Malaysian Cocoa Journal
Volume 13(1)/2021

PRELIMINARY EVALUATION ON THE SNP MARKERS FOR THE MALAYSIAN COMMERCIAL COCOA CLONES

Lea Johnsiul[1], Sairan Asim[1], Nuraziawati Mat Yazik[2]
[1]Cocoa Biotechnology Research Centre, Biotechnology Division, Malaysian Cocoa Board, Kota Kinabalu Industrial Park (KKIP), 88460 Kota Kinabalu, Sabah.
[2]CRDC Bagan Datuk, Upstream Technology Division, Malaysian Cocoa Board, Jalan Sg. Dulang, 36307 Perak.
Corresponding author: lea@koko.gov.my

Malaysian Cocoa J. (2021) 13(1): 70-79

ABSTRACT - *Accurate identification of individual genotypes is important for cocoa (Theobroma cacao L.) breeding, planting, propagation and germplasm conservation. Single nucleotide polymorphism (SNP) is the current preferred markers for DNA fingerprinting in most crops including cocoa, nevertheless, a SNP panel with a minimum number of SNPs for optimal clone verification for Malaysian cocoa commercial clones is still unavailable. The development of single nucleotide polymorphism (SNP) markers in cocoa provides an effective way to do high-throughput genotyping system for cocoa commercial clones' verification. A set of 94 SNP loci of varying sizes were assembled and assessed to determine their ability to distinguish among a set of two samples from each of 53 Malaysian cocoa commercial clones. The genotyped data obtained were of high quality with very little missing data and good coverage of all samples. Based on this study, a set of 15 SNP markers were selected to be the core SNP panel for verifying the Malaysian cocoa commercial clones with the exception of two clones which were found to be different at only one SNP loci.*

Keywords: *Theobroma cocoa*, SNP marker, Malaysian cocoa commercial clone, clone verification, planting material.

INTRODUCTION

Cocoa (*Theobroma cacao* L.), the source of cocoa powder and cocoa butter, the important materials used in making chocolate, is a tropical forest species native to the South America that is now cultivated widely in the tropical regions. Although it is only one of the small soft commodities of the world, it has a significant implication on the chocolate and confectionery industries as the primary and irreplaceable ingredient of chocolate and cocoa-based products. In Malaysia, cocoa is the third important crops cultivated after oil palm and rubber (Azhar and Lee, 2004). Malaysia used to be the third largest producer in the world in 1990, but due to the rapid rise of demand, inherent use of poor planting materials, inefficient pest and diseases management and poor technology utilisation has set the decrease on cocoa planting. In 2014, Malaysia is left with the total cocoa planting acreage of 16,102 hectares with the total export amount of RM4.795 billions (Dept. of Statistics Malaysia, 2015) and the fourth-largest producer in Asia Pacific region.

In overcoming the inherent use of poor planting materials, the Malaysian Cocoa Board has developed, evaluated and chosen 53 locally produced cocoa clones as the recommended cocoa planting materials to be used in the farmer's fields.

The Malaysian Commercial Clones are cocoa clones that has been selected based on their good agronomic characteristics recommended for farmers' planting. The 53 cocoa clones in commercial cocoa clones list are divided into 4 classes, Class I, Class II, Class III and Class IV according to their adaptability to a wide range of Malaysia agro-climatic condition, good agronomic traits, tolerant to major pests and diseases and high butter fat content and good flavour.

Cocoa is an outcrossing species (Wood and Lass, 1985) and germplasm is conserved as clonally propagated trees in field genebanks. Cocoa collections have been shown to exhibit some variety of mislabeled individuals in many cocoa germplasm collections (Motilal and Butler, 2003; Sounigo *et al*., 2006, Irish *et al*., 2010, Boza *et al*., 2013). Mislabeling occurrence can be due to frequent introduction and transfers of plant from point-of-collection to early holding sites, subsequent recollection of budwood and repropagation of materials for planting materials distribution to farmers' fields. Human errors during plot division and planting also contributes to this mislabeling problems. Misidentification of plants due to environmental effects on the plants and pods causing the cocoa plants to exhibit slightly different size and characteristics from the original clones.

Molecular markers have been used to characterize cocoa germplasm since in the 1980s (Guiltinan *et al*., 2008, Lindo *et al*., 2017). Mislabeled germplasms were distinguished using dominant markers (Sounigo *et al*., 2006) such as random amplified polymorphic DNA (RAPD) and codominant markers (Lerceteau *et al*., 1997) such as restriction fragment length polymorphisms (RFLP)). The findings on microsatellite markers (Lanaud *et al*., 2000) had great increased the efficiency and possible automation for higher capacity fingerprinting and resulted in a wide application of genotype identification (Motilal *et al*., 2010). A core set of microsatellite markers have also been proposed and used widely as the international molecular standards for DNA fingerprinting of cocoa (Saunders *et al*., 2005, De Wever *et al*., 2019).

