RESEARCH ARTICLE

SSR mining in oil palm EST database: application in oil palm germplasm diversity studies

NGOOT-CHIN TING[1†], NOORHARIZA MOHD ZAKI[1†], ROZANA ROSLI[1†], ENG-TI LESLIE LOW[1], MAIZURA ITHNIN[1], SUAN-CHOO CHEAH[1,3], SOON-GUAN TAN[2] and RAJINDER SINGH[1*]

[1]*Advanced Biotechnology and Breeding Centre, Malaysian Palm Oil Board, P.O. Box 10620, 50720 Kuala Lumpur, Malaysia*
[2]*Department of Cell and Molecular Biology, Faculty of Biotechnology and Biomolecular Sciences, University Putra Malaysia, 43400 UPM Serdang, Malaysia*
[3]*Present address: Asiatic Centre for Genome Technology, Lot L3-1-1, Enterprise 4, Technology Park, Malaysia, Bukit Jalil, 57000 Kuala Lumpur, Malaysia*

Abstract

This study reports on the detection of additional expressed sequence tags (EST) derived simple sequence repeat (SSR) markers for the oil palm. A large collection of 19243 *Elaeis guineensis* ESTs were assembled to give 10258 unique sequences, of which 629 ESTs were found to contain 722 SSRs with a variety of motifs. Dinucleotide repeats formed the largest group (45.6%) consisting of 66.9% AG/CT, 21.9% AT/AT, 10.9% AC/GT and 0.3% CG/CG motifs. This was followed by trinucleotide repeats, which is the second most abundant repeat types (34.5%) consisting of AAG/CTT (23.3%), AGG/CCT (13.7%), CCG/CGG (11.2%), AAT/ATT (10.8%), AGC/GCT (10.0%), ACT/AGT (8.8%), ACG/CGT (7.6%), ACC/GGT (7.2%), AAC/GTT (3.6%) and AGT/ACT (3.6%) motifs. Primer pairs were designed for 405 unique EST-SSRs and 15 of these were used to genotype 105 *E. guineensis* and 30 *E. oleifera* accessions. Fourteen SSRs were polymorphic in at least one germplasm revealing a total of 101 alleles. The high percentage (78.0%) of alleles found to be specific for either *E. guineensis* or *E. oleifera* has increased the power for discriminating the two species. The estimates of genetic differentiation detected by EST-SSRs were compared to those reported previously. The transferability across palm taxa to two *Cocos nucifera* and six exotic palms is also presented. The polymerase chain reaction (PCR) products of three primer-pairs detected in *E. guineensis, E. oleifera, C. nucifera* and *Jessinia bataua* were cloned and sequenced. Sequence alignments showed mutations within the SSR site and the flanking regions. Phenetic analysis based on the sequence data revealed that *C. nucifera* is closer to oil palm compared to *J. bataua*; consistent with the taxanomic classification.

[Ting N.-C., Noorhariza M. Z., Rozana R., Low E.-T., Ithnin M., Cheah S.-C., Tan S.-G. and Singh R. 2010 SSR mining in oil palm EST database: application in oil palm germplasm diversity studies *J. Genet.* **89**, 135–145]

Introduction

Oil palm is the most valuable oil crop in Southeast Asia, particularly in Malaysia and Indonesia. It is a monoecious and diploid cross-pollinating species. The haploid genome size of *E. guineensis* has been estimated to be 1700 Mbp (Rival *et al*. 1997). Both *E. guineensis* and *E. oleifera* have same number of chromosomes ($2n = 2x = 32$) and can be hybridized to produce fertile offspring. However, the oil palm breeding populations in Malaysia and Indonesia are derived from a narrow genetic pool. Most of the commercial planting materials used are derived from the Deli *dura*, which was first introduced in Indonesia and subsequently in Malaysia (Corley and Tinker 2003). This had resulted in several expeditions being mounted in central Africa and south-central America to collect germplasm (Rajanaidu *et al*. 1999).

Assessment of the oil palm genetic diversity is an important step towards effective utility in breeding programmes. In Malaysia, the use of molecular markers to screen the oil palm germplasm included isozymes (Hayati *et al*. 2004)

*For correspondence. E-mail: rajinder@mpob.gov.my.
†These authors contributed equally to this work.

Keywords. genetic variation; transferability; phenetic analysis; *Elaeis guineensis.*

and restriction fragment length polymorphism (RFLP) markers (Maizura *et al*. 2006). However, the application of new marker systems is beneficial so as to further exploit and understand their genetic structure. Billotte *et al*. (2001) had successfully developed oil palm genomic-SSRs and showed their application in characterizing germplasm materials. Due to their ease of use, SSRs appeared to be the favored marker system for studies of genetic diversity in coconut (Perera *et al*. 2000, 2001), date palm (Zehdi *et al*. 2004, 2005) and also in peach palm (Adin *et al*. 2004). Recently, Singh *et al*. (2008) reported on the development of SSR markers from a small collection of ESTs and showed that they were useful for the genetic analysis of *E. guineensis* germplasm.

In this study, we expanded our previous effort to characterize more informative SSR markers from a larger collection of ESTs, which should provide a clear picture of distribution of the EST-SSRs in oil palm. The scanning of ESTs using specific computer programs facilitates the discovery of SSRs, making it easy and relatively a cheaper process (Varshney *et al*. 2005). A small subset of the SSRs identified were utilized for studying the genetic diversity of the *E. guineensis* and *E. oleifera* germplasm. An added advantage is that EST-SSRs are present in the gene-rich regions, which are usually more conserved, thus improving its transferability upon breeding. The work reported here also includes the transferability of EST-SSRs across palm taxa.

Materials and methods

In silico mining for SSRs

The *PalmGenes* (http://palmoilis.mpob.gov.my/palmgenes.html) is an oil palm EST database that provides sequence information, functional classification, clone information, references, research findings and other information related to the gene clone of interest. An *in silico* mining for SSRs were performed as described by Singh *et al*. (2008) using 5610 ESTs from *PalmGenes*, and an additional 13633 sequences were downloaded from an in-house database, at Malaysian Palm Oil Board (MPOB) known as *PalmDNABase* (Sambanthamurthi *et al*. 2009). The functions of the ESTs were predicted by performing a similarity search against the National Centre for Biotechnology Information (NCBI) non-redundant protein database using Blastx. A Blastx score of at least 80 was considered significant to assign a putative function. An additional information on the ESTs sequences is provided by Low *et al*. (2008).

