduced from Crop Science. Published by Crop Science Society of America. All copyrights reserved.

collections are frequently subject to errors in identification. Differences occur among germplasm collections as well as within collection. Therefore, the present study only judged, in case of full SSR profile matches, which accession (comprising a group of trees) had the problem of synonymous mislabeling (or duplicates). However, we cannot confirm that the accessions having unique SSR profiles are not mislabeled.

Samples for this research were collected in 2001 from trees in the original cacao collection established in 1960 at Mayaguez. In 2002, a new replicated collection was established at Mayaguez using Amelonado as a common rootstock and many of the duplicates identified in this study were deleted (Table 3). Works continues on the 15 SSR primers used in this study to discover if clones with the same accession number are mislabeled.

Germplasm collections invariably contain duplicate accessions, which burdens the effectiveness of genebank management because these redundancies do not contribute to the diversity in the collection. The present study has provided the first step to identify the duplicates in a cacao germplasm collection, allowing the elimination of redundancy to certain extent. However, this step is not sufficient to fully correct the mislabeling in this collection. All the cacao accessions in the USDA Mayaguez repository were introduced from various original collections in Central and South America. Therefore, the correction of mislabeling in the Mayaguez collection will have to be based on the "reference multilocus" profile of the original trees in the source genebanks. Currently, we are genotyping original trees in the two international cacao collections, in Trinidad (the ICG, T) and in Costa Rica (the CATIE collection). Further activities in assessing the intra-accession variation will be necessary to fully sort out the mislabeling in the Mayaguez collection.

ACKNOWLEDGMENTS

We thank Stephen Pinney and Eric Tilson for their contributions to the genotyping. We are grateful to Ainong Shi and Lambert Motilal for review of the manuscript.

REFERENCES

Bonin, A., E. Bellemain, P. Bronken, F. Eidesen, C. Pompanon, C. Brochmann, and P. Taberlet. 2004. How to track and assess genotyping errors in population genetic studies. Mol. Ecol. 13:3261–3273.

Broquet, T., and E. Petit. 2004. Quantifying genotyping errors in noninvasive population genetics. Mol. Ecol. 13:3601–3608.

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

Coe, S.D., and M.D. Coe. 1996. p. 16–34. *In* The true history of chocolate. Thames & Hudson, Ltd., London.

Cryer, N.C., M.G.E. Fenn, C.J. Turnbull, and M.J. Wilkinson. 2006. Allelic size standards and reference genotypes to unify international cocoa (*Theobroma cacao* L.) microsatellite data. Genet. Resour. Crop Evol. Available at: http://dx.doi.org/10.1007/s10722-005-1286-9; verified 18 July 2006.

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

Evett, I.W., and B.S. Weir. 1998. Interpreting DNA evidence: Statistical genetics for forensic scientists. Sinauer, Sunderland, MA.

Griffiths, L.A. 1958. Phenolic acids and flavonoids of *Theobroma cacao* L. Separation and identification by paper chromatography. Biochem. J. 70:120–125.

Hoffman, J.I., and W. Amos. 2005. Microsatellite genotyping errors: Detection approaches, common sources and consequences for paternal exclusion. Mol. Ecol. 14:599–612.

John, M.E. 1992. An efficient method for isolation of RNA and DNA from plants containing polyphenolics. Nucleic Acids Res. 20:2381.

Kalinowski, S.T., M. Sawaya, and M. Taper. 2006. Individual identification and distributions of genotypic differences between individuals. J. Wildlife Manage. (In press).

Katterman, F.R.H., and V.I. Shattuck. 1983. An effective method of DNA isolation from the mature leaves of *Gossypium* species that contain large amounts of phenolic terpenoids and tannins. Prep. Biochem. 13:347–359.

Kennedy, A.J., and V. Mooleedhar. 1993. Conservation of cocoa in field genebanks—The International Cocoa Genebank, Trinidad. p. 21–23. *In* Proc. Int. Workshop on Conservation, Characterization and Utilization of Cocoa Genetic Resources in the 21st Century. 13–17 Sep. 1992. The University of the West Indies, St. Augustine, Trinidad and Tobago.

Lanaud, C., A.M. Risterucci, I. Pieretti, M. Falque, A. Bouet and P.J.L. Lagoda. 1999. Isolation and characterization of microsatellites in *Theobroma cacao* L. Mol. Ecol. 8:2141–2143.

Lanaud, C., J.C. Motamayor, and A.M. Risterucci. 2001. Implications of new insight into the genetic structure of *Theobroma cacao* L. for breeding strategies. p. 89–107. *In* Proceedings of the international workshop on new technologies and cocoa breeding. INGENIC, London.

Liu, K., and S.V. Muse. 2005. PowerMarker: An integrated analysis environment for genetic marker analysis. Bioinformatics 21: 2128–2129.