Previously, we have generated DNA fingerprinting profiles of all the Malaysian Commercial Cocoa Clones for reference. Recent progress in the development of cocoa genomic resources has led to the use of single nucleotide polymorphisms (SNPs) as markers for cocoa DNA fingerprinting. Compared to SSR markers, the assay of SNPs can be done without requiring separation of DNA by size, and therefore can be automated in an assay-plate format or on microchips. Diallelic nature of SNPs also generating lower error rate in allele calling and genotyping can be multiplexed, allowing quicker completion in genotyping works. Current researches on determining set of SNPs to be used as tools to determine the presence of important agronomic traits such as high yielding, pest and disease resistant and high cocoa butter content are in progress.

The objective of this present work was to develop a core SNP panel to be used in verifying the authenticity of the Malaysian Commercial Cocoa Clones. Even though the SSR DNA fingerprint profile for the Malaysian Commercial cocoa clones had been established previously, the core SNP panel is sought so that verifying and determining the correct clones and present of traits can be done effectively using one method or platform, thus can be done concurrently to save cost, time and resources.

MATERIALS AND METHODS

Healthy leaves were collected from 53 clones from the Malaysian commercial cocoa clones (2 individuals from each clone). Fifty clones were collection from Malaysian Cocoa Board Research Station in Bagan Datuk, Perak, two clones from MCB Research Station in Tawau Sabah and one clone was collected from MCB Research Station at Kota Kinabalu, Sabah. Four to five leaf discs (6 mm diameter) were punched from each cleaned leaf and placed in the collection plates from LGC Genomics, UK. These set of 106 samples represented accessions within the four classes of the Malaysian Commercial Cocoa Clones. The samples were submitted to LGC Genomics for DNA extraction and SNP genotyping.

SNP Genotyping

A total of 93 cocoa SNPs was used to fingerprint the cocoa samples. SNP genotyping was performed using KASP™ assays from LGC Genomics (http://www.lgcgroup.com/kasp). KASP genotyping assays are based on competitive, allele-specific PCR and enable high-throughput genotyping of specific SNPs. Once the KASP reaction was completed, the resulting fluorescence was measured on a BMG PHERAstar plate reader. The raw data were analyzed using LGC's proprietary Kraken™ software and scored on a Cartesian plot, also known as a cluster plot, in order to assign a

genotype to each DNA sample. Raw data was imported and organized in Microsoft Excel for each SNP locus and sample call.

Data analysis
The application SNPViewer was used to review the SNP clustering from each locus. The application Flapjack (Milne *et al*, 2010) was used to compare SNPs between clones, generate similarity matrix and cluster analysis. The core SNP panel for the 53 Malaysian Commercial Cocoa Clones was determined by calculating the average heterozygosity of the SNPs loci using Flapjack programme.

The quality of the data was evaluated by reviewing the SNP clustering from each locus in SNPViewer. All ambiguous data points were removed before further processing and treated as "missing data" which is a standard approach which does not impact on the major results and conclusions given the high quality of the remaining data.

The data were then processed in Flapjack program in several stages after the initial data cleaning. Data were converted to Flapjack format and linked with the SNP locus position in the cocoa genome to create a map file in Flapjack format. The data for each clone pair were then sorted together to assist in the loci comparison between each clone's sample pairs. The SNP data were also used to generate a similarity matrix and carry out a cluster analysis.

