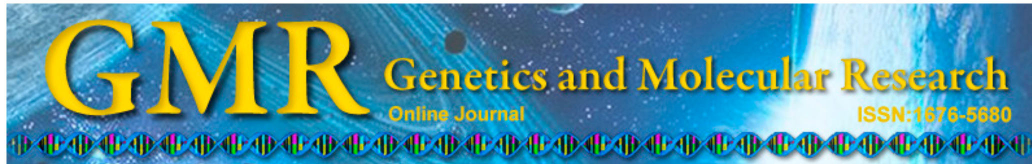

Development of SNP markers and their application for genetic diversity analysis in the oil palm (*Elaeis guineensis*)

P.W. Ong[1,2], I. Maizura[1], N.A.P. Abdullah[2], M.Y. Rafii[2], L.C.L. Ooi[1], E.T.L. Low[1] and R. Singh[1]

[1]Advanced Biotechnology and Breeding Centre, Malaysian Palm Oil Board, No. 6 Persiaran Institusi, Bandar Baru Bangi, Kajang, Selangor, Malaysia
[2]Department of Crop Science, Faculty of Agriculture, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

Corresponding author: I. Maizura
E-mail: maizura@mpob.gov.my

Genet. Mol. Res. 14 (4): 12205-12216 (2015)
Received January 30, 2015
Accepted May 25, 2015
Published October 9, 2015
DOI http://dx.doi.org/10.4238/2015.October.9.9

ABSTRACT. The genetic evaluation of oil palm germplasm collections is required for insight into the variability among populations. The information obtained is also useful for incorporating new genetic materials into current breeding programs. Single nucleotide polymorphisms (SNPs) have been widely used in many plant genetic studies due to the availability of large numbers of genomic sequences and expressed sequence tags. The present study examined 219 oil palms collected from two natural Angolan populations, a few hundred kilometers apart. A total of 62 SNPs were designed from oil palm genomic sequences and converted to cleaved amplified polymorphic sequence (CAPS). Of these, nine were found to be informative across the two populations. The nine informative SNPs revealed mean major allele frequency of 0.693. The average expected and observed heterozygosities were 0.398 and 0.400, respectively. The mean polymorphism information content

 ©FUNPEC-RP www.funpecrp.com.br

was 0.315 (ranging between 0.223 and 0.375). None of the loci deviated from Hardy-Weinberg equilibrium and no rare alleles were detected. In cluster analysis using unweighted pair group method with arithmetic, the 219 oil palms fell into two clusters. This was further supported by the population structure analysis result (K = 2), suggesting that the samples were divided into two main genetic groups. However, the two groups did not coincide with the geographic populations. Analysis of molecular variance indicated that within-population variation contributed 93% of the total genetic variation. This study showed that SNP-based CAPS markers are useful for studying the genetic diversity of oil palm and have potential application for marker-trait association studies.

Key words: Oil palm; SNP-based CAPS markers; Genetic diversity

INTRODUCTION

Oil palm (*Elaeis guineensis*) is an important commodity crop to Malaysia because of its significant contribution to the economy of the country. In 2013, the total value of Malaysian exports from oil palm products (e.g., palm oil, palm kernel oil, palm kernel cake, oleo chemicals, biodiesel, and finished products) reached USD 19.48 billion. The total oil palm planted area of the country increased by 3% in the same year, recorded at 5.23 million hectares (MPOB, 2013).

Various efforts have been initiated to ensure the sustainability of the oil palm industry. Oil palm genetic improvement through conventional breeding is one of them. However, the progress achieved is hampered by the restricted gene pool of the populations used in breeding programs. As a result, the Malaysian Palm Oil Board (MPOB) has carried out systematic collection of oil palm genetic materials at its center of origin in West Africa. The collection effort covered Nigeria, Cameroon, Zaire, Madagascar, Angola, Senegal, Gambia, Sierra Leone, Guinea, and Ghana. The aim of the expedition was to assemble a broader range of genetic materials for oil palm improvement and breeding programs. Field evaluation of these materials has resulted in identification of selected palms with economically important traits which are being incorporated into the current breeding programs; these include high oil yield, low height increment, large kernel, long stalk, low levels of lipase, and high levels of carotene, vitamin E, iodine, and oleic acid (Rajanaidu et al., 2008).

The oil palm genetic materials assembled from Angola are reported to be superior in performance. These materials were gathered in 1991 and planted at MPOB Kluang Research Station, Johore, Malaysia in 1994. They exhibit excellent bunch quality components (Kushairi et al., 2000), long stalk (Noh et al., 2008), and low height increment (Kushairi et al., 2003). In addition, the palms also exhibit high variability in fatty acid composition, iodine value, and carotene content (Noh et al., 2002). The substantial variation observed in the MPOB Angolan germplasm suggests their potential utilization in oil palm improvement programs.

