Microsatellite fingerprinting in the International Cocoa Genebank, Trinidad: accession and plot homogeneity information for germplasm management

Lambert A. Motilal*, Dapeng Zhang, Pathmanathan Umaharan, Sue Mischke, Stephen Pinney and Lyndel W. Meinhardt

1 Cocoa Research Unit, The University of the West Indies, St. Augustine, Trinidad, Rep. Trinidad and Tobago, West Indies and 2 USDA/ARS, Beltsville Agricultural Research Center, PSI, SPCL, 10300 Baltimore Avenue, Bldg. 001, Rm. 223, BARC-W, Beltsville, MD 20705, USA

Received 25 September 2010; Accepted 27 February 2011

Abstract
The International Cocoa Genebank, Trinidad, is the largest field genebank collection of cacao (*Theobroma cacao* L.) in the public domain and the correct identity of each tree is crucial for germplasm movement, evaluation and phenotypic characterization. Nine microsatellite loci were used to assess the identity of 1477 trees from 486 cacao accessions representing approximately 16.9% of the trees and 29.2% of the accessions within the genebank. Heterogeneous plots (plots containing more than one genotype group) averaged 25.1% in The International Cocoa Genebank, Trinidad, with maximal admixture (32.6%) being recorded in Field 5B. The error rate did not differ significantly among different fields. Mislabeling error could be affected by accession grouping with an average error rate of 27.4% for accession groups in the genebank. Synonymous accessions were estimated to account for 14.4% of the field genebank. The results of the present study provide essential information for the management and utilization of the germplasm collection. Single-tree genotyping of every tree in this collection is strongly recommended.

Keywords: field genebank; mislabelling; simple sequence repeat; *Theobroma cacao*; tree identity; verification

Introduction

*Theobroma cacao* L. (cacao) is a tropical understory tree, whose seeds are the raw materials for making chocolate. Cacao is predominantly an outcrossing species with recalcitrant seeds (Toxopeus, 1985). Therefore, germplasm must be maintained in living genebanks. The International Cocoa Genebank, Trinidad (ICG,T), managed by the Cocoa Research Unit, is the largest public international cacao germplasm collection, containing over 2000 accessions. Each accession in the genebank is a putatively unique genotype. Accession nomenclature follows that recommended by Turnbull and Hadley (2011). This takes an alphanumeric form, where the names are assigned according to the farm, region, germplasm type or a combination of these. The alpha code of the accession is taken to represent an accession group. The accessions AM 1/19 [POU] and AM 2/12 [POU] would therefore belong to the same accession group (AM). In contrast, the accession SCA 6 would belong to the SCA accession group. Details on accession groups and accessions can be found in studies by Wood and Lass (1985), Kennedy and Mooledhar (1993), Iwaro et al. (2003), Turnbull et al. (2004) and Bartley (2005).
Diverse cacao germplasm material was brought to Trinidad as seed or budwood from multiple collecting expeditions (1930 onwards) from Amazonian South America, Central America and the West Indies (Kennedy and Mooleedhar, 1993). The early cacao research and breeding programs at the Imperial College of Tropical Agriculture (now the St. Augustine Campus of The University of the West Indies) resulted in various progeny and other selected material being planted in various estates throughout the island. Initial cacao germplasm sites were at the Imperial College of Tropical Agriculture, Las Hermanas Estate, Marper Estate, San Juan Estate and St. Joseph Estate. The demand for land, lack of adequate management, loss of trees from natural causes and the ageing trees led to the consolidation of these cacao germplasm into one site.