RESULTS

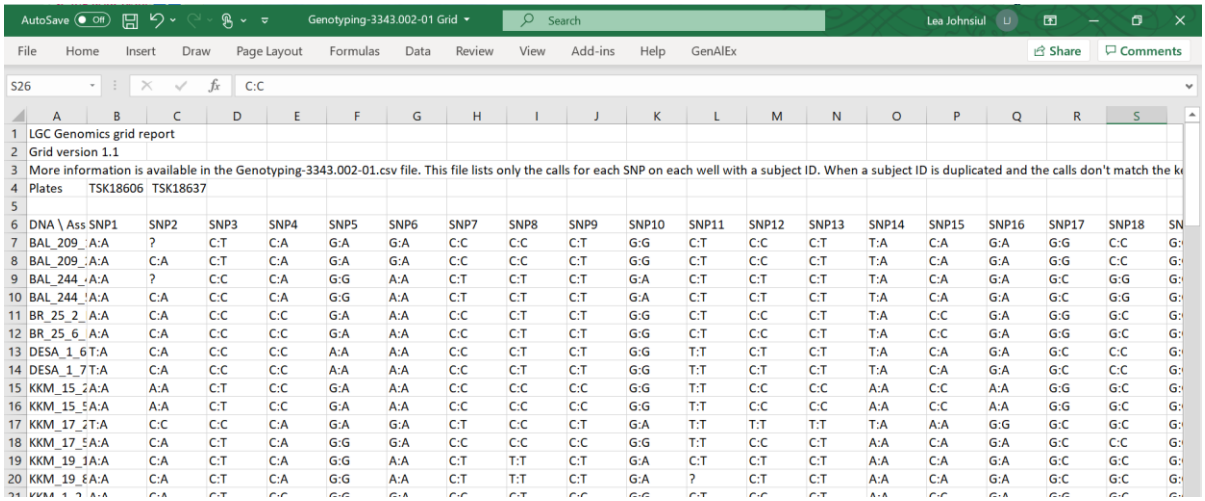
SNP Genotyping
A set of 2 samples from each of 53 Malaysian commercial cocoa clones were genotyped using KASP technology with a reference set of 94 SNP loci. The genotype data is of high quality with little missing data and good coverage of all samples. Figure 1 shows the SNP data in matrix format. Alleles in each locus combination are separated by a colon.

Data Analysis
The SNP data were overall of high quality and ideal to support the major conclusions. In all cases the data was found of high quality with little missing or ambiguous data. Figure 2 shows some of ranges of outcomes with either 3 possible genotypes, heterozygote or single homozygote in the samples viewed using SNPViewer. Clustering of signal from each of the 3 genotypes expected in a heterozygous diploid. Each locus sample combination can be highlighted from sample plate.

Figure 3 showed the map of the SNP data linked with the SNP locus position in the cocoa genome viewed in Flapjack program. The data for each clone pair were compared between each clone's sample pairs. Colours were chosen to maximize contrast for each viewing and comparison.

Figure 4 to 7 are showing the usability and the SNP data in distinguishing between clones and samples. Figure 4 is showing the five clones (KKM 3, KKM 17, KKM 5, MCBC 13 and MCBC 2) that were found to exhibit at least 30 allelic differences between two samples of the same clone name which indicating that the sample pairs were not from the same clone even though the clone name was the same.

Figure 5 is showing the clones KKM 6, KKM 22, KKM 28, KKM 4 and KKM 2 are probably identical or very closely related as all the SNP data in the first three chromosomes the SNPs were identical. The same results were observed between PBC 130 and PBC 140 clones where all the SNPs in the first 3 chromosomes were identical (Figure 6). Similarly, clone RP 1 and DESA 1 also demonstrating the same pattern as the previously mentioned clones which could be assumed that the clones were closely related (Figure 7).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	LGC Genomics grid report																		
2	Grid version 1.1																		
3	More information is available in the Genotyping-3343.002-01.csv file. This file lists only the calls for each SNP on each well with a subject ID. When a subject ID is duplicated and the calls don't match the k																		
4	Plates	TSK18606	TSK18637																
5																			
6	DNA \ Ass	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12	SNP13	SNP14	SNP15	SNP16	SNP17	SNP18	SN
7	BAL_209_	A:A	?	C:T	C:A	G:A	G:A	C:C	C:C	C:T	G:G	C:T	C:C	C:T	T:A	C:A	G:A	G:G	C:C	G:
8	BAL_209_	A:A	C:A	C:T	C:A	G:A	G:A	C:C	C:C	C:T	G:G	C:T	C:C	C:T	T:A	C:A	G:A	G:G	C:C	G:
9	BAL_244_	A:A	?	C:C	C:A	G:G	A:A	C:T	C:T	C:T	G:A	C:T	C:T	C:T	T:A	C:A	G:A	G:C	G:G	G:
10	BAL_244_	A:A	C:A	C:C	C:A	G:G	A:A	C:T	C:T	C:T	G:A	C:T	C:T	C:T	T:A	C:A	G:A	G:C	G:G	G:
11	BR_25_2_	A:A	C:A	C:C	C:C	G:A	A:A	C:C	C:T	C:T	G:G	C:T	C:C	C:T	T:A	C:C	G:A	G:G	G:C	G:
12	BR_25_6_	A:A	C:A	C:C	C:C	G:A	A:A	C:C	C:T	C:T	G:G	C:T	C:C	C:T	T:A	C:C	G:A	G:G	G:C	G:
13	DESA_1_6	T:A	C:A	C:C	C:C	A:A	A:A	C:C	C:T	C:T	G:G	T:T	C:T	C:T	T:A	C:A	G:A	G:C	C:C	G:
14	DESA_1_7	T:A	C:A	C:C	C:C	A:A	A:A	C:C	C:T	C:T	G:G	T:T	C:T	C:T	T:A	C:A	G:A	G:C	C:C	G:
15	KKM_15_2	A:A	A:A	C:T	C:C	G:A	A:A	C:C	C:C	C:C	G:G	T:T	C:C	C:C	A:A	C:C	A:A	G:G	G:C	G:
16	KKM_15_5	A:A	A:A	C:T	C:C	G:A	A:A	C:C	C:C	C:C	G:G	T:T	C:C	C:C	A:A	C:C	A:A	G:G	G:C	G:
17	KKM_17_2	T:A	C:C	C:C	C:A	G:A	G:A	C:T	C:C	C:T	G:A	T:T	T:T	T:T	T:A	A:A	G:G	G:C	G:C	G:
18	KKM_17_5	A:A	C:A	C:T	C:A	G:G	G:A	C:C	C:C	C:C	G:G	T:T	C:C	C:T	A:A	C:A	G:A	G:C	C:C	G:
19	KKM_19_1	A:A	C:A	C:T	C:A	G:G	A:A	C:T	T:T	C:T	G:A	C:T	C:T	C:T	A:A	C:A	G:A	G:C	G:C	G:
20	KKM_19_8	A:A	C:A	C:T	C:A	G:G	A:A	C:T	T:T	C:T	G:A	?	C:T	C:T	A:A	C:A	G:A	G:C	G:C	G:

