VERSITA

Biologia **63**/2: 227—235, 2008
Section Cellular and Molecular Biology
DOI: 10.2478/s11756-008-0041-z

Exploiting an oil palm EST database for the development of gene-derived SSR markers and their exploitation for assessment of genetic diversity

Rajinder Singh[1]*, Noorhariza Mohd Zaki[1], Ngoot-Chin Ting[1], Rozana Rosli[1], Soon-Guan Tan[2], Eng-Ti Leslie Low[1], Maizura Ithnin[1] & Suan-Choo Cheah[1,3]

[1] *Advanced Biotechnology and Breeding Centre, Malaysian Palm Oil Board (MPOB), P.O. Box 10620, 50720 Kuala Lumpur, Malaysia, e-mail: rajinder@mpob.gov.my*
[2] *Biology Department, Faculty of Science, University Putra Malaysia, 43400 UPM Serdang, Malaysia*
[3] *Present address: Asiatic Centre for Genome Technology, Lot L3-I-1, Enterprise 4, Technology Park Malaysia, Bukit Jalil, 57000 Kuala Lumpur, Malaysia*

Abstract: A total of 5,521 expressed sequence tags (ESTs) from oil palm were used to search for type and frequency of simple sequence repeat (SSR) markers. Dimeric repeat motifs appeared to be the most abundant, followed by tri-nucleotide repeats. Redundancy was eliminated in the original EST set, resulting in 145 SSRs in 136 unique ESTs (114 singletons and 22 clusters). Primers were designed for 94 (69.1%) of the unique ESTs (consisting of 14 consensus and 80 singletons). Primers for 10 EST-SSRs were developed and used to evaluate the genetic diversity of 76 accessions of oil palm originating from seven countries in Africa, and the standard Deli *dura* population. The average number of observed and effective alleles was 2.56 and 1.84, respectively. The EST-SSR markers were found to be polymorphic with a mean polymorphic information content value of 0.53. Genetic differentiation (F_{ST}) among the populations studied was 0.2492 indicating high level of genetic divergence. Moreover, the UPGMA (unweighted pair-group method with arithmetic mean) analysis revealed a strong association between genetic distance and geographic location of the populations studied. The germplasm materials exhibited higher diversity than Deli *dura*, indicating their potential usefulness in oil palm improvement programmes. The study also revealed that the populations from Nigeria, Congo and Cameroon showed the highest diversity among the germplasm evaluated in this study. The EST-SSRs further demonstrated their worth as a new source of polymorphic markers for phylogenetic analysis, since a high percentage of the markers showed transferability across species and palm taxa.

Key words: oil palm; EST-SSR; germplasm.

Abbreviations: EST, expressed sequence tags; MPOB, Malaysian Palm Oil Board; PIC, polymorphic information content; RAPD, random amplified polymorphic DNA; RFLP, restriction fragment length polymorphism; SSR, simple sequence repeat.

Introduction

Oil palm (*Elaeis guineensis*) is a diploid monocotyledon belonging to the family Palmae. Its diploid genome comprises 16 pairs of homologous chromosomes (Maria et al. 1995) and its nuclear DNA content has been estimated to be 2C = 3.76 pg by flow cytometry (Rival et al. 1997) with a haploid genome size of 1700 Mbp. Oil palm is currently a major economic crop in Southeast Asia (particularly Malaysia and Indonesia).

Oil palm originates from Africa, where the centre of origin is postulated to be in the tropical rain forest region of West and Central Africa (Zeven 1967). In 1848, some seeds from West Africa were planted in the Buitenzorg Botanical Gardens (now, Bogor), Indonesia (Purseglove 1972). Seeds from the palms were in turn planted in Deli, Sumatra, which evolved into the Deli *dura* population. Deli *dura* is currently the basic breeding stock for commercial planting materials used all over Malaysia and Indonesia and in other parts of the world as well. The limitations of this gene pool resulted in a prospection being made to Africa for more genetic materials. Collections were made in most of the countries where oil palm occurs naturally – Nigeria, Cameroon, Congo, Tanzania, Madagascar, Angola, Senegal, Gambia, Sierra Leone, Guinea Conakry and Ghana (Rajanaidu & Jalani 1994). These germplasm were planted in Malaysia by the Malaysian Palm Oil

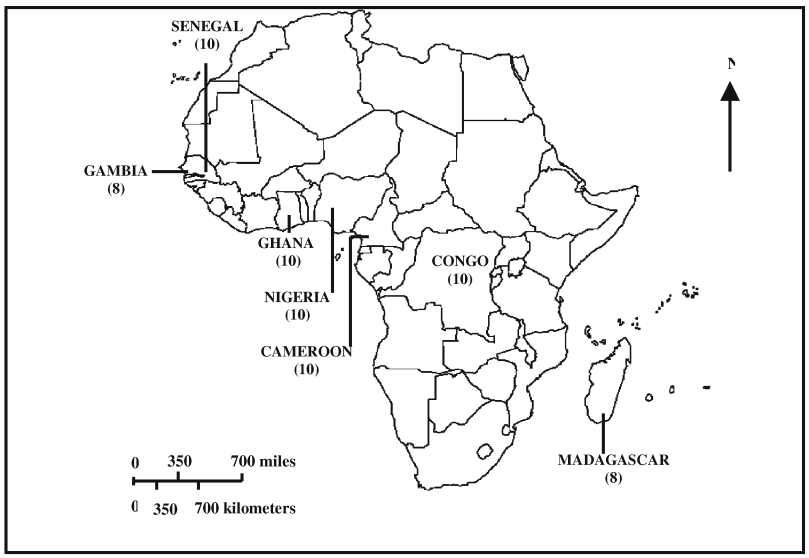

Fig. 1. Map of Africa showing the origin of the populations. The number of palms used in the analysis is indicated by the number in parentheses.

