Hindawi Publishing Corporation
Dataset Papers in Biology
Volume 2013, Article ID 670926, 7 pages
http://dx.doi.org/10.7167/2013/670926

Dataset Paper

Transcriptome Assembly and Expression Data from Normal and Mantled Oil Palm Fruit

Jeremy R. Shearman,[1] Chatchawan Jantasuriyarat,[2] Duangjai Sangsrakru,[1] Thippawan Yoocha,[1] Apichart Vannavichit,[3] Sithichoke Tangphatsornruang,[1] and Somvong Tragoonrung[1]

[1] *Genome Institute, National Center for Genetic Engineering and Biotechnology, 113 Phahonyothin Road, Klong 1, Klong Luang, Pathumthani 12120, Thailand*
[2] *Department of Genetics, Kasetsart University, 50 Phahonyothin Road, Chatuchak, Bangkok 10900, Thailand*
[3] *Department of Agronomy, Kasetsart University, Kamphaeng Saen, Nakhon Pathom 73140, Thailand*

Correspondence should be addressed to Sithichoke Tangphatsornruang; sithichoke.tan@biotec.or.th

Received 27 March 2012; Accepted 18 April 2012

Academic Editors: F. Coppedè and T. Yin

We performed RNA sequencing of fruit from three normal and three mantled (somaclonal variant affecting flower development) oil palm plants using a 454 pyrosequencer. The three normal fruit samples were combined and sequenced, generating 237 748 reads. The three mantled fruit samples were combined and sequenced giving 231 438 reads. The reads were assembled into 13 984 sequences that were clustered into 10 218 genes or gene families. This paper describes the generation of this transcriptome database and includes annotation of these genes from Blast2GO and blast results against the Arabidopsis protein database as well as identification of putative transcription factors. In addition to this, the expression values for each gene sequence of the normal samples are presented. This dataset will be of use to anyone working in oil palm genetics.

1. Introduction

Somatic embryogenesis is a desirable way of producing new oil palm plants for oil production [1, 2]. Somatic embryogenesis involves harvesting cells from a donor plant and growth of those cells in media that induces cell division to form a cell mass known as callus which can generate new plants. A somaclonal variant that affects flower development, known as mantled, is observed in oil palm derived from somatic embryogenesis in approximately 5% of plants [3–5] (for recent review see [6]). Currently the only way to identify a mantled oil palm is to wait until it reaches sexual maturity two to four years after planting. Although no data has been published with oil yields, farmers report that mantled fruits have a poor oil yield and no seed is produced.

Genomewide hypomethylation has been found in oil palm with the mantled phenotype [7] and no gross genetic abnormalities have been found [8, 9]. The current hypothesis is that the phenotype is caused by epigenetic modification, likely occurring as result of the somatic embryogenesis procedure. The mantled phenotype is expected to change the expression pattern of a number of genes responsible for early flower and fruit development.

A promising method that has recently become mainstream is high-throughput RNA sequencing to identify differential expression [10]. In organisms with a published genome, sequence reads are mapped against the genome and gene expression is given as the number of reads that map to each exon [11]. Differential expression in organisms without a genome is done by first generating a de novo assembly and then calculating read counts based on that. This makes RNA sequencing desirable for organisms that do not have an available reference genome, such as the oil palm, because the data can also be used to generate a transcriptome database. The process can be likened to large-scale qPCR with the added benefit of generating a transcriptome library [10].

With the aim to generate a transcriptome library and search for differentially expressed genes between normal and mantled oil palm fruit, we performed transcriptome sequencing on fruit samples aged 90 days after pollination (dap) from three age-matched normal and three age-matched mantled plants derived from the same clonal batch and planted in the same field on the same day.

2. Methodology

RNA extraction for 454 sequencing was as follows. We collected fruit samples at the age of 90 dap for 454 sequencing from three normal and three mantled Dura × Pisifera cross-oil palm plants. The plants were around eight years old and were planted at the same time, from the same tissue culture batch, and in the same field on the same day in Krabi, Thailand. The cell line from which these plants were derived was originally imported as a culture of Malaysian clone D and has undergone several rounds of somatic embryogenesis and planting in Thailand. We froze the fruit samples in liquid nitrogen and stored them at −80°C until RNA extraction, which we performed using the standard CTAB method. We processed total RNA using absolutely mRNA kits (Stratagene) to isolate polyadenylated mRNA and then prepared the purified mRNA using the standard Roche 454 mRNA protocol (version 2008).