Single nucleotide polymorphisms (SNPs) are defined as variations in single nucleotide positions in genome sequences. SNPs are the most abundant and widely distributed genome markers (Agarwal et al., 2008). The many available SNP detection and genotyping methods include TaqMan assay, molecular beacons, invader assay, matrix-assisted laser desorption

ionization time-of flight mass spectrometry, and DNA chip (Syvanen, 2001). Some of these methods require expensive equipment and reagents, thus limiting the application of SNPs in plant genetics and breeding studies. In contrast, the cleaved amplified polymorphic sequence (CAPS) method, also known as polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP, Konieczny and Ausubel, 1993), is a simple and reliable method for SNP detection that does not require expensive instruments or consumables.

In CAPS, a targeted DNA fragment is amplified by PCR, followed by digestion of the PCR products with an appropriate restriction enzyme. The digested PCR products are analyzed using gel electrophoresis to detect RFLPs. Efforts have been initiated to convert SNPs to CAPS markers in many crops, such as rice (Lee et al., 2009) and soybean (Shu et al., 2011). In this study, we developed a set of SNP-based CAPS markers for the molecular characterization of two natural oil palm populations collected from Angola.

MATERIAL AND METHODS

Plant materials

We sampled 219 palms from two populations of the MPOB Angola germplasm collection, designated AGO01 (61 palms) and AGO08 (158 palms). Two families from population AGO01 and five families from AGO08 were included in the analysis. In the experimental record sheets, the sites where the genetic materials were collected were redefined as populations and the seedlings raised from a single bunch harvested from a palm were defined as a family. These populations were selected due to their high genetic similarity, based on clustering analysis carried out using 57 morphological characters (data not shown). Nevertheless, these populations exhibit differences in height increment, with population AGO08 being relatively shorter than population AGO01 (Kushairi et al., 2003). Spear leaves harvested from the palms were frozen in liquid nitrogen and stored at -80°C until needed.

Development of SNP-based CAPS markers

Two types of SNP markers, random and candidate gene SNPs, were used in this study. For development of the randomly selected SNPs, oil palm genomic sequences generated from Gene Thresher sequencing technology (Low et al., 2014) were aligned by binning using BLAST (95% identity over a minimum of 50 bp) and phrap assembly (minimum match of 15, minimum score of 60). Clustering and assembly of unique identified sequences collapsed them into unique clusters and singletons. The sequence clusters were further analyzed to identify SNPs. CAPS were identified by analyzing the SNP sites for modification or introduction of restriction enzyme recognition sites.

For the development of candidate gene SNPs, several genes related to plant growth and development were selected. Among these were auxin-responsive *GRETCHEN HAGEN 3* family genes (Wang et al., 2008), such as *Dwarf in light 1*, *Dwarf in light 2*, *indole-3-acetic acid-amidosynthetase*, and *Jasmonate resistant 1*. The sequences of these genes were downloaded from NCBI database and used to perform a BLAST search against oil palm genomic sequences. Oil palm sequences with homology to the candidate genes were identified. These sequences were then aligned and assembled into contigs using BioEdit version 7.0.5.3 (Hall,

1999). SNP2CAPS was used to analyze the potential number of SNPs that could be converted to CAPS markers (Thiel et al., 2004). The SNP primers were designed by Primer3 (Rozen and Skaletsky, 2000) for consensus sequences of contigs containing SNPs with one or more restriction enzyme recognition sites.

DNA extraction and screening

Genomic DNA was extracted using the modified CTAB method (Doyle and Doyle, 1990). DNA concentration was measured with a NanoDrop ND-1000 spectrophotometer (NanoDropTechnologies, Wilmington, DE, USA). The quality of each DNA sample was further examined by digestion with *Eco*RI and *Hae*III restriction enzymes (New England Biolabs, Ipswich, MA, USA). Both digested and undigested DNA samples were electrophoresed on a 0.9% SeaKem LE agarose gel (Cambrex, East Rutherford, NJ, USA) at 100 V for 90 min. After staining with ethidium bromide, the gels were visualized under UV light. A screening panel comprising 23 samples, randomly selected from AGO01 and AGO08 populations, was established to screen the SNP-based CAPS markers that were developed.

PCR amplification and CAPS analysis

PCR amplification was carried out in a total volume of 30 µL reaction mixture containing 50 ng template DNA, 1X PCR buffer (20 mM Tris-HCl, pH 8.4, and 50 mM KCl), 1.8 mM $MgCl_2$, 0.1 mM dNTP mix, 0.2 µM of each primer, 1 U *Taq* DNA polymerase (Invitrogen, Carlsbad, CA, USA), and 0.1% Triton X-100. Amplification was performed using GeneAmp PCR System 9700 (Applied Biosystems, Foster City, CA, USA). The PCR cycling program was as follows: 94°C for 4 min; 35 cycles at 94°C for 40 s, 60 or 62°C (depending on primer) for 40 s, 72°C for 40 s; and a final step of 72°C for 10 min. PCR products were analyzed on a 2% SFR agarose gel (Amresco, Solon, OH, USA) in 1X TAE buffer and visualized on an AlphaImager 3400 (Alpha Innotech, San Leandro, CA, USA) after staining with ethidium bromide.

