ANALYSES USING GENOMIC AND TRANSCRIPTOMIC DATA OF MADS-BOX GENES IN *Elaeis guineensis*

SUHAILA SULAIMAN[1*], LEE YANG PING[1], KWAN YEN YEN[1] and
SHARIFAH SHAHRUL RABIAH SYED ALWEE[1]

*[1]Felda Global Ventures Research and Development Sdn. Bhd., FGV Innovation Centre (Biotechnology),
PT 23417 Lengkuk Teknologi, 71760 Bandar Enstek, Negeri Sembilan, Malaysia
[*]E-mail: suhaila.s@feldaglobal.com*

Accepted 27 January 2017, Published online 31 March 2017

ABSTRACT

MADS-box genes encode a group of transcription factor family, well known as key regulators of plant vegetative and development processes, including flowering. Here, we aimed at analysing MADS-box genes in African oil palm (*Elaeis guineensis*) genome, the pre-dominant source of worldwide production of vegetable oils. A total of 209 potential MADS-box genes are identified in the published 1.8 Gb *E. guineensis* draft genome. A *de novo* assembly of RNA-seq data from inflorescence tissues was constructed using Trinity software. The analysed transcriptome data has validated 36 full-length genes, inclusive of seven transcripts that were previously annotated to encode unknown proteins. Of the 36 genes discovered, 21 genes were characterised as Type I MADS-box genes and phylogenetic analysis using maximum likelihood approach further classified them into three sub-groups of Mα, Mβ and Mγ. Based on *in silico* analyses, we have successfully identified one gene annotated as unknown protein to contain a domain of "MADS-box transcription factor". Its differential expression data, comparison between normal and mantled inflorescence of oil palm, suggested the involvement of the gene in the mantling related process. The findings contribute to updated oil palm genome information which may potentially lead to future understanding of the association of MADS-box genes in mantled oil palm. This may also lead to a resource for biomarker discovery.

Key words: Genomic, transcriptomic, MADS-box, *Elaeis guineensis*

INTRODUCTION

Malaysia is the second largest producer of oil palm after Indonesia, which contributes to one of the country's main economies. Various research and developments efforts are placed to mitigate new challenges and to improve the production of palm oil. To achieve sustainable production of palm oil yield, an important aspect of focus is flowering.

MADS-box genes are well known factor regulating flowering processes (Gramzow & Theißen, 2013). The name of the MADS-box gene family was derived from the first letter of four subsequently founding members: *MINICHROMOSOMAL MAINTENANCE 1*, *MCM1* (*Saccharomyces cerevisiae*) (Passmore *et al.,* 1988), *AGAMOUS*, *AG* (*Arabidopsis thaliana*) (Yanofsky *et al.,* 1990), *DEFICIENS, DEF* (*Antirrhinum majus*) (Sommer *et al.,* 1990) and *SERUM RESPONSE FACTOR, SRF* (*Homo sapiens*) (Norman *et al.,* 1988). MADS-box genes can be classified into two main lineages, namely Type I and Type II genes, which are distinguished by the amino acid consensus sequences in their MADS-box domain. MIKC-type (Type II) was extensively studied in plants and contains four domains namely MADS (M) domain, intervening (I) domain, coiled-coil keratin-like (K) domain and C-terminal (C) domain (Díaz-Riquelme *et al.,* 2009; Gramzow & Theißen, 2013). As for Type I MADS-box genes, they are recognised based on their high similarity with the MADS domain of SRF and are still poorly characterised. Therefore, it is important to further investigate the MADS-box genes in African oil palm (*Elaeis guineensis*) to understand more on their functions in flowering process.

[*] To whom correspondence should be addressed.

MATERIALS AND METHODS

Genomic and transcriptomic data

A draft of oil palm genome with 1.8 Gb size and 16 chromosomes was used (Singh *et al*., 2013). A total of 30,752 gene models were predicted in the second version of the gene annotation. Besides, 18 libraries of RNA-seq samples (single end reads) from inflorescence tissues were sequenced using Illumina technology. Assembly was done using *de novo* approaches by performing Trinity (Grabherr *et al*., 2011) and Cufflinks package (Trapnell *et al*., 2012). Differential expression analyses were done using a Cufflinks program named Cuffdiff.

