

Journal Pre-proof

The first Engkabang Jantong (*Rubroshorea macrophylla*) genome survey data

Hung Hui Chung , Asmeralda Ai Leen Soh , Melinda Mei Lin Lau ,
Han Ming Gan , Siong Fong Sim , Leonard Whye Kit Lim

PII: S2352-3409(24)01210-1
DOI: <https://doi.org/10.1016/j.dib.2024.111248>
Reference: DIB 111248



To appear in: *Data in Brief*

Received date: 12 September 2024
Revised date: 12 December 2024
Accepted date: 17 December 2024

Please cite this article as: Hung Hui Chung , Asmeralda Ai Leen Soh , Melinda Mei Lin Lau , Han Ming Gan , Siong Fong Sim , Leonard Whye Kit Lim , The first Engkabang Jantong (*Rubroshorea macrophylla*) genome survey data, *Data in Brief* (2024), doi: <https://doi.org/10.1016/j.dib.2024.111248>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2024 Published by Elsevier Inc.

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

ARTICLE INFORMATION

Article title

The first Engkabang Jantong (*Rubroshorea macrophylla*) genome survey data

Authors

Hung Hui Chung^{1*}, Asmeralda Ai Leen Soh¹, Melinda Mei Lin Lau¹, Han Ming Gan^{2,3}, Siong Fong Sim¹ & Leonard Whye Kit Lim¹

Affiliations

¹Faculty of Resource Science and Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.

²Patriot Biotech Sdn Bhd, 47500 Subang Jaya, Selangor, Malaysia

³Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Geelong, Victoria, Australia

Corresponding author's email address and Twitter handle

hhchung@unimas.my

Keywords

engkabang jantong; *Rubroshorea macrophylla*; genomic landscape; microsatellite; Illumina

Abstract

The engkabang jantong (*Rubroshorea macrophylla*) is one of the most indispensable tree species for reforestation due to its high survival rate and rapid growth rate. Due to relatively low genetic interest of this tree species, its genomic landscape has since faced scarcity, impeding our further elucidation on genes that are involved in expressing its aforementioned superior properties. In this study, we performed genome survey and microsatellite analysis of engkabang jantong. Based on the results, the estimated genome size of this species is 312,071,515 bp with 18.43% repeated sequences and 1.16% heterozygosity. BUSCO analysis unearthed that 83.5% of the contigs are single-copy genes whereas 12.7% of them are duplicated. Only 2.8% and 1% of them are fragmented and missing respectively. The short-read sequencing results obtained from the Illumina platform in this study will be essential to complement the Nanopore long-read sequencing results in hybrid genome assembly endeavors in the near future.

SPECIFICATIONS TABLE

[]

| | |
|---------------------------------|--|
| Subject | Biological Sciences. |
| Specific subject area | Genomics |
| Type of data | Sequencing raw reads, Table and Figure. |
| Data collection | The extracted engkabang jantung DNA was sheared into 350 bp fragments using Covaris Ultrasonicator. Then, library preparation was done using NEB Ultra II library preparation kit according to manufacturer's protocol. The assembled library was then subjected to sequencing via Illumina NovaSeq 6000 platform. |
| Data source location | The collection of engkabang leaves from a single individual tree is under the permission of Sarawak Forestry Corporation (Reference Number: SFC.810-4/6/1(2022)). The engkabang leaves are provided by the ranger of the Sarawak Forestry Corporation. The collection of leaves is carried out at Semenggoh Wildlife Center, Kuching, Sarawak, Malaysia (1.402258002376039, 110.31446195505569). The preserved dry engkabang leaves were deposited in UNIMAS Herbarium with accession number HB008123. |
| Data accessibility | The sequencing reads used in the analysis are available under the NCBI BioProject PRJNA1127791 Repository name: NCBI GenBank database Data identification number: BioProject PRJNA1127791 Direct URL to data: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1127791 Instructions for accessing these data: Click on the link provided above |
| Related research article | None |

VALUE OF THE DATA

- First genome survey conducted on engkabang jantung

- Enable for genome-wide association studies
- Enable future research on genotypes contributing to high quality timber and fatty acids produced by this species

BACKGROUND

The engkabang jantong (*Rubroshorea macrophylla*) belongs to the family Dipterocarpaceae. It plays significant roles in the ecology system, aquaculture feed and reforestation due to its rapid growth rate and high survival rate [1-3]. Besides its trunk that produces high quality timber, its fruit (also known as illipe nut) is highly encompassed with high quality oil and fatty acids [4] and is natural food loved by a highly priced fish species, namely the empurau in the wild [5-7]. Its unique fatty acid content and fragrance is what makes the empurau fish taste scrumptiously palatable with unique texture [8,9].