An aliquot of the amplified product was digested with 5 U of the appropriate restriction enzyme (from New England Biolabs, Ipswich, MA, USA), following the manufacturer protocol. The digested PCR products were separated on a 3% SFR agarose gel. The digested PCR products of SNP-based CAPS markers were scored manually as absence (A) or presence (B) of restriction enzyme recognition site. For homozygous AA individuals, only one band was observed as the restriction enzyme recognition site was absent in both alleles. For homozygous BB individuals, two bands were observed due to the presence of the restriction enzyme recognition site in both alleles. In the case of heterozygous AB individuals, three bands were observed, due to the combination of alleles (**Figure S1**).

Data analysis

The genetic diversity parameters of polymorphic loci, such as major allele frequency, expected heterozygosity (H_E), observed heterozygosity (H_O), and polymorphism information content (PIC), were calculated using PowerMarker version 3.25 (Liu and Muse, 2005). Allele frequencies and genotype frequencies were calculated for each locus using

 ©FUNPEC-RP www.funpecrp.com.br

GeneAlEx version 6.41 (Peakall and Smouse, 2006). The same software was applied to carry out a chi-square (χ^2) test for each SNP marker to determine deviation from Hardy-Weinberg equilibrium (HWE). For this, P value less than 0.005 (after Bonferroni's correction) was considered statistically significant. Cluster analysis was done based on the unweighted pair group method with arithmetic (UPGMA). Genetic distances among families and individuals were estimated using the algorithm included in the PowerMarker software and viewed using molecular evolutionary genetics analysis version 5.03 (Tamura et al., 2011). Analysis of molecular variance (AMOVA) was done in Gene Al Exto partition the total variation into among- and within-population variation. Analysis of population structure was performed using STRUCTURE version 2.3.3 (Pritchard et al., 2000). A model-based clustering method was applied for inferring the population structure and for assigning individuals to populations by using multilocus genotype data. The optimum number of clusters was identified after 10 independent runs for each K value ranging from 1 to 10, using the admixture model and correlated allele frequencies. The length of the burn-in period and Markov chain Monte Carlo iterations were set to 100,000 iterations during analysis. The most likely number of populations (K) was determined via the *ad hoc* statistic ΔK (Evanno et al., 2005).

RESULTS AND DISCUSSION

Development of SNP-based CAPS markers

In mining the SNPs, sequences generated from the gene-rich hypomethylated regions of the oil palm genome were used. As a result, 29 SNP-containing sequences were randomly selected for use in this study. In addition, for the candidate gene approach, 97 homologous genes related to plant growth and development were identified from the oil palm genomic sequences. The gene sequences were subjected to cluster analysis to eliminate redundancy. A total of 16 clusters were attained (data not shown). Of these, 11 groups contained sequences from both the *Elaeis* species (*E. guineensis* and *E. oleifera*) as well as sequences from *dura*, *tenera*, and *pisifera* (different fruit forms of *E. guineensis*) palms. The 11 groups were assembled in 17 contigs, and 90 SNPs were identified in eight of the contigs. However, only 41 SNPs had restriction enzyme sites identified for development of the CAPS assays. Of these 41 SNPs, eight were excluded due to either uncommon enzymes or small restriction enzyme-digested PCR product (<20 bp). Hence, only 33 candidate gene SNP-based CAPS markers were further evaluated.

SNP and genetic diversity

The SNP-based CAPS marker is a co-dominant marker and can be used to differentiate homozygous from heterozygous genotypes. Thus, three types of banding profile (AA, AB and BB) were observed and scored. An example of an informative profile is illustrated in Figure 1. Of 62 SNP-based CAPS markers tested, nine (653_*Aci*I, 3064_*Taq*I, 5962_*Alu*I, SNPG00002_*Hpy*188I, SNPG00004_*Aci*I, SNPG00005_*Bcg*I, SNPG00006_*Fat*I, SNPG00014_*Hpy*CH4III, and SNPG00014_*Ssp*I) revealed polymorphism and were classified as informative (Table 1). Each informative SNP-based CAPS markers revealed two alleles, making a total of 18 alleles been scored and analyzed.

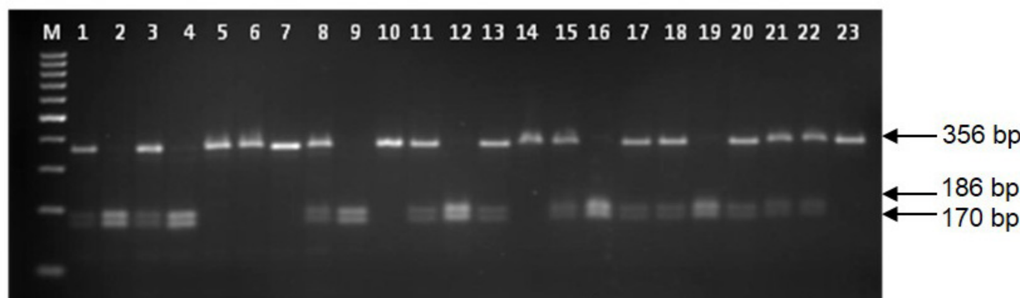

Figure 1. Example of banding pattern from digestion of PCR products generated using SNP-based CAPS marker SNPG00002_ *Hpy*188I.

