Bioinformation Volume 5 open access
www.bioinformation.net Issue 4 Hypothesis

Insights from the GC content analysis of 76 genome survey sequences (GSS) from *Elaeis oleifera*$^{\Psi}$

Subhash J. Bhore[1, 2, 3*], Amelia Kassim[1, 4], Farida H. Shah[1, 2]

[1]Molecular Biology Division, Melaka Institute of Biotechnology, Lot 7, Melaka International Trade Center City, 75450 Ayer Keroh, Melaka, Malaysia; [2]School of Biosciences and Biotechnology, National University of Malaysia, Selangor, Malaysia; [3]Current Address: Department of Biotechnology, Faculty of Applied Sciences, AIMST University, Bedong-Semeling Road, Bedong, 08100, Kedah, Malaysia; [4]Current Address: Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450, Melaka, Malaysia; Subhash J. Bhore- Email: subhashbhore@gmail.com; Phone: +60-4-429 8176; Fax: +60-4-429 8109; *Corresponding author

Received August 19, 2010; Accepted August 26, 2010; Published September 20, 2010

Abstract:
South American oil-palm (*Elaeis oleifera*) is not cultivated in tropical countries like Malaysia on large scale due to low yield of palm oil derived from its fruit mesocarp. However, its fruit mesocarp oil contains about 68.6 % oleic acid ($C_{18:1}$) which is more than double in comparison to commercially cultivated oil-palm, *E. guineensis* Jacq Tenera (hybrid of Dura (♀) x Pisifera (♂)). It is also known that *E. oleifera* is a good source of tocotrienols and carotenoids. Therefore, it is of interest to know the genome sequence of *E. oleifera*. The objective of this study is to generate genome survey sequences (GSS) to get GC content insight in the *E. oleifera* genome. The nuclear genomic DNA isolated from young leaf-tissues was digested with *Eco*RI and *Nde*I/*Dra*I restriction enzymes; and three genomic DNA libraries were constructed using Lambda *ZAP-II*, pGEM®-T Easy, and pDONR 222^{TM} as cloning vectors. Generated 76 GSSs were analyzed by using Bioinformatics tools. The analysis result indicates that the adenine, cytosine, guanine and thymine content in generated GSSs are 30%, 20%, 20%, and 30% respectively. In conclusion, based on the precise GC content analysis of the randomly isolated 76 GSSs by using Bioinformatics tools we hypothesize that GC content in *E. oleifera* genome is 40%. The hypothesized 40% GC content in *E. oleifera* genome is expected to remain close to the GC content based on the whole genome analysis.

Keywords: American oil-palm, GC content, Genomics, Genome, GSS, *in silico*, palm oil

Abbreviations: gDNA, Nuclear genomic DNA; GSSs, Genome survey sequences; SAOP, South American oil-palm

$^{\Psi}$The nucleotide sequence data reported in this paper have been submitted to dbGSS division of the international DNA database (GenBank/DDBJ/EMBL) under accession numbers: DX575945- DX575972 and EI798032-EI798079.

Background:

Lipids include vegetable oils and fats that are very important in our daily food. The major sources of vegetable oils in the world market of fats and oils is from oil-palm, palm kernel, soybean, cotton, groundnut, sunflower, rapeseed, corn, coconut, olive, castor, sesame and linseed. In year 2008, the world production of fats and vegetable oils was 160,471,000 tonnes. In Asia, Malaysia and Indonesia are the major producer and exporter of vegetable oil (palm oil). In tropical countries, SAOP (*Elaeis oleifera*) is not preferred for large scale cultivation due to low yield of palm oil derived from its fruit mesocarp **[1-2]**. But, its fruit mesocarp oil contains about 68.6 % oleic acid ($C_{18:1}$) which is more than double in comparison to palm oil obtained from commercially cultivated oil-palm, *Elaeis guineensis* Jacq Tenera (hybrid of Dura (♀) x Pisifera (♂)) **[2-3]**. *Elaeis oleifera* (2n=32) is also a good source of tocotrienols and carotenoids. As stated by Singh *et al*. (2009), in South America, preference to *E. oleifera* for cultivation is due to its resistance to bud rot disease **[4]**.

