Interspecific differences in single nucleotide polymorphisms (SNPs) and indels in expressed sequence tag libraries of oil palm *Elaeis guineensis* and *E. oleifera*

Aikkal Riju[1] and Vadivel Arunachalam[2*]

[1]Aikkal , Kanul, Kannur , Kerala - 670564,India

[2] Central Plantation Crops Research Institute, Kudlu P.O, Kasaragod – 671124, Kerala, India.

ABSTRACT

Oil palm is the second largest source of edible oil, which meets one- fifth of global demands of oils and fats. Expressed sequence tag (EST) sequencing programs have provided a wealth of information, identifying novel genes from a broad range of organisms and providing an indication of gene expression level in particular tissues. It also provides the richest source of biologically useful SNPs due to the relatively high redundancy of gene sequence, the diversity of genotypes represented within databases. EST based SNPs are potential molecular markers and aid in genetic improvement. A total of 21062 and 2053 polymorphic (SNP and Indel) sites in *E. guineensis* species and in *E. oleifera*, 4955 SNPs and 1172 Indels were detected. SNP(17.5/kbp) and Indel(4.1/kbp) frequency was higher in *E. oleifera* than *E. guineensis* species (16.8/kbp , 1.6/kbp). *E. oleifera* showed higher transition to transversion ratio (1.40) than in *E. guineensis* (1.02). The ratio of Ts vs Tv showed, the genetic divergence is occurring in this crops in different fashion and E. *guineensis* had diverged more than E. *oleifera*. We provide the results of the study as online database (http://riju.byethost31.com/oilpalm/) for use by oil palm breeders.

Keywords: Mutation, Nucleotide, Transition, Trans- version, Single Nucleotide Polymorphisms & Indel

INTRODUCTION

The Oil palm a tropical palm tree originated in West Africa but has since been planted successfully in tropical regions within 20 degrees of the equator. Oil palm is the second largest source of edible oil next only to soybean, which contributes approximately 20% of the world's production of oils and fats. Oil is extracted from both the pulp of the fruit (palm oil, edible oil) and the kernel (palm kernel oil, used mainly for soap manufacture). Both palm oil and palm kernel oil are high in olefins, a potentially valuable chemical group that can be processed into many non- food products as well. There are cultivated two species of oil palm, *Elaeis guineensis Jacq.* from tropical western Africa and *E. oleifera* from Americas. African oil palm is widely cultivated and American oil palm is important in terms of compact growth habit and resistance to diseases. Oil palm has a large diploid genome of

3400 MB distributed in 32 chromosomes.

Expressed Sequence Tags or ESTs provide researchers with a quick and inexpensive route for discovering new genes, for obtaining data on gene expression and regulation, and for the construction of the Genome maps. A single nucleotide polymorphism, or SNP (pronounced *snip*), is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome differs between members of a species (or between paired chromosomes in an individual). Single Nucleotide Polymorphisms may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions. SNPs within a coding sequence will not necessarily change the amino acid sequence of the protein that is produced, due to degeneracy of the genetic code. If DNA sequence in which any change in the base pairs do not result in the change of polypeptide sequence, then it is termed synonymous (sometimes called a silent mutation) - if a different polypeptide sequence is produced, they are non-synonymous. SNPs that are not in protein-coding regions may still have consequences for gene splicing, transcription factor binding, or the sequence of non-coding RNA. The study of single nucleotide polymorphisms is also important in crop and livestock breeding programs. Expressed sequence tags (ESTs) are an important resource for identifying polymorphisms in transcribed regions.