Table 1. Primer pairs used in the amplification of SNP-based CAPS markers.

Locus	Primer sequence (5'-3')	Ta (°C)	PCR size (bp)	Restriction enzyme	Digested PCR size (bp)	NCBI probe database ID
653_*Aci*I	F: GCTGAGACATGAAATGTGCGTAG	65.7	272	*Aci*I	175, 97	16589809
	R: ATGAACAACAACTCGGAGTCACC	66.2				
3064_*Taq*I	F: CACCCTCTCAGGCATATTGTTG	65.4	260	*Taq*I	188, 72	16589810
	R: AAAGGGAGAAAGACACAGAACCC	65.4				
5962_*Alu*I	F: CTGCGTGACTACGTGAGAGGG	67.1	269	*Alu*I	182, 87	16589811
	R: ACTTGCATTAGCCACCAACAAAC	65.8				
SNPG00002_*Hpy*188I	F: TAAGGGCTGGAGGAAGGATT	60.0	356	*Hpy*188I	186, 170	16589812
	R: CGAAGTGATCTTGGTGCTGA	59.9				
SNPG00004_*Aci*I	F: GGTCATCATTGACGGTCATC	58.7	221	*Aci*I	162, 59	16589813
	R: TCACCAACCTAAACGCAAGA	59.3				
SNPG00005_*Bcg*I	F: CGAAGCAAACACTTCAGACG	59.6	413	*Bcg*I	294, 119	16589814
	R: GCTCCCGTCATAATGCCATA	60.8				
SNPG00006_*Fat*I	F: CAGGAAGCTTGCCACTGATA	59.0	334	*Fat*I	267, 67	16589815
	R: AGCATTTCATTGGCTCGAAG	60.3				
SNPG00014_ *Hpy*CH4III	F: TCGACCTCGTTGATGTGAAG	59.8	501	*Hpy*CH4III	439, 102	16589816
	R: CTTGCCGGTACACGCTATTC	60.6				
SNPG00014_*Ssp*I	F: TCGACCTCGTTGATGTGAAG	59.8	501	*Ssp*I	451, 90	16589817
	R: CTTGCCGGTACACGCTATTC	60.6				

Ta = annealing temperature.

The summary of genetic diversity statistics is presented in Table 2. The mean value of the major allele frequency was 0.693, ranging between 0.505 and 0.849. The average H_E and H_O values were 0.398 and 0.400, respectively. The PIC ranged from 0.223 (653_*Aci*I) to 0.375 (SNPG00002_*Hpy*188I), with a mean of 0.315, which is comparable to the PIC reported in oil palm (0.293, Pootakham et al., 2013). PIC value can be classified into three classes: slightly informative (PIC < 0.25), reasonably informative (0.5 > PIC > 0.25), and highly informative (PIC > 0.5; Hayden et al., 2010). Based on this classification, seven SNPs were classified as reasonably informative, suggesting their potential use for linkage disequilibrium and association mapping studies.

The average H_E recorded in this analysis was greater than the mean value reported for tongkatali (*Eurycoma longifolia*, 0.216; Osman et al., 2003), castor bean (*Ricinus communis*, 0.220; Foster et al., 2010) and maize (*Zea mays*, 0.319; Hamblin et al., 2007), as revealed by SNP markers. However, the heterozygosity value of selected Angola populations attained from the present work was lower than that reported based on simple sequence repeat (SSR)

markers (H_E = 0.537; Zulkifli et al., 2008), but higher than those based on than isozyme (H_E = 0.194; Hayati et al., 2004) and RFLP (H_E = 0.211; Maizura et al., 2006). The higher H_E detected by SSR markers is likely due to the multi-allelic nature of the SSRs. This finding is in agreement with Hamblin et al. (2007), whose study on maize showed higher H_E values for SSR (0.801) as compared to SNPs (0.319). Nevertheless, these results suggest the suitability of the SNP markers developed for analyzing the diversity of oil palm populations.

Table 2. Summary of statistics calculated for genetic diversity based on nine informative SNP-based CAPS markers (N = 219).

Locus	Major allele frequency	Expected heterozygosity	Observed heterozygosity	Polymorphism informative content
653_*Aci*I	0.849	0.256	0.219	0.223
3064_*Taq*I	0.795	0.327	0.365	0.273
5962_*Alu*I	0.721	0.402	0.347	0.321
SNPG00002_*Hpy*188I	0.505	0.500	0.553	0.375
SNPG00004_*Aci*I	0.676	0.438	0.384	0.342
SNPG00005_*Bcg*I	0.571	0.490	0.502	0.370
SNPG00006_*Fat*I	0.822	0.293	0.338	0.250
SNPG000014_*Hpy*CH4III	0.555	0.494	0.516	0.372
SNPG00014_*Ssp*I	0.740	0.385	0.374	0.311
Mean	0.693	0.398	0.400	0.315