Board (MPOB) and form the largest collection of oil palm breeding materials in the world.

Using molecular marker to screen the germplasm has been limited to only a few studies using random amplified polymorphic DNA (RAPD; Shah et al. 1994), isozymes (Hayati et al. 2004), restriction fragment length polymorphism (RFLP; Maizura et al. 2006) and amplified fragment length polymorphism (Kularatne 2000). Recently Bakoume (2006) used the genomic simple sequence repeat (SSR) markers developed for oil palm by Billotte et al. (2001) to extensively screen the Cameroon collection. These studies have demonstrated the usefulness of molecular markers to assess the genetic variability for cultivar protection and establishment of core collections for conservation of the oil palm gene pool.

Among the molecular marker systems available for germplasm screening, SSRs appear to be the most promising for understanding the population genetic structure and gene flow. This is due to their co-dominant inheritance, multiallelic nature, relative abundance, extensive genome coverage and simple detection using the PCR. SSR primers are also easily transferable between laboratories. However, constructing genomic libraries enriched for SSRs is technically difficult. The establishment of expressed sequence tag (EST) projects for gene discovery in several plants has aided in the discovery of SSRs (usually referred to as EST-SSR or genic SSR). The scanning of the sequence data (EST, genes) available in GenBank (Benson et al. 2007) using specific computer programs has facilitated the discovery of SSR, making the process easier and cheaper (Varshney et al. 2005). An added advantage is that EST-SSR occurs in the expressed region of the genome, which is usually more conserved thus improving the transferability of the markers. The greater sequence conservation in transcribed regions has also resulted in EST-SSR being less polymorphic than genomic SSRs (Chabane et al. 2005). Nevertheless, EST-SSRs have proven to be a valuable source of polymorphic markers for crops such as *Picea* species (Rungis et al. 2004), barley (Chabane et al. 2005) and Triticeae species (Zhang et al. 2006).

In this study we present the development of EST-SSR markers for oil palm to study the genetic diversity of the *Elaeis guineensis* germplasm collections. Transferability of the EST-SSR to another species of oil palm (*Elaeis oleifera*) and across palm taxa was also investigated.

Material and methods

Plant materials

Figure 1 shows the countries in Africa from where the germplasms studied originated. Eight to 10 palms were randomly chosen from the collection from each country of origin. Ten palms from an advanced oil palm breeding population – the Deli *dura* – was included as a reference. A total of 76 *E. guineensis* palms were typed. The across-species transferability of the EST-SSR was tested on two *E. oleifera* palms prospected from Colombia (South America).

Frond-1 was harvested from each palm, and the leaflets were packed and frozen in liquid nitrogen. The samples were then stored in a freezer at −80 ℃ until use.

Across palm-taxa transferability of the EST-SSR was tested on two coconut (*Cocos nucifera*) varieties (the green and yellow dwarfs) and one *Jessenia bataua* palm. The leaflets from these other palms were similarly processed and stored until use.

DNA extraction

Genomic DNA was extracted from all the samples as described by Doyle & Doyle (1990).

Oil palm EST database and mining of SSR

A total of 5,521 EST sequences were screened for microsatellites containing mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats. The ESTs were developed from cDNA libraries representing different tissues (inflorescence and

young etiolated seedlings) as well as tissues from three different stages of oil palm tissue culture (embryogenic callus, non-embryogenic callus and embryoids). All the ESTs were derived from *E. guineensis*. Raw chromatogram reads were base-called using PHRED (Ewing & Green 1998; Ewing et al. 1998). Vector sequences were screened using Cross_Match (http://www.phrap.org/). Edited sequences of fewer than 100 bases were eliminated from further analysis. Poly A and poly T tails were removed using the program TRIMEST downloaded from EMBOSS (Rice et al. 2000). The minimum length parameter for a repeat of A and T was set at five.

Identification and localization of the EST-SSR markers were performed using the software MISA, a Perl5 script as described by Thiel et al. (2003). Flanking primers were designed using the program Primer3 (Rozen and Skaletsky 2000).

SSR analysis

One primer of each primer pair was 5'-end labelled at 37°C for 30 min using T4 polynucleotide kinase (INVITROGEN). The labelling reaction contained 50 pmoles primer, 3 μL γ-^{33}P dATP (GE Healthcare Biosciences, UK, 3000Ci/mmol) and 1 U T4 polynucleotide kinase in a total volume of 25 μL.