Formally planned in 1982, the ICG,T was established on a portion of land from the La Reunion Estate, which was once a cacao estate. The road access, bed system and intricate drainage system of the lands of the original estate were retained. Additional drains were dug as the internal drainage of the soil was moderate. The genebank consists of five adjacent but non-contiguous fields (Fields 4A, 5A, 5B, 6A and 6B) that were established continually from 1986 to 1994. The five fields are each subdivided into sections which are further split into plots. Each plot was planned to contain a maximum of 16 replicate trees of an accession, with a core group of four trees surrounded by peripheral guard rows. Tree numbering is consistent in orientation for all plots. Each tree is given a unique identifier based on its field, section, plot and tree location. For example, a tree of the accession IMC 67 may be found at Field 6B, Section A, Plot 23 and Tree number 12. An assigned accession may be present in (a) different plots within the same section of a field, (b) more than one section within the same field, (c) more than one field or (d) only one plot. The last is the most common occurrence. In the majority of plots, each accession was replicated from rooted cuttings; however, later introductions were established from grafted plants. An accession plot is therefore expected to contain clonal trees of the named accession. When the accession is present in more than one plot, all trees are expected to be identical to each other and belong to the stipulated accession group.

Genebank error can be estimated at various levels including accession and accession group heterogeneity (frequency of accessions containing mislabelling), plot heterogeneity (frequency of plots with mixed genotypes), field error (frequency of mislabelling within a field) and tree mislabelling (frequency of mislabelled trees in the entire genebank). The term genotype group is used in this study to denote equivalent multilocus profiles. Mislabelling events are considered homonymous cases when the same accession name is assigned but different multilocus profiles are present. Synonymous mislabelling is encountered when different accession names are assigned but the same multilocus profile is present.

Mislabeled plants have been identified as a serious problem in germplasm collections (Hurka et al., 2004). Errors in germplasm collections have been reported for Cicer (Shan et al., 2005), French olive (Khadari et al., 2003), grape (Leão et al., 2009), persimmon (Badenes et al., 2005) and cacao (Figueira, 1998; Risterucci et al., 2001; Motilal and Butler, 2003). DNA fingerprinting using microsatellite markers has been proved useful in resolving identity issues in cacao collections (Figueira, 1998; Risterucci et al., 2001; Saunders et al., 2004; Cryer et al., 2006; Zhang et al., 2006). The error rates in the ICG,T have been continually assessed. Christopher et al. (1999) reported a 30% mislabelling rate for the ICG,T by accession from a sample of 500 trees from 117 accessions. Motilal (2005) reported an error rate of 27.8% in 298 trees. Sounigo et al. (2001) investigated, but did not formally report, the mislabelling rate on 132 accessions in the ICG,T with the dominant marker system of randomly amplified polymorphic DNA. Examination of their results and allowing a flexibility of mistyping when only one primer differentiated trees within the same accession yielded a 40.9% mislabelling rate. Reference germplasm from which budwood was sourced for the establishment of the ICG,T contained mislabelling errors of 27.3% in 482 Refractario accessions (Zhang et al., 2008) and 29.4% in 612 Upper Amazon cacao accessions (Zhang et al., 2009a). The number of microsatellites employed has varied among studies. In cacao, nine loci were shown to be suitable for detecting mislabelling errors on a capillary sequencer system (Motilal et al., 2009).

The present study focuses on elucidating the error rate as heterogeneity at the plot, accession and field levels in the largest international field genebank of cacao.

Materials and methods

Plant material

Five hundred and twenty-five cacao (T. cacao L.) accessions comprising 1477 trees within the ICG,T were sampled (Table 1). These samples represented approximately 30% of the accessions and 17% of the trees within the genebank. Additionally, 18 reference accessions taken from three original planting sites in Trinidad and two reference samples from Peru were included. The complete list of samples can be obtained upon request.
DNA extraction, amplification and fragment analysis

Leaf genomic DNA was extracted with a modified protocol from Kobayashi et al. (1998), as described earlier (Motilal et al., 2009), or with the DNeasy plant system (Qiagen Inc., Valencia, CA, USA), according to Saunders et al. (2004). Nine microsatellite primer pairs (mTcCIR12, 15, 26, 33, 37, 42, 57, 243 and 244) were assessed. Characteristics of these primers can be found in studies by Lanaud et al. (1999), Saunders et al. (2004) and Pugh et al. (2004). Microsatellite amplification, separation and binning were carried out, as described by Motilal et al. (2009), on a Beckman Coulter capillary electrophoresis system (Fullerton, CA, USA).

