Genetic Resources and Crop Evolution (2006) 53:1643–1652
DOI 10.1007/s10722-005-1286-9

Allelic size standards and reference genotypes to unify international cocoa (*Theobroma cacao* L.) microsatellite data

N.C. Cryer*, M.G.E. Fenn, C.J. Turnbull and M.J. Wilkinson
*School of Biological Sciences, The University of Reading, Whiteknights, P.O. Box 221, Reading, RG6 6AS, UK; *Author for correspondence: E-mail: n.c.cryer@reading.ac.uk; phone: +44-1183786085; fax +44-1183788160*

Received 10 February 2005; accepted in revised 25 June 2005

Key words: Germplasm, Microsatellite, Quarantine, SSR, *Theobroma cacao*

Abstract

Standardisation of microsatellite allele profiles between laboratories is of fundamental importance to the transferability of genetic fingerprint data and the identification of clonal individuals held at multiple sites. Here we describe two methods of standardisation applied to the microsatellite fingerprinting of 429 *Theobroma cacao* L. trees representing 345 accessions held in the worlds largest Cocoa Intermediate Quarantine facility: the use of a partial allelic ladder through the production of 46 cloned and sequenced allelic standards (AJ748464 to AJ48509), and the use of standard genotypes selected to display a diverse allelic range. Until now a lack of accurate and transferable identification information has impeded efforts to genetically improve the cocoa crop. To address this need, a global initiative to fingerprint all international cocoa germplasm collections using a common set of 15 microsatellite markers is in progress. Data reported here have been deposited with the International Cocoa Germplasm Database and form the basis of a searchable resource for clonal identification. To our knowledge, this is the first quarantine facility to be completely genotyped using microsatellite markers for the purpose of quality control and clonal identification. Implications of the results for retrospective tracking of labelling errors are briefly explored.

Introduction

The drive to increase the genetic base of important crop plants has led to an expansion in the number of *ex situ* germplasm repositories. There have been over 1000 new collections created since 1970 containing in excess of 6 million accessions (Cooper et al. 2001). The consequences of accession misidentification in these collections are profound for both breeding and research applications, and ultimately undermine the value of any collection to its users.

Verification of accession identity is to some extent reliant upon the structure of the collection: those based on seed populations must accommodate for some variation between individuals within an accession, whereas accessions within clonal collections should be genetically invariant. It follows that the task for ensuring accession fidelity is considerably simpler for clonal crops.

Cocoa, (*Theobroma cacao*) is a tree crop of huge importance to the economies of many developing world 'producer' countries and also to the predominantly first world and temperate 'consumer' countries. Germplasm collections maintain varieties of cacao as clones for distribution as breeding material. Many of these varieties are barely improved from wild, ancestral lines (Wood and Lass

1985; Motamayor et al. 2002) and yet are usually susceptible to a wide range of important pests and diseases (Wilson 1999). Different geographic regions typically vary in the profile of pests and diseases that attack the crop. This has led to growing fears that the movement of diseases could threaten continuity of cocoa supply on a regional or even global basis. The identification and distribution of elite resistant material will provide valuable insurance against the spread of disease. In spite of this, a large proportion of the genetic diversity of *Theobroma cacao* deposited in germplasm collections has not yet been exploited in breeding programmes (Eskes 2001). This can be attributed in part by the need to avoid inadvertent transmission of disease when potentially useful germplasm is exchanged between cocoa germplasm collections. For this reason, three international cacao quarantine collections were established at The University of Reading, Reading (UK), CIRAD, Montpellier (France) and USDA, Miami (USA) to facilitate transfer of material whilst minimising the probability of disease proliferation. The University of Reading Intermediate Cocoa Quarantine Facility maintains individual type specimens for each genotype it handles and is by far the largest and most active of these facilities, representing the major conduit for the distribution of disease-free stock around the world. Clearly, there is considerable scope for mislabelling to occur before, during and after the passage through quarantine. Such errors have great significance since they will transfer to the receiving germplasm and breeder's collections. Thus, mislabelling events before or during quarantine will tend to spread and accumulate. Such events will confound current efforts to establish replicated field trial data from sites across the globe. There is therefore a strong need to establish a reliable system for the routine screening of material for genetic fidelity during the quarantine period.