The SAOP is an important edible-oil producing plant, but not much research is done on this plant. As of July 13, 2010, search in all databases available at the National Center for Biotechnology Information (NCBI) shows only 18 publications, 198 core-nucleotides and 3251 expressed sequence tag (ESTs) records (our unpublished data). In this era of modern biotechnology, it is possible to enhance the nutritional value of palm oil and productivity of the oil-palm in short time compared to traditional methods of crop improvement **[2, 3, 5]**. However, traditional breeding remains equally important though the long-term nature of oil-palm breeding mostly impair its improvement.

In perennial plants like oil-palm, genomic science holds promise of their improvements and in the *E. oleifera* the development of genomic tools is crucial for its understanding and exploitation in palm oil yield improvement. In this paper, we report the guanine-cytosine (GC) content in *E. oleifera* genome based on a glimpse by analysis of 76 randomly isolated GSSs for GC content using free online Bioinformatics tools. In double stranded DNA, the

GC pair is bound by three hydrogen bonds, while adenine-thymine (AT) pairs are bound by two hydrogen bonds. The level of GC content is an important attribute of plant (and other living organisms) genomes. The GC content is strictly controlled and moderately balanced across the genome **[6]**. To determine GC content, flow cytometry is most commonly used for large number of samples **[7]**. However, if the coding or non-coding DNA or cDNA molecules under investigation are sequenced then by using the free online Bioinformatics tools (GC calculator) the GC-content can be precisely calculated **[8, 9]**.

Methodology:

Plant material:

Young leaves from the field grown *E. oleifera* at the United Plantation Berhad, Perak, Malaysia were collected and brought to the laboratory. Collected young leaves were washed with plenty of running tap water and then in 70 % ethanol for 5 minute followed by washing with deionized water for 2 minute to avoid surface contamination. Leaves samples were frozen in liquid nitrogen, and stored at -80 °C until required. Two (2) gram leaf-tissue was used to isolate total gDNA for gDNA library construction.

Genomic DNA preparation:

Two (2) gram leaf-tissue was ground into powder with the help of liquid nitrogen using mortar-pestle. The gDNA was isolated from powdered leaf-tissue using a method described by Sambrook *et al.* **[10]** with some minor modifications. Leaves tissue powder was added to 15 ml extraction buffer (0.1 M Tris HCl, 0.05 M EDTA, 0.5 M NaCl, 1% PVP, 1.4% SDS, and 10 mM 2-mercaptoethanol). The suspension was incubated in a water bath at 37 °C for an hour. Potassium acetate was added to each tube and again incubated on ice for an hour. The phases were separated by centrifugation and isopropanol was added to the new suspension. The mixtures were then allowed to precipitate for an hour at –20 °C. After centrifugation, pellets were dried and dissolved in TES buffer. RNase (10 mg/ml) was added to each tube and incubated for an hour at 37 °C. The gDNA was purified by using phenol/chloroform/isoamyl-alcohol extraction method, followed by precipitation with isopropanol and sodium acetate. Finally, DNA pellet was washed with 70 % ethanol twice. The dried pellet was dissolved in 1X TE buffer. The quantitative estimation was carried out using spectrophotometer. DNA samples were stored at –80 °C till its use in DNA library construction.

Genomic DNA library construction and GSS clones isolation:

Three libraries of gDNA were constructed to generate GSSs. The first gDNA library was constructed using lambda *ZAP-II* as the cloning vector. The genomic DNA (10 µg) was digested with *Eco*RI restriction enzyme and generated *Eco*RI fragments were cloned into Lambda *ZAP-II* cloning vector. Cloning of the DNA fragments and packaging was completed following guidelines given by the Lambda *ZAP-II* vector supplier (Promega). Recombinant (white) plaques were selected from plates for the *in vivo* excision. The gDNA inserts were excised to pBluescript vector with the aid of helper phage (Stratagene). The second gDNA library was constructed using pGEM®-T Easy cloning vector (self-ligated and predigested with *Eco*RI) (Promega). Ampicillin resistant recombinant colonies were selected by blue/white colony selection method. Plasmid DNA was isolated and purified using Wizard®Plus SV minipreps DNA purification system (Promega). The third gDNA library was constructed by using gDNA fragments generated by *Nde*I/*Dra*I restriction enzymes and components from the CloneMiner™ cDNA library construction kit (Invitrogen). The quality of all three gDNA libraries was evaluated by random analysis and % of recombinant clones was calculated. The gDNA insert carrying entry clones (pDNAs) were used in transformation of *E. coli* DH5α and Kanamycin resistant *E. coli* colonies were selected, cultivated and GSSs were generated.