Due to the relatively low genetic interest placed into this tree species, the only genomic data found on *R. macrophylla* was that from the complete chloroplast genome sequencing performed by [10]. In this study, we conducted a genome survey on the engkabang jantong whole genome before performing BUSCO, k-mer and microsatellite analyses to characterize the whole genome. Furthermore, we also further characterize the genome with analyses such as functional annotation and phylogenetic tree construction. It is hoped that the short-read sequencing results obtained from the Illumina platform in this study will be essential to complement the Nanapore long-read sequencing results in hybrid genome assembly endeavors in the near future.

DATA DESCRIPTION

The engkabang jantong genome contigs were filtered and screened directly after sequencing, unraveling a sum of 29.85 Gb of clean reads generated with ~68.6x coverage. A total of 96,606 contigs with total contig length 435,158,746 bp was reported in this study. The lengthiest contig is 799,893 bp long while the genome size was estimated at 312,071,515 bp (~312 Mbp) (Table 1). This genome size is close to that of its genus counterparts such as *Rubroshorea robusta*, *Rubroshorea leprosula*, *Rubroshorea henryana* and *Rubroshorea roxburghii* with genome size estimated at 357.11 Mbp, 323.6 Mbp, 302.6 Mbp and 306.2 Mbp [11,12] respectively. The engkabang jantong genome GC content documented in this study is 33.38%, which is very similar to that of *R. roxburghii* (33.2%) [11], *R. leprosula* (33.4%) [11], and *R. robusta* (33.69%) [12]. This demonstrated the high conservation of GC content and genome size across the *Rubroshorea* genus. BUSCO analysis unearthed that 83.5% of the engkabang jantong genome contigs are complete and single-copy protein-coding genes, whereas 12.7% of them are complete and duplicated genes. Only 2.8% and 1%

of them are fragmented and missing respectively, according to the BUSCO analysis in this study. The heterozygosity of the engkabang jantung recorded in this study is $\sim 1.16\%$, which is considered high when compared to other plant species like Satsuma (0.435%), sago palm (0.63%), pummelo (0.022%), date palm (0.46%), sweet orange (0.716%) and Clementine (0.462%) [13]. The repeated sequence percentage of the engkabang jantung genome is 18.43%, which is very much lower compared to that of its genus counterpart, *R. leprosula* (33%) (Ng et al., 2021) as well as other plant species, for instance, *Oryza sativa* (39.5%), oil palm (57%), sago palm (35.7%), *Vitis vinifera* (41.4%), date palm (38.41%) and *Populus trichocarpa* (42%) [13]. The heterozygosity ratio and repeated sequence percentage are responsible for the fabrication and splicing in the genome, they also reflect the heterogeneity of the plant habitat, contributing to its biodiversity [14].