Single nucleotide polymorphisms (SNPs) have been shown to be the most abundant source of DNA polymorphism in human, animal and plant genomes. SNPs are the most common type of alleles found within and between varieties of a crop species. Single Nucleotide polymorphisms (SNPs) possess desirable properties as molecular markers. Biallelism makes them easy to score in high throughput genotyping assays. Molecular genetic markers developed from ESTs can be used to examine a group of individuals or populations to estimate various diversity measures and genetic distances, infer genetic structure and clustering patterns, test for Hardy-Weinberg equilibrium and multi-locus equilibrium, and to test polymorphic loci for evidence of selective neutrality. They are useful to plant breeders, germplasm managers, and population geneticists. The use of EST sequence data for the identification of SNPs has many advantages that can be exploited to facilitate the development of highly dense genetic maps and markers assisted breeding programs [Varshney *et. al.*, 2005]. SNPs can be used to saturate genetic maps in plants [Bhattramakki and Rafalski, 2001].

Recently EST resources of oil palm are being developed at IRD, Montpellier France and MPOB, Kuala Lumpur Malaysia (http://palmoils.mpob.gov.my/palm-genes.html) and deposited at dbEST. Expressed sequence tag (EST) sequencing programs have provided a wealth of information, identifying novel genes from a broad range of organisms and providing an indication of gene expression level in particular tissues [Adams *et. al.*, 1995]. EST sequence data may provide the richest source of biologically useful SNPs due to the relatively high redundancy of gene sequence, the diversity of genotypes represented within databases, and the fact that each SNP would be associated with an expressed gene [Picoult-Newberg *et. al.*, 1999]. Re-

cent study of Oil palm EST libraries for SNP detection showed that the genome has the SNP frequency 1.36/100bp [Riju *et. al.,* 2007]. We have used updated EST libraries of Oil palm species for this analysis to find the Interspecific differences in SNP / Indel polymorphisms. SNP detection perl script AutoSNP version.1.0 were used to find the SNP site information and transition vs transversion analysis [Barker *et. al.,* 2003]. EST-SNP can be detected by using other programs or servers such as SEAN [Huntley *et. al.,* 2006], PolyPhred [Nickerson *et. al.,* 1997], PolyBayes [Marth *et. al.,* 1999], TRACE_DIFF [Bonfield *et. al.,*1998], HaploSNPer [Tang *et. al.,* 2008] and HarvEST(http://harvest.ucr.edu) but AutoSNP provides user friendly approach and interpretable result as html file. SNPs can be classified based on their nucleotide substitution as either transition (G↔A or C↔T) or transversion(C↔G, A↔T, C↔A or T↔G). Indel sites can classified to four groups based on the nucleotide involved (A/T/C/G). Thus there are ten kinds of SNP/indel (two types of transition, four types of transversion and four groups of indels) are possible in genomes.

MATERIALS AND METHODS

EST database, dbEST of NCBI release 121407 contains 16,999 *Elaeis guineensis* (GenBank: EL680966 - EL695503 and EE593287 - EE593337) and 3,205 *Elaeis oleifera* (GenBank: EL563704 - ES414798 and BM402088, BM402089 and EB643519 - EB643628) expressed sequence tags. We have used five (root, mature flower, shoot apical meristems, suspension cells and young flower) EST libraries of *Elaeis guineensis* [Jouannic *et. al.,* 2005], [Chai-Ling *et. al.,* 2007] and one EST library (mesocarp) of *Elaeis oleifera*. Zygotic embryo and Lambda Zap libraries hold very few ESTs, so we ignored those libraries for this study. Majority of the sequences represents root, mesocarp tissue, young flower and mature flower. We have used the contig building package Cap3 [Huang and Madan, 1999] server (http://mobyle.pasteur.fr/cgibin/MobylePortal/portal.py#) to cluster the ESTs and made necessary ace files of each tissue. AutoSNP version.1.0 was used to find the candidate SNPs from these contigs. The transition(Ts) vs transversion(Tv) ratio of all the libraries to find the DNA substitution in Oil palm genome was also calculated.