For a better determination of allele distribution, allele frequencies were classified using three approaches: a two-class system (Marshall and Brown, 1975); a three-class system (Zhao et al., 2009); and a four-class system (Buchert et al., 1997). The two-class system uses only rare and common allele categories, with frequencies <0.05 and >0.05, respectively. The three-class system groups alleles into rare (frequency <0.05), intermediate (0.05-0.50) and abundant (>0.50). The four-class system classifies alleles as rare (frequency <0.01), low (0.01-0.25), intermediate (0.25-0.75) and high (>0.75). The distribution of allele frequencies according to the four-, three-, and two-class systems are shown in Figure 2. No rare alleles were detected. Most alleles fell within the intermediate and common classes. The genotypic and allelic frequencies of the SNPs are presented in Table 3. None of the loci deviated significantly from HWE after Bonferroni's correction (P < 0.005), which further indicates their appropriateness for population genetic studies.

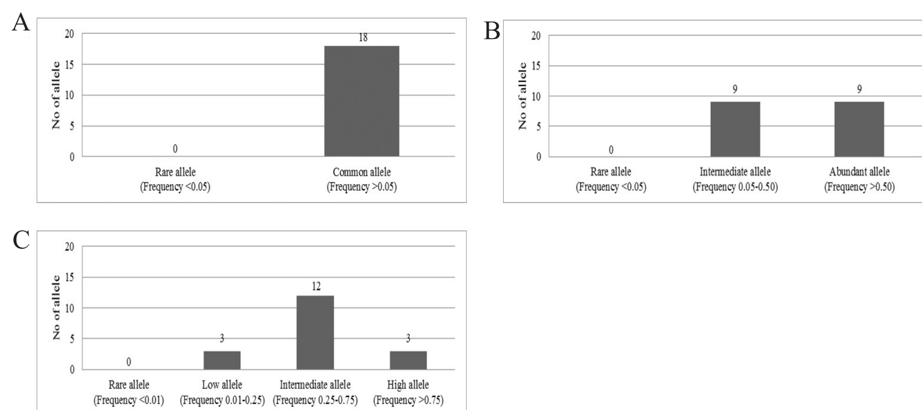

Figure 2. Classification of allele frequencies according to three approaches: **A.** 2-class system. **B.** 3-class system. **C.** 4-class system.

Table 3. Genotype distribution and allele frequencies of each informative locus (N = 219).

Locus	Observed genotypes			Allele frequencies		P value
	AA	AB	BB	A	B	
653_*Aci*I	9	48	162	0.151	0.849	0.059[ns]
	0.041	0.219	0.740			
3064_*Taq*I	5	80	134	0.206	0.795	0.110[ns]
	0.023	0.365	0.612			
5962_*Alu*I	120	76	23	0.722	0.279	0.027[ns]
	0.548	0.347	0.105			
SNPG00002_*Hpy*188I	50	121	48	0.505	0.495	0.146[ns]
	0.228	0.553	0.219			
SNPG00004_*Aci*I	29	84	106	0.324	0.676	0.043[ns]
	0.132	0.384	0.484			
SNPG00005_*Bcg*I	70	110	39	0.571	0.429	0.793[ns]
	0.320	0.502	0.178			
SNPG00006_*Fat*I	143	74	2	0.822	0.178	0.008[ns]
	0.653	0.338	0.009			
SNPG000014_*Hpy*CH4III	65	113	41	0.555	0.445	0.499[ns]
	0.297	0.516	0.187			
SNPG00014_*Ssp*I	121	82	16	0.740	0.260	0.729[ns]
	0.553	0.374	0.073			

[ns]No significant deviation from HWE after Bonferroni's correction (P value > 0.005).

The genetic distance values calculated among families from AGO01 and AGO08 populations within the Angola germplasm exhibited a close genetic relationship (Table 4). Dendrograms revealed that families from different populations and individual palms are mixed within clusters. A UPGMA dendrogram showed that seven families from the two populations were divided into two clusters: two families in cluster I and five families in cluster II (Figure 3). Families AGO0105 and AGO0812 were classified into cluster I, whereas cluster II was made up of the families AGO0811, AGO0801, AGO0808, AGO0104, and AGO0810. Cluster II was further divided into two sub-clusters (II-A and II-B) containing three and two families, respectively. The dendrogram shows that in some cases, families in the same population were grouped together (e.g., AGO0811, AGO0801 and AGO0808 in Sub-cluster II-A), while in most instances the families from the two populations were intermixed (e.g., AGO0105, AGO0812, AGO0104 and AGO0810 in cluster I and Sub-cluster II-B). The dendrogram did not specifically distinguish the families by population.

Table 4. Estimates of mean Nei's genetic distance among seven families from two populations of Angolan oil palm germplasm.

