UTILIZATION OF MOLECULAR MARKERS TO DETECT THE AUTHENTICITY OF COCOA CLONES

Lea Johnsiul[1]
Malaysian Cocoa Board, Cocoa Biotechnology Research Centre,
Commercial Zone 1, Kota Kinabalu Industrial Park,
88460 Kota Kinabalu, Sabah, Malaysia.

Azwan Awang[2]
Faculty of Sustainable Agriculture,
Universiti Malaysia Sabah, Sandakan Campus,
Locked Bag No. 3, 90509 Sandakan, Sabah, Malaysia

ABSTRACT

Cocoa (Theobroma cacao L.), the main source of cocoa-based products and chocolate, must be conserved live in situ or ex situ as its seeds do not remain viable for more than a couple of weeks once the pod has been harvested. The Malaysian Cocoa Germplasm Collection (MCGC) is one of the ex situ collection of cocoa clones, the fourth largest in the world and regularly importing new clones from other cocoa genebank as well as exporting clones to other countries. The MCGC has been established since 1992 and currently holding more than two thousands imported and local cocoa clones. As in many germplasm collection centres, mislabeling is a critical known problem and correction of the problem is crucial to improve the information reliability and efficient management of germplasm. Using microsatellite DNA markers, the germplasm collection was assessed in its amount of synonymies and homonymies. Comparison of homonymous plants across the collection revealed a significant misidentification rate estimated to be 37.3% and 10.87% synonymous errors. The microsatellite DNA markers amplified a total of 182 alleles with mean allelic richness of 18.2 alleles per locus and average polymorphism information content (PIC) value of 0.9948. The observed heterozygosity (H_{obs}) is 0.6855, indicate a high allelic diversity in this collection.

Keywords: Cocoa, Germplasm, Microsatellite Markers, DNA Fingerprinting, Mislabeled clones

Introduction

Theobroma cacao L. or commonly known as cocoa is an important tropical crop native to South America (Dias *et al.* 2003). It contributes significantly to the economy of many regions of the world with its beans as the source of a multibillion dollar industries for the production of cocoa butter, chocolates as well as cosmetics and antioxidants source.

The germplasm collections of MCGC possesses around 1500 accessions of *T. cacao* from several countries around the world and are located at four separate areas in Malaysia; Lower Perak, Perak, Jengka, Pahang, Madai Baturong and Tawau in Sabah. They are maintained in field collection of living plants due to difficulty working with its seeds. Due to its large and extensive areas occupancy, they are burdensome and difficult to manage appropriately, demanding a lot of works involving various cultivation practices.

As the case with many other plant germplasm collections, MCGC invariably contains duplicate accessions, synonymy problems, homonymous and identification mistakes, which make the transfer of results and recommendations among different breeding programs quite difficult and increase the level of time and resources needed for evaluating genetic potential of the accessions. The use of germplasm for crop improvement is also often interfered by the unavailability of information on the clones' origin, genetic identity and genetic diversity.

As part of the cocoa germplasm collection centres in the world, it is important to know the passport of each cocoa clone in the collection in order to ensure conservation and selection of clones for interplanting and breeding are correct and would produce expected desirable agronomic traits especially. These issues have brought up the awareness on the need to genotype cocoa germplasm all over the world to address the issues and in ensuring efficient utilization of cocoa germplasm held in many germplasm collections around the world by using a standard set of molecular markers (Lanaud *et al*., 1999 and Saunders *et al.* 2004).

The study was implemented in order to identify and access the mislabelling errors in the MCGC. The correct identification of the genetic resources used in the cocoa plant breeding programme is a critical in ensuring the success of the programme in producing desirable planting materials.

At the end of the study, the duplicates and mislabelling identification using the pairwise comparison based on the fifteen international cocoa microsatellite standard (Saunders *et al.* 2004) were implemented. The rate of three type of mislabeling errors were identified and presented.