For RNA sequencing and annotation, we formed two RNA groups by mixing RNA from fruit of three normal plants and RNA from fruit of three mantled plants. We sequenced each sample group on 1/4 of a picotitre plate using 454 titanium chemistry and following the standard protocol (Roche). We assembled the resulting RNA-sequence data from each lane using the cDNA option in Newbler de novo assembly software version 2.3 (Roche). We performed an assembly combining the two sample groups to obtain a sequence set representative of all samples.

We annotated the full set of isotigs (contig combinations representing full mRNAs) from the combined assembly using Blast2GO against the plant nonredundant database [12]. We also performed a blastx against the TAIR Arabidopsis protein database using a stringent E-value cut-off of 1E-30. We annotated isogroups (groups of isotigs composed from the same subset of contigs representing genes or gene families) with more than one isotig using the most common isotig annotation within that group. Gene ontology distribution was assigned using the Blast2GO function. We reduced isogroups with more than one isotig to a single isotig representative of the group for GO analysis to avoid overrepresentation of isogroups with multiple isotigs. We identified transcription factors based on GO terms and by performing a blastx against the amino acid sequence of all transcription factors in the plant transcription factor database [13] using a stringent E-value cut-off of 1E-30 and a minimum alignment of 90 amino acids.

Isotigs that did not match any sequence in the database were screened against known plant miRNAs using the program C-mii (D. Wichadakul, unpublished program), which identifies miRNA sequence homology and then checks if flanking sequence is capable of forming the required hairpin structure.

We calculated the gene expression of each isotig by summing the read counts per contig from the 454ReadStatus file (output file of assembly software Newbler), which lists the alignment of the 3′ and 5′ ends of each read to the contig that it formed. Thus each read was counted twice, which takes advantage of the long read lengths obtained by the 454 platform. We obtained the read counts per isotig and isogroup by summing the reads from constituent contigs and used this to identify differential expression using the DESeq package in R [14].

RNA sequencing and sequence annotation were as follows. Total RNA-sequence output was 237 748 reads totaling 82.47 Mb for the three combined normal samples and 231 438 reads totaling 80.15 Mb for the three combined mantled samples. Assembling each sample group separately produced 9799 contigs that formed 8481 isotigs (representative of mRNAs) from 6566 isogroups (representative of genes or gene families) for the normal samples and 9929 contigs that formed 8566 isotigs from 6467 isogroups and for the mantled samples. Since oil palm does not have a reference genome available, we performed an assembly using both sample pools to produce a set of isotigs representative of all samples. The combined assembly produced 16 452 contigs and 13 984 isotigs from 10 218 isogroups and is listed in (Dataset Items 1–7 (Tables)) to be used with the transcriptome assembly. The increase in isotig and contig number of 1.5x despite a doubling of sequence reads indicates that the transcriptional profiles of the samples were similar. The increase most likely represents singleton reads from each sample overlapping to produce a contig. However, it also indicates that the sequencing depth was not high enough to cover all transcripts. Combining the mantled and normal samples for assembly was suitable in this case since all palms were from the same clone and thus should be genetically identical.

Annotation using Blast2GO against the entire GenBank plant nonredundant database assigned a description to 90.5% of the isotigs (Dataset Items 1–7 (Tables)). The GO term distribution for the molecular function category was similar between the two sample sets, as expected considering that all samples were from fruit of genetically identical palms (Table 1). A total of 1036 isotigs (from 954 isogroups) were identified as transcription factors (Dataset Items 1–7 (Tables)). From the oil palm fruit transcription factors published by Tranbarger et al. [15], 23 were not found among the isotigs, suggesting that their expression was too low to detect in fruit samples aged 90 dap. The expression of the transcription factors identified here appears to be consistent with what Tranbarger et al. [15] found.

Three miRNAs were identified: miR-164, miR-396, and miR-398 (Table 2). The hairpin structure of each miRNA is shown in Figures 1, 2, and 3. They contain the microRNA identification work that was performed on isotigs that did not match any proteins in the Blast2GO search. Five isotigs matched to a microRNA from three isogroups. The minimum free energy (MFE) is listed and the expression of each isotig is given as raw read count. They also show the precursor miRNA

Table 1: Comparison of molecular function gene ontology term distribution between mantled and normal oil palm fruit samples.