RESULTS AND DISCUSSION

We found 21062 SNPs and 2053 Indels sites in *E. guineensis* (Table 1). *In E. oleifera* 4955 SNPs and 1172 Indels were found (Table 2). Considering those SNP site by nucleotide substitution, , a total of 10635 transitions(Ts), 10427 transversions(Tv) and 2053 indel polymorphisms were detected in *E. guineensis*, where as in *E. oleifera* a total of 2891 transition, 2064 transversion and 1172 indel polymorphisms were detected. SNPs occurred at a frequency of 16.8/1kbp in *E. guineensis* species and 17.5/1kbp in *E. oleifera* species. Indel frequency 1.6/1kbp and 4.1/1kbp observed in *E. guineensis* and *E. oleifera* respectively. While finding the substitution comparison, tissues like mature flower and suspension cell libraries shows higher transversion sites than transition. The Indel polymorphism ratio (4.1/1kbp) found to be high in mesocarp tissue than other tissue libraries observed. The overall transition vs transversion ratio in *E. oleifera* is 1.4, which shows that the relative increase of transition over transversion. Though in case of *E. guineensis* the ratio is 1.02, which shows that the transition as well

as transversion occurring in a similar fashion. Transversion was high in *E. guineensis* and Indels were reported higher in *E. oleifera*. Estimation of the Ts Vs Tv rate is important to understand the pattern of DNA sequence evolution. It has been repeatedly noted that at low level of genetic divergence, ts/tv appears to be high and at high levels of genetic divergence, ts/tv appears to be low. By concluding the transition and transversion bias result, in our case the transversion have occurred approximately similar manner in E. *guineensis* and slightly variable in E. *oleifera*. It may because of the higher level of genetic divergence has occurred so far in E.*guineensis* than *E.olefera*. AFLP and RFLP markers revealed that genetic divergence between the two species is of the same magnitude as that among provenances of *E.olefera* (Barcelos *et. al.*, 2000). The genetic diversity relationship between the two *Elaeis* species by AFLP and RFLP suggest that AFLP markers, is more important than the divergence detected by RFLP markers (Barcelos et al., 2002). EST analysis shows an insight of intrerspecific molecular genetic variations within oil palm. These transitions to transversions ratio showed the molecular evolution happening in this crop with different fashion. Only fewer EST resources were generated on E. *oleifera* species yet. Even though with available source has to believe that the molecular divergence is happening in E. *guineensis* species than E. *oleifera*.

Recent SNP report on beetroot expressed gene showed the SNP frequency as 1 every 130bp [Schneider *et. al.*, 2001] and ESTs [Batley *et.al.*, 2003] of maize also reported relative increase of transition sites over transversion. Germano and Klein (1999) identified 5 SNPs in 1kbp of cDNA of *Picea rubens* and *Picea mariana*, and also discovered SNPs in the chloroplasts of these species. In soyabean(*Glycine max*), 5 SNPs were found approximately every 1kbp [Coryell *et. al.*, 1999],[Van *et. al.*, 2004]. In maize (*Zea mays*), SNPs occur at frequently, with one SNP approximately every 48 bp and every 130 bp in 3' untranslated regions and coding regions, respectively [Tenaillon *et. al.*, 2001], [Rafalski, 2002]. SNPs occur at very low frequency of .07 every 1kbp in apple (*Malus domestica*) ESTs [Newcomb *et. al.*, 2006]. Comparing these crops the SNPs in oil palm is more frequent as observed in maize and agree with earlier study [Riju *et.al.*, 2007] on this crop with fewer EST libraries (SNP frequency as 1.36 SNPs per 100 bp with a higher rate of transition over transversion). *In slico* approach to ESTs revealed the SNPs in oil palm as more abundant hence a valuable tool to find the genetic diversity and breeding programs along with AFLP and RFLP. Our study will help oil palm researchers about the single nucleotide polymorphism and nucleotide substitutions. The tissue wise EST clusters and their SNP and Indel site information is made available at www.riju.byethost31.com/oilpalm.