Plant materials

A total of 105 *E. guineensis* and 30 *E. oleifera* accessions were analysed, 8–16 palms were chosen from each country as shown in table 1. In addition, an advanced oil palm breeding material, Deli *dura* was included as a reference. Across palm taxa, transferability of the EST-SSR was tested on two

Table 1. Hundred and five *E. guineensis* palms, 30 *E. oleifera* palms, three coconut palms and seven exotic palms used in the current study for screening by 15 EST-SSR primer pairs.

Family	Tribe	Subtribe	Species	Origin	No. of individuals
Palmae	Cocoeae	Eleaeidinae	*Elaeis guineensis*	Deli *dura*	10
				Madagascar	12
				Gambia	8
				Ghana	16
				Congo	16
				Cameroon	13
				Nigeria	16
				Senegal	14
				Subtotal	105
Palmae	Cocoeae	Eleaeidinae	*Elaeis oleifera*	Colombia	10
				Panama	10
				Costa Rica	10
				Subtotal	30
Palmae	Cocoeae	Butiinae	*Cocos nucifera*	Solomon Island	3
Palmae	Areceae	Euterpeinae	*Euterpe oleracea*	South America	2
Palmae	Areceae	Euterpeinae	*Jessenia bataua Mart.*	South America	1
Palmae	Areceae	Euterpeinae	*Oenocarpus multicaulis Spruce*	North-western South America	1
Palmae	Areceae	Ptychospermatinae	*Ptychosperma macarthurii*	Not native to North America	1
Palmae	Areceae	Cyrtostachydinae	*Cyrtostachys renda Blume*	Malaysia, Indonesia	1
Palmae	Areceae	Iguanurinae	*Dictyosperma album*	Mauritius	1
				Subtotal	10
Total					145

coconut *(Cocos nucifera)* and six exotic palms (*E. oleracea, J. bataua, O. multicaulis, P. macarthurii, C. renda* and *D. album*). The germplasm used in this study are currently grown and maintained at the MPOB Research Station at Kluang, Johor, Malaysia.

SSR analysis

Genomic DNA was extracted from young leaves. The DNA extraction and SSR analysis were performed as described by Singh *et al*. (2008) with slight modification. The forward primer was 5′ end labelled at 37°C for 1 h using T4 polynucleotide kinase (Invitrogen, Carlsbad, USA). The labelling reaction contained 4.5 μM forward primer, 0.1 μL dATP (γ^{-33}P)-3000Ci/mmol (GE Healthcare Biosciences, Buckinshamshire, UK) and 1 U T4 polynucleotide kinase in a total volume of 1 μL. The PCR reaction was subsequently carried out in a final volume of 10 μL consisting of 1 μL of 10× PCR buffer (buffer composition -$MgCl_2$), 15 mM $MgCl_2$, 1 mM dNTPs, 5 μM unlabelled reverse primer, 1 μL labelled forward primer, 0.5 U *Taq* DNA polymerase and 50 ng template DNA. PCR was performed in a Perkin Elmer 9600 thermocycler as follows: denaturation at 95°C for 3 min; 35 cycles of 95°C for 30 s, 52–56°C for 30 s (depending on the primer) and 72°C for 30 s; and a final extension at 72°C for 5 min. The PCR reaction was stopped by the addition of 10 μL of formamide dye. Five μL of the mixture was denatured at 90°C for 3 min, chilled on ice, and separated on 6.0% polyacrylamide gel.

Only fragments that could be clearly scored were used in the data analysis. The genotype data were analysed using POPGENE version 1.32 (http://www.ualberta.ca/~fyeh/pr01.htm). The genetic variability measures calculated include mean number of alleles observed per locus (A_o), percentage of polymorphic loci (P), expected and observed heterozygosity (H_e and H_o) and inbreeding coefficient (F_{IS}). To test whether the F_{IS} differed significantly from Hardy–Weinberg equilibrium (HWE), chi-square analysis, $\chi^2 = F_{IS}^2 N(k-1)$ was carried out. The degrees of freedom was determined as follows: df $= (k(k-1))/2$, where k is the number of alleles and N is the number of palms sampled in each collection (Jorgensen *et al*. 2002). The genetic distance between populations was computed according to Nei (1978) using BIOSYS-1, Release 1.7 (Swofford and Selander 1989). These values were used to generate a dendrogram using the unweighted pair-group with arithmetic averaging (UPGMA) cluster analysis.

Sequencing of cloned SSR-PCR products for alignment and phenetic analysis

Selected loci amplified by primers sEg00126, sEg00127 and sEg00080 from *E. guineensis* (Cameroon origin), *E. oleifera* (Colombia origin), *C. nucifera* 1 (yellow coconut), *C. nucifera* 2 (green coconut) and *J. bataua* (exotic palm) were excised from the agarose gel. The excised fragments were cloned into pCR2.1-TOPO vector (Invitrogen, Carlsbad, USA) and sequenced using the ABI PRISM 377 automated DNA sequencer (Perkin Elmer-Applied Biosystems, Foster City, USA). The sequences were aligned using ClustalW Multiple Alignment in BioEdit Sequence Alignment Editor (version 7.0.0) (www.mbio.ncsu.edu). Subsequently, an unrooted phylogenetic tree was constructed using the Phylip's drawtree program via Biology Workbench 3.2 (http://workbench.sdsc.edu). The distance was calculated based on the Neighbour Joining method (Saitou and Nei 1987). Plot options such as orientation of tree and branches, label sizes, angles and margin were adjusted to get the desired plot.

Results and discussion

Frequency and distribution of SSRs

The 19243 ESTs examined represent ~8.8 Mb of the oil palm genome. Of these, 17599 sequences have been deposited in GenBank (accession nos. EY396120–EY413718). The cluster analysis identified 6809 singletons and 3591 consensus to give a total of 10400 unique sequences. A total of 142 sequences were eliminated due to their short length (< 100 bp) and were suspected to be vector sequences. The remaining 10258 unique sequences were mined for SSRs which is almost double the number reported by Singh *et al*. (2008). The present analysis revealed 722 SSRs in 629 ESTs. The percentage of SSR uncovered (7.0%) is comparable to barley (Thiel *et al*. 2003) but greater than that reported for sugarcane, where only 2.9% of the ESTs were found to contain SSRs (Cordeiro *et al*. 2001). This suggests that oil palm ESTs are indeed a valuable source of markers. Dinucleotide repeats formed the largest group (45.6%) consisting of 66.9% AG/CT, 21.9% AT/AT, 10.9% AC/GT and 0.3% CG/CG motifs. Trinucleotide was the second largest group (34.5%) consisting of AAG/CTT (23.3%), AGG/CCT (13.7%), CCG/CGG (11.2%), AAT/ATT (10.8%), AGC/GCT (10.0%), ACT/AGT (8.8%), ACG/CGT (7.6%), ACC/GGT (7.2%), AAC/GTT (3.6%) and AGT/ACT (3.6%) motifs. This was followed by mononucleotides (17.7%), tetranucleotides (14 SSRs) and pentanucleotides (two SSRs) (table 2).