GO Term	Normal Number of Sequences*	Mantled Number of Sequences*
Molecular function	5732	5763
Binding	3905	3942
Catalytic activity	3365	3368
Nucleotide binding	1280	1289
Hydrolase activity	1103	1109
Protein binding	995	1001
Transferase activity	1256	1265
Nucleic acid binding	1017	1036
Kinase activity	563	571
DNA binding	441	448
Transporter activity	467	463
Transferase activity, transferring phosphorus-containing groups	563	571
Structural molecule activity	305	308
RNA binding	241	250
Transcription regulator activity	258	258
Signal transducer activity	187	191
Transcription factor activity	147	148
Receptor activity	121	123
Translation factor activity, nucleic acid binding	109	108
Molecular transducer activity	187	191
Enzyme regulator activity	67	68
Carbohydrate binding	52	52
Nuclease activity	49	49
Lipid binding	42	42
Hydrolase activity, acting on ester bonds	49	49
Motor activity	19	18
Nucleoside-triphosphatase activity	19	18
Chromatin binding	11	11
Pyrophosphatase activity	19	18
Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	19	18
Hydrolase activity, acting on acid anhydrides	19	18
Oxygen binding	2	2
Translation regulator activity	1	1
Receptor binding	1	1

* Number of sequencees to which the GO term was assigned.

Table 2: miRNAs identified from isotigs that did not match a protein.

Isotig	Isogroup	Best Hit	MFE (kcal/mol)
isotig03681	isogroup01067	ath-miR164a	−35.7
isotig03073	isogroup00770	ath-miR396b	−38.7
isotig03074	isogroup00770	ath-miR396b	−38.7
isotig03075	isogroup00770	ath-miR396b	−38.7
isotig12595	isogroup08829	ath-miR398a	−35.8

sequence and fold structure, and mature miRNA is indicated by the use of capital letters.

The gene expression of each isotig was counted using the raw reads and is listed for the normal samples in Dataset Items 1–7 (Tables). The expression is listed as number of raw reads and also as the number of raw reads per 1 kb of sequence to allow direct comparison between isotigs of different lengths. It is possible for an isotig to have a read count of zero because the assembly includes the mantled samples.

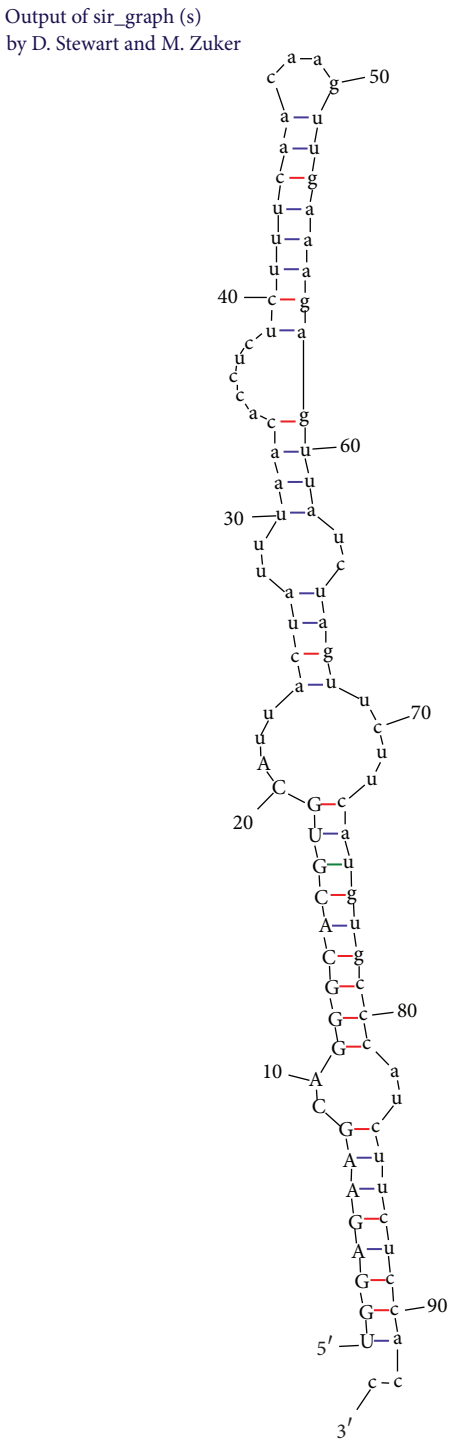

Figure 1: Precursor ath-miR164a for isotig03681. Mature sequence is indicated by capital letters.