Family	AGO0104	AGO0105	AGO0801	AGO0808	AGO0810	AGO0811	AGO0812
AGO0104	0						
AGO0105	0.088	0					
AGO0801	0.060	0.079	0				
AGO0808	0.048	0.054	0.009	0			
AGO0810	0.050	0.124	0.092	0.068	0		
AGO0811	0.080	0.100	0.018	0.031	0.088	0	
AGO0812	0.115	0.035	0.064	0.046	0.099	0.097	0

The 219 oil palms formed two clusters, I and II (**Figure S2**). Cluster I consisted of eight palms (four from AGO01 and four from AGO08) and cluster II included 211 palms. Sub-

cluster II-A comprised 14 palms (five from AGO01 and nine from AGO08. Sub-cluster II-B comprised 22 palms from AGO01 and 30 from AGO08. Sub-cluster II-C included 31 palms, seven from AGO01 and 24 from AGO08. Sub-cluster II-D was the largest group, which was made up of 23 palms from AGO01 and 91 palms from AGO08. The dendrogram indicated that palms from the same population did not necessarily occur in the same cluster, signifying considerable genetic similarity among the AGO01 and AGO08 populations, despite a geographical separation of a few hundred kilometers. Among the Angolans, palm oil and palm wine are important in traditional culture (Carrere, 2010). These products are occasionally harvested non-commercially by Angolans from oil palm trees grown not far from their residences (e.g., their backyards). As such, it is unsurprising that when moving from one area to another, people bring along oil palm seeds for planting in the new settlement (Koyame, 2005). The civil war in 1975 may have also expedited the movement of people in Angola. In this manner, seeds may have been transported across distances of hundred of kilometers, thus explaining the substantial genetic similarity among palms from the AGO01 and AGO08 populations.

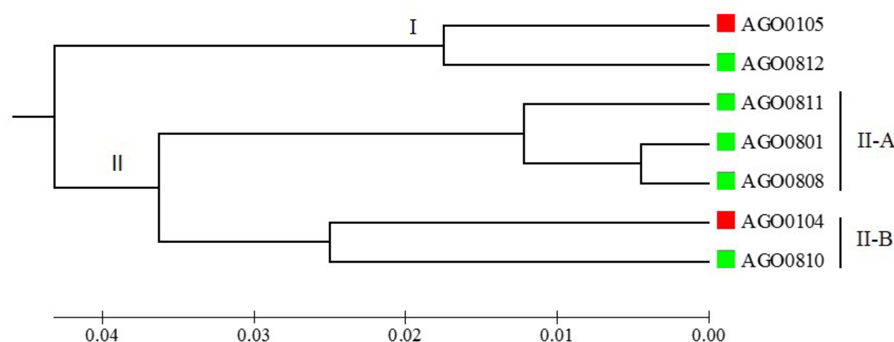

Figure 3. UPGMA dendrogram at the family level, based on Nei's genetic distance (red: AGO01, green: AGO08).

AMOVA

AMOVA revealed that only 7% of the total genetic variation was explained by the variation between populations (AGO01 and AGO08). The remaining 93% was attributed to variation within the populations themselves (Table 5).

Table 5. AMOVA of two populations based on nine SNP-based CAPS markers.

Source	d.f.	Sum of squares	Mean squares	Estimate variation	Percentage	Statistics	Value	P value
Among populations	1	27.135	27.125	0.269	7	PhiPT	0.072	0.001
Within populations	217	754.802	3.478	3.478	93			
Total	218	781.927		3.747	100			

d.f. = degrees of freedom.

Similarly, random amplified polymorphic DNA analysis of the Brazilian oil palm (*E. oleifera*) reveals that 81.70% of the genetic variation is attributable to within-population differences (Moretzsohn et al., 2002). Cochard et al. (2009) employed SSR markers on selected

African oil palm (*E. guineensis*) populations and found that 90.16% of the variation also came from within-population differences. Levels of genetic variation in plants are directly associated with breeding system (Olmstead, 1990). Cross-pollinated and long-lived perennial species, such as oil palm, have high genetic variation within populations. This information can be applied in selecting a sampling strategy for genetic conservation. The AMOVA results obtained here suggest that future germplasm expeditions should collect more individuals within a limited number of populations.

Population structure

The model-based approach described by Pritchard et al. (2000) was applied in the population structure analysis. The *ad hoc* statistic (ΔK) was calculated according to Evanno et al. (2005) to determine K value. The highest value of ΔK was found at K = 2, suggesting that the samples consisted of two main genetic groups (Figure 4). The two populations (AGO01 and AGO08) did not show a strong subpopulation structure, although ΔK suggested a subpopulation number of 2. This supports the AMOVA result of only 7% of the variation occurring between populations. Odong et al. (2011), relying on real and simulated marker data on over 1000 coconut accessions, predicted that populations with low differentiation levels, as observed in this study, will have an optimum of two clusters.