A total of 405 SSR primer pairs were successfully designed from 722 unique EST-SSRs. Subsequently, 15 of the primers representing a variety of motifs (eight, di-; five, tri- and two, tetra- repeats) were randomly selected to analyse the oil palm germplasm. The 15 selected primers included 10 primers that were reported previously by Singh *et al*. (2008) and the primer information is shown in table 3. Six out of eight sequences with dinucleotide repeats did not show any similarity to genes in the database. It is likely that these di-repeats are mostly found in the UTR instead of the coding regions of the ESTs. Similarly, Tang *et al*. (2009) reported that most of the dinucleotide EST-SSRs for Iris was indeed present in the UTR.

Table 2. Oil palm non-redundant SSRs discovered from 10258 unique ESTs.

	Number of repeats												
Repeat type	5	6	7	8	9	10	11	12	13	14	15	> 15	Total
Mono-repeats (N)	–	–	–	–	–	44	27	10	9	2	9	27	128
Di-repeats (NN)	–	–	96	75	61	16	20	11	10	10	8	22	329
Tri-repeats (NNN)	121	67	28	12	12	3	2	2	–	1	–	1	249
Tetra-repeats (NNNN)	7	5	–	1	–	1	–	–	–	–	–	–	14
Penta-repeats (NNNNN)	2	–	–	–	–	–	–	–	–	–	–	–	2
Grand total													722

Table 3. EST-SSR primer information, putative function of the ESTs and the SSR motif.

SSR ID	Primer pair sequence (5′–3′)	Accession no.	Putative ID[a]	Annealing temperature (°C)	Motif
Di-repeats					
sEg00009	F: TCCACTGACAACAGGACTCA R: AAAAACGCATCTCAGAGAGA	9947958^{+}	No significant similarity	52	$(AG)_{12}$
sEg00036	F: GGACCCTTTTTGTTACTGTTT R: AGCCTACCACAACTTCCTTT	9947959^{+}	No significant similarity	52	$(AG)_{9}$
sEg00066	F: ACTGATGCAGGAAAGAGGAA R:GAAGTACACAAGGTAAGTTCATAG	ES324079^{++}	No significant similarity	52	$(AT)_{8}$
sEg00076	F:GTGCTATATACTGTCACTTAAGATT R: GGTAGAATATCTTCGTTTCGATT	9947961^{+}	No significant similarity	52	$(TG)_{9}$
sEg00077	F: TTACAAGCCACCTCACAAGC R: ATACCAGCATCAAGTCAAAAT	ES324081^{++}	No significant similarity	52	$(TA)_{8}$
sEg00090	F: TCAGAAATGCCTACATCAAAC R: AGGGACACGAGAATACATACA	ES324083^{++}	L-ascorbate peroxidase (*Arabidopsis thaliana*)[b]	52	$(AT)_{9}$
sEg00113	F: GTCACCGAACCCTAATAAAAT R: ATGCAGTTGAGGACAAAAAG	9947962^{++}	No significant similarity	54	$(CT)_{15}$
sEg00140	F: TAGAAAGTGAGACGGTGGAT R: GTAATATTCTCAAGCTGGCAGT	ES324087^{++}	Hypothetical protein (*Vitis vinifera*)[b]	53	$(GA)_{10}$
Tri-repeats					
sEg00038	F: ATCAAGCGGCAGTTATGAGAT R: ATACATTATTCCCACCACCA	9947960^{++}	No significant similarity	52	$(AAT)_{9}$
sEg00080	F: AAGAACTATGACCTCACCAAAA R: AACTCTATGCTATTGCTACACGA	ES324082^{++}	No significant similarity	52	$(TCA)_{6}$
sEg00125	F: TACCCTTTTCCCTCCCTCCATA R: CATCATCTCCGTTGCCAGTATT	ES324084^{++}	Hypothetical protein (*Vitis vinifera*)[b]	52	$(GCG)_{6}$
sEg00126	F: CCGTCTCAAAAGCCCTAAAC R: TTGTTGTCCCACTCCCTCTT	ES324085^{++}	Predicted protein (*Populus trichocarpa*)[b]	52	$(CGC)_{7}$
sEg00127	F: CTAAAATTCCCTCATCGTCTC R: CTCGAAGCTCATCGTCTCTC	ES324086^{++}	Syntaxin (Glycine max)[b]	52	$(TTC)_{9}$
Tetra-repeats					
sEg00032	F: CTGTTGAGCTGGAGAGACCC R: CCAACCAGGATCAGTTTGGT	ES324078^{++}	Predicted protein (*Populus trichocarpa*)[b]	55	$(CTTT)_{5}$
sEg00067	F: GATTAAGTCCCAACCGTCTC R: TAAGAGAGCACGCAGTTCAG	ES324080^{++}	No significant similarity	52	$(TGTA)_{6}$

[a]Putative gene ID was inferred from significant homology search (score ≥ 80) using Blastx. No significant similarity represents sequences that do not have any homology to sequences in GenBank at score of ≥ 80. $^{+}$ProbeDB UID.$^{++}$GenBank.

Genetic variations in E. guineensis and E. oleifera

The number of alleles for each of the 15 SSRs across all the 11 collections analysed are summarized in table 4. A total of 101 alleles were detected and of these, 63.3% and 14.9% were unique for *E. guineensis* and *E. oleifera*, respectively, while 21.8% were common in both species. The presence of a high percentage (78.2%) of unique and specific alleles revealed genetic variation in the two species. The results established a set of polymorphic EST-SSRs that can describe

the two oil palm species. For example, sEg00038 detected numerous unique alleles in *E. guineensis* when compared to only two monomorphic alleles in *E. oleifera* (figure 1).