To date, there is no genome wide microsatellite analysis done onto any Dipterocarpaceae family members. The genome-wide microsatellite analysis conducted in this study unearthed that a sum of 54607 short sequence repeats (SSRs) were discovered within the engkabang jantung genome. The dinucleotide repeats cover majority (22502, 41%) of the short sequence repeats identified in this study with minimum four repeats (Figure 2A). The most least found short sequence repeats are the pentanucleotide repeats (5584, 10%). A similar phenomenon was seen in sago palm genome whereby the most abundant SSRs is the dinucleotide repeats (62.24%) whereas the most least discovered SSRs is the pentanucleotide repeats (5.91%) [13]. The most abundant SSR species found is the AT/TA with 18165 (33.26%) found within the engkabang jantung genome, which made them the top among the dinucleotide repeats identified in this study (Figure 2B). The top three trinucleotide microsatellites of the engkabang jantung genome are AAT/ATT (6.21%), TTA/TAA (4.42%) as well as TAT/ATA (2.91%). The AAAT/ATTT, TTTA/TAAA, and TATT/AATA topped the engkabang jantung genome tetranucleotide SSRs chart with compositions of 5.96%, 4.87% and 2.56% correspondingly. The most abundant engkabang jantung genome pentanucleotide microsatellite is the AAAAT/ATTTT (1.33%), followed by AAAAG/CTTTT (0.86%) and TTTTA/TAAAA (0.8%). Interestingly, the top three engkabang jantung genome tetranucleotide and pentanucleotide microsatellites mirrored to that of the sago palm genome with differing population sizes [13]. These microsatellite data generated in this study lay imperative groundwork for the DNA fingerprinting and barcoding endeavor for species identification across the Dipterocarpaceae family members in the future.

The functional annotation of Gene Ontology (GO) terms using EggNOG mapper v2 [15] revealed 41,061 genes associated with GO term IDs within the engkabang jantung genome. Upon further functional annotation using TbTools II [16], 18,173 genes were found to have link with 94 parent GO terms. Almost half (48%) of the identified genes are associated with the biological process parent GO term. A quarter of them (25%) are grouped under cellular component parent Go term while the remaining (27%) are housed under the molecular function parent GO term (Figure 3A). Zooming into each parent GO terms identified in the engkabang jantung genome in this study (Figure 3B), we depicted the top three hits for each parent GO term. Under the biological process parent GO term, the cellular process topped the chart with 14,056 hits, followed by metabolic process (11,156 hits) as well as biosynthetic

process (6,363 hits). Under the cellular component parent GO term, the intracellular anatomical structure is the leading hit with 12,709 hits, the second leading hit is the cytoplasm (9,543 hits) and the third one is the membrane (6,447 hits) within the engkabang jantung genome. Under the engkabang jantung genome molecular function parent GO term, the catalytic activity is the most frequently found hit with 8,234 hits, while the second and third ones are binding (6,032) and transferase activity (3,558) respectively. The abundant metabolic and biosynthetic genes found within the engkabang jantung genome is postulated to have associated with its capability to produce huge amount of high quality cell wall, secondary cell wall (timber), fatty acids (fragrant oleoresin) as well as for defense mechanism [11].

A phylogenetic tree was constructed to include a sum of 18 plant species with two outgroups (*M. sagu* and *A. thaliana*) with 1000 bootstrap replications (Figure 4). All the 18 plant species form monophyletic clade with their respective genus counterparts with varying bootstrap values from 55 to 100. The engkabang jantung formed a strong clade with its genus counterpart, *R. leprosula* with the maximum bootstrap value of 100. This similar phenomenon was seen in other phylogenetic tree constructed such as the sago palm [13] whereby the genus counterparts shared the same clade with maximum bootstrap values. This is essential for accurate and precise species identification and also beneficial for the discovery of hybrid species in the future if there is any occurrence happening in the future. The results from this study also offer the engkabang jantung genome as a reference genome for future genome sequencing and assembly of other currently yet to be sequenced Dipterocarpaceae family members genomes.

EXPERIMENTAL DESIGN, MATERIALS AND METHODS

Sampling, DNA extraction and genome sequencing

The leaf tissues of *Rubroshorea macrophylla* were obtained from Semenggoh Wildlife Centre (1°23'59"N 110°19'27"E) with authorization from the Sarawak Forestry Corporation (Reference Number: SFC.810-4/6/1(2022)). The DNA extraction was done emulating that from [17]. The DNA extracted was quality checked using DeNovix DS-11+ spectrophotometer (DeNovix, USA) and agarose gel electrophoresis analysis. The extracted engkabang jantung DNA was sheared into 350 bp fragments using Covaris Ultrasonicator (Covaris, United Kingdom). Then, library preparation was done using NEB Ultra II library preparation kit (NEB, United Kingdom) according to manufacturer's protocol. The assembled library was then subjected to sequencing via Illumina NovaSeq 6000 platform. Adapter- and quality-trimming was performed using fastp v. 0.18 onto the raw pair-end reads.