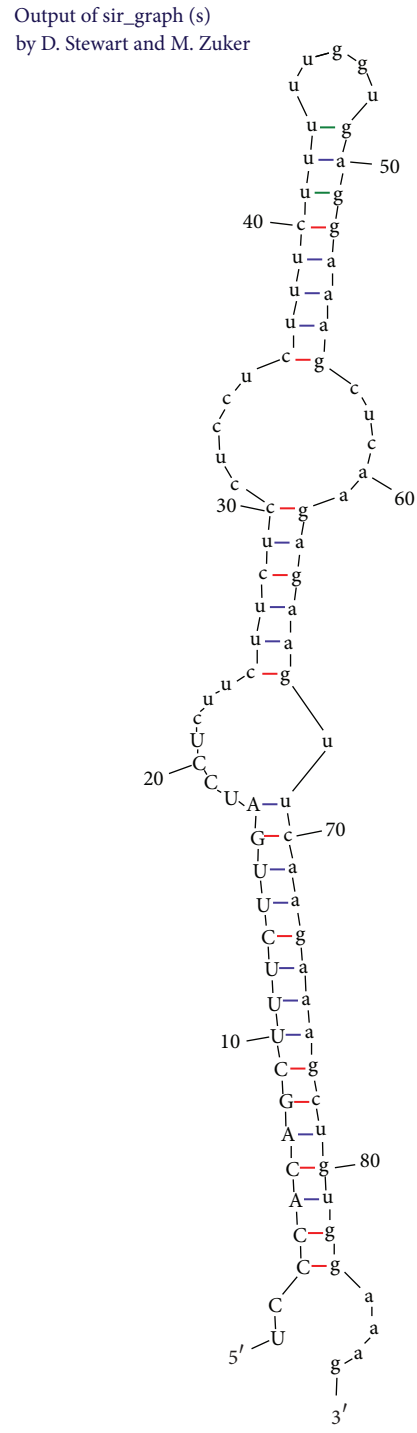

Figure 2: Precursor ath-miR396b for isotig03073, isotig03074, and isotig03073. Mature sequence is indicated by capital letters.

3. Dataset Description

The dataset associated with this Dataset Paper consists of 7 items which are described as follows.

Dataset Item 1 (Table). A list of isotigs (which represent mRNAs), their Blast2GO results, expression results, and number of differences identified. The column Sequence Name presents the isotig number designated by the assembly program; Normal Read Count, the sum of number of sequences that mapped to each contig in the isotig derived

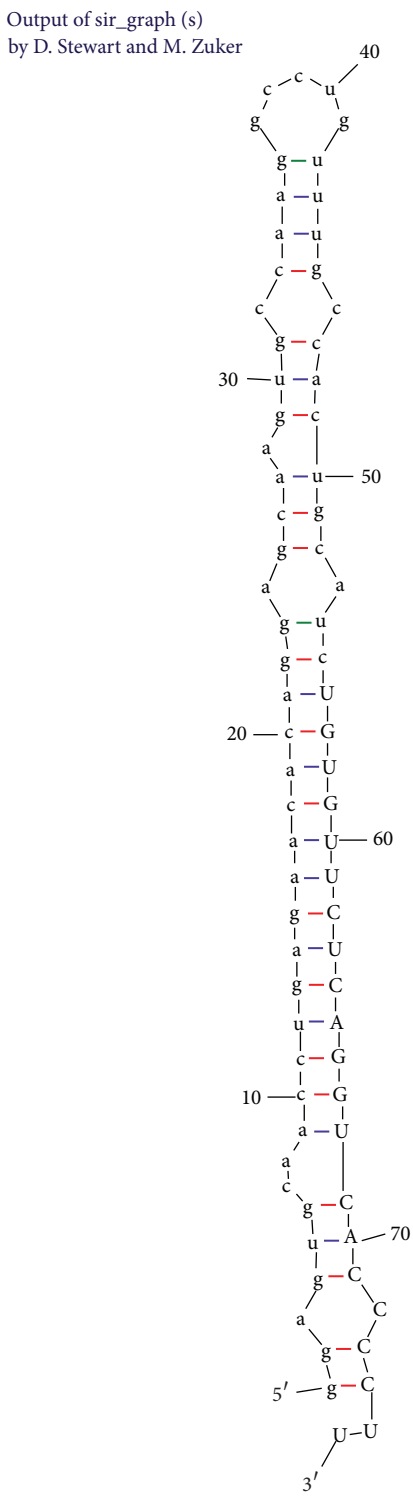

FIGURE 3: Precursor ath-miR398a for isotig12595. Mature sequence is indicated by capital letters.