Subsequently the PCR reaction was carried out in a 25 μL reaction volume containing 1 U *Taq* polymerase (INVITROGEN), 50 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 μM of each primer, 0.2 mM dNTPs (INVITROGEN) and 50 ng template DNA. PCR was performed in a Perkin Elmer 9600 thermocycler essentially as described by Billotte et al. (2001). The PCR reaction was stopped by the addition of 25 μL of formamide buffer (0.3% bromophenol blue; 0.3% xylene cyanol; 10 mM EDTA, pH 8.0; 97.5% deionized formamide). Each PCR reaction was subjected to electrophoresis on a 6% denaturing acrylamide gel containing 7 M urea, using 0.5×TBE buffer (1 × TBE: 90 mM Tris-borate, 2 mM EDTA, pH 8.0) at constant power of 40 W for 3 h. The gels were then dried and exposed to X-ray film (Kodak) for 3–4 days at −80°C. Sizing of each allele was done using an amplified fragment length polymorphism molecular weight ladder (INVITROGEN).

Data analysis

Only fragments that could be clearly scored were used in the data analysis. Polymorphism information content (PIC) was calculated by applying the formula given by Anderson et al. (1993):

$$\mathrm{PIC} = 1 - \sum_{i=1}^{k} P_i^2$$

where k is the total number of alleles detected for an EST-SSR marker and P_i is the frequency of i^{th} allele in a set of data analyzed.

The EST-SSR data were analysed using the computer program POPGENE version 1.32 (Yeh and Boyle 1999). The genetic variability measures calculated included allelic frequencies, mean number of alleles per locus, effective number of alleles per locus, percentage of polymorphic loci, expected and observed heterozygosity in the collections used, genetic differentiation (F_{ST}) and fixation indices (F_{IS}). The genetic distance between the populations was computed according to Nei (1978). These values were used to generate a dendrogram using the unweighted pair-group method with arithmetic mean (UPGMA) cluster analysis as described by Sneath & Sokal (1973).

Results and discussion

PalmGenes – an oil palm EST database

A total of 6,787 ESTs were sequenced resulting in 5,521 high-quality trimmed sequences, which were deposited with the PalmGenes (accessible at http://palmoilis.mpob.gov.my/palmgenes.html). PalmGenes is the first oil palm DNA database available on the internet containing oil palm ESTs. The MPOB houses the database which, apart from the gene sequences, also contains information on their similarities to known genes and putative functional characterization. The sequences have a length of between 300–500 bp each and PHRED quality value of at least 20.

Frequency and distribution of EST-SSR

The 5,521 EST sequences searched for SSR represented approximately 2.2 Mb of the oil palm genome. The search found 230 SSRs with mono-, di-, tri- and tetra-nucleotide repeats in 221 ESTs. Penta- and hexa-nucleotide repeats were not identified in the present EST collection. Seven of the ESTs contained two SSR, while one EST contained three SSR. The criteria described by Thiel et al. (2003) were used to identify the SSRs that are monomers with at least 10 repeats, dimers with 7 repeats or more, trimers with at least 5 repeats and tetramers with over 4 repeats. This corresponded to approximately one SSR containing EST for every 25 ESTs and an average distance between SSRs of approximately 9.6 kb. Comparatively, the frequency of SSRs in sets of redundant ESTs has been estimated to be 3.4 kb for rice, 6.3 kb (barley), 11.1 kb (tomato), 13.8 kb (*Arabidopsis*) and 20.0 kb (cotton) (Cardle et al. 2000; Thiel et al. 2003). As such, it would appear that the frequency of SSRs in the expressed genes of oil palm is relatively high in comparison with the other species.

Assembly of the SSR containing EST sequences into unique genes was performed to reduce the redundancy of the sequences and identify those coding for the same protein. The assembly was performed using StackPACK 2.2 (Miller et al. 1999), producing 114 singletons and 22 clusters for a total of 136 SSR containing unique genes. Identifying the non-redundant SSR containing ESTs reduced overestimation of the specific SSR types and, more importantly, it also reduced the chance of developing a redundant set of SSR markers.

The occurrence of individual SSR motifs among the non-redundant set of 145 SSRs in 136 EST sequences is summarized in Table 1. The largest group of repeats were di-nucleotides (51 or 35.2%), followed by tri-nucleotides (47 or 32.4%), mono-nucleotides (44 or 30.3%) and tetra-nucleotides (3 or 2.1%). Among the dimeric repeats, AG/CT (90%) was by far the most common whereas AT (10%) was present only in a low abundance. The low abundance of AT SSRs in EST collections has also been reported for barley (Thiel et al. 2003), *Arabidopsis* (Cardle et al. 2000) and rice (Temnykh et al. 2000). With the trimeric SSR, AAG/CTT (32%) was the most common motif, as is also the case in

Table 1. Non-redundant SSRs in a set of 5,521 oil palm ESTs.

SSR Motif	Number of repeats												Total
	5	6	7	8	9	10	11	12	13	14	15	>15	
A/T	–	–	–	–	–	18	5	4	2		5	8	42
C/G	–	–	–	–	–	2						0	2
AG/CT	–	–	10	2	2	6	1	3	1		1	20	46
AT/AT	–	–	2	2	1							0	5
AAC/GTT	1											0	1
AAG/CTT	8	2	3		2							0	15
AAT/ATT	1											0	1
ACC/GGT	1	2										0	3
ACG/CTG	5		1									0	6
ACT/ATG	3	1		1								0	5
AGC/CGT	5		1	1								0	7
AGG/CCT		1										0	1
AGT/ATC	1	2										0	3
CCG/CGG	1	3	1									0	5
AAAG/CTTT	1											0	1
ACAT/ATGT	1	1										0	2
N													44
NN													51
NNN													47
NNNN													3
												Total	145

Arabidopsis (Cardle et al. 2000). With respect to mononucleotide repeats, A/T (95%) was the most common. Three ESTs containing two different tetra-nucleotide repeat motifs (AAAG/CTTT and ACAT/ATGT) were also identified in this study.