Figure 1: SNP Data Matrix File extracted and organized in Microsoft Excel according to each clone samples.

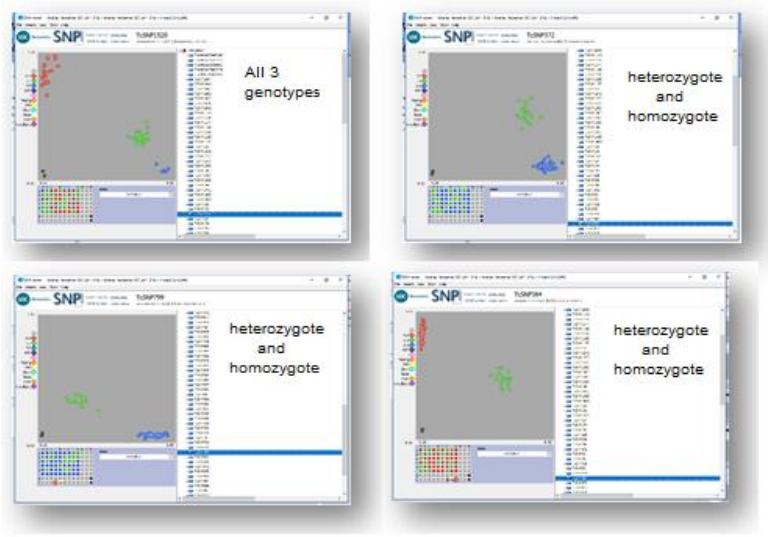


Figure 2: SNPViewer views of samples from 6 different SNP loci showing a range of outcomes with either all 3 possible genotypes or a heterozygote and a single homozygote.