from the readstatus.txt output file; Normal Expression, the reads per 1 kb sequence; Sequence Description, Blast2GO sequence description; At the highest match *Arabidopsis thaliana* protein from a blastx against the TAIR Arabidopsis database with *P* value cut-off of 1E-30 (E-value not shown); Minimum E-value, the E-value of the highest matching sequence in Blast2GO; Mean Similarity, the mean similarity of the highest matching sequence from Blast2GO; #GOs, the number of gene ontology terms of the highest matching sequence from Blast2GO; GOs, the gene ontology terms assigned to the highest matching sequence from Blast2GO; and Enzyme Codes, the enzyme code of the highest matching sequence from Blast2GO.

Column 1: Sequence Name
Column 2: Length
Column 3: Normal Read Count
Column 4: Normal Expression Corrected
Column 5: Sequence Description
Column 6: At
Column 7: Minimum E-value
Column 8: Mean Similarity
Column 9: #GOs
Column 10: GOs
Column 11: Enzyme Codes

Dataset Item 2 (Table). A list of isogroups (which represent genes), the most common Blast2GO description, expression results, and number of differences identified. The column Gene presents the isogroup number designated by the assembly program; Normal Read Count, the sum of number of sequences that mapped to each contig in the isogroup derived from the readstatus.txt output file; and Description(s), the most common sequence description (multiple descriptions are listed for isogroups without a consensus description).

Column 1: Gene
Column 2: Normal Read Count
Column 3: Description(s)

Dataset Item 3 (Table). A list of the isotigs that were identified as transcription factors. The column Isotig presents the isotig identified as being a transcription factor; Normal Read Count, the sum of number of sequences that mapped to each contig in the isogroup derived from the readstatus.txt output file; Sequence Description, the description of the isotig from Blast2GO; #GOs, the number of GO terms; GOs, the GO terms; and Enzyme Codes, the enzyme code for the closest matched protein.

Column 1: Isotig
Column 2: Isogroup
Column 3: Identification Method
Column 4: Normal Read Count
Column 5: Sequence Description
Column 6: #GOs

Column 7: GOs

Column 8: Enzyme Codes

Dataset Item 4 (Table). The isotigs that were identified by blasting the protein sequences from the plant transcription factor database against the isotig sequences using a tblastx. The column Isotig presents the isotig identified as being a transcription factor; Species, the species with the highest match for that isotig; Database Match, the protein ID with species name appended for the highest match transcription factor; E-value, the E-value of the highest match transcription factor; Percentage Identity, the percentage identity of the highest match transcription factor; Alignment Length, the alignment length of the highest match transcription factor; Mismatches, the number of mismatches in the alignment of the highest match transcription factor; Gap Openings, the number of gap opening in the alignment of the highest match transcription factor; Description, the description of the isotig from Blast2GO; and Normal Read Count, the sum of number of sequences that mapped to each contig in the isogroup derived from the readstatus.txt output file.

Column 1: Isotig

Column 2: Identification Method

Column 3: Species

Column 4: Database Match

Column 5: E-value

Column 6: Percentage Identity (%)

Column 7: Alignment Length

Column 8: Mismatches

Column 9: Gap Openings

Column 10: Description

Column 11: Normal Read Count

Dataset Item 5 (Table). The isotig match to the sequences described by Tranbarger et al. (2011) as transcription factors/regulators. The column Tranbarger presents the sequence from Tranbarger et al. (2011) dataset; Isotig Match, the closest isotig match identified by Blast; Normal Read Count, the sum of number of sequences that mapped to each contig in the isogroup derived from the readstatus.txt output file; and Description, the description of the isotig from Blast2GO.

Column 1: Tranbarger

Column 2: Isotig Match

Column 3: Normal Read Count

Column 4: Description

Dataset Item 6 (Table). The accession for each sequence uploaded to GenBank.

Column 1: Sequence

Column 2: Accession

Dataset Item 7 (Table). The sequence information of 34 sequences that were too short for GenBank (<200 bp) to accept.