The high number of di-nucleotide repeats identified in the oil palm EST collection is consistent with what has been reported for coffee (Aggarwal et al. 2007). However, in cereals (Varshney et al. 2002) and barley in particular (Thiel et al. 2003), tri-nucleotide repeats are more common than di-nucleotide ones. The difference could be due to the different number of EST sequences screened and analysed for SSRs and also to the different sources of the DNA sequences.

SSR-marker development

The 136 non-redundant EST-SSRs were used for primer design. For the generation of PCR primers, multiple SSRs separated by less than 100 bp in an EST were defined as 'compound' and treated as a single marker loci. Primers were successfully designed for 94 (69.1%) of the ESTs (14 consensus and 80 singletons). Primer design failed for the other sequences because they had either too little flanking sequences for primer design or the sequences did not match the minimum criteria required by the primer design software.

Twelve of the primer pairs representing the four-repeat motifs were synthesized and tested for their ability to amplify oil palm DNA. Ten primer pairs (83%) could amplify fragments during PCR with oil palm genomic DNA (data not shown), and are listed in Table 2. The amplicon size for five of the pairs deviated from expectation. Such deviation is frequently encountered in SSRs derived from ESTs (Varshney et al. 2005). Three primers (sEg00066, sEg00067 and sEg00077) produced amplification products smaller than expected, suggesting that deletions may have occurred in their genomic sequences. Two primers (sEg00090 and sEg00140) produced amplification products larger than expected, suggesting the simultaneous amplification of introns.

Screening of oil palm germplasm material

The EST-SSR primers listed in Table 2 were used to screen 76 palms representing germplasm collected from seven countries in Africa (Fig. 1). The reference population included in this analysis, the Deli *dura*, was used for comparison as the first oil palm material introduced to Malaysia. Furthermore, this family of palms still forms the basis for most of the mother palms in producing commercial planting materials in Malaysia (Maizura et al. 2006). The mean PIC value for the EST-SSR markers was 0.53 (Table 2). The value reported here is generally in agreement with other studies on EST-SSR (Varshney et al. 2007; Wang et al. 2007) and is also higher than the values observed for RAPD analysis of another perennial palm, coconut (ranging from 0.031 to 0.392) (Manimekalai & Nagarajan 2006). This suggests that EST-SSRs are a good source of markers for evaluating the oil palm germplasm.

A total of 48 alleles were detected at 10 loci across the 7 germplasm collections and one Deli *dura* population surveyed (Table 2). The mean number of alleles per locus (A_o) ranged from 2.20 to 3.20 (mean = 2.56) (Table 3). The number of alleles observed was lower than that reported for *E. guineensis* using genomic SSR (average 5.25; Billotte et al. 2001). Similar results were also observed for wild and cultivated barley (Chabane et al. 2005). The effective number of alleles ranged from 1.50 to 2.23 (mean = 1.84). The percentage of polymorphic loci (P) in the natural populations stud-

Table 2. SSR markers derived from oil palm ESTs.

No.	ID	Primer pair sequence (5'-3')	GenBank Acc. No.	Ta (°C)	Repeat type	Amplicon size (bp) (expected)	Amplicon size (bp) (observed)	Number of alleles	PIC[a]	Putative function[b]
1	sEg00032	F: CTGTTGAGCTGGAGAGACCC R: CCAACCAGGATCAGTTTGGT	ES324078	55	$(CTTT)_5$	269	260–270	2	0.50	unnamed protein product (CAO23368.1)[c]
2	sEg00066	F: TTGCTCCAACTGACTGATGC R: ACATTCCAGATCCCAGCAAG	ES324079	52	$(AT)_8$	245	192–215	10	0.84	No hit
3	sEg00067	F: GTCAGCCCGTAGAAGATTGC R: CTTTCGGATAGCCAAAACGA	ES324080	52	$(TGTA)_6$	254	150–170	4	0.59	No hit
4	sEg00077	F: GGACCTTACAAGCCACCTCA R: TGTGCTAGCAAAGCCAGAAA	ES324081	52	$(TA)_8$	261	160–180	5	0.73	No hit
5	sEg00080	F: AAGAACTATGACCTCACCAAAA R: AACTCTATGCTATTGCTACACGA	ES324082	52	$(TCA)_6$	153	145–155	2	0.19	No hit
6	sEg00090	F: TATGCGGGTGATCAAGTGAA R: CCACCATGGTTCTCAGGAAA	ES324083	52	$(AT)_9$	149	200–210	8	0.74	L-ascorbate peroxidase (NP_194958.2)
7	sEg00125	F: TACCCTTTTCCCTCCCTCCATA R: CATCATCTCCGTTGCCAGTATT	ES324084	52	$(GCG)_6$	152	150–170	4	0.32	unnamed protein product (CAO48553.1)
8	sEg00126	F: CCGTCTCAAAAGCCCTAAAC R: TTGTTGTCCCACTCCCTCTT	ES324085	52	(CGC)7	216	212–215	2	0.44	Unknown protein (AAF02880.1)
9	sEg00127	F: CTAAAATTCCCTCATCGTCTC R: CTCGAAGCTCATCGTCTCTC	ES324086	52	$(TTC)_9$	157	145–165	6	0.45	Syntaxin (AAN03474.1)
10	sEg00140	F: AAGTGAGACGGTGGATTTGG R: GTTCCAGTTGTCCTCGCATT	ES324087	53	$(GA)_{10}$	194	210–226	5	0.51	Unknown protein (ABK21888.1)

[a] PIC: polymorphic information content (Average = 0.53).
[b] Putative function is based on BLASTx search of GenBank sequences (January 2008). Sequence similarity was considered significant at a BLAST E value of $\leq 10^{-2}$.
[c] Figures in parentheses are the GenBank numbers for the most significant hits.