Materials And Methods

Germplasm Collection

A total of 1000 cocoa leaves were collected from the MCGC in Tawau and Madai Baturong Kunak in Sabah and Hilir Perak and Jengka in Peninsular Malaysia. The germplasm samples included 500 clones, each represented by 2 replicate trees. The samples used for DNA fingerprinting profiles included fully expanded cocoa leaves from various ages collected from individual cocoa clones leaves. Fully-expanded cocoa leaves were harvested from each tree and shipped to the MCB Centre for Cocoa Biotechnology Researches in Kota Kinabalu, Sabah for analysis. Leaf materials were kept at surrounding temperature during shipment (took 1 - 2 days) and kept at -20^{o}C upon arrival until DNA extractions were conducted.

DNA Isolation, Ssr Markers And Polymerase Chain Reaction (Pcr) Amplification

Genomic DNA was extracted from 100mg of leaf sample using DNeasy Plant Mini Kit (Qiagen, Hilden, DE) according to manufacturer's instruction. The study used 15 SSR loci (Table 1) identified as the international standard set for cocoa germplasm characterization as reported by Saunders *et al.* (2004). Primers were synthesized by First Base (Malaysia) and forward primers were 5' labelled using HEX or 6-FAM dyes. PCR reaction mixtures consisted of 1µl containing 10-200ng of genomic DNA template, 1.6µl of a stock solution containing both forward and reverse primers (10µM each), 2µl of 2.5mM dNTP mix, 1x PCR buffer, 1.6µL of 2.5mM $MgCl_2$, 0.5µL of 0.5unit/µl Taq DNA Polymerase and sterile distilled water for a total volume of 20µL.

Table 1: Description for the 15 SSR loci genotyped in the cocoa germplasm collection.

Locus Name	Linkage group	Tm (^{o}C)	Estimated size range (bp)
mTcCIR1	8	56	120-140
mTcCIR6	6	52	220-260
mTcCIR7	7	56	150-170
mTcCIR8	9	52	300-320
mTcCIR11	2	50	280-320
mTcCIR12	4	56	170-260
mTcCIR15	1	56	230-270
mTcCIR18	4	56	330-350
mTcCIR22	1	50	280-300
mTcCIR24	9	60	180-210
mTcCIR26	8	46	280-310
mTcCIR33	4	52	270-350
mTcCIR37	10	50	140-190
mTcCIR40	3	52	270-300
mTcCIR60	2	52	190-220

PCR amplification were carried out in a GeneAmp PCR System 9700 (PE Applied Biosystems) and MyIQ Single Color Real-Time PCR Detection System iCycler (BioRad) with following profile: 94^{o}C for 4 minutes, followed by 35 cycles of 94^{o}C 30 seconds, 46^{o}C to 60^{o}C (depending upon the annealing temperature of the primer pair) for 1 minute; followed by a hold at 72^{o}C for 5 minutes.

The amplified microsatellite loci were separated by capillary electrophoresis and analysed on an ABI Prism 3730 Genetic Analyzer (fragment analysis service outsourced to First Base Sdn Bhd, Malaysia). A dataset of multi-locus genotype was generated as a result.

Identification Of Duplicates And Mislabeling

For the purpose of this study, three types of mislabelling in the collection were identified. The first mislabelling error type was homonymous or intraplant error, where the trees that had same name in the collection but difference multilocus profiles. The second was synonymous mislabelling or duplicate error, whereby the clone had different names but shared the same microsatellite profiles. The third type of error is referred as nonmatching error when the clones had unique microsatellites profiles but do not match with the established genebank database.

Duplicates were identified by using pair-wise comparisons among all the 1000 individuals based on their multilocus SSR profile. The program GenAlEX 6.5 (Peakall and Smouse, 2006, 2012) was used for genotype matching. Accessions with different names that were fully matched at all 15 loci were declared as duplicates or synonymously mislabelled accessions.