Genome completeness, functional annotation, microsatellite and phylogenetic analysis

BUSCO V5.7.1 [19] was employed to assess the genome completeness of the engkabang jantung genome. Various genomic matrices were examined utilizing QUAST v5.2.0 [20]. The k-mer frequency distribution data was generated using Jellyfish v2.3.0 [21] at k-mer size of 31. GenomeScope webserver [22] was used for the visualization of various genomic matrices based on parameters, namely read length 150 bp and K-mer coverage of 1000. AUGUSTUS v3.4.0 [23] was utilized to predict the protein-coding genes from the engkabang jantung whole genome with reference to pre-trained model species *Theobroma cacao*. Functional annotation of all the predicted protein sequences was conducted using EggNOG mapper v2 [15] on the eggNOG 5 database, with E-value set at 0.001. These genes were further functionally annotated using TbTools II [16]. Microsatellite analysis was also performed using Kmer-SSR [24]. Only microsatellites with more than four repeats were selected for further analysis. A total of 20 plant species was selected for phylogenetic tree construction with *Metroxylon sagu* and *Arabidopsis thaliana* as outgroups. The ‘complete and single-copy’ protein sequences of the 20 plant species obtained from BUSCO analysis were subjected to multiple sequence alignment using MAFFT v7.471 [25]. MEGA 11 [26] was employed to construct Neighbor-Joining (NJ) tree with 1000 bootstrap replications.

LIMITATIONS

The limitation of this study is that the Illumina sequencing platform only generates short reads. The number of complete and single copy BUSCO of 83.5%, this number can be further increased with long-read Nanopore sequencing platform and then subsequently conducting a hybrid genome assembly [18].

ETHICS STATEMENT

The authors have read and follow the [ethical requirements](#) for publication in Data in Brief and confirming that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

CRedit AUTHOR STATEMENT

Hung Hui Chung: Conceptualization, Funding acquisition, Writing –review & editing; Asmeralda Ai Leen Soh: Data curation, Writing – original draft; Melinda Mei Lin Lau: Data curation; Han Ming Gan: Methodology, Conceptualization, Writing –review & editing; Siong Fong Sim: Writing –review & editing; Leonard Whye Kit Lim: Writing – original draft, Writing –review & editing.

ACKNOWLEDGEMENTS

This research is funded by Tun Zaidi Chair (UNI/F07/TZC/85864/2024) awarded to H.H. Chung.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

- [1] Perumal, M., Wasli, M.E., Ying, H.S., Lat, J. and Sani, H. (2017). Survivorship and Growth Performance of *Shorea macrophylla* (de Vriese) after Enrichment Planting for Reforestation Purpose at Sarawak, Malaysia. *OnLine Journal of Biological Sciences*, 17(1), 7–17.
- [2] Lim LWK, Chung HH, Ishak SD, Waiho K (2021a) Zebrafish (*Danio rerio*) ecotoxicological ABCB4, ABCC1 and ABCG2a gene promoters depict spatiotemporal xenobiotic multidrug resistance properties against environmental pollutants. *Gene Reports*, 23, 101110.
- [3] Lim, LWK. (2024). Implementation of Artificial Intelligence in Aquaculture and Fisheries: Deep Learning, Machine Vision, Big Data, Internet of Things, Robots and Beyond. *Journal of Computational and Cognitive Engineering*, 3(2), 112-118.
- [4] Lim, LWK (2022) Eco-Economically Indispensable Borneo-Endemic Flora and Fauna: Proboscis Monkey (*Nasalis larvatus*), Malaysian Mahseer (*Tor tambroides*), Engkabang (*Shorea macrophylla*), Sarawak Rasbora (*Rasbora sarawakensis*) and Sago Palm (*Metroxylon sagu*). *International Journal of Zoology and Animal Biology*, 5(3), 000381.
- [5] Lau MML, Lim LWK, Chung HH, Gan HM (2021a) The first transcriptome sequencing and data analysis of the Javan mahseer (*Tor tambra*). *Data in Brief*, 39, 107481.
- [6] Lau MML, Lim LWK, Ishak SD, Abol-Munafi A, Chung HH (2021b) A Review on the Emerging Asian Aquaculture Fish, the Malaysian Mahseer (*Tor tambroides*): Current Status and the Way Forward. *Proc Zool Soc*, 74, 227-237.
- [7] Lim, LWK. (2023). Cultivated Meat in Singapore: The Road to Commercialization. *International Journal of Zoology and Animal Biology* 6(4), 1-5.
- [8] Lim LWK, Chung HH, Lau MML, Aziz F, Gan HM (2021b) Improving the phylogenetic resolution of Malaysian and Javan mahseer (Cyprinidae), *Tor tambroides* and *Tor tambra*:

