



Data Article

Transcriptome dataset of sago palm in peat soil



Wei-Jie Yan^{a,*}, Hasnain Hussain^{a,*}, Hung Hui Chung^b,
Norzainizul Julaihi^c, Rina Tommy^c

^a Centre for Sago Research (CoSAR), Faculty of Resource Science and Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

^b Faculty of Resource Science and Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

^c Land Custody and Development Authority, Level 4, 8 & 12, Wisma Satok, Jalan Satok, 93400 Kuching, Sarawak, Malaysia

ARTICLE INFO

Article history:

Received 14 December 2021

Revised 20 January 2022

Accepted 31 January 2022

Available online 3 February 2022

Keywords:

Sago palm

Metroxylon sagu

RNA sequencing

Transcriptome

Non-trunking

ABSTRACT

Sago palm (*Metroxylon sagu* Rottb.) is an important agricultural starch-producing palm that contributes to Malaysia's economics, especially in the State of Sarawak, Malaysian Borneo. In this palm tree, the central part of the plant storage-starch. Under normal condition, sago palm develop its trunk after 4–5 years being planted. However, sago palms planted on deep-peat soil failed to develop their trunk even after 17 years of being planted. This phenomenon is known as 'non-trunking', which eliminates the economic value of the palms. Numerous research has been done to address the phenomenon, but the molecular mechanisms of sago palm responding toward the responsible stresses are still lacking. Therefore, in this study, leaf samples were collected from trunking (normal) and non-trunking sago palms planted on peat soil for total RNA extraction, followed by next-generation sequencing using the BGISEQ-500 platform. The raw reads were cleaned, and *de novo* assembled using TRINITY software package. A total of 40.11 Gb bases were sequenced from the sago palm leaf samples. The assembled sequence produced 102,447 unigenes, with N50 score 1809 bp and GC ratio of

* Corresponding authors.

E-mail addresses: 17010084@siswa.unimas.my (W.-J. Yan), hhasnain@unimas.my (H. Hussain).

Social media: [@DrMd9](https://twitter.com/DrMd9) (H. Hussain)

44.34%. The alignment of unigenes with seven functional databases (NR, NT, GO, KOG, KEGG, SwissProt and InterPro) resulted in the annotation of 65,523 (63.96%) unigenes. Functional annotation results in the detection of 46,335 coding DNA sequences by Transdecoder. A total of 30,039 simple-sequence repeats distributed on 21,676 unigenes were detected using Primer3 software, and 2355 transcription factor coding unigenes were predicted using getorf and hmmscan software. This work is registered under NCBI BioProject PRJNA781491. The raw RNA sequencing data are available in Sequence Read Archive (SRA) database with accession numbers SRX13165895, SRX13165896, SRX13165897, SRX13165898, SRX13165899, and SRX13165900. Gene expression and annotation information are accessible in public functional genomics data repository Gene Expression Omnibus (GEO) with accession number GSE189085.

© 2022 The Authors. Published by Elsevier Inc.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

Specifications Table

Subject	Biological sciences; Omics: Transcriptomics
Specific subject area	Trunk development of sago palm under stress
Type of data	Transcriptomics data (raw RNA sequence reads, gene expression and sequence annotation)
How the data were acquired	BGISEQ-500 platform
Data format	Raw: *fastq.gz files Assembly: *Unigene.fa.gz files Processed Data: *gene.fpkms.txt.gz files
Description of data collection	Total RNA was extracted from trunking and non-trunking sago palm (<i>M. sago</i>) leaf tissue, mRNA library preparation and then was sequenced using the BGISEQ-500 platform
Data source location	Dalat Sago Plantation, Mukah, Sarawak, Malaysia GPS location are listed in Table 1)
Data accessibility	Repository name: NCBI's Gene expression omnibus (GEO) Data identification number: GSE189085 Direct URL to data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE189085 Repository name: NCBI's Sequence Read Archive (SRA) Sample ID: GSM5694359 (ST1: Trunking Sample 1) Data identification number: SRX13165895 Direct URL to data: https://www.ncbi.nlm.nih.gov/sra/SRX13165895[accn] Repository name: NCBI's Sequence Read Archive (SRA) Sample ID: GSM5694360 (ST4: Trunking Sample 4) Data identification number: SRX13165896 Direct URL to data: https://www.ncbi.nlm.nih.gov/sra/SRX13165896[accn] Repository name: NCBI's Sequence Read Archive (SRA) Sample ID: GSM5694361 (ST5: Trunking Sample 5) Data identification number: SRX13165897 Direct URL to data: https://www.ncbi.nlm.nih.gov/sra/SRX13165897[accn] Repository name: NCBI's Sequence Read Archive (SRA) Sample ID: GSM5694362 (NT7: Trunking Sample 7) Data identification number: SRX13165898 Direct URL to data: https://www.ncbi.nlm.nih.gov/sra/SRX13165898[accn] Repository name: NCBI's Sequence Read Archive (SRA) Sample ID: GSM5694363 (NT8: Trunking Sample 8) Data identification number: SRX13165899

(continued on next page)