Previously, the identification of accessions held in germplasm collections has been achieved using various approaches including morphological descriptors, and molecular techniques such as RAPD, AFLP, ISSR, and microsatellites (also known as Simple Sequence Repeats or SSR's). It is vital that the methodology selected for global use generates consistent results across all users. Reference to similar large-scale fingerprinting problems in other organisms indicates that there has been a broad consensus among users towards the use of microsatellite analysis for this purpose. For example, microsatellite analysis that has emerged as the method of choice for the assemblage of genetic identification profiles for human forensic purposes (Goldstein and Schlotterer 1999) and have been extensively used for the identification of small to modest sized collections of plant species, including potato (McGregor et al. 2000), maize (Pejic et al. 1998), wheat (Dograr et al. 2000), and clover (Kolliker et al. 2001). Whilst the methodology is widely acknowledged to be reproducible, difficulty can lie in the standardization of profile scoring between laboratories, particularly when alleles differ only by one or two bases, as is the case for the dinucleotide motifs that dominate in studies of plants. Here, we address this problem by first presenting a comprehensive data set of microsatellite profiles for all material currently held in The University of Reading Intermediate Cocoa Quarantine Facility and then presenting a set of allelic and genotype standards to allow consistent allele assignment between laboratories. We believe that this is the first example of the systematic characterisation of an entire quarantine collection using microsatellite markers, although several other cacao germplasm collections are currently being characterised as part of an international effort to compile a global data set of genetic profiles for important cocoa germplasm. Once complete, the information generated will greatly improve the maintenance and exploitation of *ex situ* collections for the improvement of the crop by establishing records of the genetic identity of material held, defining the level of redundancy between collections, and allowing identification and correction of mislabelled accessions, thereby reducing the likelihood of errors during the transfer of material through quarantine. We believe that the approach adopted here is appropriate for the global genetic characterization of germplasm collections of other important clonal crops.

Methods

Plant material and DNA isolation

Total genomic DNA was extracted from 429 accessions of *Theobroma cacao* L. held in The University of Reading Intermediate Cocoa Quar-

antine Facility, using the DNeasy 96 Plant Kit (Qiagen Ltd, UK) according to the manufacturer's instructions. DNA was aliquoted and stored at −20 °C until use. A complete list of accessions analysed in this study is available at www.icgd.rdg.ac.uk.

Microsatellite markers

A panel of 17 microsatellite markers were selected on the basis of polymorphism between genotypes and yielding products with size ranges that allowed for multiplex analysis (Table 1). HPLC purified fluorescently labelled oligonucleotide primers were supplied by Sigma-genosys and Applied Biosystems.

Polymerase chain reaction and fragment analysis

PCR was performed as described by Lanaud et al. (1999) with the addition of fluorescent labels to primers as described in Table 1. PCR amplification was performed using a MJ Research PTC 100 thermal cycler in a reaction volume of 20 μL comprised of: 0.5 U *taq* polymerase 10 ng genomic DNA; 2.25 mM $MgCl_2$; 16 mM NH_4SO_4; 67 mM

Table 1. List of international microsatellite markers for global cacao fingerprinting project.

Marker	Primer sequence 5′ → 3′	Tm (°C)	Dye	Range
mTcCIR1	GCAGGGCAGGCTCAGTGAAGCA TGGGCAACCAGAAAACGAT	51	NED	110–155
mTcCIR6	TTCCCTCTAAACTACCCTAAAT TAAAGCAAAGCAATCTAACATA	46	NED	225–255
mTcCIR7	ATGCGAATGACAACTGGT GCTTTCAGTCCTTTGCTT	51	NED	145–170
mTcCIR8	CTAGTTTCCCATTTACCA TCCTCAGCATTTTCTTTC	46	HEX	275–315
mTcCIR11	TTTGGTGATTATTAGCAG GATTCGATTTGATGTGAG	46	NED	280–325
mTcCIR12	TCTGACCCCAAACCTGTA ATTCCAGTTAAAGCACAT	46	FAM	160–265
mTcCIR15	CAGCCGCCTCTTGTTAG TATTTGGGATTCTTGATG	46	HEX	225–268
mTcCIR17*	AAGGATGAAGGATGTAAGAGAG CCCATACGAGCTGTGAGT	51	FAM	256–296
mTcCIR18	GATAGCTAAGGGGATTGAGGA GGTAATTCAATCATTTGAGGATA	51	FAM	330–375
mTcCIR22	ATTCTCGCAAAAACTTAG GATGGAAGGAGTGTAAATAG	46	NED	251–301
mTcCIR24	TTTGGGGTGATTTCTTCTGA TCTGTCTCGTCTTTTGGTGA	46	NED	170–215
mTcCIR25*	CTTCGTAGTGAATGTAGGAG TTAGGTAGGTAGGGTTATCT	46	HEX	124–170
mTcCIR26	GCATTCATCAATACATTC GCACTCAAAGTTCATACTAC	46	FAM	275–315
mTcCIR33	TGGGTTGAAGATTTGGT CAACAATGAAAATAGGCA	51	NED	268–349
mTcCIR37	CTGGGTGCTGATAGATAAT ATTTGTGTGGAGGGTATT	46	FAM	130–190
mTcCIR40	AATCCGACAGTCTTTAATC CTTAAATGTTATGTGTATGC	51	HEX	247–287
mTcCIR60	CTTAAATGTTATGTGTATGC AGAGCAACCATCACTAATCA	51	NED	188–214