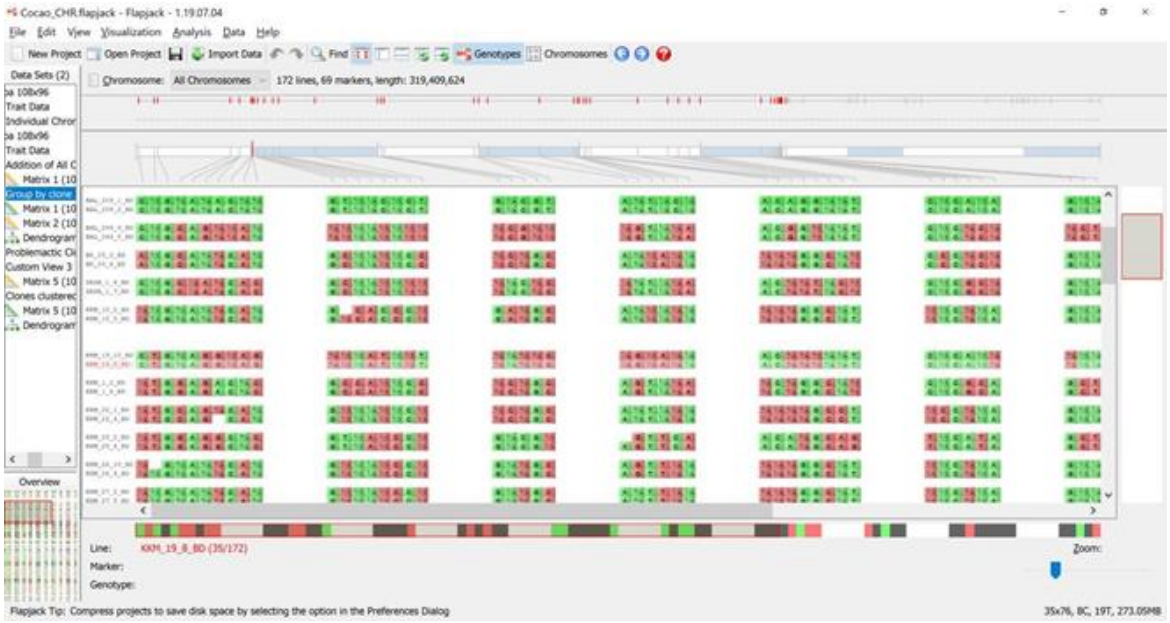


Figure 3: View of the SNP data in Flapjack with lines grouped by clone and SNP loci arranged in chromosome order.

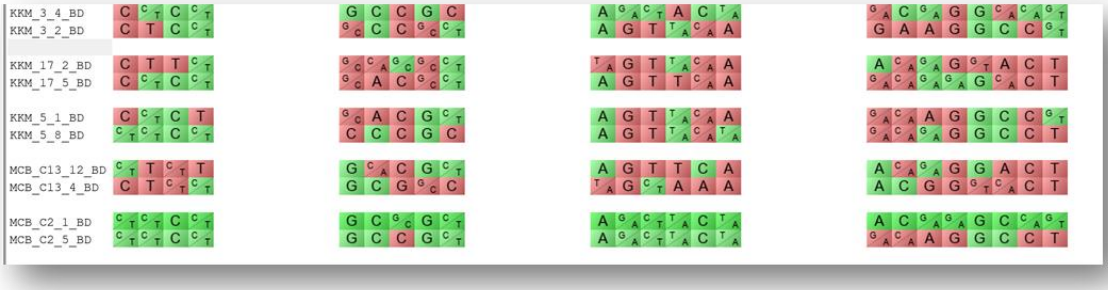


Figure 4: SNP Data from the first 4 chromosomes showing unambiguous examples of mismatched SNPs which shows the samples are not from the same clones.

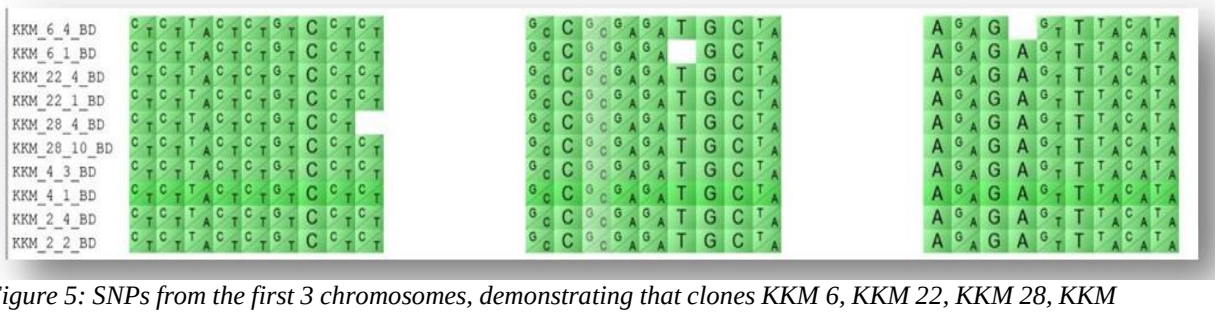


Figure 5: SNPs from the first 3 chromosomes, demonstrating that clones KKM 6, KKM 22, KKM 28, KKM 4 and KKM 2 are probably identical or very closely related.

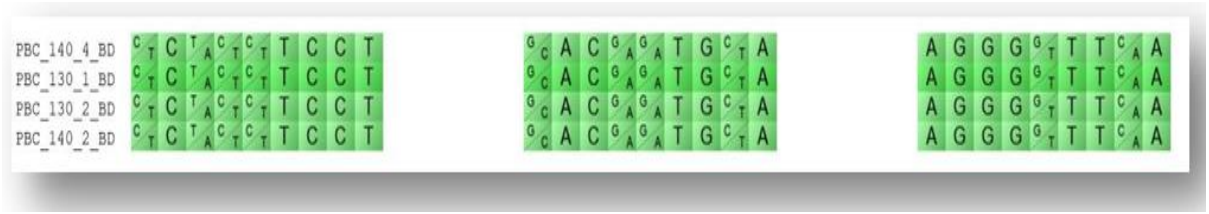


Figure 6: SNPs from the first 3 chromosomes demonstrating that clones PBC 130 and PBC 140 are the same.