Microsatellite typing error

Sixteen DNA samples were typed at each locus 3–20 times. The allele dropout (ADO) rate and false allele rate were assessed with GIMLET (Valière, 2002). The frequency of mistyping by a shift of two base pairs and ADO at the first allele or second allele of a heterozygote were calculated.

Multilocus matching

The allelic dataset was checked for binning errors with The Excel Microsatellite Toolkit v.3.1.1. add-in (Park, 2001). Match declaration (no flexibility) was performed using the regroup option in the software GIMLET (Valière, 2002). Declarations were given some flexibility by allowing one locus mismatch with CERVUS v3.0.3 (Kalinowski et al., 2007). Final declarations were guided by the outcome of the frequency estimate from the previous section. Mismatching arising from few loci, which exhibited the highest ADO or frequency of base pair shift, was discounted and the samples were deemed equivalent. Probabilities of identity (Waits et al., 2001) were determined using the software GIMLET (Valière, 2002).

Mislabelling error estimation

Designated accessions containing at least two trees were examined for heterogeneity from the output of the previous section. The number of heterogeneous cases was determined for (a) accessions present in more than one plot in the same field, (b) accessions present in more than one field, (c) plots over all fields, (d) accession groups and (e) the entire genebank. Contingency tables were constructed and the distribution was subjected to chi-square and Spearman’s correlation tests using the Contingency table programs v3.0 (Chang, 2001), according to the methodology of Siegel and Castellan (1988).

Mislabelling within accessions groups was assessed by utilizing accession groups that had more than one tree/accession. Five accession groups with a total of six trees were discarded yielding a dataset of 480 accessions. The AM, B, CL, JA, LP and NA accessions groups contained at least seven accessions exhibiting errors. This satisfied the chi-square association test of a minimum value of 5 in any cell. The remaining accession groups that contained less than five accessions with errors were therefore randomly assigned into three groups (Other 1, Other 2 and Other 3). Contingency analysis on these nine accession groupings was then performed as before.

Accessions containing at least three trees were categorized for heterogeneity as containing one, two or at least three genotype groups.

Synonymy in the ICG,T

To assess synonymy, the full dataset was reduced by (a) taking only one tree to represent a homogenous plot, (b) keeping trees that exhibited differing profiles

<p>| Table 1. Total numbers of accessions and trees present in five fields in the ICG,T and number of accessions and trees fingerprinted with nine microsatellite loci |</p>
<table>
<thead>
<tr>
<th>Field</th>
<th>Number of unique accessions present/field</th>
<th>Number of accessions analysed&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Number of trees present</th>
<th>Number of trees analysed&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Average number of trees/accession analysed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field 4A</td>
<td>548</td>
<td>168 (30.7)</td>
<td>1273</td>
<td>357 (28.0)</td>
<td>2.1</td>
</tr>
<tr>
<td>Field 5A</td>
<td>351</td>
<td>63 (18.5)</td>
<td>1408</td>
<td>160 (11.4)</td>
<td>2.5</td>
</tr>
<tr>
<td>Field 5B</td>
<td>636</td>
<td>190 (29.9)</td>
<td>3699</td>
<td>658 (17.8)</td>
<td>3.5</td>
</tr>
<tr>
<td>Field 6A</td>
<td>100</td>
<td>44 (44.0)</td>
<td>401</td>
<td>105 (26.2)</td>
<td>2.4</td>
</tr>
<tr>
<td>Field 6B</td>
<td>374</td>
<td>58 (15.8)</td>
<td>1939</td>
<td>197 (10.2)</td>
<td>3.4</td>
</tr>
<tr>
<td>Total</td>
<td>1765</td>
<td>525 (29.7)</td>
<td>8720</td>
<td>1477 (16.9)</td>
<td>2.8</td>
</tr>
</tbody>
</table>