DNA Sequencing:

GSS clones obtained from gDNA-Lambda *ZAP-II* Library were sequenced by using T3 (5'-AATTAACCCTCACTAAAGGG-3') and T7 (5'-GTAATACG ACTCACTATAGGGC-3') primers. The GSS clones obtained from gDNA-pGEM®-T Easy cloning vector library were sequenced by using T7 (5'-TAATACGACTCACTATAGGG-3') and SP6 (5'-TATTTAGGTGACAC TATAG-3') primers ; whereas GSS clones obtained from gDNA Entry Library (CloneMiner) were sequenced only from one end using universal primer M13 (Forward) (5'-GTAAAACGACGGCCAG-3').

DNA sequences data analysis:

All GSSs were processed by using VecScreen program **[11]** available at NCBI to identify nucleic acid sequence of cloning vector. After processing and trimming single reads of GSSs, sequences were blasted by using blasn and blastx and putative identities were given to GSSs. The annotated GSSs nucleotide sequence data was submitted to GSS database of GenBank/DDBJ/EMBL. Guanine and cytosine content (GC %) calculation was carried out using 'DNA/RNA base composition calculator', a free online Bioinformatics tool that calculates the molecular mass, elemental composition, base composition, and percent AT and GC content for DNA and RNA sequences [http://www.currentprotocols.com/tools/dnarna-base-composition-calculator].

Results and Discussion:

Elaeis oleifera nuclear genomic DNA was isolated in good quality from young leaf-tissue. The total yield of the isolated intact nuclear genomic DNA was 350 µg from 2 gram leaf tissue. Three libraries of gDNA were successfully constructed by using 10 µg gDNA. The quality analysis of three libraries by evaluation of randomly selected clones suggests that lambda *ZAP-II* library, pGEM®-T Easy library and CloneMiner library was with 80%, 60% and 90% recombinant clones. It suggests that gDNA library constructed using CloneMiner™ DNA library construction kit method is most effective out of three different methods used to construct gDNA libraries to generate GSS in this study (**Figure 1**).

In order to generate GSSs on a small scale, altogether 14, 5 and 39 recombinant clones were isolated randomly from gDNA-lambda *ZAP-II* library, gDNA-pGEM®-T Easy library and gDNA-CloneMiner library respectively. By sequencing each recombinant clone from both sides with T3 and T7 primers, 28 GSSs (GSS#: EoGSS0001-EoGSS0028; accession numbers: DX575945- DX575972) were generated from gDNA-lambda *ZAP-II* library. Altogether, 9 GSSs (GSS#: EoGSS0029-EoGSS0037; accession numbers: EI798032- EI798040) were generated from 5 recombinant clones obtained from gDNA-pGEM®-T Easy library. Whereas, all 39 recombinant clones from CloneMiner library were sequenced only from one side, hence only 39 GSSs (GSS#: EoGSS0038-EoGSS0076; accession numbers: EI798041-EI798079) were generated from 39 recombinant clones. Based on blastn and blastx analysis, putative identities were assigned to GSSs. From isolated 76 GSSs sequences, 13% GSSs were with partial sequences of known genes in it; and 87% GSSs were identified as unknown. The annotated GSSs nucleotide sequence data was finally submitted to GSS database of GenBank/DDBJ/EMBL. The assigned accession numbers to GSSs are DX575945-DX575972 and EI798032-EI798079.

