



Data Article

Dataset from *de novo* transcriptome assembly of *Myristica fatua* leaves using MinION nanopore sequencer



Deden Derajat Matra^{a,*}, M Adrian^b, Karmanah^c, Jakty Kusuma^{d,e}, Jérôme Duminil^e, Sobir^a, Roedhy Poerwanto^a

^a Department of Agronomy and Horticulture, Faculty of Agriculture, IPB University, Bogor, Indonesia

^b Agronomy and Horticulture Study Program, Graduate School of IPB University, Bogor, Indonesia

^c Agrotechnology Study Program, Faculty of Agriculture, Nusa Bangsa University, Bogor, Indonesia

^d Department of Plant Science, Politeknik Negeri Lampung, Lampung, Indonesia

^e DIADE, IRD, Univ Montpellier, CIRAD, Montpellier, France

ARTICLE INFO

Article history:

Received 21 September 2022

Revised 7 December 2022

Accepted 12 December 2022

Available online 17 December 2022

Dataset link: [RujakBase project - Myristica database for Whole Genome and Transcriptome Studies \(Original data\)](#)

Keywords:

Conservation

Genetic diversity

Long-reads sequencing

Spices

Full-length transcript

ABSTRACT

Myristica fatua is a tropical fruit tree species originating from Indonesia. Very few genomic resources are available for the species. We developed a full-length transcriptome assembly using long-read sequencing (MinION Nanopore technology) and produced 4.3 million reads (3.5 G of bases). The assembled full-length transcript was constructed using the RAT-TLE program and assembled 21,098 transcripts. The transcript ranged from 201 – 14,174 bp, and N50 was 2,017 bp. The transcripts were annotated with the UNIPROT database using BlastX. The functional annotation was performed using Blast2go software. The 8,445 microsatellite motif-containing contigs were identified. The raw reads are deposited in the ENA (European Nucleotide Archive) with ENA experiment accession number ERX6798613.

© 2022 The Author(s). Published by Elsevier Inc.

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

* Corresponding author.

E-mail address: dedenmatra@apps.ipb.ac.id (D.D. Matra).

Social media: [@matradeden](#) (D.D. Matra)

Specifications Table

Subject area	Agronomy and Crop Science
More specific subject area	Spices
Type of data	RNA sequencing Data
How data was acquired	MinION Nanopore Sequencer
Data format	Raw Sequencing reads, Table, Figure
Description of data collection	RNA sequencing was performed by using MinION Nanopore on leaves from the seedling stage
Data source location	Dramaga, Bogor, West Java, Indonesia
Data accessibility	The data have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number ERX6798613 (https://www.ebi.ac.uk/ena/browser/view/ERX6798613) or (https://identifiers.org/insdc.sra:ERX6798613)
Related research article	Kusuma, J., Scarcelli, N., Couderc, M. et al. Microsatellite markers development for Indonesian nutmeg (<i>Myristica fragrans</i> Houtt.) and transferability to other Myristicaceae spp.. Mol Biol Rep 47, 4835–4840 (2020). https://doi.org/10.1007/s11033-020-05535-y [1]

Value of the Data

- This data provides *Myristica fatua* coding sequence (CDS) as the first transcriptome reference using Oxford Nanopore Technologies of long-read sequencing
- This data could benefit studies to identify full-length transcripts related to flavonoid biosynthesis for molecular biologists that are used for downstream analysis in *Myristica fatua* and related genera.
- This data provides datasets of EST-microsatellite molecular markers for the breeder to improve crop breeding programs in *Myristica fatua*-related genera.
- The raw sequencing data may be carried out further in differential expressed gene study.

Objective

This plant is one of the best potential spices sources from Indonesia. However, genetic information such as transcriptome data is not yet available. Therefore, these data were used to obtain transcriptome information from leaves in the seedling phase of *M. fatua*. The transcripts' results were obtained using long-read sequencing from oxford nanopore technology. This data is able to provide full-length transcripts that are useful for studying gene expression analysis.

1. Data Description

In this data, full-length transcripts were sequenced from *Myristica fatua* using long-read sequencing. The total RNA was extracted from the leaves on the seedling stage with high-quality total RNA. The full length was obtained with raw data produced 4.3 million reads (3.5G of bases) [2]. The raw reads are deposited in the ENA database with the accession number ERX6798613 [3]. The clean reads were filtered by pychopper and cutadapt programs. The de novo assembly was constructed using the RATTLE program and produced 21,098 transcripts [4]. All statistics of reads and assembled transcripts were analyzed (Table 1). The transcripts were annotated with a filtered-UNIPROT database using the BLAST+ v.2.7.1 program [5] and processed by Blast2go software (Table 2) [6–8]. An overview of *Myristica fatua* Gene Ontology (GO) classification is presented in Fig. 1a for Biological Process, Fig. 1b for Molecular Function, and Fig. 1c for Cellular Component [9] and KEGG pathways [10]. Open reading frames (ORFs) from transcripts were determined using the TransDecoder program (Table 3) [11]. The distribution of the identified EST-SSRs in transcripts was performed using the MISA program (Table 4) [12].