Column 1: Isotig

Column 2: Isogroup

Column 3: Length

Column 4: Sequence

4. Concluding Remarks

The transcriptome assembly described here is to be used in conjunction with (Dataset Items 1–7 (Tables)) to easily and rapidly identify any gene of interest from within the oil palm fruit transcriptome and obtain the expression data and sequence for that gene.

Dataset Availability

The dataset associated with this Dataset Paper is dedicated to the public domain using the CC0 waiver and is available at http://dx.doi.org/10.7167/2013/670926/dataset.

Conflict of Interests

The authors declare that there is no conflict of interests.

Acknowledgments

The authors would like to thank Mr. A. Limsrivilai, Golden Tenera, for contributing the oil palm samples that allowed this work to take place. This work was funded by the Agricultural Research Development Agency, Bangkok, Thailand, and the National Science and Technology Development Agency, Thailand.

References

[1] J. B. Teixeira, M. R. Söndahl, and E. G. Kirby, "Somatic embryogenesis from immature inflorescences of oil palm," *Plant Cell Reports*, vol. 13, no. 5, pp. 247–250, 1994.

[2] J. Schwendiman, C. Pannetier, and N. Michaux-ferriere, "Histology of somatic embryogenesis from leaf explants of the oil palm *Elaeis guineensis*," *Annals of Botany*, vol. 62, no. 1, pp. 43–52, 1988.

[3] A. Rival, J. Tregear, J. L. Verdeil et al., "Molecular search for mRNA and genomic markers of the oil palm "mantled" somaclonal variation," *ISHS Acta Horticulturae*, vol. 461, pp. 165–172, 1998.

[4] A. Rival, E. Jaligot, T. Beule, J. L. Verdeil, and J. Tregear, "DNA methylation and somaclonal variation in oil palm," in *Proceedings of the International Symposium on Methods and Markers for Quality Assurance in Micropropagation*, A. C. Cassells, B. M. Doyle, and R. F. Curry, Eds., vol. 1, pp. 447–454, International Society Horticultural Science, Leuven, Belgium, 2000.

[5] R. H. V. Corley, C. H. Lee, I. H. Law, and C. Y. Wong, "Abnormal flower development in oil palm clones," *Planter*, vol. 62, article 7, 1986.

[6] E. Jaligot, S. Adler, É. Debladis et al., "Epigenetic imbalance and the floral developmental abnormality of the in vitro-regenerated oil palm *Elaeis guineensis*," *Annals of Botany*, vol. 108, article 9, 2011.

[7] E. Jaligot, A. Rival, T. Beule, S. Dussert, and J. L. Verdeil, "Somaclonal variation in oil palm (*Elaeis guineensis* Jacq.): the DNA methylation hypothesis," *Plant Cell Reports*, vol. 19, no. 7, pp. 684–690, 2000.

[8] A. Rival, L. Bertrand, T. Beule, M. C. Combes, P. Trouslot, and P. Lashermes, "Suitability of RAPD analysis for the detection of somaclonal variants in oil palm (*Elaeis guineensis* Jacq)," *Plant Breeding*, vol. 117, no. 1, pp. 73–76, 1998.

[9] A. Rival, T. Beule, P. Barre, S. Hamon, Y. Duval, and M. Noirot, "Comparative flow cytometric estimation of nuclear DNA content in oil palm (*Elaeis guineensis* jacq) tissue cultures and seed-derived plants," *Plant Cell Reports*, vol. 16, no. 12, pp. 884–887, 1997.

[10] B. J. Haas and M. C. Zody, "Advancing RNA-seq analysis," *Nature Biotechnology*, vol. 28, no. 5, pp. 421–423, 2010.

[11] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[12] S. Götz, J. M. García-Gómez, J. Terol et al., "High-throughput functional annotation and data mining with the Blast2GO suite," *Nucleic Acids Research*, vol. 36, no. 10, pp. 3420–3435, 2008.

[13] P. Pérez-Rodríguez, D. M. Riaño-Pachón, L. G. G. Corrêa, S. A. Rensing, B. Kersten, and B. Mueller-Roeber, "PlnTFDB: updated content and new features of the plant transcription factor database," *Nucleic Acids Research*, vol. 38, pp. D822–D827, 2010.

[14] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, article R106, 2010.

[15] T. J. Tranbarger, S. Dussert, T. Joët et al., "Regulatory mechanisms underlying oil palm fruit mesocarp maturation, ripening, and functional specialization in lipid and carotenoid metabolism," *Plant Physiology*, vol. 156, no. 2, pp. 564–584, 2011.