Figure 7: SNPs from the first 3 chromosomes demonstrating that clones RP 1 and DESA 1 samples were probably the same.

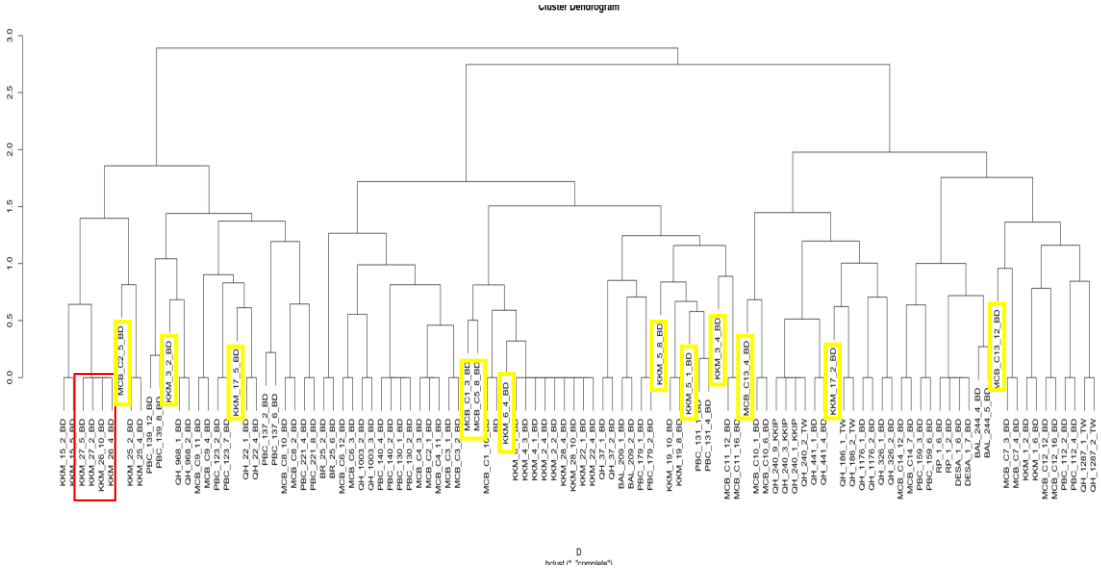


Figure 8: The dendrogram showing the four major clusters in the Malaysian Commercial Cocoa Clones.

In order to further test whether the clones with identical SNP loci at the first three chromosomes were the same clone or closely related, the cluster analysis was employed using the Flapjack program. Figure 8 is showing the cluster dendrogram of all the Malaysian Commercial Cocoa clones samples used in this study. Based on this cluster analysis, the Malaysian Commercial Cocoa clones were divided into 4 major clusters and there were several clones clustered as identical samples. The clones in yellow boxes were clones that deviated between

their sample pairs which indicates that they were putatively mislabeled samples. The samples in red box are KKM 27 and KKM 26 which based on the SNP loci used in this study, there two clones are identical. A comparison on each SNP locus between clone KKM 27 and KKM 26 shows that they were only differed by two SNP loci (Figure 9).

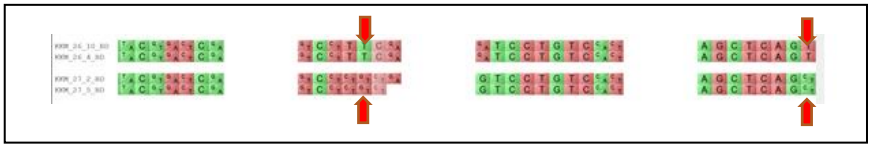


Figure 9: Comparison of each SNP locus between clone KKM 26 and KKM 27 shown that they were only differed by two SNP loci as indicated by the red arrows.

Choice of Subset of Marker Loci as the Core SNP Panel for Verification of the Malaysian Commercial Cocoa Clones

Marker loci chosen were based on their average heterozygosity value in the samples studied whereby the maximum possible value is 0.5. The distribution of average heterozygosity of all the SNP loci used in this study were shown in Figure 9. Based on the average heterozygosity values, a set of thirteen SNP loci which have the highest level of average heterozygosity were chosen as they are the ones that would have a good diagnostic power for all the clones used in this study i.e. the Malaysian Commercial Cocoa Clones. In addition to the thirteen marker loci chosen, two additional loci were added to the list. The two additional loci are the two loci that were identified as the only loci that can differentiate the two very similar clones, KKM 26 and KKM 27 (Figure 8). The final subset of marker loci chosen as the core SNP panel for verification of the Malaysian Commercial Cocoa Clones were as listed in Table 1.