Table 3. Summary of the mean number of observed and effective alleles (A_o and A_e), percentage of polymorphic loci (0.99 criterion) (P), observed and expected heterozygosity (H_o and H_e) and fixation index (F_{IS}) for 7 oil palm germplasm collections and one Deli *dura* breeding population.[a]

Country of origin	N	A_o	A_e	P	H_o	H_e	F_{IS}
Deli *dura*	10	2.20	1.50	70.0	0.240 (0.366)	0.243 (0.244)	0.0123
Madagascar	8	2.20	1.66	80.0	0.213 (0.312)	0.340 (0.222)	0.3735
Gambia	8	2.30	1.75	80.0	0.438 (0.379)	0.347 (0.257)	−0.2622
Ghana	10	2.70	1.97	70.0	0.388 (0.369)	0.385 (0.285)	0.0078
Congo	10	3.20	2.23	100.0	0.390 (0.360)	0.406 (0.247)	0.0394
Cameroon	10	2.50	1.65	100.0	0.338 (0.296)	0.328 (0.208)	−0.0305
Nigeria	10	2.80	2.12	100.0	0.460 (0.378)	0.442 (0.205)	−0.0407
Senegal	10	2.60	1.85	90.0	0.420 (0.388)	0.353 (0.259)	−0.1899
Mean	–	2.56	1.84	86.25	0.361 (0.356)	0.356 (0.241)	−0.0113

[a] N = Number of palms assayed. Values in parentheses are the standard deviations. H_e = Nei's 1973 expected heterozygosity. F_{IS} = Wright's inbreeding coefficient $\{1 - (H_o/H_e)\}$.

ied was relatively high and ranged from 70% (Ghana) to 100% (Congo, Cameroon and Nigeria) (Table 3). The higher percentage of polymorphic loci observed for the germplasm from Congo, Cameroon and Nigeria, compared to collections from the rest of Africa was in agreement with Maizura et al. (2006). Generally P revealed by EST-SSR was higher than that reported by RFLP (Maizura et al. 2006) or isozyme analysis (Hayati et al. 2004).

The highest diversity was found in the germplasm from Nigeria, Cameroon and Congo, which decreases somewhat moving to the Western bulge of Africa (consisting of Ghana, Gambia and Senegal). Although Ghana, situated at the Southern coast of the Western bulge, is located close to Nigeria, the percentage of polymorphism observed in the germplasm was much lower, possibly due to its natural oil palms being lost (and with it the diversity) to the cultivation of cocoa (Maizura et

Table 4. F-statistics (F_{ST}) for all loci across 7 *E. guineensis* germplasm collections excluding Deli *dura* (Nei's genetic distances).

Locus	sEg00077	sEg00066	sEg00126	sEg00127	sEg00090
F_{ST}	0.3644	0.2037	0.1486	0.1799	0.3155
Locus	sEg00080	sEg00067	sEg00125	sEg00032	sEg00140
F_{ST}	0.1708	0.4543	0.1661	0.0000	0.2613
Mean	0.2492				

Table 5. Estimates of mean Nei's 1978 genetic distances between populations of *E. guineensis* from 7 provenances and one Deli *dura* family.

Population identification	Deli *dura*	Madagascar	Gambia	Ghana	Congo	Cameroon	Nigeria	Senegal
Deli *dura*	****							
Madagascar	0.360	****						
Gambia	0.460	0.370	****					
Ghana	0.355	0.378	0.182	****				
Congo	0.132	0.271	0.332	0.231	****			
Cameroon	0.058	0.299	0.351	0.240	0.101	****		
Nigeria	0.133	0.232	0.311	0.230	0.100	0.062	****	
Senegal	0.338	0.257	0.065	0.135	0.215	0.237	0.190	****

al. 2006) and for making 'down-wine' (alcoholic beverage, the juice extracted from the terminal cabbage of felled palms and fermented) (Hartley 1988). Generally the results from this study using the EST-SSRs are in agreement with Maizura et al. (2006) using the RFLP analysis that the area covering Nigeria and Congo may represent the centre of diversity for wild oil palm.

Genetic variability and fixation indices

The mean observed (H_o) and expected (H_e) heterozygosity across populations were 0.361 and 0.356, respectively (Table 3). The genetic diversity in the populations examined in this study ($H_e = 0.356$) using EST-SSRs was higher than those reported for oil palm using RFLP (H_e ranging from 0.168 to 0.232) (Maizura et al. 2006) and isozymes (mean $H_e = 0.184$) (Hayati et al. 2004). The differences could be due to the different marker systems used and also to the differences in the number of accessions analyzed per country. Generally, the genetic diversity observed using EST-SSRs for oil palm was higher than that reported for monocots ($H_e = 0.144$) (Hamrick & Godt 1989) and predominantly allogamous plants ($H_e = 0.214$) (Loveless & Hamrick 1984). The observed differences can also be affected by the number of loci analyzed, as observed in other crop plants; as such, care must be taken when comparing different studies (Mantovani et al. 2006). The EST-SSR markers revealed that the germplasm from Africa had higher genetic variability than Deli *dura*. This was not surprising, as the Deli *dura* population has often been selfed. The results point to the African germplasm being a rich source of new genes for oil palm improvement in Malaysia.