Identification of MADS-box genes

MADS-box gene sequences were obtained from previous study (Gramzow & Theißen, 2013) and reformatted as a BLAST database using *formatdb* program in BLAST tool. Oil palm gene models were compared with MADS-box genes database using *tblastn* program with initial restriction to E-value 1e^{-5} and percentage identity at least 30%. A PERL script was used to parse the BLAST output according to desired parameters. The gene ontology of the genes was obtained from Gene Ontology (GO) database.

Validation of MADS-box gene candidates using RNA-seq data

The assembled RNA-seq data was mapped against MADS-box gene candidates from oil palm genome. A full-length gene is denoted by 100% coverage between subject match and query sequences. The data handling was performed using UNIX command.

Phylogenetic analysis

A collection of Type I MADS-box genes from *Arabidopsis thaliana* were obtained from previous study by Gramzow and Theißen (2013). Multiple sequence alignment of *A. thaliana* and selected oil palm Type I MADS-box genes were performed with ClustalX. Phylogenetic analysis was carried out using PHYLIP package to produce a Maximum Likelihood (ML) tree by invoking the *proml* programme with the Jones-Taylor-Thornton substitution model. The robustness of the tree was evaluated by bootstrap analysis of 1,000 random replicates using *seqboot*, while *consense* was used to generate the consensus tree. Subsequently, MEGA4 program was used to view and edit the generated phylogenetic trees.

RESULTS AND DISCUSSION

The genome mining was conducted by similarity search of oil palm gene models against a database of 2,062 MADS-box genes with a cut off E-value of 10^{-5}. After filtering the coverage of subject match and query sequences to be at least 30% identical, a total of 209 gene models were identified as potential MADS-box genes. The same classification in Gramzow and Theißen (2013) was used in this study to classify the potential MADS-box genes in oil palm (Figure 1). There were 88 and 68 genes identified as Type I and Type II MADS-box genes, respectively, while seven genes were unassigned and 46 genes were ambiguous as they were similar to both types of MADS-box genes. There are some cases of previously unassigned genes that were successfully assigned to either Type I or Type II class in this study, suggesting an update of genes annotation in oil palm. Albeit more Type I MADS-box genes were identified, they are known to be poorly characterised as compared to Type II.

The mapping of 209 genes to *de novo* assembled transcripts using *tblastn* program exhibited validation of 36 full length genes, while a lower coverage of gene and transcript for at least 90% has validated an additional of 52 genes. Looking in depth into these 36 full length validated genes, seven of them were annotated to encode

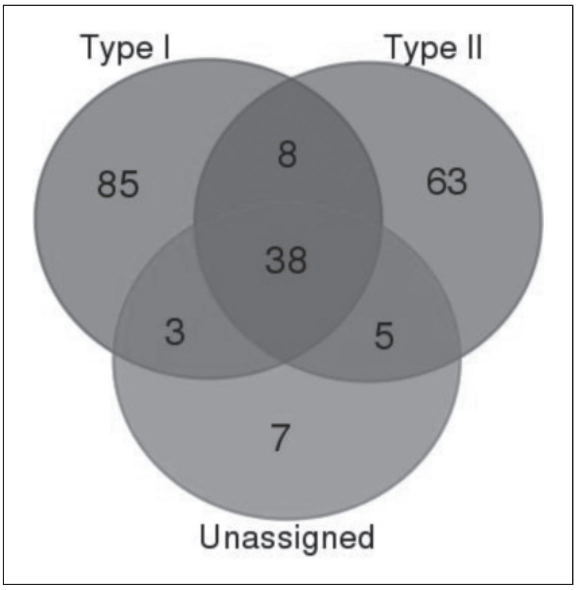

Fig. 1. Classification of MADS-box genes in oil palm. The study has successfully classified eight previously unassigned genes into Type I (three genes) and Type II (five genes).

unknown protein by Singh *et al* (2013). They were either annotated as uncharacterised protein (4 genes), predicted protein (2 genes) or conserved hypothetical protein (1 gene).