Primer sequences and map positions reported by Lanaud et al. (1999) and Risterucci et al. (2000). Tm, annealing temperature (°C); Dye, florescent label used in this study; Range, corrected size range of alleles in bp. * markers additional to the international agreed set for cacao.

Tris–HCl pH 8.8 @ 25 °C; 0.4 mM dNTP mix. The thermal cycling protocol consisted of 2 min at 96 °C followed by 35 cycles of 15 s at 96 °C, 15 s at either 46 or 51 °C and an elongation time of 2 min at 72 °C. A final elongation of 5 min at 72 °C preceded cooling to 10 °C for 10 min.

Amplified microsatellite products were diluted (1:100) using HPLC grade water (Aldrich) and fractionated on an ABI PRISM 3100 genetic analyser. Fragment sizes were determined relative to the 500-ROX internal size standard (Applied Biosystems) by the local southern method using Genotyper version 3.7 (Applied Biosystems). Allele binning was achieved by the method of Idury and Cardon (1997).

Cloning and sequencing of microsatellite alleles

Microsatellite products were amplified from genomic DNA using unlabelled primers, purified using Nucleofast 96 PCR clean-up plates (Macherey–Nagel) and cloned into pDrive vector as recommended by the manufacturer (Qiagen Ltd, UK). Single colonies were removed from LB plates, supplemented with 100 μg mL^{-1} ampicillin, and transferred into 10 μL sterile water. Aliquots of the resulting suspension (5 μL) were used to inoculate 5.0 mL LB supplemented with 100 μg mL^{-1} kanamycin, and grown overnight at 37 °C with 225 rpm orbital shaking. Plasmid DNA was then harvested from 3.0 mL of the culture using QIAprep kit (Qiagen Ltd, UK). The remaining 5 μL of cell suspension was subjected to PCR using the appropriately labelled primers and fractionated by capillary electrophoresis on a ABI 3100 genetic analyser (Applied Biosystems) to select clones containing microsatellite alleles of known size. The culture corresponding to these clones were harvested and the plasmid sequenced on both strands from the universal forward and reverse primer binding sites of pDRIVE using dye terminator cycle sequencing (Applied Biosystems). Sequence data was viewed using the Chromas software (Technelysium Pty Ltd, Queensland, Australia). The identity of cloned fragments was verified by BLASTn (Altschul et al. 1990) and alignments against other representatives of the same microsatellite marker produced using clustalw software (Higgins et al. 1994). Allele sizes as inferred from capillary fractionation (Genescan and Genotyper software, Applied Biosystems), were adjusted by comparison to allele sizes derived directly from the sequencing of cloned fragments.

Data analysis

Corrected allele sizes were recorded in a binary, presence/absence, notation and transformed to yield Jaccards coefficient. This was plotted using UPGMA to give a graphical representation of the data using MSVP software (version 3.11 h, Kovach Computing Services, Anglesey, Wales, UK). Allele frequencies (Supplementary Table 1, http://www.personal.rdg.ac.uk/~abs00ncc/cacao/Supplementary_Table1.htm) were reported by Cervus software (Marshall et al. 1998).

Results

The 17 microsatellite loci used generated 359 unique profiles from the 429 trees and 345 accession names held in The University of Reading Intermediate Cocoa Quarantine Facility. There was considerable redundancy between microsatellites in their ability to distinguish clones held. Indeed, it was possible to affect the same number of identifications on the basis of just 13 of the 17 polymorphic loci used. An exact match search function for the internationally agreed core 15 markers (Table 1) was created to allow new profiles to be compared to existing data at www.icgd.rdg.ac.uk/public_html/microsat_search_index.htm. Allele frequency data and corrected sizes of recorded alleles are reported in supplementary Table 1. Several microsatellite loci used in this study were complex and comprised of two or more repeat motifs. Sequence analysis confirmed the presence of more than one variable repeat motif contributed to the polymorphism in locus size. As it is not possible to derive which repeat unit is variable from capillary fractionation, we report the actual size of each locus rather than attempting to derive the repeat number from this value.