For nonmatching errors, microsatellites profiles were compared to an established reference genotype in the International Cocoa Germplasm Database (ICGD) based in University of Reading, United Kingdom.

Results

Identification Of Duplicates And Mislabeling

Fingerprint profiles of all the 1000 trees were generated with all the 15 microsatellite primers. Genotype data was used to identify homonymous and synonymous errors in the collection. Pairwise comparison using GenAlEX 6.0 programme among the samples revealed a significant misidentification rate estimated to be 37.3% homonymous and 10.87% synonymous errors. A total of 62.7% of the collection studied matched with the established reference genotypes in the ICGD.

Figure 1: Pairwise comparison using GenAlEX 6.0 programme among the samples revealed a significant misidentification rate estimated to be 37.3% homonymous and 10.87% synonymous errors

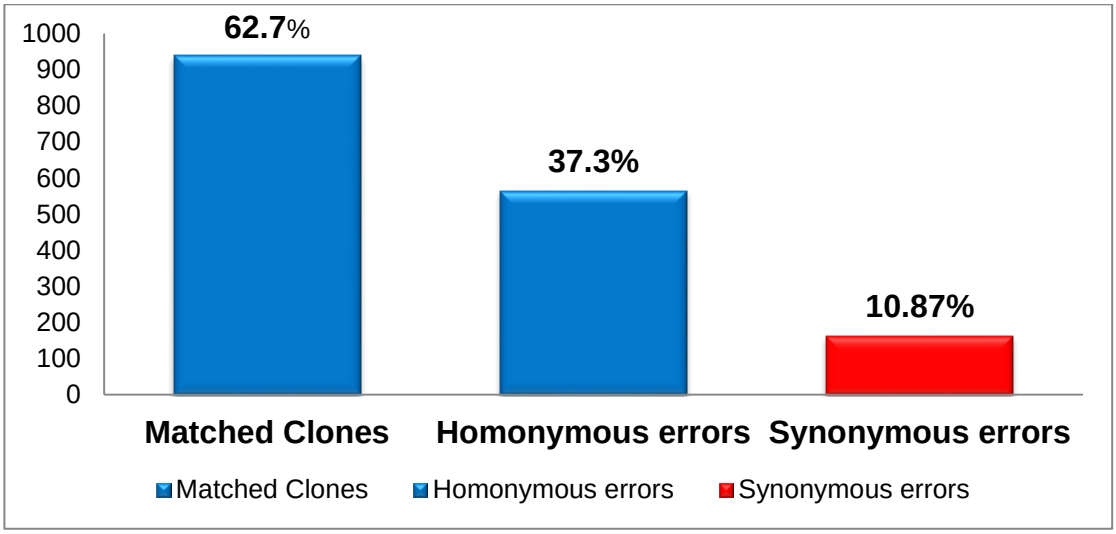

Table 2: Description and summary statistics for 15 loci genotyped in the MCB germplasm collection

Locus name	Linkage group	Size range (bp)	No of alleles	H_{obs}	H_{exp}	PIC Value
mTcCIR26	10	282-307	15	0.7311	0.7344	0.999983541
mTcCIR37	4	133-185	18	0.7133	0.838	0.999999853
mTcCIR6	4	222-247	16	0.674	0.7699	0.999997742
mTcCIR8	1	388-304	17	0.5869	0.691	0.999477681
mTcCIR11	8	288-317	20	0.6974	0.7626	0.999956055
mTcCIR24	2	185-203	17	0.5073	0.4685	0.926325058
mTcCIR40	2	259-284	15	0.5957	0.7684	0.999962893
mTcCIR7	4	155-296	16	0.6613	0.7284	0.997449701
mTcCIR12	6	188-251	23	0.7634	0.769	0.999999905
mTcCIR18	3	331-355	17	0.7622	0.7703	0.999996335
mTcCIR22	9	279-290	23	0.6314	0.6046	0.99952181
mTcCIR1	1	127-144	18	0.5934	0.5487	0.999781305
mTcCIR15	9	232-256	22	0.8418	0.8389	0.99999998
mTcCIR33	8	264-346	15	0.7886	0.8322	0.999999987
mTcCIR60	2	187-223	18	0.7344	0.7944	0.99999998
Mean			18.2	0.6855	0.728	0.994830124

