Data in Brief 47 (2023) 109029

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Complete chloroplast genome data of *Shorea macrophylla* (Engkabang): Structural features, comparative and phylogenetic analysis

Ivy Yee Yen Chew[a], Hung Hui Chung[a,*], Leonard Whye Kit Lim[a], Melinda Mei Lin Lau[a], Han Ming Gan[b,c], Boon Siong Wee[a], Siong Fong Sim[a]

[a] *Faculty of Resource Science and Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia*
[b] *Patriot Biotech Sdn Bhd, 47500 Subang Jaya, Selangor, Malaysia*
[c] *Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Geelong, Victoria, Australia*

ARTICLE INFO

Article history:
Received 6 December 2022
Revised 19 February 2023
Accepted 24 February 2023
Available online 4 March 2023

Dataset link: Shorea macrophylla chloroplast, complete genome (Original data)

Keywords:
Shorea macrophylla
Dipterocarpaceae
Chloroplast genome
Phylogenetic analysis
Monophyletic

ABSTRACT

Shorea macrophylla belongs to the *Shorea* genus under the Dipterocarpaceae family. It is a woody tree that grows in the rainforest in Southeast Asia. The complete chloroplast (cp) genome sequence of *S. macrophylla* is reported here. The genomic size of *S. macrophylla* is 150,778 bp and it possesses a circular structure with conserved constitute regions of large single copy (LSC, 83,681 bp) and small single copy (SSC, 19,813 bp) regions, as well as a pair of inverted repeats with a length of 23,642 bp. It has 112 unique genes, including 78 protein-coding genes, 30 tRNA genes, and four rRNA genes. The genome exhibits a similar GC content, gene order, structure, and codon usage when compared to previously reported chloroplast genomes from other plant species. The chloroplast genome of *S. macrophylla* contained 262 SSRs, the most prevalent of which was A/T, followed by AAT/ATT. Furthermore, the sequences contain 43 long repeat sequences, practically most of them are forward or palindrome type long repeats. The genome structure of *S. macrophylla* was compared to the genomic structures of closely related species from the

* Corresponding authors.
E-mail address: hhchung@unimas.my (H.H. Chung).
Social media: @IvyChew971012 (I.Y.Y. Chew)

https://doi.org/10.1016/j.dib.2023.109029

same family, and eight mutational hotspots were discovered. The phylogenetic analysis demonstrated a close relationship between *Shorea* and *Parashorea* species, indicating that *Shorea* is not monophyletic. The complete chloroplast genome sequence analysis of *S. macrophylla* reported in this paper will contribute to further studies in molecular identification, genetic diversity, and phylogenetic research.

Specifications Table

Subject	Biological Sciences
Specific subject area	Omics: Chloroplast Genome
Type of data	Sequencing raw reads, assembly, Table, Figure, Graph
How data were acquired	Sequencing
Data format	Raw Reads
Description of data collection	The leaf pieces were washed in sorbitol buffer (0.35 M sorbitol, 1% PVP 40,000, 100 mM Tris-HCL pH 8, 5 mM EDTA pH 8) before bringing resuspended in 500 μL homogenization buffer (150 mM NaCl, 50 mM Tris-HCL pH 8, 25 mM EDTA). Approximately 100 ng of the gDNA, as quantified by the Denovix high sensitivity kit (Denovix) was fragmented to 350 bp using a Bioruptor. Then, library preparation was performed using the NEB Ultra II Illumina library preparation kit. On a NovaSEQ6000, the generated library was sequenced with a read configuration of 2 $\times$ 150 bp, yielding about 1 Gb of sequencing data.
Data source location	The collection of engkabang leaves 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).
Data accessibility	All relevant data are included in this manuscript. The chloroplast genome of *S. macrophylla* were deposited on GenBank with accession number ON321899 (https://www.ncbi.nlm.nih.gov/nuccore/ON321899.1).

Value of Data

- Complete chloroplast genome is important for improving our understanding of chloroplast biology and for engineering of chloroplast transgenes to enhance plant agronomic features or to develop high-value agricultural or biomedical products.
- As the distributions of *Shorea macrophylla* is limited to some rainforests in Southeast Asia, the genomic data availability of *Shorea* spp. is very valuable for future analyses.
- The identification and analysis of the chloroplast genome of *S. macrophylla* could help researchers better understand the species' variety and evolutionary links.

1. Objectives

The primary objective of this study is to sequence, assemble and annotate the complete chloroplast genome of the *Shorea macrophylla* which is deemed an important species for reforestation purpose and provide canopy for many primary rainforests in Southeast Asia region. The improved understanding of the chloroplast genome allows deeper understanding into their photosynthesis capacity and may potentially serves as template for genetic markers identification which further our knowledge on the species distribution and evolution.