Cloned allele standard

The calling of allele sizes is hindered by an offset inherent to capillary electrophoresis and by

non-integer changes in measured allele size. Allele sizes have been corrected by reference to standard cloned alleles of known size for each marker, in addition to the internal ROX size standard (Applied Biosystems). The sequences of these cloned standards have been submitted to EMBL (Accession numbers AJ748464 to AJ48509). These standards act as exact size markers to assist with the allele calling of unknown material. This reference material is freely available upon request.

Allele calling

For some microsatellite markers, increases in locus size when estimated from capillary electrophoresis run time did not correspond exactly to associated increments in repeat number. Thus, there is potential for the miscalling of some alleles when a simple offset is used against the ROX size standard alone. This problem has been overcome by the use of several alleles of known size, which act as a calibration series when used in combination with the ROX size standards. Application of this approach allowed all genotypes to be distinguished without anomalies between replicate fractionation events. Such accuracy is essential in allowing datasets produced on different platforms and at distant locations to be compatible, and is also a prerequisite for accurate fingerprinting of the base collection held in the quarantine facility.

Standard genotype DNA standard

Genotype standards offer a less expensive alternative to the cloned standards and if used with the cloned standards, broaden the allele range of the reference standards. The allele sizes of all microsatellites across a selection of proposed standard genotypes are shown in (Table 2). Clearly, it is important that material used as a genotype standard is not taken from a mislabelled tree and so DNA or plant material from all genotype standards is also available upon request.

Table 2. Microsatellite allele sizes of selected CFC/ICCO/IPGRI clones.

Accession	RUQ	mTc-CIR11		mTc-CIR12		mTc-CIR15		mTc-CIR17		mTc-CIR18		mTc-CIR22		mTc-CIR33		mTc-CIR40		mTc-CIR60	
Linkage group 1–5 (Risterucci et al. 2000)																			
AMAZ 15-15	1	307	307	199	211	234	234	272	272	340	340	287	287	297	297	273	273	195	197
APA 4	843	317	317	250	252	234	234	272	272	344	344	289	289	297	297	266	266	212	212
CATIE 1000	844	287	317	203	252	234	234	272	272	344	354	287	287	297	297	266	266	195	212
eet 59	791	287	315	203	211	234	234	272	272	335	344	279	287	295	303	278	284	193	193
IMC 47	849	287	307	199	203	250	250	270	272	333	335	287	287	295	295	284	284	193	210
MAN 15-2	86	293	317	199	207	250	250	272	272	333	346	279	287	295	297	278	278	212	212
P 7	851	307	315	199	211	250	250	272	272	333	333	279	287	295	297	278	284	208	210
PA 120	852	315	315	199	199	201	252	272	272	337	344	279	287	303	305	278	279	191	212
PA 150	853	293	309	199	199	234	250	272	274	344	344	279	287	303	303	278	281	189	189
PLAYA ALTA 2	232	299	315	187	250	234	234	274	282	331	344	279	287	303	303	278	280	199	212
SNK 413	854	315	315	187	250	234	234	272	272	344	344	279	287	297	303	271	271	193	212
T 85/799	855	307	307	201	203	250	252	272	272	333	354	279	287	303	303	278	278	210	210
Linkage group 6–10 (Risterucci et al. 2000)																			
Accession	RUQ	mTc-CIR1		mTc-CIR6		mTc-CIR7		mTc-CIR8		mTc-CIR24		mTc-CIR25		mTc-CIR26		mTc-CIR37			
AMAZ 15-15	1	127	140	239	239	155	161	287	291	185	201	131	131	294	300	155	161		
APA 4	843	140	140	249	249	161	161	289	289	185	185	163	163	288	288	157	157		
CATIE 1000	844	140	140	249	256	155	161	287	289	185	201	157	157	288	294	157	164		
EET 59	791	140	140	247	251	155	159	287	295	185	201	140	146	294	296	153	164		
IMC 47	849	140	140	231	237	148	159	287	287	185	201	132	145	294	296	146	158		
MAN 15-2	86	127	140	241	251	161	161	287	293	197	197	146	159	296	298	139	146		
P 7	851	127	140	241	249	159	161	287	289	185	185	131	137	294	296	139	146		
PA 120	852	140	140	253	253	155	159	315	315	185	185	129	148	300	303	139	160		
PA 150	853	140	140	243	243	159	159	287	287	185	185	146	146	296	296	144	144		
PLAYA ALTA 2	232	127	140	233	233	157	161	287	287	185	185	151	159	296	303	164	176		
SNK 413	854	127	140	249	251	155	157	287	304	185	185	146	146	296	296	164	176		
T 85/799	855	127	140	233	237	159	161	287	287	185	201	131	133	294	303	151	153		