whole mitogenomes sequencing, phylogeny and potential mitogenome markers. *Gene*, 791, 145708.

[9] Lau MML, Kho CJY, Lim LWK, Sia SC, Chung HH, et al. (2022) Microbiome Analysis of Gut Bacterial Communities of Healthy and Diseased Malaysian Mahseer (*Tor tambroides*). *Malaysian Society for Microbiology*, 18(2), 170-191.

[10] Chew IYY, Chung HH, Lim LWK, Lau MML, Gan HM, et al. (2022) Complete chloroplast genome of *Shorea macrophylla* (engkabang): Structural features, comparative and phylogenetic analysis. *Data in Brief*, 47, 109029.

[11] Tian, Z., Zeng, P., Lu, X., Zhou, T., Han, Y., Peng, Y., ... Cai, J. (2022). Thirteen Dipteroocarpoideae genomes provide insights into their evolution and borneol biosynthesis. *Plant Communications*, 3(6), 100464.

[12] Mishra, G., Meena, R.K., Kant, R., Pandey, S., Ginwal, H.S., & Bhandari, M.S. (2023). Genome-wide characterization leading to simple sequence repeat (SSR) markers development in *Shorea robusta*. *Funct Integr Genomics*, 23(1), 51.

[13] Lim LWK, Chung HH, Hussain H, Gan HM (2021c) Genome survey of sago palm (*Metroxylon sagu* Rottboll). *Plant Gene*, 28, 100341.

[14] Katayama, N., Amano, T., Naoe, S., Yamakita, T., Komatsu, I., Takagawa, S., Sato, N., Ueta, M. and Miyashita, T. (2014). Landscape Heterogeneity–Biodiversity Relationship: Effect of Range Size. *PLoS ONE*, 9(3), e93359.

[15] Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P. and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38(12), 5825– 5829.

[16] Chen, C., Wu, Y., Li, J., Wang, X., Zeng, Z., Xu, J., Liu, Y., Feng, J., Chen, H., He, Y. and Xia, R. (2023). TBtools-II: A ‘One for All, All for One’ Bioinformatics Platform for Biological Big data Mining. *Molecular Plant*. doi:https://doi.org/10.1016/j.molp.2023.09.010.

[17] Lim LWK, Chung HH, and Hussain H (2020) Complete chloroplast genome sequencing of sago palm (*Metroxylon sagu* Rottb.): molecular structures, comparative analysis and evolutionary significance. *Gene Rep*, 19, 100662.

[18] Lim, LWK, Lau, MML, Chung HH, Hasnain H, and Gan HM. (2022). First high-quality genome assembly data of sago palm (*Metroxylon sagu* Rottboll). *Data in Brief*, 40, 107800.

[19] Manni, M, Berkeley, MR, Seppely, M, and Zdobnov, EM. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols*, 1, e323.

[20] Gurevich A, Saveliev V, Vyahhi N, and Tesler G. (2013). QUASt: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072-1075.

- [21] Marçais, G, and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6), 764–770.
- [22] Vurture GW, Sedlazeck, FJ, Nattestad, M, Underwood, CJ, Fang, H, Gurtowski, J, and Schatz, MC. (2017). GenomeScope: fast reference-free genome profiling from short reads, *Bioinformatics*, Volume 33(14), 2202–2204.
- [23] Stanke M, and Morgenstern B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.*, 33(Web Server issue):W465-7.
- [24] Pickett, BD, Miller, JB, and Ridge, PG. (2017). Kmer-SSR: a fast and exhaustive SSR search algorithm, *Bioinformatics*, 33(24), 3922–3928.
- [25] Katoh K, and Standley DM. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.*, 30(4), 772-780.
- [26] Tamura K, Stecher G, and Kumar S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11, *Molecular Biology and Evolution*, Volume 38(7), 3022–3027.