CONCLUSION

Available est resources at dbest were analysed for the putative snp and indels in oil palm. Snps occurred at a frequency of 16.8/1kbp in *e. Guineensis* species and 17.5/1kbp in *e. Oleifera* species. Indel frequency 1.6/1kbp and 4.1/1kbp observed in *e. Guineensis* and *e. Oleifera* respectively. There are very few ests are available for *e. Oleifera* when comparing with the *e. Guineensis* species. Snp frequency was observed

approximately equal in both species but a higher frequency of indels were observed in *e. Oleifera* species (4.1/1kbp). Transition to transversion ratio was lower (1.02) in *e. Guineensis* species than *e. Oleifera* (1.4). This transition to transversion ratio shows the different level of molecular evolution happening in this crop. Only fewer est resources were generated on e. *Oleifera* species yet. Even though with available source has to believe that the molecular divergence is happening in e. *Guineensis* species than e. *Oleifera*. This *in silico* analysis on oil palm shows the potential snp markers for use in oil palm breeding and the database we created would help to use the information in designing new primers and develop more markers and saturate the linkage maps. The study also highlights the nucleotide substitution analysis in available oil palm EST resources.

REFERENCES

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D. and White ,O (1995). Initial assessment of human gene diversity and expression patterns based upon 83-million nucleotide of cDNA sequence. Nature. **377**: 3.

Barceló's, E., Amblard, P., Berthaud, J. and Seguin, M(2002). Genetic diversity and relationship in American and African oil palm as revealed by RFLP and AFLP molecular markers. Pesq. agropec. bras., Brasília, v. **37**, n. 8, 1105-1114.

Barker, G., Batley, J., O'sullivan, H., Edwards, K.J., and Edwards, D.(2003). Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics*. **19**: 421-422.

Batley, J., Barker, G., Helen, O'Sullivan., Edwards, K. J. and Edwards, D.(2003). Mining for Single Nucleotide Polymorphisms and Insertions/Deletions in Maize Expressed Sequence Tag Data *Plant Physiology*. **132**, 84–91.

Bhattramakki, D. and Rafalski, A(2001). Discovery and application of single nucleotide polymorphism markers in plants. In: Henry R.J. (ed.) Plant Genotyping: the DNA Fingerprinting of plants. CABI Publishing, Oxon, UK, pp. 179-191.

Bonfield, J.K., Rada, C. and Staden, R.(1998). Automated detection of point mutations using fluorescent sequence trace subtraction. Nucleic Acids Res., 26, 3404–3409.

Chai-Ling, H., Yen-Yen,K., Mei-Chooi, C., Sue-Sean, T., Wai-Har, N., Kok-Ang, L., Yang-Ping, L., Siew-Eng, O., Weng-Wah, L., Jin-Ming, T., Siang-Hee, T., Kulaveerasingam, H., Sharifah, S.R.S.A. and Meilina, O.A. (2007). Analysis and functional annotation of expressed sequence tags (ESTs) from multiple tissues of oil palm (Elaeis guineensis Jacq.) *BMC Genomics* 8:381.

Coryell, V.H., Jessen, H., Schupp, J.M., Webb, D. and Keim, P.(1999). Allele-specific hybridisation markers for soybean. Theor Appl Genet. **101**: 1291–1298.

Germano, J. and Klein, A.S. (1999). Species specific nuclear and chloroplast single nucleotide polymorphisms to distinguish *Picea glauca*, *P. mariana* and *P. rubens*. Theor Appl Genet. **99**, 37–49.

Huang, X. and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res*. **9**, 868–877.

Huntley, D., Baldo, A., Johri, S. and Sergot, M.(2006). SEAN: SNP prediction and display program utilizing EST sequence clusters. Bioinformatics. **22(4)**: 495- 496.

Jouannic, S., Argout,X., Lechauve, F., Fizames, C., Borgel, A., Morcillo, F., Aberlenc- Bertossi, F., Duval, Y. and Treger, J.(2005). Analysis of expressed sequences tags from oil palm (Elaeis guineensis) . FEBS . **579** , 2709 – 2714.