Cluster analysis

As expected, UPGMA dendrogram of a Nei genetic distance (Nei 1978) for the material analysed largely confirmed the genetic affinities of material listed as closely related (e.g. siblings and half-siblings or parents and offspring) and clearly distinguished potentially duplicated accessions (Supplementary Figure 1 http://www. personal.rdg.ac.uk/~abs00ncc/cacao/Supplementary_ Figure1.htm). A probability threshold of $p = 0.01$ was set to define genotype duplication among trees sharing the same profile using CERVUS software. Possible duplications were then verified by manually inspecting the relevant data.

Mislabelled material

Some trees sharing identical clone names nevertheless produced microsatellite profiles that were dissimilar and so deemed to include mislabelled material. In such circumstances, the assignment of a name to the appropriate tree or trees was usually achieved by reference to passport data, to microsatellite profiles of clones known to be related (e.g. parents or siblings) and when there were many representatives with the same name, to the trees sharing the most common profile. A list of material thereby identified as mislabelled is given in Table 3.

There were several interesting anomalies. For instance, clone IMC 67 is reported to possess resistance to *Phytophthora*, Vascular Streak Dieback, and Cocoa Swollen Shoot Virus (Wadsworth et al. 2003) and has been extensively requested from the quarantine collection. This clone is held in triplicate, with two introductions originating from the International Cocoa Genebank (ICG), Trinidad, in 2000 and a third tree donated by Royal Botanic Gardens (RBG), Kew, UK in 1985. Whilst the two trees from Trinidad produce identical microsatellite profiles, the individual from RBG produced a widely divergent profile. However, reference to the passport information of this accession revealed that the IMC clones were originally deposited in ICG and so the material received from RBG was presumed to be mislabelled. Additional information on allele frequency distribution within related groups of accessions support the designation of IMC 67 (RUQ 13) as being mislabelled (Motamayor pers. com.). Where uncertainty over any clone identity exists, both genotypes are designated mislabelled according to an internationally agreed protocol (Turnbull et al. 2004). This

Table 3. Duplicated accessions with dissimilar microsatellite profiles.

Accession	RUQ	Source	Outcome
ICS 46	967	ICG, T	Comparison with samples from other sources remains ongoing
	938	ICG,T	
IMC 20	985	ICG,T	RUQ 985 matches further ICG, T samples and is considered true to type. RUQ 10 remains unidentified
	10	KEW, UK	
IMC 67	1056	ICG, T	RUQ 1056 and RUQ 1078 are identical. RUQ 13 is distinct and considered an off-type.
	1078	ICG, T	
	13	KEW, UK	
NA 149	801	ICG, T	Comparison with samples from other sources remains ongoing.
	652	ICG, T	
PA 137	1081	ICG, T	Comparison with samples from other sources remains ongoing.
	36	KEW, UK	
SCA 19	1065	ICG, T	Both Genotypes have corresponding plants of the same name in ICG, T. Ambiguity remains over the identification of the true type.
	46	KEW, UK	
IMC 11	1077	ICG, T	Comparison with samples from other sources remains ongoing.
	8	KEW, UK	

Seven accessions relating to 15 plants were identified as ambiguously labelled. Efforts to determine the true identity of the plants is detailed. Accession, ICGD preferred name for clone. RUQ, Unique identification number for plants held in the University of Reading Intermediate Quarantine Facility; Source, ICG,T, International Cocoa Genebank, Trinidad, KEW, UK, Royal Botanic Garden, Kew, UK.

rule has been applied to all clones showing dissimilar profiles for duplicate entries.

Previous inspection of the Reading collection by specialist consultants (Bartley pers. com.) cast doubt on the naming of 32 accessions based on their atypical morphology. In these instances, difficulty rests with the possibility that abnormal growth characteristics could also arise from the local environment in which the trees were grown. For this reason, microsatellite profiles from these plants were compared with those of authentic material taken from the original source collection (Cocoa Research Unit, Trinidad). In this way, 14 clones held in RUQ were identified as off-types but a further five found to be morphological atypical representatives of the named clone (Table 4). There are still thirteen comparisons instances where morphological characterisation suggests mislabelling but where it has not been possible to source suitable reference material.