Guanine and cytosine content calculation for all 76 GSSs carried out using a free online Bioinformatics tool **[9]** and data analysis suggested that the rage of the GC content in isolated GSSs was spanning from 22% to 63%. Maximum GSSs (13%) were showing 37% GC content, where as 12% GSSs were showing 41% GC content. The precisely calculated CG content (%) in isolated 76 GSSs is depicted in **Figure 2**. The calculated AT%, GC% and individual nitrogenous base content in % in each of 76GSSs was compiled using Microsoft Excel spread sheet and average % of individual nitrogenous bases, AT, and GC was calculated. The A, T, G, C, AT and GC content analysis in isolated GSSs and their average % calculation suggests that AT content (%) was 60%. This calculation also clearly gives a glimpse into *E. oleifera* genome, indicating that it contains 40%GC. The average content of the individual nitrogenous bases based on analysis of 76 GSSs is shown in **Figure 3**.

The GC content of genome varies in different plant species. For instance, the GC content in *Arabidopsis thaliana* genome is reported 36 % **[12]**. Oil-palm is a perennial plant species. The GC contents reported in some other perennial plants such as Scots Pine (*Pinus sylvestris*) and yellow-poplar (*Liriodendron tulipifera*) is 39.5% and and 41% respectively **[13 & 14]**. The GC content reported by other researchers in some other plants is shown in **Table 1(see supplementary material)**. The reported GC content in Solanaceae species (coding regions) such as *Nicotiana tabacum* (40.4%), *Solanum tuberosum* (39.0%), and *Solanum esculentum* (39.8%), or the Fabaceae *Pisum sativum* (39.2%) **[15]** is also close to the GC content depicted in this study for SAOP. In the recent past (November 2009), Malaysian Palm Oil Board (MPOB) has announced that they have completed sequencing of oil-palm (*E. oleifera* and *E. guineensis*) genomes **[16]**. However, genome sequence and or its annotation information is not available in the public domain.

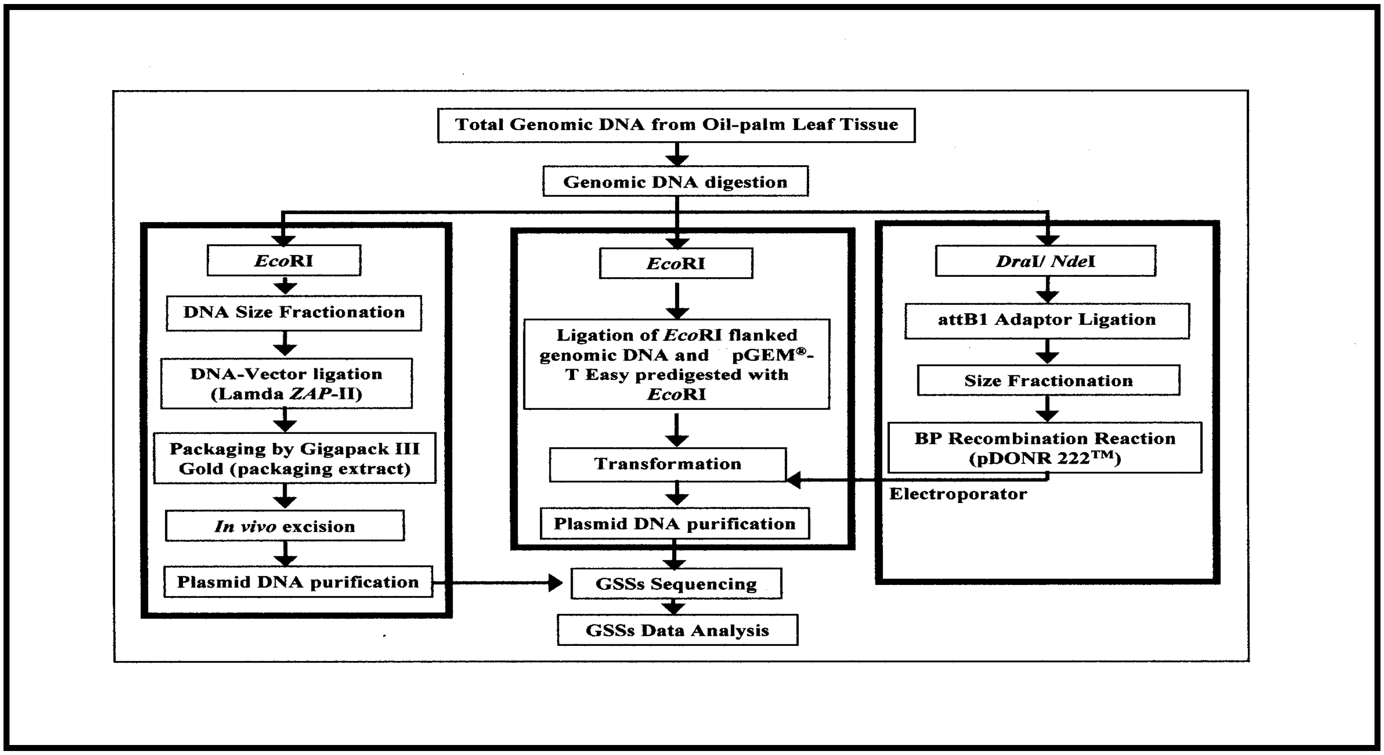
Total Genomic DNA from Oil-palm Leaf Tissue
Genomic DNA digestion
EcoRI
DNA Size Fractionation
DNA-Vector ligation (Lamda ZAP-II)
Packaging by Gigapack III Gold (packaging extract)
In vivo excision
Plasmid DNA purification
EcoRI
Ligation of EcoRI flanked genomic DNA and pGEM®-T Easy predigested with EcoRI
Transformation
Plasmid DNA purification
DraI/ NdeI
attB1 Adaptor Ligation
Size Fractionation
BP Recombination Reaction (pDONR 222™)
Electroporator
GSSs Sequencing
GSSs Data Analysis