The fixation index (F_{IS}) for most of the germplasm collections ranged from −0.0407 (Nigeria) to 0.3735 (Madagascar) (Table 3). F_{IS} indicates a deviation of genotype frequencies from HWE in a single population (Soltis & Soltis 1989). The high and positive F_{IS} value for the Madagascar germplasm (0.3735) indicates an excess of homozygotes. Similar results were observed using isozyme analysis (Hayati et al. 2004). This is expected as Madagascar is an island off Southeast Africa (Fig. 1), and the oil palm there would have been an isolated population. Furthermore, the oil palms in Madagascar do not occur as densely as in the other African countries with the palm (Rajanaidu 1985). These factors favoured inbreeding and restricted the gene flow, contributing to the lower heterozygosity. The negative F_{IS} values for some of the germplasms (Gambia, Cameroon, Nigeria and Senegal) generally indicate that out-crossing prevails in these populations.

Genetic differentiation and genetic relatedness

Significant differences were also observed in genetic differentiation (F_{ST}) for the EST-SSR loci, with values ranging from 0.1486 (sEg00126) to 0.4543 (sEg00067) (Table 4). The high values for sEg00077 (0.3644), sEg00090 (0.3155) and sEg00067 (0.4543) indicate the ability of the EST-SSR markers identified in this study to differentiate the germplasm by their provenance. The overall degree of genetic differentiation (F_{ST}) among the oil palm populations was 0.2492, lower than that reported for the analysis of oil palm populations from Africa using isozyme analysis (0.301) (Hayati et al. 2004).

The mean genetic distance between the populations is summarized in Table 5, the lowest being between Nigeria and Cameroon (0.062), and the highest between Ghana and Madagascar (0.378). The results are not surprising as Nigeria and Cameroon are neighbouring countries while Ghana and Madagascar are far apart (Fig. 1). Generally, there was a strong association between the genetic distance and geographical location. This is further illustrated in Figure 2, which shows a dendrogram of the genetic relatedness between the 7 populations and Deli *dura*. The populations from

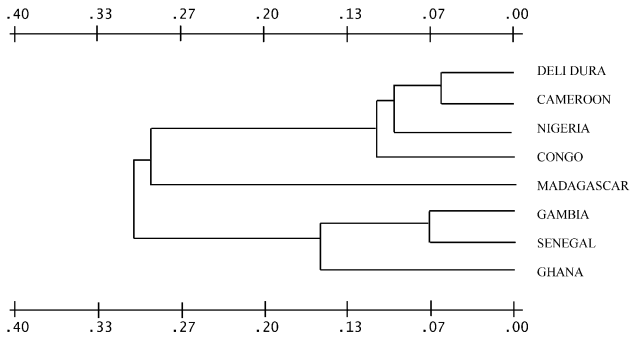

Fig. 2. UPGMA clustering of 7 populations of *E. guineensis* and one Deli *dura* population based on the genetic distances according to Nei (1978).

Congo, Cameroon and Nigeria are in one cluster, while the populations from the Western bulge of Africa (Gambia, Ghana and Senegal) form a separate cluster. The Madagascar population clearly stood apart from the other collections. This was expected as the isolated island palms have developed their own unique traits, such as shorter trunk resulting from slower growth rate (Hartley 1988). The results also indicate that despite the continuous distribution of oil palm all over Africa, there is a poor gene flow across the continent. The high genetic differentiation might be attributed to the pollinators (wind and weevil) being effective only over short distances, in effect minimizing the gene spread.

Transferability of the EST-SSRs

The 10 EST-SSR primers could also amplify in the second species of oil palm, *E. oleifera* (Table 6). However, transferability of the SSR loci varied when tested on coconut and *Jessenia* samples. Six of the loci also yielded clear amplicons in coconut and *Jessenia* samples (see Table 6 and Figure 3). At this stage it is not known whether the primer pairs are amplifying the same (orthologous) gene in the different samples. Nevertheless, the results indicate that the EST-SSR markers are good candidates for the development of molecular markers across the two species of oil palm and even for genetic analysis across the palm taxa.

Conclusions

This paper reports on the development of SSR markers from an oil palm EST database. The search revealed that the oil palm ESTs contained mono-, di-, tri- and tetra-nucleotide motifs. All the SSRs were found in ESTs and they are expected to become useful tools for the oil palm ecological, genetic and evolutionary studies. Dimeric repeats, especially AG/CT, were the most abundant. A total of 94 primer pairs were designed from

Table 6. Summary of transferability of 10 SSR loci across the different species and taxa in the Palmae family.