Interestingly, only 41 alleles were detected in Deli *dura* where A_o was 2.7 (as shown in figure 2) compared to the *E. guineensis* germplasm (2.8–3.9). These confirmed earlier suspicions that some alleles found in the native populations were lost in Deli *dura*. Comparing the two palm species, fewer alleles per locus were generally detected in *E. oleifera* where the A_o ranged from 1.87 to 2.20. This is expected as the accessions were collected from a relatively narrow region, near the central part of the continent. In contrast, the *E. guineensis* accessions originated from a wider range of geographical area on the African continent. The A_o level is also affected by the number of markers and sample size analysed. Singh *et al*. (2008) previously observed few alleles per locus in *E. guineensis* (A_o = 2.2 – 3.2). The A_o has been increased in the current study by using a larger set of SSRs to analyse a wider pool of germplasm. Furthermore, different marker types also affect A_o. Figure 2 show that RFLP (Maizura *et al*. 2006) and isozyme (Hayati *et al*. 2004; Purba *et al*. 2000) revealed a smaller A_o($\leq$ 2.0). Comparatively, EST-SSRs were clearly able to reveal more alleles.

Fourteen of the 15 assayed SSR primers detected polymorphism in at least one of the collections. Among the repeats, dinucleotide detected more alleles (mean = 8.5 alleles) compared to tri-repeats and tetra-repeats which detected 5.2 and 3.5 alleles, respectively. This likely contributed to a higher PIC value for di-repeats (mean = 0.80) compared to tri-repeats (0.45) and tetra-repeats (0.60). The polymorphism level revealed by EST-SSRs was relatively high in *E. guineensis* ranging from 86.7% to 100.0%, while in

Table 4. Number of alleles for 15 EST-SSRs across 11 oil palm germplasm collections including the advanced planting material (Deli *dura*).

			E. guineensis								*E. oleifera*		
Locus	PIC	No. of alleles	*dura*	Mad	Gam	Gha	Con	Cam	Nig	Sen	Col	Cos	Pan
Di-repeats													
sEg00009	0.86	8	5	5	3	5	5	4	6	4	5	2	3
sEg00036	0.78	10	4	4	3	2	3	3	3	5	3	3	4
sEg00066	0.85	10	5	5	4	6	10	6	7	4	4	3	3
sEg00076	0.74	5	2	3	3	3	4	4	4	2	2	1	2
sEg00077	0.81	6	3	2	3	3	3	2	3	3	3	3	3
sEg00090	0.82	12	2	3	3	3	5	4	4	3	3	4	4
sEg00113	0.87	11	4	6	4	7	7	5	9	6	1	1	1
sEg00140	0.64	6	3	3	2	4	2	2	2	3	2	1	1
Tri-repeats													
sEg00038	0.79	9	3	3	6	6	5	3	4	6	2	2	2
sEg00080	0.13	2	1	1	1	1	2	2	2	1	1	1	2
sEg00125	0.57	6	1	1	2	4	3	3	3	2	1	2	2
sEg00126	0.36	2	2	2	2	1	2	2	2	2	1	1	1
sEg00127	0.40	7	2	2	3	6	3	2	3	3	1	1	1
Tetra-repeats													
sEg00032	0.50	2	2	2	2	2	2	2	2	2	2	2	2
sEg00067	0.69	5	2	3	1	2	3	3	3	2	2	1	0
Total	–	101	41	45	42	55	59	47	57	48	33	28	30

dura, Deli *dura*; Mad, Madagascar; Gam, Gambia; Gha, Ghana; Con, Congo; Cam, Cameroon; Nig, Nigeria; Sen, Senegal; Col, Colombia; Cos, Costa Rica; Pan, Panama.

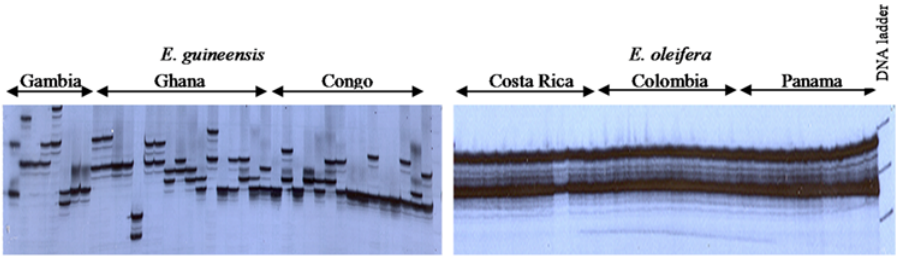

Figure 1. EST-SSR polymorphism revealed by sEg00038 in *E. guineensis* and *E. oleifera* visualized on 6.0% polyacrylamide gel.

E. oleifera it ranged from 53.3% –66.7% (table 5). This could be due to ascertainment bias where relatively long repeats derived from *E. guineensis* would have a higher chance of being polymorphic in *E. guineensis*. The highest *P* value was observed in the Congo, Cameroon and Nigeria collections, results which were similar to those reported by Singh *et al*. (2008). However, the *P* value observed in *E. guineensis* germplasm on an average was higher. The *P* value observed in this study was also higher compared to isozyme (Hayati *et al*. 2004) and RFLP (Maizura *et al*. 2006).

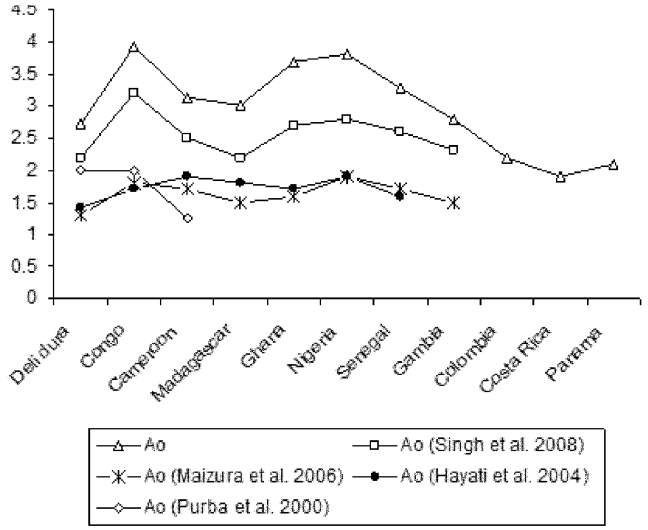

Figure 2. SSRs allelic variation observed in the current study compared to the previous reports by using SSRs (Singh *et al*. 2008), RFLPs (Maizura *et al*. 2006) and isozymes (Purba *et al*. 2000; Hayati *et al*. 2004).