Discussion

Standard reference material

Two types of standard reference material to aid allele calling have been produced to allow harmonisation of the interpretation of cocoa microsatellite data in the continuing efforts to fingerprint all of the world's *ex situ* cacao germplasm.

Cloned standard alleles

The use of conventional internal size standards for microsatellite allele calling has value in ensuring reproducibility within laboratories but can lead to inconsistencies in the assignment of allele size when applied across different laboratories (Rossouw et al. 2002). It is therefore good practice when attempting to assemble internationally transferable microsatellite fingerprint profiles to use cloned amplicons to calibrate the size of microsatellite alleles. The distribution of the allelic standards generated by this study should therefore allow all cocoa microsatellite data produced using this internationally agreed set of markers to be directly comparable regardless of which platform is used to generate, resolve or analyse genotypes. The validity and robustness of this strategy has been long-established for human forensic applications (Leclair et al. 2004)

Table 4. Clones identified as off-types based on morphology.

Clone name	RUQ	Received	Donor	Match to CRU, T
AMAZ 15/15 [CHA]	RUQ 1	1985	Kew, UK	Yes
CBO 177 [VEN]	RUQ 681	1996	ICG, T	Yes
EET 58 [ECU]	RUQ 383	1993	ICG, T	No
EET 272 [ECU]	RUQ 6	1985	Kew, UK	No
ICS 27	RUQ 143	1988	ICG, T	No
IMC 14	RUQ 9	1985	Kew, UK	No
IMC 20	RUQ 10	1985	Kew, UK	No
IMC 67	RUQ 13	1985	Kew, UK	No
IMC 96	RUQ 16	1985	Kew, UK	No
MO 20	RUQ 233	1989	ICG, T	No
PA 88 [PER]	RUQ 34	1985	Kew, UK	Yes
POUND 4/A [POU]	RUQ 22	1985	Kew, UK	No
SC 1 [?]	RUQ 41	1985	Kew, UK	Yes
SC 3 [?]	RUQ 42	1985	Kew, UK	No
SC 4 [?]	RUQ 43	1985	Kew, UK	No
SPA 4 [COL]	RUQ 341	1993	ICG, T	Yes
SPA 9 [COL]	RUQ 235	1989	ICG, T	No
UF 705	RUQ 57	1985	Kew, UK	No

A comparison of microsatellite profiles for clones identified as displaying atypical morphology against verified accessions obtained from CRU, T. RUQ, Reading unique accession number; CRU, T., Cocoa Research Unit, Trinidad; KEW, UK, Royal Botanic Garden, Kew, UK.

Genotype standards

Larger numbers of cocoa microsatellite markers are now available (Clement et al. 2004; Kuhn et al. 2003; Pugh et al. 2004) for use in fingerprinting, QTL studies and Marker Assisted Breeding (MAB) programmes. Whilst allele standards are the method of choice for fingerprinting applications, it is often impractical to produce sufficient cloned and sequenced allele standards for QTL and MAB purposes. In these cases, it is generally regarded as sufficient to have standard DNA from well-characterised *cacao* clones. Selecting standard DNA from the CFC/ICCO/IPGRI clone list that were given priority for movement through quarantine (Eskes et al. 1998) provides a source of material that is already widely distributed and known to be true to type. The allele sizes of the 15 reference microsatellite loci for these genotypes have been calculated using the allelic standards generated above and are freely available. When used in combination with the allelic standards, this allows local sources of standard trees to be verified prior to use as a standard for additional loci.

Germplasm management

The microsatellite characterisation of The University of Reading Intermediate Cocoa Quarantine Facility has allowed for better management of the material held in this facility. Forty-four accessions were identified as being held in duplicate or triplicate. Where trees of the same accession have identical microsatellite profiles, the provision of microsatellite fingerprints has consequently allowed for duplicates to be removed, thereby reducing the redundancy of the collection. In this way, space is created for the acquisition of fresh material and so the genetic diversity (and therefore usefulness) of the collection is improved without incurring costs for additional infrastructure. Indeed, overall, there were 28 accessions (6.65% of the collection) found to have more than one copy represented in the collection but that produced identical microsatellite profiles. Such rationalisations are of huge importance for clonal collections of trees, where growing space usually limits collection size.