SSR locus	Family Palmae Tribe Cocoeae *Elaeis guineensis*	*Elaeis oleifera*	*Cocos nucifera* (Green)	*Cocos nucifera* (Yellow)	Tribe Areceae *Jessenia bataua*
sEg00032	260–270	260–270	260	260	260
sEg00066	192–215	201–210	NC[a]	NC	167–179
sEg00067	150–170	169–174	>152[b]	>152[b]	NC
sEg00077	160–180	170–184	NC	NC	>170[b]
sEg00080	145–155	145–155	152–161	154–161	143–152
sEg00090	200–210	228–250	200–260	200–260	–[c]
sEg00125	150–170	155	147–162	147–162	155–163
sEg00126	212–215	215	224	224	205
sEg00127	145–165	161	132–161	132–161	148–161
sEg00140	210–226	>213	202–243	202–243	>165[b]

[a] NC Banding pattern not clear. [b] Further optimization needed because of the presence of non specific bands. [c] No amplification observed.

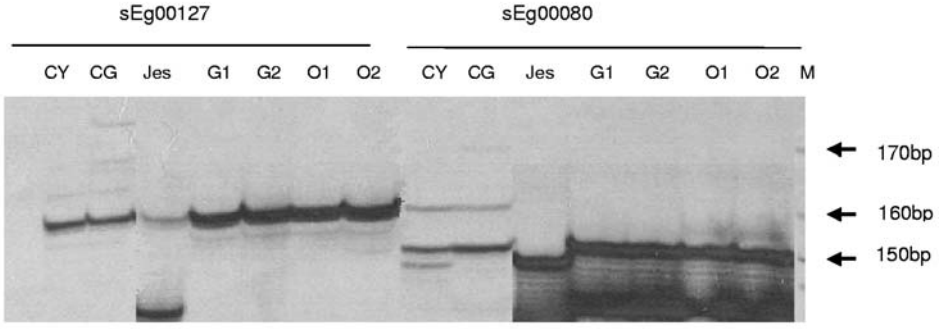

Fig. 3. Autoradiograph showing transferability of 2 microsatellite loci across different samples; *Cocos nucifera* (Yellow) (CY), *Cocos nucifera* (Green) (CG), *Jessenia bataua* (Jes), *Elaeis guineensis* (G1 and G2), *Elaeis oleifera* (O1 and O2). M is the AFLP molecular weight ladder.

the non-redundant set of SSR containing ESTs, for possible use as genic markers. The EST-SSR were able to estimate the genetic diversity as well as determine the genetic relationships between the different germplasm collections although only a limited set of primer pairs were used. The EST-SSRs further confirmed that there is a sufficient genetic variation in the MPOB African germplasm collections for further oil palm improvement. The genetic diversity profile revealed by the EST-SSRs in the MPOB African germplasm collections was generally similar to that exposed by RFLP and isozyme analysis carried out previously. The centre of diversity for wild oil palm appears to be the area covering Nigeria and Congo. More importantly, the EST-SSRs can help to manage the germplasm collections in Malaysia effectively by identifying populations with high levels of diversity for which more palms should be conserved than for populations with lower diversity. The EST-SSRs also showed high transferability across palm taxa, indicating their potential application in comparative genomic studies.

Acknowledgements

The authors wish to thank the Director-General of MPOB for permission to publish this manuscript. Special thanks to Mr. Andy Chang Kwong Choong and Dr. N. Rajanaidu for their valuable comments on the manuscript.

References

Anderson J.A., Churchill G.A., Autrique J.E., Tanksley S.D. & Sorrells M.E. 1993. Optimizing parental selection for genetic linkage maps. Genome **36:** 181–186.

Aggarwal R.K., Hendre P.S., Varshney R.K., Bhat P.R., Krishnakumar V. & Singh L. 2007. Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. Theor. Appl. Genet. **114:** 359–372.

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J. & Wheeler D.L. 2007. GenBank. Nucleic Acids Res. **35 (Database Issue):** D21–D25.

Bakoume C.R. 2006. Genetic diversity of natural oil palm (*Elaeis guineensis* Jacq.) populations using microsatellite markers. PhD. Thesis, Universiti Kebangsaan Malaysia, Kuala Lumpur.

Billotte N., Risterucci A.M., Barcelos E., Noyer J.L., Amblard P. & Baurens F.C. 2001. Development, characterisation, and across-taxa utility of oil palm (*Elaeis guineensis* Jacq.) microsatellite markers. Genome **44:** 413–425.

Cardle L., Ramsay L., Milbourne D., Macaulay M., Marshall D. & Waugh R. 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics **156:** 847–854.

Chabane K., Ablett G.A., Cordeiro G.M., Valkoun J. & Henry R.J. 2005. EST versus genomic derived microsatellite markers for genotyping wild and cultivated barley. Genet. Resour. Crop Evol. **52:** 903–909.

Doyle J.J. & Doyle J.L. 1990. Isolation of plant DNA from fresh tissue. Focus **12:** 13–15.

Ewing B. & Green P. 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res**. 8:** 186–194.

Ewing B., Hillier L., Wendl M.C. & Green P. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res. **8:** 175–185.

Hamrick J.L. & Godt M.J.W. 1989. Allozyme diversity in plant species, pp. 43–63. In: Brown A.H.D, Clegg M.J, Kahler A.L & Weir B.S (eds), Plant Population Genetics, Breeding and Genetic Resources, Sinauer Associates Inc., Sunderland.

Hartley C.W.S. 1988. The Oil Palm (*Elaeis guineensis* Jacq.). Longman Scientific and Technical Publication, New York, 761 pp.