Elaeis oleifera collections had a lower genetic variation with values ranging from 0.213 to 0.323 (mean H_e = 0.286). In contrast, the average H_e within *E. guineensis* accessions was 0.437. The Deli *dura* exhibited the lowest H_e value (0.340) while the Nigeria germplasm exhibited the highest (0.534). The individual *E. guineensis* germplasm could be ranked based on the H_e measurements as follows: Nigeria > Congo > Ghana > Cameroon > Senegal > Gambia > Madagascar > Deli *dura*. This trend is similar to that reported by Singh *et al*. (2008) and Maizura *et al*. (2006) supporting the postulation that Nigeria is the centre of diversity for oil palm. This result also showed that heterozygosity decreases from the central regions (Nigeria, Congo, Ghana and Cameroon) towards the marginal regions (Senegal, Gambia and Madagascar). In fact, the natural palm groves in the central regions of Africa are also much denser than those found in the marginal regions. The overall genetic diversity observed for *E. guineensis* was lower than that reported using genomic SSRs (0.680) (Billotte *et al*. 2001). Unlike genomic SSRs, most of the EST-SSRs are located in the coding regions which would naturally possess greater conservation, thus explaining the lower level of polymorphism detected. However, EST-SSRs revealed higher diversity compared to that observed using AFLP (Purba *et al*. 2000), isozyme (Hayati *et al*. 2004) and RFLP (Maizura *et al*. 2006) where the H_e values reported were 0.300, 0.184 and 0.199, respectively. As such, EST-SSR is considerably more efficient in revealing heterozygosity.

The conformity to HWE was also determined. F_{IS} is a measure of the excess or reduction in heterozygosity of an individual due to nonrandom mating within the germplasm. Overall, F_{IS} values are low except for the Madagascar,

Table 5. Summary of the percentage of polymorphic loci (*P*), observed and expected heterozygosity (H_o and H_e) for all loci across 10 germplasm collections and the Deli *dura* (standard deviation in parentheses).

Germplasm	*P* (%)	H_o (SD)	H_e (SD)	F_{IS}
Elaies guineensis				
Deli *dura*	86.7	0.262 (0.340)	0.340 (0.241)	0.229*
Madagascar	86.7	0.194 (0.251)	0.374 (0.193)	0.481*
Gambia	86.7	0.437 (0.320)	0.400 (0.247)	−0.093
Ghana	86.7	0.441 (0.338)	0.461 (0.293)	0.043
Congo	100.0	0.468 (0.332)	0.495 (0.235)	0.055
Cameroon	100.0	0.444 (0.329)	0.452 (0.195)	0.018
Nigeria	100.0	0.518 (0.311)	0.534 (0.187)	0.030
Senegal	93.3	0.434 (0.355)	0.436 (0.280)	−0.005
Mean	92.5	0.400	0.437	0.095
Elaies oleifera				
Colombia	66.7	0.268 (0.368)	0.323 (0.268)	0.170
Costa Rica	53.3	0.181 (0.340)	0.213 (0.240)	0.150
Panama	62.5	0.240 (0.348)	0.323 (0.308)	0.257*
Mean	60.8	0.230	0.286	0.192

F_{IS}, Wright's inbreeding coefficient ($1 - H_o/H_e$).
*Significant deviation from HWE at $P < 0.01$.

Deli *dura* and Panama collections. The germplasm from these three origins were found to be significantly ($P < 0.01$) deviated from HWE, showing heterozygosity deficit. This is not surprising for the Madagascar germplasm, as the distribution of palms was very sparse where only a small number of samples with limited coverage (seven samples each from four sites) were collected (Rajanaidu *et al*. 1999). Twelve of these were used in the current study. This could have resulted in inbreeding, which is clearly demonstrated by the large difference between H_o and H_e due to non-HW mating.

The genetic differentiation among the germplasm (F_{ST}) ranged from 0.142 to 0.645 with a mean value of 0.373. In comparison with Singh *et al*. (2008), the use of a larger set of EST-SSRs improved the power of discrimination. Locus sEg00067, in particular had amplified distinctive alleles not only across the two species but it was also able to distinguish the Gambia collection from the other *E. guineensis* germplasm. The relatively high divergence among the germplasm had been attributed to several factors such as restriction of gene flow (e.g. pollen and seed dispersal) and ecotypic selection (e.g. geographical distance) as explained by Hayati *et al*. (2004).

Genetic relatedness and structure of the genus Elaeis

Nei's (1978) genetic distances revealed divergence within and among species ranging from 0.039 to 1.272. The results obtained is similar to Singh *et al*. (2008) where the highest distance among *E. guineensis* was observed between Ghana and Madagascar (0.562) and the lowest was between Nigeria and Cameroon (0.039). It is obvious that genetic divergence is associated with geographical distances. These can be attributed to environmental variations (and associated selective effects) becoming more heterogeneous with large distances and the low migration rates encouraging divergence by random drift (Cho and Tiedje 2001). Similarly in *E. oleifera*; Panama, which is located next to Costa Rica, showed the lowest genetic distance (0.060) and the highest distance was between Colombia and Costa Rica (0.166).

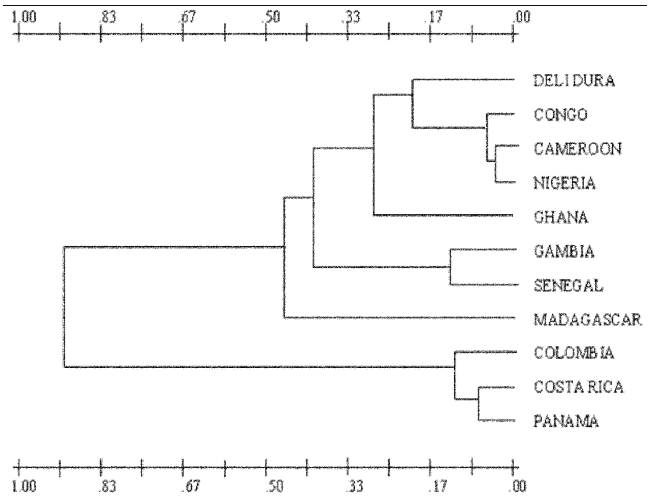

Figure 3. UPGMA clustering of seven *E.guineensis* and three *E. oleifera* populations with one Deli *dura* family based on Nei's (1978) genetic distances.