<sup>a</sup>Numbers in parentheses refer to percentages of the total.
<sup>b</sup>Represent total accessions without replicates across fields.
within an accession, (c) obtaining a consensus genotype from samples bearing the same accession name and attributed to the same genotype. A reduced dataset of 613 trees inclusive of ten unique reference accessions was assessed for multilocus matches with GIMLET (Valière, 2002) and with CERVUS v3.0.3 (Kalinowski et al., 2007). A mismatch at one locus was allowed for the latter. The output was further refined by discarding pairwise matches, in which only one locus differed but with differential heterozygotes at the said locus. The number of distinct accessions that could occur in the entire genebank based on this subsample was estimated following van Hintum (2000):

\[
N_{\text{dist}} = f_{\text{dist}} N_{\text{acc}}, \quad \text{where } f_{\text{dist}} = \sum_{i} f_i / i,
\]

where \(N_{\text{acc}}\) is the total number of accessions in the collection (set as 2000), \(f_i\) is the fraction of accessions which appears \(i\) times in the collection and \(f_{\text{dist}}\) is the fraction of distinct accessions in the collection.

The variance of \(f_{\text{dist}}\) is \(\sigma^2 = \frac{(\ln N_{\text{acc}} - f_{\text{dist}} N_{\text{acc}} - k)}{N_{\text{acc}} - 1}\), where \(k\) is the sample size (603) and the standard error of \(N_{\text{dist}}\) is given by \(\sigma_{N_{\text{dist}}} = N_{\text{acc}} \sqrt{\frac{\sigma^2_{f_{\text{dist}}}}{N_{\text{acc}}}}\) (Sokal and Rohlf, 1981).

### Results

#### Locus error rate

Mistyping by two basepairs occurred at a frequency of 0–0.02 across loci and 0–0.04 across samples. False alleles were absent. The ADO rate was estimated in GIMLET (Valière, 2002) as 0.053 across loci and ranged from 0.00 to 0.15 with four samples contributing to the maximum rate. Error as ADO ranged from 0 to 0.06 and 0 to 0.17 over samples. The ADO ranged from 0 to 0.03 and 0 to 0.08 over loci at the first and second alleles, respectively. For heterozygous cases, average ADO was estimated as 0.01 and 0.02 at the first and second alleles, respectively.

#### Plot heterogeneity

Heterogeneous plots (plots containing more than one genotype) averaged 25% in the ICG,T, with maximal admixture (33%) being recorded in Field 5B (Table 2). However, the field identity did not significantly influence...
error scores \( (\chi^2 = 4.2, \text{d.f.} = 4, P = 0.38\text{; } r_s = 0.04, \text{d.f.} = 435, P = 0.20) \). Heterogeneous plots ranged from 9.1–53.8% by field section. When sections were pooled to obtain valid size classes, chi-square analysis showed that the error score was not influenced by section groupings (Table 3; \( \chi^2 = 4.3, \text{d.f.} = 7, P = 0.74\text{; } r_s = 0.15 \)). Further analysis using randomly combined field sections returned a similar result (\( \chi^2 = 6.2, \text{d.f.} = 7, P = 0.51\text{; } r_s = 0.05, \text{d.f.} = 411, P = 0.14 \)).

### Accession heterogeneity

Four accessions (JA 5/47 [POU], LCT EEN 162/S-1010, LP 1/21 [POU] and NA 471) were each represented by two plots in one field. One accession (LCT EEN 162/S-1010) exhibited differential genotypes between plots. Thirty-eight accessions were present in two fields and 55% of these were different between the fields. In this study, the sub-sample of the ICG,T had 40 accession groups, which contained at least two trees/accession. A range of 0–100% heterogeneity levels was observed in these groups. Analysis of a constructed dataset with appropriate class sizes revealed that mislabelling error may be affected by the accession groups (Table 4). Chi-square testing returned a non-significant result (\( \chi^2 = 8.1, \text{d.f.} = 8, P = 0.42 \); Spearman’s \( r_s = -0.12, \text{d.f.} = 423, P = 0.01 \)). Approximately, 29% (486) of the accessions of the ICG,T were fingerprinted (Table 1) and, of these, 332 accessions contained at least two putative clonally propagated trees (Table 4). In the latter subset, 28% contained mislabelling errors. Two hundred and seven accessions contained at least three putatively clonally propagated trees and, of these, 35% were heterogeneous (Fig. 1).