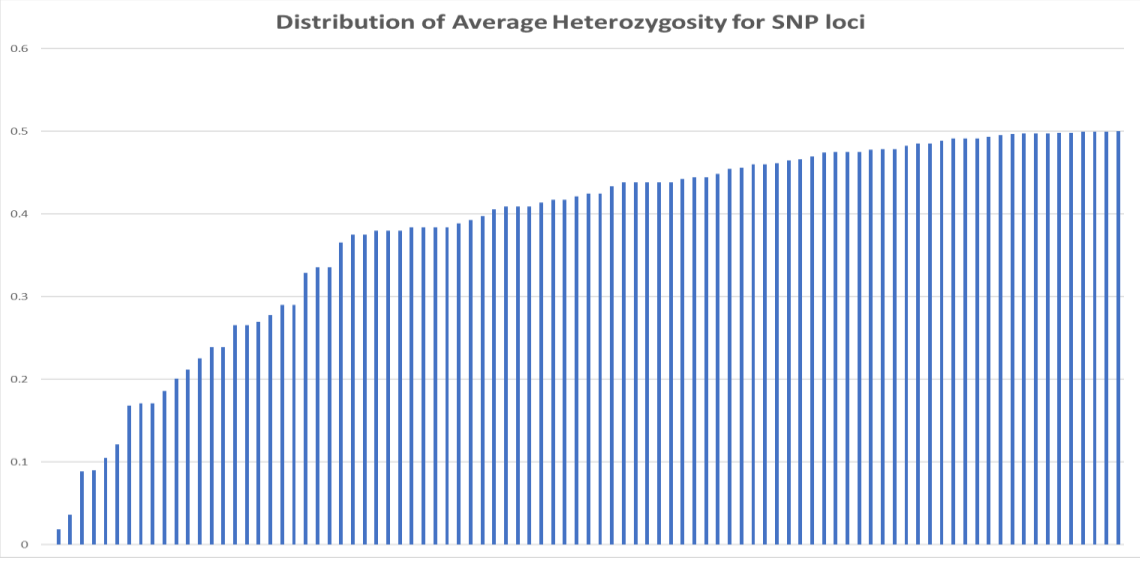


Figure 9: Distribution of Average Heterozygosity for 92 SNP loci used in this study.

Table 1: Set of SNP markers with Average Heterozygosity >0.46 chosen as the Core SNP Panel for Verification of the Malaysian Commercial Cocoa Clones

No	SNP Name	Chromosome	Physical position
1	MCCCSNP561	1	11352078
2	MCCCNP1165	2	1733750
3	MCCCSNP316	2	3946690
4	MCCCSNP122	2	6913492
5	MCCCSNP560	2	7557544
6	MCCCSNP929	3	226748
7	MCCCSNP645	4	27426892
8	MCCCSNP1175	4	29632187
9	MCCCSNP32	4	31580452
10	MCCCSNP602	6	20159544
11	MCCCSNP1484	6	25702944
12	MCCCSNP547	6	26034597
13	MCCCSNP1144	6	26188764
14	MCCCSNP144	10	94961
15	MCCCSNP944	7	401980

DISCUSSION

Virtually all the KASP-based SNP markers that have been utilized show high levels of discrimination between the Malaysian commercial cocoa clone samples and it should be straightforward to develop a subset of markers that will uniquely identify any clone in the Malaysian commercial cocoa clones set. Out of 108 samples, 88 samples of 44 clones showed complete matching between the 2 sample pairs from each clone. Eighteen samples were found to differ between their sample pairs and some even paired with another sample with different clone names (Figure 7). Figure 8 also shown that several clones were very closely related and almost identical such as KKM 26 and KKM 27, PBC 130 and PBC 140 and also the group as indicated in Figure 5.

Five clones as shown in Figure 4 were found to differ by over 30 SNP alleles which indicating that it is highly unlikely that the samples came from the same clone even though the clone identifiers were the same. In principle a single high-quality SNP allele difference between the two samples would indicate they are not from the same clone. In this study for all the five pairs of samples (KKM 3, KKM 17, KKM 5, MCBC 13 and MCBC 2) there were at least 30 SNPs showing allele difference. There are 2 possible explanations for this finding. The first

explanation is the two samples are not from the same original clonal line. In this case there is no simple way to indicate which of the 2 samples, if any, is a true representative of the clone. The second explanation is that samples have been mislabeled at some stage in the sample collection and analysis workflow. If this were the case, we would expect to be able to match even the erroneous samples in pairs. This is not the case. However, one of the 10 samples that did not match its documented partner, matched the pair of samples from another clone i.e. MCBC 2 matched with MCBC 4 (Figure 8).