The UPGMA analysis categorized the *E. guineensis* and *E. oleifera* germplasm into two distinct clusters (figure 3). The *E. guineensis* cluster was further divided into two distinguishable sub-clusters, Madagascar and the African mainland. The separation of Madagascar might indicate the uniqueness of the oil palm found on the island such as short trunk and small fruits, not usually observed in *E. guineensis* on mainland (Rajanaidu *et al*. 1999). Subsequently, the continental cluster could be further separated into those of the western (Congo, Cameroon and Nigeria) and the northern (Gambia and Senegal) regions, respectively. The Deli *dura* was clustered with palms from the western region of Africa particularly close to Congo. This is in accordance with the early records where the Deli *dura* used in Malaysia and Indonesia could have their origin in West Africa (Purba *et al*. 2000).

Transferability of E. guineensis EST-SSRs to other palm species and their phenetic relationship

A total of 10 loci with clear banding profiles in both *E. guineensis* and *E. oleifera* were selected for transferability analysis (table 6). All the primers could amplify the DNA of coconut and of at least one of the exotic palms tested. Five primers (sEg00032, sEg00066, sEg00080, sEg00126 and sEg00127) exhibited clear banding profiles in all the coconut and exotic palms. Two primers, sEg00125 and sEg00140, also amplified clear bands in all samples except for *O. multicaulis* and *C. renda*, respectively. The remaining three primers sEg00067, sEg00077 and sEg00090 amplified only 28.6%, 71.4% and 42.9% of the samples screened, respectively.

The PCR products of sEg00080, sEg00126 and sEg00127 in *E. guineensis, E. oleifera, C. nucifera* 1, *C. nucifera* 2 and *J. bataua* were cloned and sequenced. Sequences were aligned and compared (figure 4). The sEg00080-amplicon revealed 100.0% similarity between *E. guineensis, E. oleifera, C. nucifera* 1 and *C. nucifera* 2 demonstrating a distinct cluster from *J. bataua* (figure 5a). However, the divergence was small (distance = 0.0272) and this was mostly due to changes in a few sequences in *J. bataua* which included point deletions, single nucleotide polymorphisms (SNPs), short insertions and loss of one (TCA) repeat. Oil palm and coconut, both belong to tribe Cocoeae thus appear to have been well differentiated from the tribe of Areceae (*J. bataua*) in the Palmae family. Similarly, sEg00126 (figure 5b) also showed a high degree of sequence conservation across the palms. The phenogram showed that *E. oleifera* was genetically closer to *E. guineensis* compared to *C. nucifera* 1 and *J. bataua*. Sequence differences were found within and outside the (CGC) repeats (figure 4b). For sEg00127, the amplicon from *C. nucifera* 2 revealed that except for the priming

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

	Palmae									
	Cocoeae			Areceae						
SSR locus	*E. guineensis*	*E. oleifera*	*C. nucifera*	*E. oleracea*	*J. bataua*	*O. multicaulis*	*P. macarthurii*	*C. renda*	*D. album*	*D. dura*
Di-repeats										
sEg00066	192–214	202–210	NC	164–180	166–180	166–180	184–188	204–208	188–190	192–208
sEg00077	164–166	164	NC	170	163	163	156–164	154*	NC	164–166
sEg00090	202–210	220–250	268–270	NC	–	270	NC	170	NC	202
sEg00140	218–220	216	216	210	214	214–216	224	–	232	218
Tri-repeats										
sEg00080	146–154	146–154	154	154	144–152	144–152	154–168	146–158	154–168	146–154
sEg00125	154	156	154–156	154	156	–	148–160	154	148–156	154
sEg00126	214–216	214	216	206	204	204*	206–222	196 - 214*	206	216
sEg00127	160	160	160	148–154	148	148	152	160	154	160
Tetra-repeats										
sEg00032	260–272	260–272	260	228–272	260	260–272	292	260–272	292–310	260–272
sEg00067	152–164	168–174	152-174	NC	–	152 - 174*	NC	NC	NC	152

NC, banding pattern not clear; *further optimization needed because of the presence of unspecific bands; –, missing sample.

site, the entire stretch of the expected sequence was different from what had been amplified in the other samples. The TTC repeat was not observed and it is possible that the primer had an alternative binding site in *C. nucifera* 2, which was amplified and cloned in this study. This also probably explains the phenogram in figure 5c where *C. nucifera* 2, at a distance of 0.56135, is clustered separately from the other samples.

Overall, the sequence data showed that polymorphism had been detected at the SSR sites and the flanking regions, which were also reported by Billotte *et al*. (2001). The polymorphism involved mostly SNPs (except sEg00127) which resulted only in small differences in the PCR product size revealing their potential for cross transferability. The transferability rate of EST-SSRs has been reported to be higher than the genomic SSRs in barley (Castillo *et al*. 2008) and sugarcane (Cordeiro *et al*. 2001). It will be interesting to compare the transferability rate for oil palm EST-SSRs and genomic SSRs using the same set of germplasm materials as reported in this study.

Conclusion

This study made use of an efficient and cost effective method for mining genic SSRs from a large ESTs collection. This study with a relatively large set of EST-SSRs characterizes the genetic variability measures observed in a large set of germplasm from both palm species. The measures such as A_o, P and H_e were improved compared to the preliminary work done by Singh *et al*. (2008). The high PIC value (average = 0.65) indicates that the EST-SSR markers are useful for genetic diversity analysis. The high percentage of unique alleles (78.0%) demonstrated that the EST-SSRs are preferentially conserved among the oil palm species. This study estimated the absence of 45 alleles in the Deli *dura* reflecting bottleneck of the gene pool. Therefore, the native palms provide a good source of interesting genes to improve the agronomic value of the existing planting materials. The genic SSRs can also be used as tools for oil palm genetic resource collection, characterization and conservation, which is of importance as the natural palm grown in Africa and probably in South and Central America are disappearing at an unexpectedly rapid rate due to human disturbance. The EST-SSR markers also proved useful in estimating the natural heterozygosity deficit. Significant heterozygosity deficit has been noticed in the Madagascar and Panama germplasm. Thus, germplasm collections and conservation in these countries are essential to avoid the risk of genetic erosion. The transferability of EST-SSRs provides candidate markers for synteny studies among oil palm, coconut and exotic palms although transferability is somewhat dependent on the individual locus being analysed.