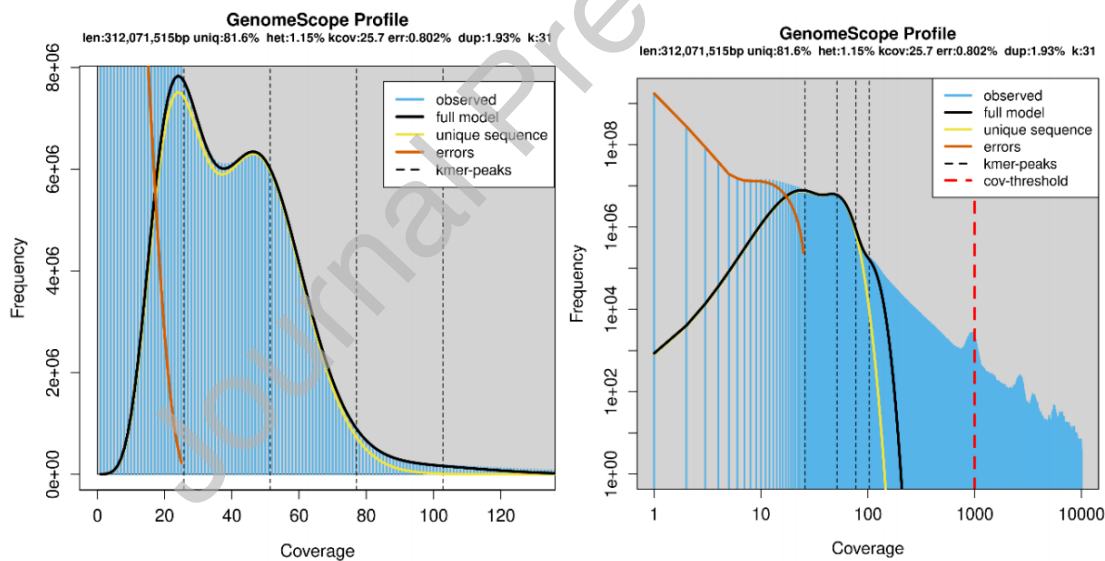
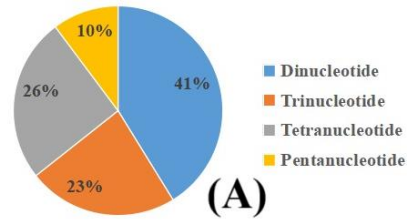
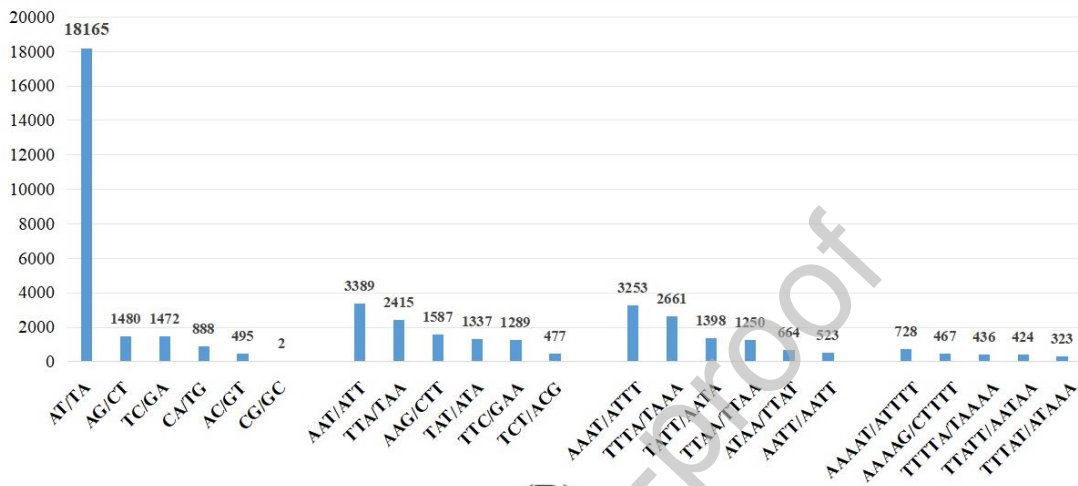