Lockwood, C., and M. End. 1993. History, technique and future needs for cacao collection. p. 1–14. *In* Proc. Int. Workshop on Conservation, Characterization and Utilization of Cocoa Genetic Resources in the 21st Century. 13–17 Sep. 1992. The University of the West Indies, St. Augustine, Trinidad and Tobago.

McKelvey, K. S., and M.K. Schwartz. 2004. Genetic errors associated with population estimation using non-invasive molecular tagging: Problems and new solutions. J. Wildlife Manage. 68:439–448.

Motamayor, J.C., A.M. Risterucci, P.A. Lopez, C.F. Ortiz, A. Moreno, and C. Lanaud. 2002. Cacao domestication I: The origin of the cacao cultivated by the Mayas. Heredity 89:380–386.

Motamayor, J.C., A.M. Risterucci, M. Heath, and C. Lanaud. 2003. Cacao domestication II: Progenitor germplasm of the Trinitario cacao cultivar. Heredity 91:322–30.

Paetkau, D. 2003. An empirical exploration of data quality in DNA-based population inventories. Mol. Ecol. 12:1375–1387.

Paetkau, D. 2004. The optimal number of markers in genetic capture-mark-recapture studies. J. Wildlife Manage. 68:449–452.

Perry, M.D., M.R. Davey, J.B. Power, K.C. Lowe, H.F.J. Bligh, P.S. Roach, and C. Jones. 1998. DNA isolation and AFLP genetic fingerprinting of *Theobroma cacao* (L.). Plant Mol. Biol. Rep. 16:49–59.

Powell, W., M. Morgante, C. Andre, M. Hanafey, J. Vogel, S. Tingey, and A. Rafalski. 1996. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol. Breed. 2:225–238.

Pugh, T., O. Fouet, A.M. Risterucci, P. Brottier, M. Abouladze, C. Deletrez, B. Courtois, D. Clement, P. Larmande, J.A.K. N'Goran, and C. Lanaud. 2004. A new cacao linkage map based on codominant markers: Development and integration of 201 new microsatellite markers. Theor Appl. Genet. 108:1151–1161.

Saunders, J.A., A.A. Hemeida, and S. Mischke. 2001. USDA DNA fingerprinting programme for identification of *Theobroma* cacao accessions. p. 108–114. *In* F. Bekele et al (ed.) Proc. Int. Workshop on new technologies and cocoa breeding. Kota Kinabalu, Sabah, Malaysia. 16–17 Oct. 2000. INGENIC Press, London.

Saunders, J.A., S. Mischke, E.A. Leamy, and A.A. Hemeida. 2004. Selection of International molecular standard for DNA fingerprinting. Theor. Appl. Genet. 110:41–47.

Schnell, R.J., C.T. Olano, J.S. Brown, A.W. Meerow, C. Cervantes-

Martinez, C. Nagai, and J.C. Motamayor. 2005. Retrospective determination of the parental population of superior cacao (*Theobroma cacao* L.) seedlings and association of microsatellite alleles with productivity. J. Amer. Soc. Hort. Sci. 130:181–190.

Sounigo, O., Y. Christopher, F. Bekele, V. Mooleedhar, and F. Hosein. 2001. The detection of mislabelled trees in the International Cocoa Genebank, Trinidad (ICG,T) and options for a global strategy for identification of accessions. p. 34–39. *In* F. Bekele et al (ed.) Proc Int. Workshop on new technologies and cocoa breeding. Kota Kinabalu, Sabah, Malaysia. 16–17 Oct. 2000. INGENIC Press, London.

Taberlet, P., S. Griffin, B. Goossens, S. Questiau, V. Manceau, N. Escaravage, L. Waits, J. Bouvet. 1996. Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Res. 24:3189–3194.

Taberlet, P., and G. Luikart. 1999. Non-invasive genetic sampling and individual identification. Biol. J. Linn. Soc. 68:41–55.

Turnbull, C.J., D.R. Butler, N.C. Cryer, C. Lanaud, D. Zhang, A.J. Daymond, and P. Hadley. 2004. Tackling mislabelling in cocoa germplasm collections. INGENIC Newsl. 9:8–11.

Valière, N. 2002. GIMLET: A computer program for analysing genetic individual identification data. Mol. Ecol. Notes 2:377–379.

Valière, N., P. Berthier, D. Mouchiroud, D. Pontier. 2002. GEMINI: Software for testing the effects of genotyping errors and multitube approach for individual identification. Mol. Ecol. Notes 2:83–86.

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

Young, A.M. 1994. The chocolate tree, A natural history of cacao. Smithsonian Institution Press, Washington, DC.

Reproduced from Crop Science. Published by Crop Science Society of America. All copyrights reserved.