Morphological off-types

The majority of clones putatively identified as morphological 'off types' by specialist consultants were also found to exhibit divergent microsatellite profiles from authentic material from CRU. On the one hand, this finding confirms the value of morphological assessment as a provisional screen for the provisional identification of mislabelling. On the other, the limitation of morphological assessment is clearly illustrated by the fact that trees with matching microsatellite profiles commonly exhibited unusual phenotypes (e.g. AMAZ 15/15) and that several genetically divergent trees were not identified as being morphologically unusual for the assigned name (e.g. NA 149). Nevertheless, in the absence of an international reference database of microsatellite profiles, morphological assessment of material currently represents the only means of ensuring external consistency in the naming of clones.

Distribution of mislabelled clones

The distribution of incorrectly labelled material has profound implications for the breeding of improved cacao based on the selection of breeding material with desirable traits. In spite of careful management, however, mislabelled accessions have nevertheless been distributed from The University of Reading Intermediate Cocoa Quarantine Facility (Table 5). Our study has shown that at least 13 incorrectly labelled trees have been historically distributed from the collection. Whilst this is obviously a matter of some concern, the establishment of definitive fingerprint profiles for all material currently being handled by the facility, provides a useful tool with which to manage the problem in future. On the other hand, it is perhaps encouraging to note that in the vast majority of cases of documented duplication (on the basis of passport data) microsatellite profiles indicated trueness to type, at least internally within the collection. The true problem lies in labelling incongruence between sites. In the longer term, the establishment and comparison of reference microsatellite profiles of all genotypes and from all repositories will address this problem.

A striking example of the problems inherent to the distribution of misidentified clones is given by

Table 5. Distribution of morphologically atypical clones confirmed as false by microsatellite analysis.

Accession	RUQ number	Source	Sites
EET 272 [ECU]	RUQ 6	Kew UK	(8)
EET 58 [ECU]	RUQ 383	ICG T	(4)
ICS 27	RUQ 143	ICG T	(5)
IMC 14	RUQ 9	Kew UK	(7)
IMC 20	RUQ 10	Kew UK	(9)
IMC 67	RUQ 13	Kew UK	(4)
IMC 96	RUQ 16	Kew UK	(3)
MO 20	RUQ 233	ICG T	(10)
POUND 4/A [POU]	RUQ 22	Kew UK	(4)
SC 3 [?]	RUQ 42	Kew UK	(5)
SC 4 [?]	RUQ 43	Kew UK	(7)
SPA 9 [COL]	RUQ 235	ICG T	(5)
UF 705	RUQ 57	Kew UK	(4)

RUQ, Reading University quarantine number; Sites, number of collections to which these accessions have been distributed to since 1985.

SCA 6, probably the most important cocoa clone in breeding for resistance to *Crinipellis perniciosa*: being held in 39 national and international collections; father to 77 clones; mother to 105 clones. (Wadsworth et al. 2003). The original genotype is known to be located at Trinidad, Reading, and CEPEC/CEPLAC, whereas off-types are reported at Miami (Kuhn et al. 2003) and Penn State university (data not shown). Unusual shaped pods are also reported at Papua New Guinea that may indicate a mislabelled clone (Wadsworth et al. 2003). This uncertainty undermines the selection of specific clones based on their desirable characteristics.

Concluding remarks

We believe that this is the first systematic microsatellite fingerprinting of an entire international quarantine collection. Furthermore, the provision of a set of allelic size standards and genotype standards represents an important advance towards the ultimate goal of establishing a global database of microsatellite fingerprints of cocoa germplasm held in *ex situ* collections. Such a facility will eventually improve maintenance of collections and enable quality assurance procedures to be established to ensure the efficient exchange of material between breeding efforts across the world. Furthermore, the approach adopted here for cocoa is equally well suited for clonal collections of other tree crops.

Acknowledgements

This work was funded by Cocoa Research UK Ltd. Additional financial support was given by Masterfoods. We thank The University of Reading Intermediate Cocoa Quarantine Facility for supplying plant material. Reference cocoa genotype samples were kindly supplied by Dr David Butler, Cocoa Research Unit, University of West Indies. We offer our grateful thanks to Dr A.J. Daymond and Professor Paul Hadley for critical reading of the manuscript.

References

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Clément D., Lanaud C., Sabau X., Fouet O., Le Cunff L., Ruiz E., Risterucci A.M., Glaszmann J.C. and Piffanelli P. 2004. Creation of BAC genomic resources for cocoa (*Theobroma cacao* L.) for physical mapping of RGA containing BAC clones. Theor. Appl. Genet. 108: 1627–1634.