Next, further analysis of domain using Interproscan showed that one of the genes (p5_sc00148.V1.gene100) consists of a domain named 'Transcription factor, MADS-box' (IPR002100) derived from SRF-like family, inferring that they are Type I MADS-box genes, a subfamily of MADS-box genes that still remained largely unexplored. It was reported that the MADS-box has been shown to be involved in DNA-binding and dimerization (Glover, 2014). Interestingly, DNA binding (GO:0003677) and protein dimerization activity (GO:0046983) were associated with this domain. The result might give a clue on the function of MADS-box genes as flower development-related genes.

A set of 35 Type I MADS-box genes from *A. thaliana* (Gramzow & Theißen, 2013) were further classified into three sub-groups namely Mα (20), Mβ (4) and Mγ (11) that were clustered based on their phylogenetic relationships between MADS-box genes (Parenicová, 2003). A distinct feature that distinguishes Type I MADS-box genes from Type II is shorter gene size and encoded by single exon. In contrast, Type II MADS-box genes are longer and consist of multiple exons (Glover, 2014).

As the characterisation of Type I MADS-box genes is still lagging behind Type II MADS-box genes, an attempt to classify potential Type I MADS-box genes in oil palm was conducted. Therefore, Maximum Likelihood (ML) phylogenetic tree was generated using PHYLIP package to investigate the relationship of Type I MADS-box genes identified in oil palm and *A. thaliana* (Figure 2). In this study, 21 oil palm MADS-box genes were selected based on their full length validation with transcripts data. Therefore, a more reliable gene selection was subjected to subsequent analysis.

Based on ML tree, three major clades were generated, distinguishing the sub-groups of Type I MADS-box genes. Seven oil palm genes were grouped in the first clade displayed in blue, jointly Mα MADS-box genes from *A. thaliana*. Three genes (formed a cluster with) shown to closer to *A. thaliana* Mα genes (p5_sc00152.V1.gene163, p5_sc00031.V1.gene399 and p5_sc00019.V1.gene438), while five genes were shown to be slightly distant as they formed their own sub-group in the clade (p5_sc00043.V1.gene537, p5_sc00004.V1.gene473, p5_sc00033.V1.gene404, p5_sc00220.V1.gene138 and p5_sc02524.V1.gene2).

Besides that, four oil palm genes (p5_sc00675.V1.gene5, p5_sc00116.V1.gene83, p5_sc00009.V1.gene307 and p5_sc00004.V1.gene535) were clustered in the second clade coloured in blue, along with *A. thaliana* Mβ MADS-box genes, suggesting that these genes were Mβ Type I MADS-box genes. While the rest nine oil palm genes (p5_sc00205.V1.gene94, p5_sc00189.V1.gene102, p5_sc00148.V1.gene100, p5_sc00147.V1.gene64, p5_sc00086.V1.gene299, p5_sc00060.V1.gene345,