Figure 1. Estimation of genome size, repeat content, and heterozygosity was visualized using the GenomeScope webserver, based on k-mer 31 (read length = 150 bp; k-mer max coverage at 1000). The y-axis displayed the number of k-mers while the x-axis depicted the coverage.



(A)



(B)

Figure 2. (A) The summary of microsatellite populations in the engkabang jantong genome. (B) The top six highly abundant microsatellite species for each microsatellite population in the engkabang jantong genome.

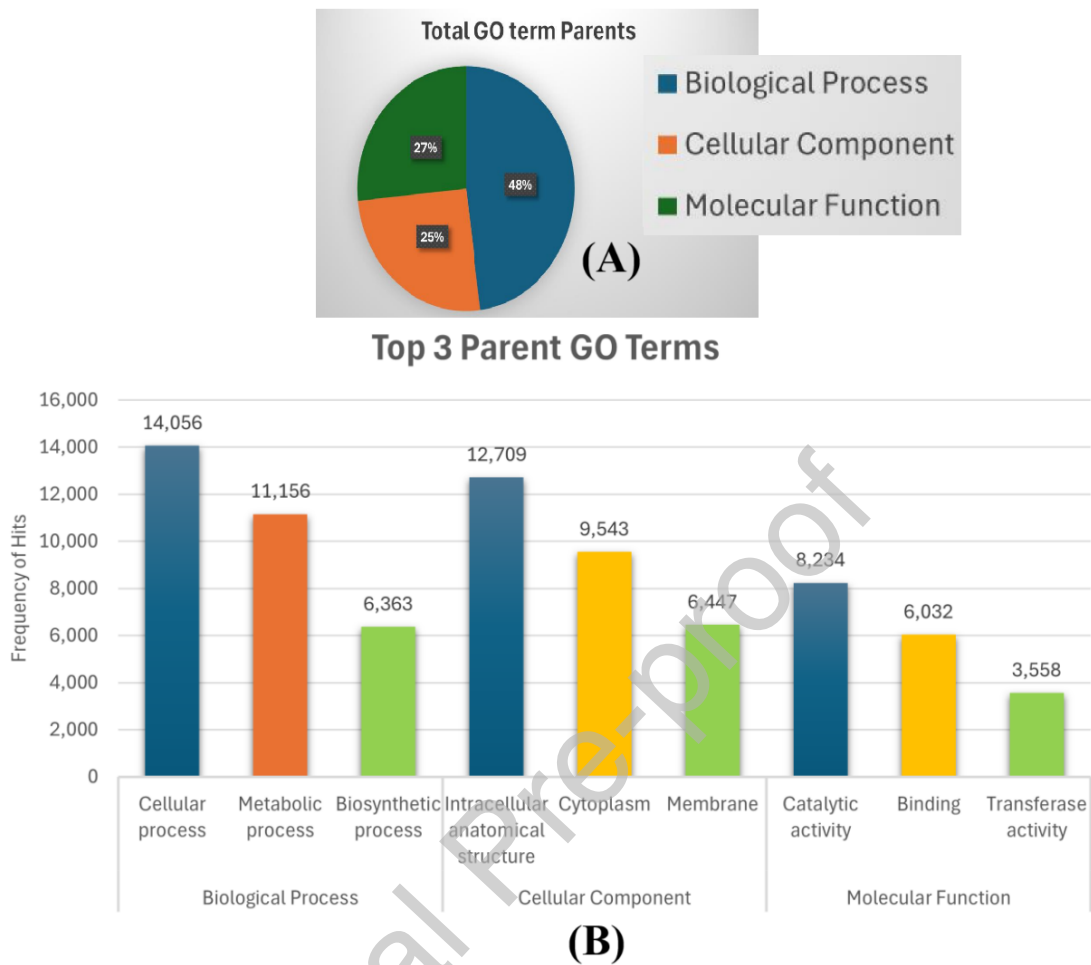


Figure 3. (A) The summary of parent GO terms found within the engkabang jantung genome. (B) The top three GO IDs for each of the parent GO terms identified in the engkabang jantung genome.