The second finding became apparent both from a study of the data in Flapjack and from the cluster analysis. Several clone pairs matched perfectly with other clones (Figure 5 -7) indicating that either the same clone has come from several sources or there has been mislabeling at some occasion in the past.

CONCLUSION

The loci that have been used are, in most cases, highly discriminating. The ideal locus for such a purpose is one in which the two alleles are equally frequent. On this basis it will be possible to use chosen core SNP panel of 15 SNP loci to provide discrimination between all the clones that have been sampled here. However, this core SNP panel would not necessarily discriminate between all possible cocoa clones and thus can only be used in verification of samples within the Malaysian commercial cocoa clones.

REFERENCE

Azhar, I. and Lee, M. T. (2004). Perspective for Cocoa Cultivation in Malaysia: Relook at the Economic Indicators. *Malaysian Cocoa Journal* 1: 1-18.

Boza, E. J., Irish, B.M,, Meerow, A.W., Tondo, C. L., Rodriguez, O. A., Ventura-Lopez, M., Gomez, J. A., Moore, J. M., Zhang, D., Motamayor, J. C., Schnell, R, J. (2013) Genetic Diversity, Conservation and Utilization of *Theobroma cacao* L: Genetic Resources in the Dominican Republic. *Genet. Resour Crop Evol* 60:605-619

Department of Statistics Malaysia (2015) http://www.statistics.gov.my/dosm/uploads/files/3_Time%20Series/Malaysia_Time_Series_2015/12Koko.pdf.

De Wever, J., Everaert, H., Coppieters, F., Rottiers, H., Dewettinck, K., Lefever, S., and Messens, K. (2019). The Development of a Novel SNP Genotyping Assay to Differentiate Cacao Clones. *Scientific Reports*. 9. 10.1038/s41598-019-45884-8.

Guiltinan, M.J., Verica, J., Zhang, D., and Figueira, A. (2008) Genomics of *Theobroma cacao*, "the Food of the Gods". In: Moore P.H., Ming R. (eds) *Genomics of Tropical Crop Plants. Plant Genetics and Genomics: Crops and Models*, Vol 1. Springer, New York, NY

Irish, B. M., Goenaga, R., Zhang, D., Schnell, R. J., Brown, J. S., Motamayor, J. C. (2010) Microsatellite Fingerprinting of the USDA-ARS Tropical Agricultural Research Station Cacao (L.) Germplasm Collection. *Crop Sci* 50:656 - 667

Lanaud, C., Risterucci, A., Pieretti, I., Falque, M., Bouet, A., Lagoda, P. (2000). Isolation and Characterization of Microsatellites in *Theobroma cacao* L. *Molecular Ecology*. 8. 2141-3. 10.1046/j.1365-294x.1999.00802.x.

Lerceteau E, Robert T, Pétiard V, Crouzillat D. (1997) Evaluation of The Extent of Genetic Variability Among *Theobroma cacao* L. Accessions Using RAPD and RFLP Markers. *Theoretical and Applied Genetics*. 95. 10-19. 10.1007/s001220050527.

Lindo, A., Robinson, D., Tennant, P., Meinhardt, L., and Zhang, D. (2018). Molecular Characterization of Cacao (*Theobroma cacao*) Germplasm from Jamaica Using Single Nucleotide Polymorphism (SNP) Markers. *Tropical Plant Biology*. 10.1007/s12042-018-9203-5.

Motilal, L., and Butler, D. (2003). Verification of Identities in Global Cacao Germplasm

Collections. *Genetic Resources and Crop Evolution*. 50. 799-807. 10.1023/A:1025950902827.

Motilal, L., Zhang, D., Umaharan, P., Mischke, S., Mooleedhar, V., and Meinhardt, L. (2010). The Relic Criollo Cacao in Belize - Genetic Diversity and Relationship with Trinitario and Other Cacao Clones Held in the International Cocoa Genebank, Trinidad. *Plant Genetic Resources*. 8. 106 - 115. 10.1017/S1479262109990232.

Saunders, J., Mischke, S., Leamy, E., and Hemeida, A. (2005). Selection of International Molecular Standards for DNA Fingerprinting of *Theobroma cacao*. TAG. *Theoretical and Applied Genetics* 110. 41-7. 10.1007/s00122-004-1762-1.

Sounigo, O., Umaharan, R., Christopher, Y., Sankar, A. and Ramdahin, S. (2006) Assessing the Genetic Diversity in the International Cocoa Genebank, Trinidad (ICGT) Using Isozyme Electrophoresis and RAPD. *Genetic Resources and Crop Evolution* 52:1111-1120

Wood, G. A., and Lass, R. A. (1985). Cocoa. Longman Group Ltd.