The microsatellite DNA markers amplified a total of 182 alleles with mean allelic richness of 18.2 alleles per locus and average polymorphism information content (PIC) value of 0.9948 (Table 2). It shows that the H_{obs} is slightly lower than the H_{exp} indicating probable inbreeding system of mating (redundancy of alleles) in the sample. The observed heterozygosity (H_{obs}) of 0.6855 also indicate a high allelic diversity in this collection.

Discussions

Mislabeled clones have been a concern for cocoa germplasm management and cocoa breeding programmes both in MCB and international cocoa germplasm collection. Genotypes can be difficult to distinguish morphologically and identification relies mostly on plant labels and field maps. Over the years of cocoa germplasm maintenance, a significant numbers of the clones may have been mislabeled or labeled incorrectly due to uncertain phenotypic assessment. Saunder *et al.* (2004) has demonstrated that a set of fifteen standardized microsatellite primers can characterize all the *T. cacao* germplasm collection and was effective for the assessment of genetic identity of cocoa germplasm.

The comparison of multilocus microsatellite profiles had led to the discovery of clones that were intraplant errors. They were defined as intraplant errors because they did not matched at all the 15 loci but labeled with the same names. Some clones were identified as synonymously mislabeled as they shared the same exact alleles across all the 15 microsatellite loci but were labeled with different names.

Based on this initial finding, we recognize the urgent need to genotype the all the MCGC to ensure that any misidentification/mislabeling can be identified to prevent hindrances in conserving and utilizing the germplasm efficiently and reduce waste of land/space due to genetic redundancy in the germplasm collection.

With the availability of microsatellite fingerprinting profiles for each cocoa accessions, information of each accession's passport data (origin, history, genetic identity etc.) and genetic diversity between individuals and the whole collection can be identified and used in the potential exploitation of germplasm in cocoa improvement.

References

DIAS, L. A. D. S., MARITA, J., CRUZ, C. D., BARROS, E. G. and SALOMÃO, T. M. F. (2003). Genetic Distance and Its Association with Heterosis in Cacao. Brazilian Archives of Biology and Technology 46 (3):507 - 514

LANAUD, C., RISTERUCCI, A. M., PIERETTI, I., FALQUE, M., BOUET, M. and LAGODA P. J. L. (1999) Isolation and Characterization of Microsatellites in *Theobroma cacao* L. Molecular Ecology **8**:2141-2143

SAUNDERS, J. A., MISCHKE, S., LEAMY, E. A., and HEMEIDA, A. A. (2004) Selection of International Molecular Standards for DNA Fingerprinting of *Theobroma cacao.* Theoretical and Applied Genetics **110(1):** 41-47

PEAKALL, R. and SMOUSE P. E. (2006). GENALEX 6: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research. Molecular Ecology Notes. **6**:288-295.

PEAKALL, R. and SMOUSE P. E. (2012). GenAlEx 6.5: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research-An Update. Bioinformatics **28**: 2537-2539.

ACKNOWLEDGEMENTS

The authors thank the Ministry of Science, Technology and Innovation (MOSTI) for supporting this project (ScienceFund grant no. 02-03-13-SF0081) and the Director General of the Malaysian Cocoa Board , Datin Noraini Udin, for permission to publish these results. We also thank Mr Haya Ramba, Ms. Sairan Asim, Ms. Rafiah Abdul Karim, Ms. Nuraziawati Mohd Yazik and Hj Zailaini Md. Jamil, for their excellent experience in cocoa breeding, cocoa germplasm, technical and samples collection assistance.