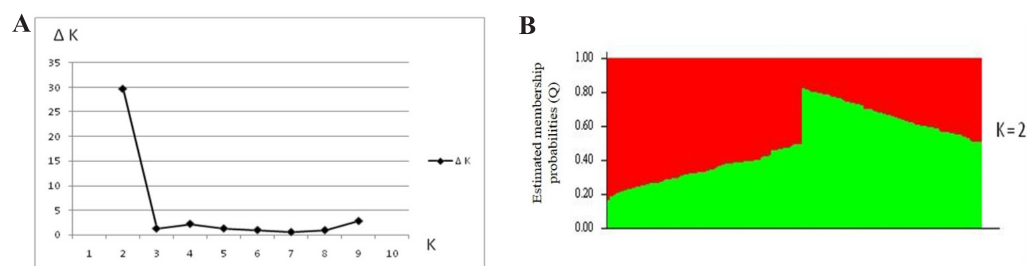

Figure 4. Model-based simulation of population structure among 219 oil palm samples determined using STRUCTURE version 2.3.3. **A.** ΔK was estimated over 10 iteration runs for a given K, ranging from 1 to 10, and used to determine the true K values of the two groups (K = 2). **B.** Genetic composition of individuals based on estimated membership probability (Q). The 219 samples were individually assigned to the red or green group.

Indeed, our analysis revealed that only common alleles with frequencies >0.05 were observed, and no lower-frequency alleles were detected. According to de Oliveira Borba et al. (2010), a rare allele tends to cause bias in covariance between markers and the population structure and increase the chance of a type I error in marker-trait association. The appropriate population structure, with no rare alleles, along with the use of a combination of randomly chosen and candidate gene SNP markers, provides a favorable platform for marker-trait association, even with a relativity small SNP panel. This will be exploited in future studies.

The results obtained in the present study signify the successful application of nine informative SNPs to analyses of the diversity of two natural oil palm populations collected from Angola. The present study also demonstrates the efficacy of a simple, reliable, and low-technology SNP genotyping assay suitable for laboratories having limited resources, especially in agricultural research. This set of SNP-based CAPS markers is useful for oil palm genetic

studies, such as diversity assessments, determinations of population structure, and linkage disequilibrium studies, as well as for marker-trait associations.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The authors would like to thank the Director-General of MPOB for permission to publish this research article. We also thank Mr. Andy Chang Kwang Choong for critically reviewing the manuscript. Research supported in part by grant from the Ministry of Science, Technology and Innovation (MOSTI) of Malaysia (eScience Project #02-03-11SF0010).

Supplementary material

REFERENCES

Agarwal M, Shrivastava N and Padh H (2008). Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Rep.* 27: 617-631.

Buchert GP, Rajora OP, Hood JV and Dancik BP (1997). Effects of harvesting on genetic diversity in old-growth eastern white pine in Ontario, Canada. *Conserv. Biol.* 11: 747-758.

Carrere R (2010). Oil palm in Africa: Past, present and future scenarios. World Rainforest Movement, Montevideo.

Cochard B, Adon B, Rekima S, Billotte N, et al. (2009). Geographic and genetic structure of African oil palm diversity suggest new approaches to breeding. *Tree Genet. Genomes* 5: 493-504.

de Oliveira Borba TC, Brondani RPV, Breseghello F, Coelho ASG, et al. (2010). Association mapping for yield and grain quality traits in rice (*Oryza sativa* L.). *Genet. Mol. Biol.* 133: 515-524.

Doyle JJ and Doyle JL (1990). Isolation of plant DNA from fresh tissue. *Focus* 12: 13-15.

Evanno G, Regnaut S and Goude J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14: 2611-2620.

Foster JT, Allan GJ, Chan AP, Rabinowicz PD, et al. (2010). Single nucleotide polymorphisms for assessing genetic diversity in castor bean (*Ricinus communis*). *BMC Plant Biol.* 10: 13.

Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Window 95/98/NT. *Nucleic Acids Symp. Ser.* 41: 95-98.

Hamblin MT, Warburton ML and Buckler ES (2007). Empirical comparison of simple sequence repeats and single nucleotide polymorphism in assessment of maize diversity and relatedness. *PLoS One* 2: e1367.

Hayati A, Wickneswari R, Maizura I and Rajanaidu N (2004). Genetic diversity of oil palm (*Elaeis guineensis* Jacq.) germplasm collections from Africa: Implications for improvement and conservation of genetic resources. *Theor. Appl. Genet.*108: 1274-1284.

Hayden MK, Tabone TL, Nguyen TM, Coventry S, et al. (2010). An informative set of SNP markers for molecular characterisation of Australian barley germplasm. *Crop Pasture Sci.* 61: 70-83.

Konieczny A and Ausubel FM (1993). A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. *Plant J.* 4: 403-310.

Koyame M (2005). United Nations resolutions and the struggle to curb the illicit trade in conflict diamonds in sub-Saharan Africa. *Afr. J. Legal Stud.* 1: 80-101.

Kushairi A, Rajanaidu N, Jalani BS, Din MA, et al. (2000). Evaluation of Tanzanian and Angolan oil palm genetic collections. Proceedings of the International Symposium on Oil Palm Genetic Resources and Their Utilization. Malaysian Palm Oil Board, Selangor, 157-187.