### Synonymies

Summary statistics with The Excel Microsatellite Toolkit add-in (Park, 2001) on the 613 accessions with nine loci revealed a mean number of 13.9 ± 4.1 alleles and an unbiased gene diversity of 0.75 ± 0.02. Polymorphism estimates (Botstein et al., 1980) per loci ranged from 0.64 to 0.79 and averaged 0.72 over loci. Probabilities of identities as full siblings ranged from 6.1 \times 10^{-3} to 1.18 \times 10^{-5}. Implementing the regroup option in GIMLET (Valière, 2002) detected 582 groups, resulting in an estimated 5.1% synonymy in the dataset of 613 accessions. Flexibility matching in CERVUS v3.0.3 (Kalinowski et al., 2007) identified similar multilocus profiles in the dataset of 613 accessions for 20 couplets, two triplets and four quadruplets. Full concordance or ADO at the second position for one locus was observed for these groups. Nine couplets, one triplet and two quadruplet groups were matched with possible ADO at the first position. A mixture of these two profiles was observed and was

### Table 4. Heterogeneity levels within cacao accession groups in the ICG,T

<table>
<thead>
<tr>
<th>Accession group</th>
<th>Total number of accessions in ICG,T</th>
<th>Number of fingerprinted accessions with at least one tree</th>
<th>Number of accessions analysed with at least two trees</th>
<th>Number of analysed accessions with errors</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>73</td>
<td>30</td>
<td>22</td>
<td>7</td>
<td>31.8</td>
</tr>
<tr>
<td>B</td>
<td>80</td>
<td>38</td>
<td>26</td>
<td>13</td>
<td>50.0</td>
</tr>
<tr>
<td>CL</td>
<td>90</td>
<td>32</td>
<td>22</td>
<td>8</td>
<td>36.4</td>
</tr>
<tr>
<td>JA</td>
<td>141</td>
<td>65</td>
<td>42</td>
<td>14</td>
<td>33.3</td>
</tr>
<tr>
<td>LP</td>
<td>79</td>
<td>28</td>
<td>20</td>
<td>7</td>
<td>35.0</td>
</tr>
<tr>
<td>NA</td>
<td>200</td>
<td>47</td>
<td>27</td>
<td>8</td>
<td>29.6</td>
</tr>
<tr>
<td>Other 1</td>
<td>383</td>
<td>57</td>
<td>42</td>
<td>7</td>
<td>16.7</td>
</tr>
<tr>
<td>Other 2</td>
<td>196</td>
<td>65</td>
<td>47</td>
<td>11</td>
<td>23.4</td>
</tr>
<tr>
<td>Other 3</td>
<td>406</td>
<td>117</td>
<td>84</td>
<td>16</td>
<td>19.1</td>
</tr>
<tr>
<td>Total</td>
<td>332</td>
<td>91</td>
<td></td>
<td></td>
<td>27.4</td>
</tr>
</tbody>
</table>

\( \chi^2 = 8.1, \text{d.f.} = 8, P = 0.42; r_s = -0.12, \text{d.f.} = 423, P = 0.01. \)

### Fig. 1. Degree of admixture as number of multilocus profiles (genotype groups) within cacao accessions in the International Cocoa Genebank, Trinidad. Values are numbers of accessions with corresponding percentages.
present in three (in three groups), four (in one group) or five (in one group) samples.

With the ten reference DNAs removed, 498 accessions were uniquely identified and 41 groups containing more than one accession were observed (29 couplets, 6 triplets, 7 quadruplets and 1 quintuplet). A synonymous rate of 10.6 and 17.4% was estimated for accession grouping and tree sampling, respectively. The number of distinct accessions ($N_{\text{dist}}$) in the ICG,T based on this study with $N_{\text{acc}}$ set at 2000 was estimated as $1713 \pm 24$ accessions from the formula of van Hintum (2000). Hence, a synonymous error rate of 14.4% was modelled for the entire genebank collection.