Marth, G.T., Korf, I., Yandell, M. D., Yeh, R. T., Zhijie, G., Zakeri,H., Stitziel, N.O., Hillier, L., Kwok,P.Y. and Gish, W. R.(1999). A general approach to single- nucleotide polymorphism discovery.*Nat. Genet*. **23**, 452–456.

Newcomb R. D., Crowhurst, R.N., Gleave, A.P., Rikkerink, E.H.A., Allan, A.C., Beuning, L.L., Bowen, J.H., Gera, E., Jamieson, K.R., Janssen, B.J., Laing, W.A., McArtney,A., Nain, B., Ross, G.S., Snowden, K.C., Souleyre, E.J.F., Walton, E. F. and Yauk, Y.K. (2006) Analyses of Expressed Sequence Tags from Apple. Plant Physiology, **141** , 147–166.

Nickerson, D.A., Tobe, V.O. and Taylor, S. L. (1997). PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence- based resequencing. *Nucleic Acids Res*. 25, 2745–2751.

Picoult- Newberg, L., Idenker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson D.A. and Boyce- Jacino, M. (1999). Mining SNPs from EST databases. Genome Res. **9**: 167- 174.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol. **5**, 94–100.

Riju, A., Chandraseker, A. and Arunachalam, V.(2007). Mining for single nucleotide polymorphisms and insertions / deletions in expressed sequence tag libraries of oil palm. Bioinformation. **2(4)**, 128- 131.

Schneider, K., Weisshaar, B., Borchardt, D.C. and Salamini, F.(2001). SNP frequency and allelic haplotype structure of Beta vulgaris expressed genes. Mol. Breed. 8, 63- 74.

Tang, J.,Leunissen, J.AM., Voorrips, R.E., Linden, C.G. and Vosman, B.(2008). HaploSNPer: a web- based allele and SNP detection tool. BMC Genetics, **9**:23.

Tenaillon, M.I., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F. and Gaut, B.S. (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp mays L.). Proc Natl Acad Sci USA. **98**: 9161–9166.

Van K, Hwang EY, Kim MY, Kim IH, Cho YI, Cregan PB, Lee SH. Discovery of single nucleotide polymorphisms in soybean using primers designed from ESTs. Euphytica. 2004;139:147–157.

Varshney, R. K., Graner, A. and Sorrells, M. E. (2005). Genic microsatellite markers in plants: features and applications. Trends in Biotechnology. **23**: 48- 55.

Table 1. A total of 6332 consensus EST sequence are used to predict the SNP site from *E. guineensis* species, which made 1922 cluster groups and found 23115 SNP-Indel sites.

Tissue Name	No of ESTs	No of contigs	Consensus size (bp)	SNP site	Transitions (Ts)	Transversions (Tv)	Indels	Ts / Tv	Frequency of indels / kbp	Frequency of SNPs / kbp
Root	2230	719	415106	6070	3132	2938	736	1.066	1.8	14.6
Mature flower	1317	376	255349	4736	2320	2416	322	0.960	1.3	18.5
Shoot apex	684	215	141729	3249	1685	1564	268	1.077	1.9	22.9
Suspension cell cultures	539	174	124035	2498	1217	1281	173	0.950	1.4	20.1
Young flower	1562	438	317548	4509	2281	2228	554	1.024	1.7	14.2
Total	6332	1922	1253767	21062	10635	10427	2053	1.02	1.6	16.8

Table 2. A total of 1403 consensus EST sequence are used to predict the SNP site form *E.Oleifera* species, which made 404 cluster groups and found 6127 SNP-Indel sites.

Tissue Name	No of ESTs	No of contigs	Consensus size (bp)	SNP Site	Transitions(Ts)	Transversions (Tv)	Indels	Ts / Tv	Frequency of indels per kbp	Frequency of SNPs per kbp
Mesocarp tissue	1403	404	283721	4955	2891	2064	1172	1.401	4.1	17.5
Total	1403	404	283721	4955	2891	2064	1172	1.401	4.1	17.5