Kushairi A, Rajanaidu N, Mohd Din A, Isa ZA, et al. (2003). Performance of Angola Genetic Materials. Proceeding of the Seminar on Progress of PS1 and PS2 Planting Materials and Release of Elite Germplasm to the Industry. Malaysian Oil Palm Board, Selangor, 75-90.

Lee GA, Koh HJ, Chung HK, Dixit A, et al. (2009). Development of SNP-based CAPS and dCAPS markers in eight different genes involved in starch biosynthesis in rice. *Mol. Breed.* 24: 93-101.

Liu K and Muse SV (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21: 2128-2129.

Low ETL, Rosli R, Jayanthi N, MohdAmin AH, et al. (2014). Analyses of hypomethylated oil palm gene space. *PLoS One* 9: e86728.

Maizura I, Rajanaidu N, Zakri AH and Cheah SC (2006). Assessment of genetic diversity in oil palm (*Elaeis guineensis* Jacq.) using restriction fragment length polymorphism (RFLP). *Genet. Resour. Crop Evol.* 53: 187-195.

Marshall DR and Brown AHD (1975). Optimum sampling strategy in genetic conservation. In: Crop Genetics Resources for Today and Tomorrow (Frankel OH and Hawkes JG, eds.). Cambridge University Press, London, 53-80.

Moretzsohn MC, Ferreira MA, Amaral ZPS, Coelho PJA, et al. (2002). Genetic diversity of Brazilian oil palm (*Elaeis oleifera* H.B.K) germplasm collected in the Amazon Forest. *Euphytica* 124: 35-45.

MPOB (Malaysian Palm Oil Board) (2013). Malaysian Oil Palm Statistics 2013. MPOB, Selangor.

Noh A, Rajanaidu N, Kushairi A, MohdRafii Y, et al. (2002). Variability in fatty acid composition, iodine value and carotene content in MPOB oil palm germplasm collection from Angola. *J. Oil Palm Res.* 14: 18-23.

Noh A, Kushairi A, Mohd Din A, Maizura I, et al. (2008). Genetic variation for long stalk and high protein kernel in oil palm germplasm. Proceedings of the 3rd Seminar on Performance of MPOB PS1 and PS2 Materials and Elite Germplasm. Malaysian Palm Oil Board, Selangor, 150-167.

Odong TL, Van Heerwaarden J, Jansen J, Van Hintum TJL, et al. (2011). Determination of genetic structure of germplasm collections: Are traditional hierarchical clustering methods appropriate for molecular marker data? *Theor. Appl. Genet.* 123: 195-205.

Olmstead RG (1990). Biological and historical factors influencing genetic diversity in the *Scutellaria angustifolia* complex (Labiatae). *Evolution* 44: 54-70.

Osman A, Jordan B, Lessard PA, Muhammad N, et al. (2003). Genetic diversity of *Eurycoma longifolia* inferred from single nucleotide polymorphisms. *Plant Physiol.* 131: 1294-1301.

Peakall R and Smouse PE (2006). GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* 6: 288-295.

Pootakham W, Uthaipaisanwong P, Sangsrakru D, Yoocha T, et al. (2013). Development and characterization of single-nucleotide polymorphism markers from 454 transcriptome sequences in oil palm (*Elaeis guineensis*). *Plant Breed.* 132: 711-717.

Pritchard JK, Stephens M and Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.

Rajanaidu N, Kushairi A, Mohd Din A, Maizura I, et al. (2008). A review on utilization and performance of MPOB PS series genetic materials. Proceedings of the 3rd Seminar on Performance of MPOB PS1 and PS2 Materials and Elite Germplasm. Malaysian Palm Oil Board, Selangor, 3-42.

Rozen S and Skaletsky H (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* 132: 365-386.

Shu Y, Li Y, Zhu Z, Bai X, et al. (2011). SNPs discovery and CAPS marker conversion in soybean. *Mol. Biol. Rep.* 38: 1841-1846.

Syvanen AC (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.* 2: 930-942.

Tamura K, Peterson D, Perterson N, Stecher G, et al. (2011). MEGA5: Molecular Evolutionary Genetic Analysis using maximum likelihood, evolutionary distance and maximum parsimony methods. *Mol. Biol. Evol.* 28: 2731-2739.

Thiel T, Kota R, Grosse I, Stein N, et al. (2004). SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development. *Nucleic Acids Res.* 32: e5.

Wang H, Tian CE, Duan J and Wu K (2008). Research progress on *GH3*s, one family of primary auxin-responsive genes. *Plant Growth Regul.* 56: 225-232.

Zhao W, Chung JW, Ma KH, Kim TS, et al. (2009). Analysis of genetic diversity and population structure of rice cultivars from Korea, China and Japan using SSR markers. *Genes Genomics* 31: 283-292.

Zulkifli Y, Maizura I and Singh R (2008). Molecular screening of oil palm germplasm collections (*Elaeis guineensis*) using EST-SSR markers. Proceedings of the 3rd Seminar on Performance of MPOB PS1 and PS2 Materials and Elite Germplasm. Malaysian Palm Oil Board, Selangor, 168-185.

 ©FUNPEC-RP www.funpecrp.com.br