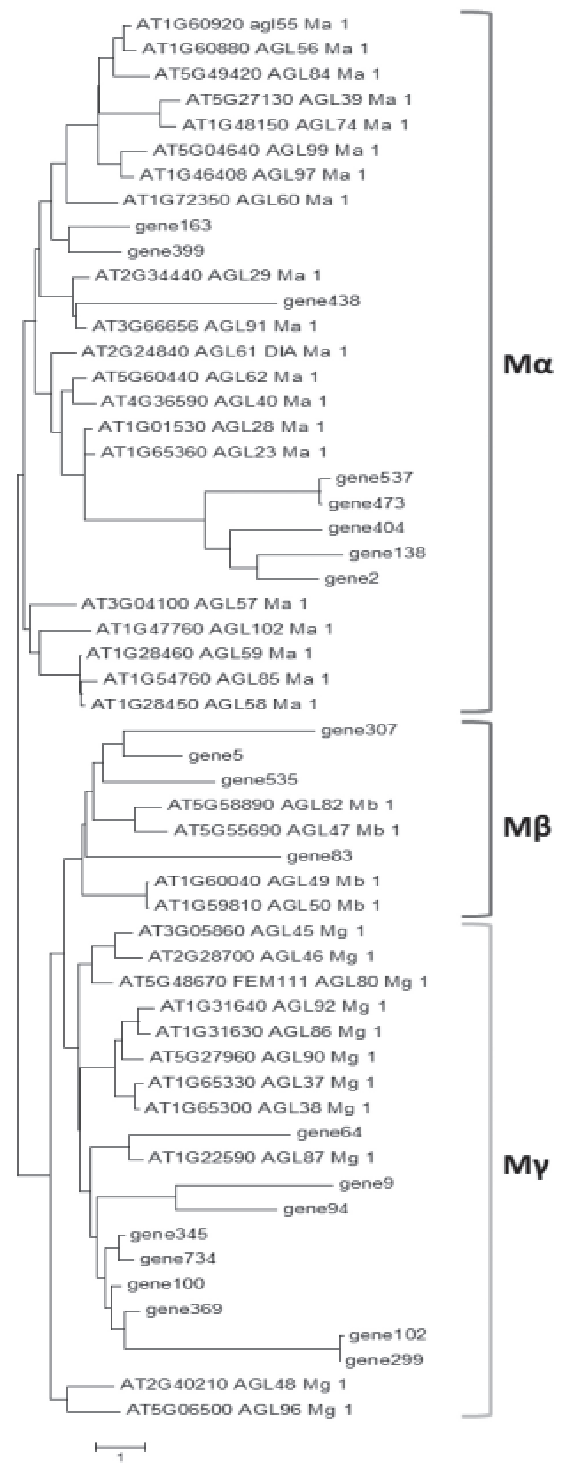

Fig. 2. Maximum Likelihood (ML) tree of Type I MADS-box genes in oil palm.

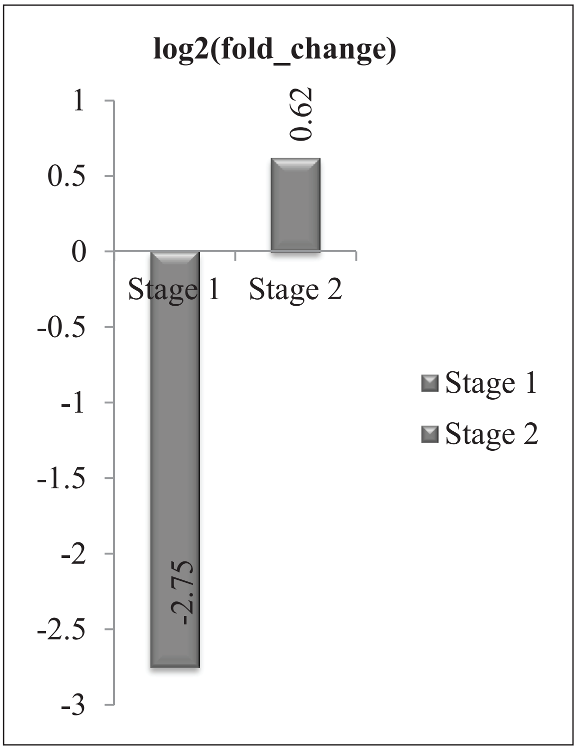

Fig. 3. Expression of p5_sc00148.V1.gene100 gene in stage 1 and 2 of female inflorescence tissues.

p5_sc00042.V1.gene9, p5_sc00025.V1.gene369 and p5_sc00019.V1.gene734) congregated in the third clade in green with the member of Mã MADS-box genes from *A. thaliana.*

The p5_sc00148.V1.gene100 that has been previously described to possess domain of 'Transcription factor, MADS-box' was clustered in the third clade, suggesting that it is Mγ MADS-box gene. On top of that, the significant gene expression analysis showed that this p5_sc00148.V1.gene100 gene was down-regulated (with log2 fold change of -2.74797) in mantled oil palm from stage 1 female inflorescence tissue. The fact that p5_sc00148.V1.gene100 gene is not differentially expressed when we compare at the stage 2 suggests that p5_sc00148.V1.gene100 might function in the early stage of flowering, such as during gamete and seed development (Figure 3).