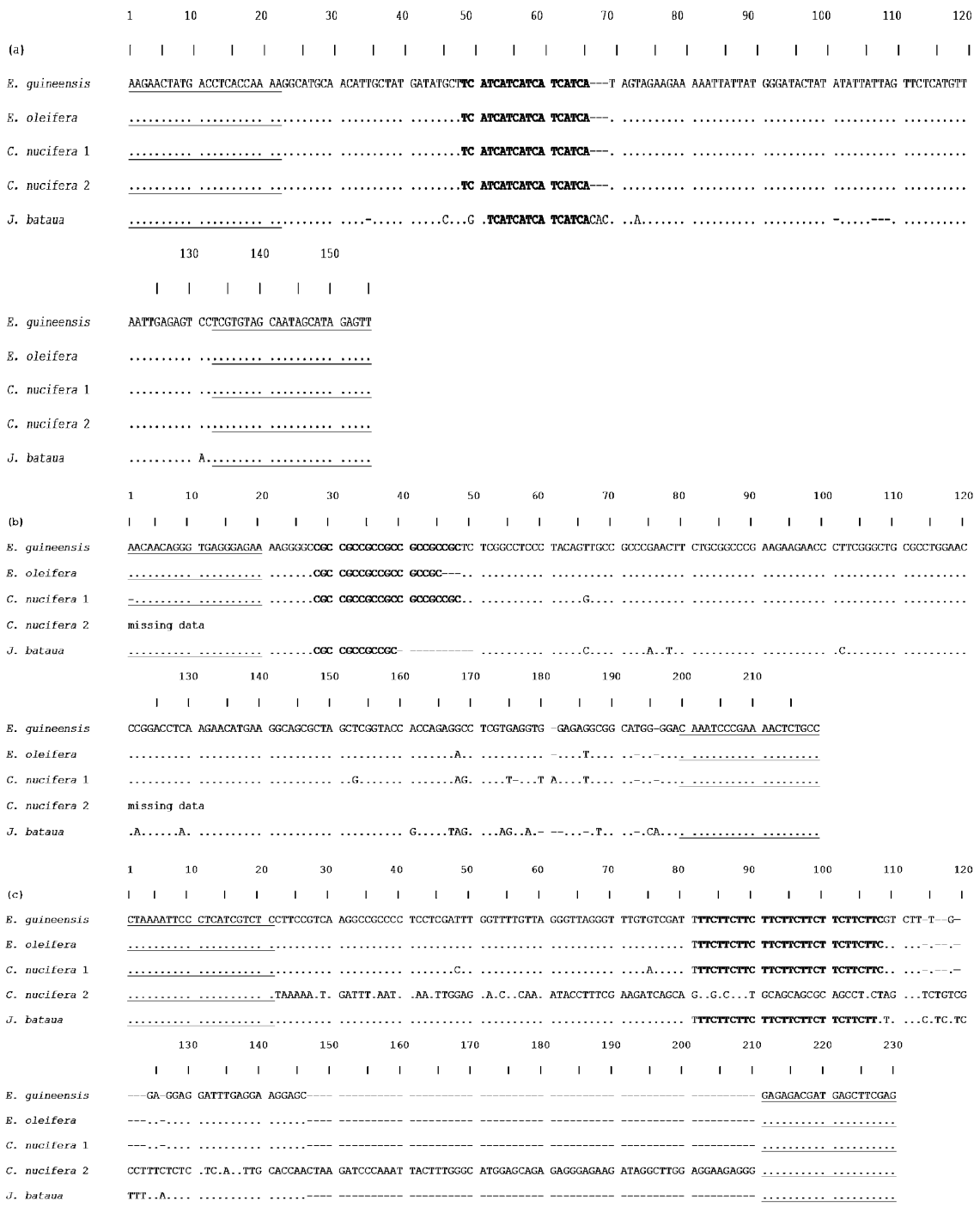

Figure 4. Comparision of PCR amplicons between *E. guineensis* (Cameroon origin), *E. oleifera* (Colombia origin), *C.nucifera 1* (yellow coconut). *C. nucifera 2* (green coconut) and *J. bataua* (exotic palm). Three EST-SSR primers-pair as indicated by underlined letters: (a) sEg00080, (b) sEg00126 and (c) sEg00127 were used to amplify SSRs (bold letters). Alignment to *E. guineensis* shows homology sequences (dots) as well as insertions (normal letters) and deletions (dashes).