**Discussion**

Mislabelling within the ICG,T, the largest public domain field genebank for cacao, was estimated at an overall rate of 28% by accessions and 25% by plots. Although error rates varied among fields, the distribution was non-significant at both the entire field and subsection groupings. This suggested that random errors were the main cause of mislabelling. The error rate varied depending on the accession grouping (Table 4), indicating that batch jobs during planting could have had inadvertent admixture. Several reasons were advanced to account for mislabelling error (Turnbull et al., 2004). Another factor is that during the establishment phase of the genebank, more than one tree designated as a particular accession was available for budwood collection. At that time, molecular methods were unavailable and full confidence was placed on the identity of these trees, provided that the fruit morphology was compliant with the accession nomenclature. Thus, faithful propagation, greenhouse establishment and field planting may have occurred. However, if the trees, from which the budwood was collected, were dissimilar, then admixture within a plot or accession would result. Erroneous budwood collection from overlapping branches would also be a contributing factor.

The mislabelling rate by accession (27%) represents the level of homonymous cases within the genebank. The level of synonymies was estimated between 10.6 and 17.4% when flexibility to match declarations was given, a twofold increase compared with that without flexibility. An estimate from modelling set the value at 14.4% redundancy. This may be an upper limit as increasing the number of discriminating microsatellite loci would (a) confirm the separation of accessions which differ at only one locus from ADO or mis-typing, (b) split accession groups into individuals and (c) decrease the likelihood of multilocus matches.

The error rate reported (28% by accession) here is lower than that reported earlier (59.3%) by Motilal et al. (2009) for the same genebank. This may be ascribed to sample size and composition effects. The smaller sample size in the previous study leads to biased reporting as it does not adequately capture the genebank. Higher error values will result when accessions with mislabelling events are predominantly represented. When larger subsamples of the ICG,T are examined (Christopher et al., 1999; Soungio et al., 2001; Motilal, 2005; Zhang et al., 2008, 2009a), similar error levels were observed. Excluding the works by Motilal et al. (2009), an overall average mislabelling error of 30.6% for the ICG,T was estimated from these workers and the present study.

<table>
<thead>
<tr>
<th>Location</th>
<th>Sample size</th>
<th>Error estimate</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malaysia and Brazil</td>
<td>11 accessions</td>
<td>27.3%</td>
<td>Figueira (1998)</td>
</tr>
<tr>
<td>Nine germplasm collections</td>
<td>28 accessions</td>
<td>30.0%</td>
<td>Risterucci et al. (2001)</td>
</tr>
<tr>
<td>Global cacao genebanks</td>
<td>335 accessions</td>
<td>43.9%</td>
<td>Motilal and Butler (2003)</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>285 offspring of 9 bi-parental crosses</td>
<td>52.3%</td>
<td>Takrama et al. (2005)</td>
</tr>
<tr>
<td>University of Reading Intermediate</td>
<td>345 accessions; 429 trees</td>
<td>5.2%</td>
<td>Cryer et al. (2006)</td>
</tr>
<tr>
<td>Quarantine facility</td>
<td>141 accessions</td>
<td>2% Homonymyb</td>
<td>Zhang et al. (2006)</td>
</tr>
<tr>
<td>Puerto Rico</td>
<td>143 accessions</td>
<td>18.4% Synonymyb</td>
<td>Zhang et al. (2009)</td>
</tr>
<tr>
<td>Trinidad and Costa Rica</td>
<td>688 accessions</td>
<td>5.6% Synonymyb</td>
<td>Johnson et al. (2009)</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>154 accessions; 924 trees</td>
<td>19.5% Synonymyb</td>
<td>Zhang et al. (2009b)</td>
</tr>
<tr>
<td>ICG,T</td>
<td>c. 2000 accessions</td>
<td>12.3% Homonymyb</td>
<td>Irish et al. (2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20.1% Synonymyb</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.6%</td>
<td>Present study</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>24.7%</td>
<td></td>
</tr>
</tbody>
</table>

a Error estimates are as quoted in reference or examination of reported data.
b Estimates not used in determining average.
Various error rates within other cacao germplasm collections have been encountered (Table 5). An average mislabelling error of 24.1% is suggested from these results. Incorporation of the ICG,T mislabelling error rate results in a conservative mean estimate of 24.7% mislabelling within cacao germplasm collections. This study therefore supports Motilal et al. (2009) in recommending verification of identities of single trees rather than pooling DNA from multiple trees of an accession.