CONCLUSION

In this study, genomic and transcriptomic data was integrated to improve the information from the published oil palm genome. We have successfully sequenced the transcriptome of oil palm inflorescence tissue using Illumina technology. Using various bioinformatics analyses, 36 full length potential MADS-box genes were validated by RNA-seq data, in which 21 of them belongs to Type I MADS-box genes. Further phylogenetic analysis revealed the classification of the Type I MADS-box genes in oil palm into three categories: Mα (8), Mβ (4) and Mγ (9). By all of 36 validated MADS-box genes, seven were formerly annotated to encode unknown protein.

The discovery of a MADS-box transcription factor domain in one of the previously unannotated genes (p5_sc00148.V1.gene100) indicating an additional update to the existing oil palm genome draft annotation. Moreover, the down-regulation of this gene at the early stage of flowering process may suggest the low expression of p5_sc00148.V1.gene100 gene is possibly one of the factors that trigger mantling. Therefore, this study suggests a comprehensive study on this pathway should be conducted in the future to eradicate mantling in the field.

REFERENCES

Díaz-Riquelme, J., Lijavetzky, D., Martínez-Zapater, J.M. & Carmona, M.J. 2009. Genome-Wide Analysis of MIKC^{C}-Type MADS Box Genes in Grapevine. *Plant Physiology*, **149(1)**: 354–369.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Ray Chowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. & Regev, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29(7)**: 644–652.

Gramzow, L. & Theißen, G. 2013. Phylogenomics of MADS-Box Genes in Plants - Two Opposing Life Styles in One Gene Family. *Biology*, **2(3)**: 1150–1164.

Glover, B. 2014. *Understanding flowers and flowering*. 2nd ed. Oxford University Press, Oxford. 292 pp.

Norman, C., Runswick, M., Pollock, R. & Treisman, R. 1988. Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the c-fos serum response element. *Cell,* **55(6)**: 989–1003.

Parenicová, L. 2003. Molecular and Phylogenetic Analyses of the Complete MADS-Box Transcription Factor Family in *Arabidopsis*: New Openings to the MADS World. *The Plant Cell Online*, **15(7)**: 1538–1551.

Passmore, S., Maine, G.T., Elble, R., Christ, C. & Tye, B.K. 1988. *Saccharomyces cerevisiae* protein involved in plasmid maintenance is necessary for mating of MAT alpha cells. *Journal of Molecular Biology*, **204(3)**: 593–606.

Singh, R., Ong-Abdullah, M., Low, E.L., Manaf, M.A.A., Rosli, R., Nookiah, R., Ooi, L.C., Ooi, S., Chan, K., Halim, M.A., Azizi, N., Nagappan, J., Bacher, B., Lakey, N., Smith, S.W., He, D., Hogan, M., Budiman, M.A., Lee, E.K., DeSalle, R., Kudrna, D., Goicoechea, J.L., Wing, R.A., Wilson, R.K., Fulton, R.S., Ordway, J.M., Martienssen, R.A. & Sambanthamurthi, R. 2013. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature,* **500(7462)**: 335–359.

Sommer, H., Beltran, J.P., Huijser, P., Pape, H., Lonnig, W.E., Saedler, H. & Schwarz-Sommer, Z. 1990. Deficiency, a homeotic gene involved in the control of flower morphogenesis in *Antirrhinum majus*: the protein shows homology to transcription factors. *The EMBO Journal*, **9(3)**: 605–613.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. & Pachter, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with Top Hat and Cufflinks. *Nature Protocols,* **7(3)**: 562–578.

Yanofsky, M.F., Ma, H., Bowman, J.L., Drews, G.N., Feldmann, K.A. & Meyerowitz, E.M. 1990. The protein encoded by the *Arabidopsis* homeotic gene agamous resembles transcription factors. *Nature*, **346(6279)**: 35–39.