Figure 1: Three different methods used in *Elaeis oleifera* nuclear genomic DNA library construction to generate genome survey sequences (GSSs) by random method.

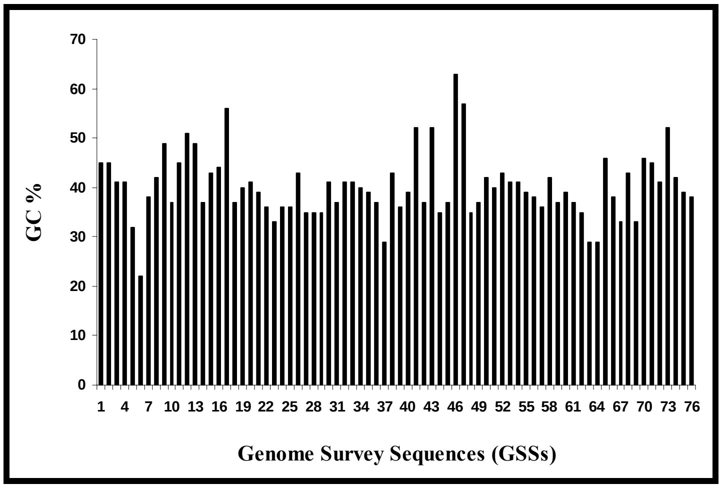

Figure 2: The guanine-cytosine (GC) content (%) in randomly isolated 76 genome survey sequences (GSSs) from *Elaeis oleifera* genome.

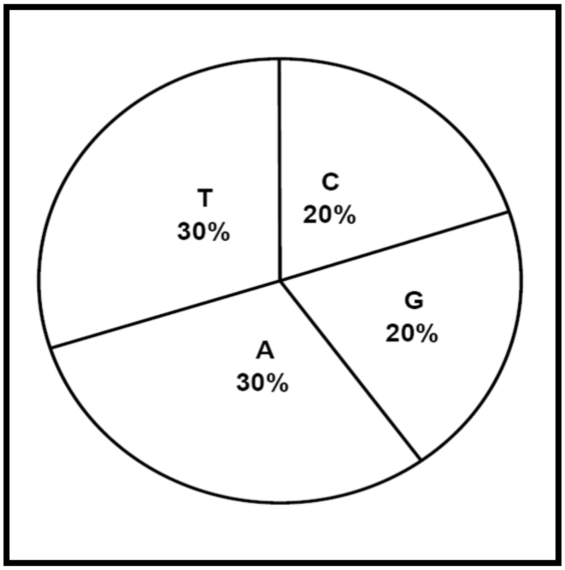

Figure 3: A diagrammatic representation of the adenine (A), cytosine (C), guanine (G) and thymine (T) content (%) in *Elaeis oleifera* genome as depicted by 76 randomly isolated genome survey sequences (GSSs).