Mislabelling estimates in other germplasm collections have been reported as 20% for apple cultivars (Baric et al., 2009); 21.7% for French olives (Khadari et al., 2003); 37.2% for Iranian olives (Noormohammadi et al., 2009); 31.9% for Mangifera indica (Duval et al., 2009); 27.8% for Moroccan fig (Khadari et al., 2005) and 33.0% for Nordic oat (Diederichsen, 2009). Data from this study and the references contained herein support the view that germplasm collections harbour substantial erroneous nomenclatures (van Hintum, 2000; Hurka et al., 2004).

Curators of cacao germplasm collections must therefore place the identification of distinct accessions as a priority. Several recommendations to deal with this issue have already been outlined (Motilal and Butler, 2003; Turnbull et al., 2004). Fingerprinting of every tree within a plot and of every tree of an alleged accession becomes an ongoing mission for many and has already been completed in one case (Irish et al., 2010). Van Hintum and Van Treuren (2002) raised the question of cost for the routine application of molecular markers for germplasm management and genebank efficiency. At the present time, running costs are the main concern as microsatellite markers have already been developed for cacao (Lanaud et al., 1999; Pugh et al., 2004). Furthermore, these costs are being reduced especially with the advent of single-nucleotide genotyping, which may be outsourced by genebank curators. Additionally, since cacao field genebanks are maintained as living trees originating as clonal replicates, the issue of an accession identity becomes more straightforward than for accessions maintained as seeds. Duplication issues and nomenclature errors in cacao collections can be more easily identified with high rigour with molecular markers.

Curators may seek to clarify redundancies within their own collection before addressing duplication issues between collections. This would facilitate autonomy. However, a true-type tree of every accession must sooner or later be identified. If possible, the most original material conforming to published descriptions and falling within the appropriate population group should be ascertained. If there is failure in the selection of a true-type tree from historical records, then a tree with characteristics agreed upon by the international cacao scientific community should be designated the true-type tree for that accession. For many internationally distributed accessions, the source material originated from the ICG,T. Reference profiles of the ICG,T material is therefore an important task to be completed.

The inclusion of true-type trees within a dataset would facilitate match declarations and alignment of multilocus profiles from different genotyping platforms. Cryer et al. (2006) recommended the use of reference genotypes to accurately compare multilocus microsatellite fingerprints. The advent of single-nucleotide polymorphism detection will, however, allow for a more reliable dataset as the mistyping level is expected to be decreased. The difference detected between any two samples will be due to actual sequence differences instead of fragment length polymorphism and will therefore have a greater potential for separation.

In addition to the management of the collection, users must be aware of the level of mislabelling that is present not only within the genebank as a whole, but within an accession group, among the trees of an accession and within plots of an accession. The permanent unambiguous labelling of all trees within the genebank, together with up-to-date accurate maps, is indispensable to users of a field genebank. However, it cannot be overemphasized that any sampling, whether for budwood for propagation or distribution, for phenotypic evaluations or for molecular determinations must always be accompanied by the full tree-location details. In addition, data collected over multiple trees of an accession should be reviewed and recoded in order to prevent the combining of data from different genotypes.

In conclusion, this is the first comprehensive study to use microsatellite multilocus profiles to estimate the mislabelling within the largest universal public domain collection of cacao. A collaborative fingerprinting project between the Cocoa Research Unit and the United States Department of Agriculture is underway to generate a DNA fingerprint from a reference tree of each accession. A future study is planned to utilize the full complement of microsatellite primers to allow for accurate accession assignment. The present study, in conjunction with that of Irish et al. (2010), Zhang et al. (2009b) and the ongoing fingerprinting within the ICG,T, will be useful examples of molecular management of field genebanks. The results of this study and the recommendations contained herein will direct researchers and users of the ICG,T in their ongoing evaluations and characterization of germplasm material.

Acknowledgements

Thanks to Ms. Alisha Omar-Ali for assisting with DNA extractions. Two anonymous reviewers are thanked...
for critiquing the manuscript. The research was made possible in part by a grant from the Government of Trinidad and Tobago Research Development Fund.

References