Hayati A.,Wickneswari R., Maizura I. & 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:** 274–1284.

Kularatne R.S. 2000. Assessment of genetic diversity in natural oil palm (*Elaeis guineensis* Jacq.) populations using amplified fragment length polymorphism markers. PhD. Thesis, Universiti Kebangsaan Malaysia, Kuala Lumpur.

Loveless M.D. & Hamrick, J.L. 1984. Ecological determinants of genetic structure in plant population. Annu. Rev. Ecol. Syst. **15:** 65–95.

Maizura I., Rajanaidu N., Zakri A.H. & Cheah S.C. 2006. Assessment of genetic diversity in oil palm (*Elaeis guineensis* Jacq.) using Restriction Fragment Length Polymorphism (RFLP). Genet. Res. Crop Evol. **53:** 187–195.

Mantovani A., Morellato L.P.C. & Reis M.S. 2006. Internal genetic structure and outcrossing rate in natural population of *Araucaria angustifolia* (Bert) O. Kuntze. J. Hered. **97:** 466–472.

Manimekalai R. & Nagarajan P. 2006. Interrelationships among coconut (*Cocos nucifera* L.) accessions using RAPD technique. Genet. Res. Crop Evol. **53:** 1137–1144.

Maria M., Clyde M.M. & Cheah S.C. 1995. Cytological analysis of *Elaeis guineensis* (*tenera*) chromosomes**.** Elaeis **7:** 122–134.

Miller R.T., Christoffels A.G., Gopalakrishnan C., Burke J., Ptitsyn A.A., Broveak T.R. & Hide W.A. 1999. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. Genome Res**. 9:** 1143–1155.

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

Purseglove J.W. 1972. Tropical Crops, Monocotyledons. London, Longman, 607 pp**.**

Rajanaidu N. 1985. The oil-palm (*Elaeis guineensis*) collections in Africa, pp 59–83. In: International Workshop on Oil Palm Germplasm and Utilization, PORIM, Bangi, Selangor, Malaysia.

Rajanaidu N. & Jalani B.S. 1994. Oil palm genetic resources – collection, evaluation, utilization and conservation. In: PORIM Colloquium on Oil Palm Genetic Resources, 13 September 1994, PORIM, Bangi, Malaysia.

Rice P., Longden I. & Bleasby A. (2000) EMBOSS: the European molecular biology open software suite. Trends Genet. **16:** 276–277.

Rival A., Beule T., Barre P., Hamon S., Duval Y. & Noirot M. 1997. Comparative flow cytometric estimation of nuclear DNA content in oil palm (*Elaeis guineensis*, Jacq.) tissue cultures and seed derived plants. Plant Cell Reports **16:** 884–887.

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

Rungis D., Berube Y., Zhang J., Ralph S., Ritland C.E., Ellis B.E., Douglas C., Bohlmann J. & Ritland K. 2004. Robust simple sequence repeat markers for spruce (*Picea* spp.) from expressed sequence tags. Theor. Appl. Genet. **109:** 1283–1294.

Shah F.H., Rasid O., Simons A.J. & Dunsdon A. 1994. The utility of RAPD markers for the determination of genetic variation in oil palm (*Elaeis guineensis*). Theor. Appl. Genet. **89:** 713–718.

Sneath P.H.A. & Sokal R.R. 1973. Numerical Taxonomy: The Principles and Practice of Numerical Classification. Freeman, San Francisco, CA.

Soltis D.E. & Soltis P.S. 1989. Polyploidy, breeding systems and genetic differentiation in homosporous pteridophytes, pp.

241–258. In: Soltis D.E & Soltis P.S. (eds), Isozymes in Plant Biology, Dioscorides Press, Portland, Ore.

Temnykh S., Park W.D., Ayres N., Cartinhour S., Hauck N., Lipovich L., Cho Y.G., Ishii T. & McCouch S.R. 2000. Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). Theor. Appl. Genet**. 100:** 698–712.

Thiel T., Michalek W., Varshney R.K. & Graner A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor. Appl. Genet**. 106:** 411–422.

Varshney R.K., Chabane K., Hendre P.S., Aggrawal R.K. & Graner A. 2007. Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. Plant Sci. **173:** 638–649.

Varshney R.K., Sorrells M.E. & Graner A. 2005. Genic microsatellite markers in plants: features and applications. Trends Biotechnol**. 23:** 48–55.

Varshney R.K., Thiel T., Stein N., Langridge P.& Graner A. 2002. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell. Mol. Biol. Lett. **7:** 537–546.

Wang H.Y., Wei Y.M., Yan Z.H. & Zheng Y.L. 2007. EST-SSR DNA polymorphism in durum wheat (*Triticum durum* L.) collections. J. Appl. Genet. **48:** 35–42.

Yeh F.C & Boyle T. 1999. POPGENE version 1.32. The user-friendly software for population genetic analysis. University of Alberta and CIFOR, Calgary.

Zeven A.C. 1967. The semi-wild oil palm and its industry in Africa. Agricultural Research Report 698. Agricultural University, Wageningen, The Netherlands.

Zhang L.Y., Ravel C., Bernard M., Balfourier F., Leroy P., Feuillet C. & Sourdille P. 2006. Transferable bread wheat EST-SSRs can be useful for phylogenetic studies among Triticeae species. Theor. Appl. Genet**. 113:** 407–418.

Received August 13, 2007
Accepted January 18, 2008