Conclusion:

In conclusion, based on the precise GC content analysis of the randomly isolated 76 GSSs by using online Bioinformatics tools we hypothesize that GC content in *E. oleifera* genome is 40%. Despite the consideration that the percentage of genome surveyed by using only 76 GSS in this study which is extremely less in comparison to the whole genome size of *E. oleifera*, the glimpse provided by GC % calculation in these GSS is in range with the GC content reported in other plants. Nonetheless, when the whole *E. oleifera* genome study findings will be available the GC content in genome is expected to be close to the 40%.

Acknowledgements:

Authors are grateful to the Ministry of Science, Technology and Innovation (MOSTI) of Malaysian Government for research funding [Grant Code: IRPA: 01-02-02-0014/PR0015/07-07], to the United Plantation Berhad, Perak, Malaysia for supplying fresh leaf samples of *Elaeis oleifera* for this study, and to Ms. Naqsya S.M. for her assistance in plasmid DNA purification.

References:

[1] J Meunier, D Boulin, *Oleagineux*. **30**:5 (1975)
[2] GK Parveez *et al*. *Biochem Soc Trans*. **28**: 969 (2000) [PMID: 11171275].
[3] GK Parveez, *AgBiotechNet*. **5**: 1 (2003)
[4] R. Singh *et al*. *BMC Plant Biol*. **9**:114 (2009) [PMID: 19706196]
[5] DJ Murphy, *Adv Exp Med Biol*. **464**:21 (1999) [PMID: 10335383]
[6] SC Parker *et al*. *Genome Inform*. **20**:199 (2008) [PMID: 19425134]
[7] AE Vinogradov, *Cytometry* **16**: 34 (1994) [PMID: 7518377]
[8] J Terol, *BMC Genomics* **9**:423 (2008) [PMID: 18801166]
[9] http://www.currentprotocols.com/tools/dnarna-base-composition-calculator
[10] J Sambrook *et al*. *Molecular Cloning: A Laboratory Manual* (1989)
[11] http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html
[12] A Barakat *et al*. *Proc. Natl. Acad. Sci. USA*. **95**: 10044 (1998) [PMID: 9707597]
[13] F Bogunic *et al*. *Plant Cell Reports* **22**:59 (2003) [PMID: 12835994]
[14] H Liang *et al*. *Tree Genetics & Genomes* **3**:215 (2007)
[15] N Carels *et al*. *J Mol Evol*. **6**: 45 (1998) PMID: 9419224]
[16] http://www.palmoilhq.com/PalmOilNews/mpob-achieves-another-first-in-oil-palm-genome-development/
[17] C Blake *et al*. *Genome Res*. **11**: 1660 (2001) [PMID: 11591643]
[18] SM Petr *et al*. *Ann Bot*. **101**: 421 (2008) [PMID: 18158307]
[19] CL Ho *et al*. *BMC Genomics* **8**:381 (2007) [PMID: 17953740]
[20] Q Yuan *et al*. Plant Physiology **138**:18 (2005) [PMID: 15888674]

Edited by P. Kangueane
Citation: Bhore *et al.* Bioinformation 5(4): 141-145 (2010)

Supplementary material:

Table 1: The GC content (%) as depicted in *Elaeis oleifera* GSS and reported GC content in other plants

No	Plant	GC content in Genome	Reference
1	*Elaeis oleifera*	40%	This study
2	*Arabidopsis thaliana*	36 %	**[12]**
3	*Zea mays*	49%	**[17]**
4	*Festuca glauca* Vill	44 %	**[18]**
5	*Liriodendron tulipifera*	41%	**[14]**
6	*Elaeis guineensis* Jacq.	48 % (in exons)	**[19]**
7	*Pinus sylvestris*	39.5%	**[13]**
9	*Oryza sativa*	47 %	**[20]**