Cooper H.D., Spillane C. and Hodgkin T. 2001. Broadening the genetic base of crop plants: an overview. In: Cooper H.D., Spillane C. and Hodgkin T. (eds), Broadening the Genetic Base of Crop Production. CABI Publishing, Wallingford, UK.

Dograr N., Akin-Yalin S. and Akkaya M.S. 2000. Discriminating durum wheat cultivars using highly polymorphic simple sequence repeat DNA markers. Plant Breed. 119: 360–362.

Eskes A.B., Engels J.M.M., and Lass R.A. 1998. The CFC/ICCO/IPGRI project: a new initiative on cocoa germplasm utilisation and conservation. Plantations Recherche Développment 5: 412–422.

Eskes B. 2001. Introductory notes. In: Proceedings of the international workshop on new technologies and cocoa breeding. Kota Kinabalu, Malaysia, 16–17 October 2000, pp. 8–11.

Goldstein D.B. and Schlotterer C. 1999. Microsatellites: Evolution and Applications. Oxford university press, UK.

Higgins D., Thompson J., Gibson T., Thompson J.D., Higgins D.G. and Gibson T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673–4680.

Idury R.M. and Cardon L.R. 1997. A simple method for automated allele binning in microsatellite markers. Genome Res. 7: 1104–1109.

Kolliker R., Jones E.S., Jahufer M.Z.Z. and Forster J.W. 2001. Bulked AFLP analysis for the assessment of genetic diversity in white clover (*Trifolium repens* L.). Euphytica 121: 305–315.

Kuhn D.N., Heath M., Wisser R.J., Meerow A., Brown J.S., Lopes U. and Schnell R.J. 2003. Resistance gene homologues in *Theobroma cacao* as useful genetic markers. Theor. Appl. Genet. 107: 191–202.

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

Leclair B., Fregeau C.J., Bowen K.L. and Fourney R.M. 2004. Precision and accuracy in fluorescent short tandem repeat DNA typing: assessment of benefits imparted by the use of allelic ladders with the AmpF/STR (R) Profiler Plus (TM) kit. Electrophoresis 25: 790–796.

Marshall T.C., Slate J., Kruuk L.E.B. and Pemberton J.M. 1998. Stastical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7: 639–655.

McGregor C.E., Lambert C.A., Greyling M.M., Louw J.H., and Warnich L. 2000. A comparative assessment of DNA fingerprinting techniques (RAPD, ISSR, AFLP and SSR) in tetraploid potato (*Solanum tuberosum* L.) germplasm. Euphytica 113: 135–144.

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

Nei M. 1978. Estimation of average heterozygosity and genetic distance for a small number of indiviuals. Genetics 89: 583–590.

Pejic I., Ajmone-Marsan P., Morgante M., Kozumplick V., Castiglioni P., Taramino G. and Motto M. 1998. Comparative analysis of genetic similarity among maize inbred lines detected by RFLPs, RAPDs, SSRs, and AFLPs. Theor. Appl. Genet. 97: 1248–1255.

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

Risterucci A.M., Grivet L., N'Goran J.A.K., Pieretti I., Flament M.H. and Lanaud C. 2000. A high-density linkage map of *Theobroma cacao* L. Theor. Appl. Genet. 101: 948–955.

Rossouw M., Warren R. and Hoal E.G. 2002. Accurate microsatellite typing and inter-study comparison: Pitfalls and solutions using interferon-gamma (ifng) and natural resistance-associated macrophage protein 2 (nramp2) genes as examples . Clin. Chem. Lab. Med. 40: 926–929.

Turnbull C.J., Butler D.R., Cryer N.C., Zhang D., Lanaud C., Daymond A.J., Ford C.S., Wilkinson M.J. and Hadley P. 2004. Tackling Mislabelling in Cocoa Germplasm Collections. INGENIC Newslett. 9: 8.

Willson K.C. 1999. Pests, diseases and weed control. In: Coffee, Cacao and Tea. Crop Production Science in Horticulture, Series 8. CABI, New York, pp. 142–166.

Wood G.A.R. and Lass R.A. 1985. Cocoa Longman. London, UK.

Wadsworth R.M., Ford C.S., Turnbull C.J. and Hadley P. 2003. International Cocoa Germplasm Database version 5.2. Euronext.life/University of Reading, UK.