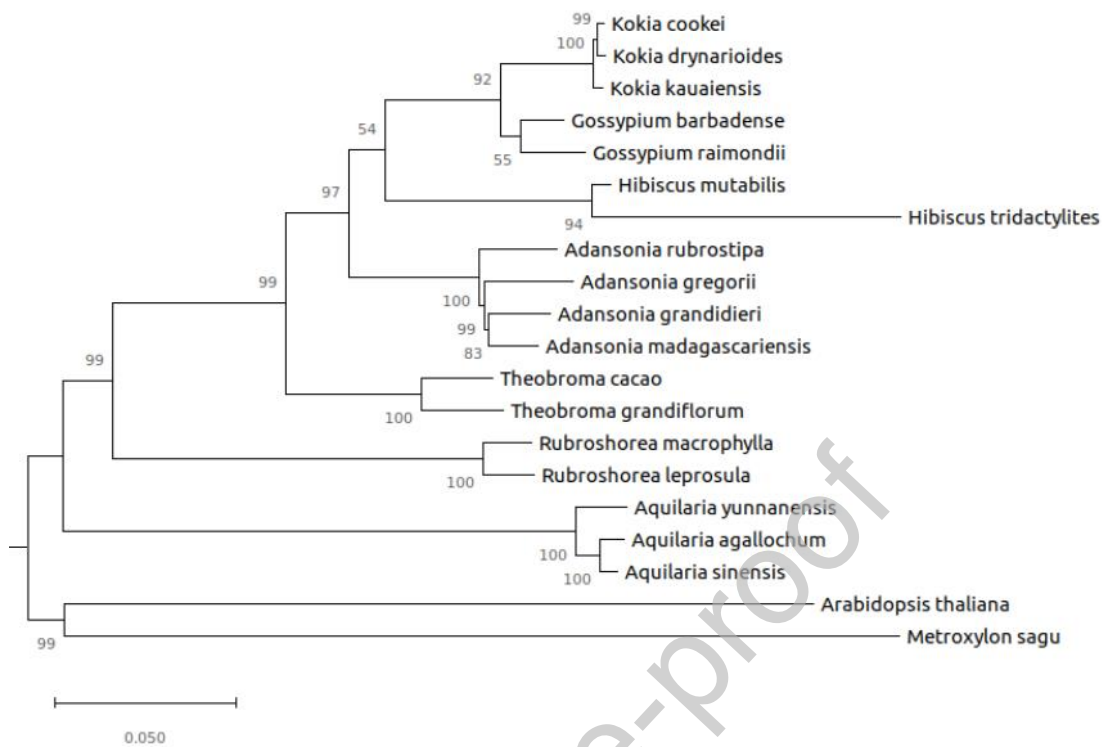


Figure 4. The phylogenetic tree constructed across 18 plant species with two outgroups based on 1000 bootstrap replications.

Table 1. The genomic data summary of the engkabang jantong genome.

| | |
|---------------------------------|--------------------------------|
| Organism | <i>Rubroshorea macrophylla</i> |
| Bioproject | PRJNA1127791 |
| Biosample | SAMN42019830 |
| Sequence Read Archive (SRA) | SRS21753026 |
| Complete and single copy BUSCOs | 83.5% (355) |
| Complete and duplicated BUSCOs | 12.7% (54) |
| Fragmented BUSCOs | 2.8% (12) |
| Missing BUSCOs | 1.0% (4) |
| Scaffold N50 | 22 kb |
| Contig N50 | 15 kb |
| Number of scaffolds | 70,981 |
| Number of contigs | 96,606 |
| Genome heterozygosity | 1.16% |
| Genome repetitiveness | 18.43% |
| Predicted genome size | 312,071,515 bp |
| GC content (%) | 33.38% |