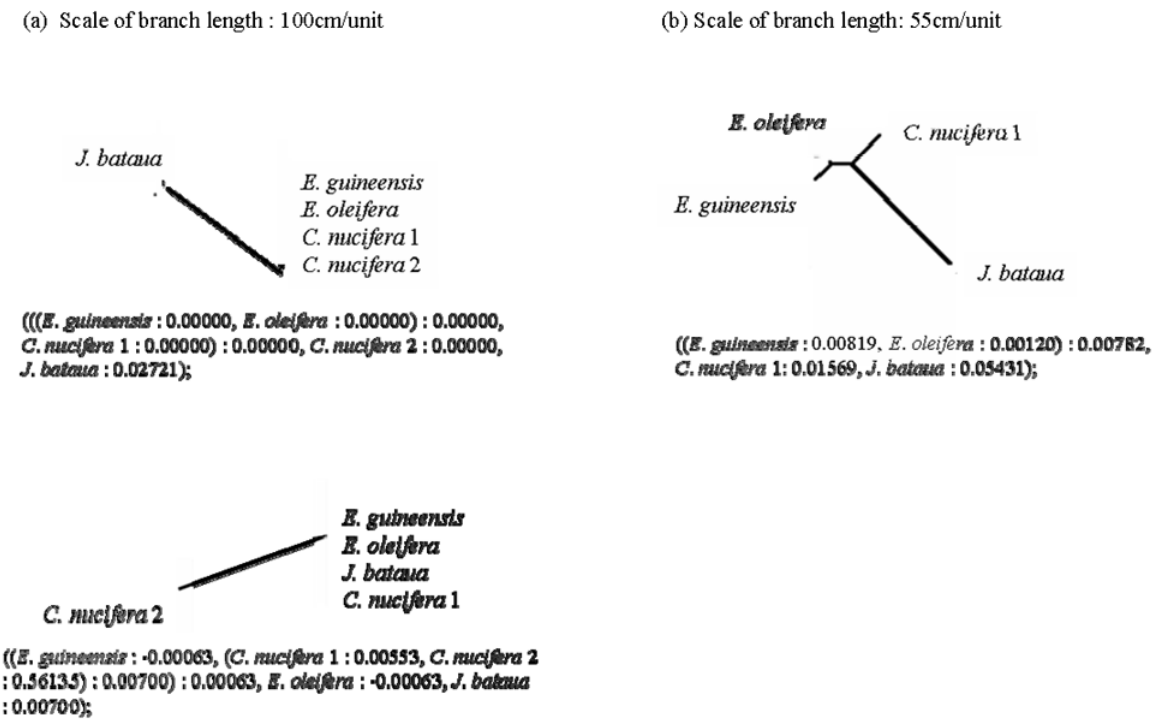

Figure 5. Unrooted phenogram of alleles amplified in *E. guineensis*, *E. oleifera*, *C. nucifera* 1, *C. nucifera* 2 and *J. bataua* observed using EST-SSR primers: (a) sEg00080, (b) sEg00126 and (c) sEg00127. The genetic distances is presented in the PHYLIP format where the tree ends with a semicolon.

Acknowledgements

The authors wish to thank the Director-General of Malaysian Palm Oil Board for giving permission to publish this manuscript. This work was funded by Malaysian Palm Oil Board (MPOB), Project code: BD348-1999.

References

Adin A., Weber J. C., Montes C. S., Vidaurre H., Vosman B. and Smulders M. J. M. 2004 Genetic differentiation and trade among populations of peach palm *(Bactris gasipaes Kunth)* in the Peruvian Amazon - implications for genetic resource management. *Theor. Appl. Genet*. **108**, 1564–1573.

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

Castillo A., Budak H., Varshney R. K., Dorado G., Graner A. and Hernandez P. 2008 Transferability and polymorphism of barley EST-SSR markers used for phylogenetic analysis in *Hordeum chilense. BMC Plant Biol*. **8**, 97–105.

Cho J. C. and Tiedje J. M. 2001 Biogeography and degree of endemicity of fluorescent *Pseudomonas* strains in soil. *Appl. Environ. Microbiol*. **66**, 5448–5456.

Cordeiro G. M., Casu R., McLntyre C. L., Manners J. M. and Henry R. J. 2001 Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. *Plant Sci*. **160**, 1115–1123.

Corley R. H. V. and Tinker P. B. 2003 *The oil palm*, 4th edition. Wiley-Blackwell, Oxford, UK.

Hayati A., Wickneswari R., Maiura 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.

Jorgensen S., Hamrick J. L. and Well P. V. 2002 Regional patterns of genetic diversity in *Pinus Flexilis* (pinaceae) reveal complex species history. *Am. J. Bot*. **89**, 792–800.

Low E. T. L., Halimah A., Boon S. H., Elyana M. S., Tan C. Y., Ooi L. C. L. *et al*. 2008 Oil palm (*Elaeis guineensis* Jacq.) tissue culture ESTs: identifying genes associated with callogenesis and embryogenesis. *BMC Plant Biol*. **8**, 62.

Maizura I., Rajanaidu N., Zakri A. H. and Cheah S. C. 2006 Assesment of genetic diversity in oil palm (*Elaeis guineensis* Jacq.) using restriction fragment length polymorphism (RFLP). *Genet. Res. Crop Evol*. **53**, 187–195.

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

Perera L., Russell J. R., Provan J. and Powell W. 2000 Use of microsatellite DNA markers to investigate the level of genetic diversity and population genetic structure of coconut (*Cocos nucifera* L.). *Genome* **43**, 15–21.

Perera L., Russell J. R., Provan J. and Powell W. 2001 Level and distribution of genetic diversity of coconut (*Cocos nucifera* L., var. *Typica* form *typica*) from Sri Lanka assessed by microsatellite markers. *Euphytica* **122**, 381–389.

Purba A. R., Noyer J. L., Baudouin L., Perrier X., Hamon S. and Lagoda P. J. L. 2000 A new aspect of genetic diversity of Indonesian oil palm (*Elaeis guineensis* Jacq.) revealed by isoenzyme and AFLP markers and its consequences for breeding. *Theor. Appl. Genet*. **101**, 956–961.

Rajanaidu N., Jalani B. S., Kushairi A. and Rao V. 1999 Oil palm genetic resources-collection, evaluation, utilization and conser-

vation. In *Proceeding of the symposium on the science of oil palm breeding* (ed. N. Rajanaidu and B. S. Jalani), pp. 219–255. PORIM, Bangi, Malaysia.

Rival A., Buele T., Barre P., Hamon S., Duval Y. and 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 Rep*. **16**, 884–887.

Saitou N. and Nei M. 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. **4**, 406–425.

Sambanthamurthi R., Singh R., Parvee G. K. A., Ong-Abdullah M. and Kushairi 2009 Opportunities for the oil palm via breeding and biotechnology. In *Breeding plantation tree crops* (ed. S. M. Jain and P. M. Priyadarshan), pp. 377–421. Springer, New York, USA.

Singh R., Noorhariza M. Z., Ting N. C., Rozana R., Tan S. G., Low E. T. L. *et al*. 2008 Exploiting an oil palm EST database for the development of gene-derived and their exploitation for assessment of genetic diversity. *Biologia* **63**, 227–235.

Swofford D. L. and Selander R. B. 1989 BIOSYS-1 A computer program for the analysis of allelic variation in population genetics and biochemical systematics. Release 1.7. Illinois Natural History Survey, Champaign, Illinois, USA.

Tang S. X., Okashah R. A., Cordonnier-Pratt M. -M., Pratt L. H., Johnson V. E., Taylor C. A. *et al*. 2009 EST and EST-SSR marker resources for Iris. *BMC Plant Biol*. **9**, 72.

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

Zehdi S., Trifi M., Billotte N., Merrakchi M. and Pintaud J. C. 2005 Genetic diversity of Tunisian date palms (*Phoenix dactylifera* L.) revealed by nuclear microsatellite polymorphism. *Hereditas* **141**, 278–287.

Zehdi S., Sakka H., Rhouma A., Ould Mohamed Salem A., Marrakchi M. and Trifi M. 2004 Analysis of Tunisian date palm germplasm using simple sequence repeat primers. *Afr. J. Biotechnol*. **3**, 215–219.

Received 24 November 2009, in revised form 20 January 2010; accepted 24 January 2010

Published on the Web: 25 June 2010