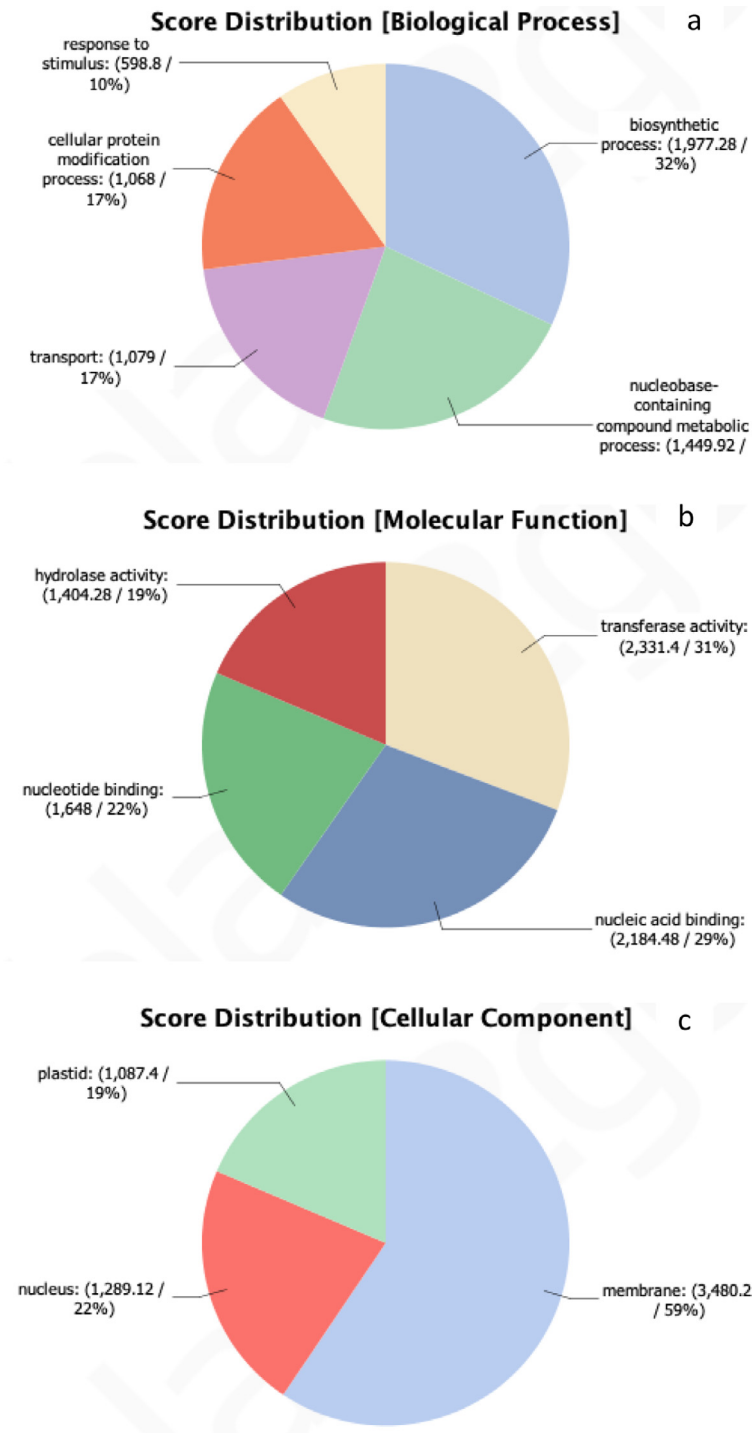


Fig. 1. Gene Ontology (GO) classification of *Myristica fatua* for (a) Biological Process, (b) Molecular Function, (c) Cellular Component.

Table 1
Read and assembly statistics of *Myristica fatua* leaves.

Features	Numbers
Number of reads	4,379,800
Number of bases	3,467,934,331
Mean read length (bp) and quality	791.8 / 12.2
Read length N50 (bp)	1,010
Number and bases total (bp) of transcripts	21,098/ 34,073,111
Length range, average (bp), and N50 (bp) of transcripts	207 – 14,174 / 1,614.99 / 2,017

Table 2
Functional annotation summary of *Myristica fatua* transcripts.

Database Source	Number of transcripts (percentage)
- UniProt	17,038 (81%)
- GO Mapping	2,654 (13%)
- GO Annotation	11,004 (52%)
- GO EnzymeCode	5,351 (25%)
- KEGG	130 pathways

Table 3
Open Reading Frames (ORFs) prediction characteristics of *Myristica fatua* transcripts.

Features	transcripts Number (percentage)
- ORF transcripts	22,665
- ORFs Type:	
a. 5'prime_partial	3,855 (17.01%)
b. 3'prime_partial	684 (3.02%)
c. Internal	64 (0.28%)
d. Complete	18,062 (79.69%)

Table 4
Number of microsatellite regions observed in *Myristica fatua* transcripts.

Characteristics	Number
- Total number of identified SSRs	8,445
- Number of SSR containing transcripts	6,644
- transcripts containing more than 1 SSR	1,175
- SSRs present in compound motif	1,015
Motif	
- Dinucleotide	7,034
- Trinucleotide	1,285
- Tetranucleotide	1,194
- Pentanucleotide	81
- Hexanucleotide	16
- Heptanucleotide	29

2. Experimental Design, Materials and Methods

2.1. Sample collection

Myristica fatua leaves were collected from Karmanah's garden (S6°33'15.372'' E106°42'19.734'', Bogor, West Java, Indonesia).

2.2. Total RNA extraction

The total RNA from young leaves was extracted using the RNeasy PowerPlant Kit (Qiagen) following the manufacturer's protocol. The quality and quantity of RNA were checked by Nanophotometer NP-80 (Implen) and Qubit™ RNA Broad Range (BR) assay on Qubit® Fluorometer (Invitrogen).

2.3. Transcriptome sequencing and De novo assembly

The total RNA was subjected to RNA sequencing using PCR-cDNA Barcoding - SQK-PCB109 (PCB_9092_v109_revB_10Oct2019). The sequencing was performed on a Flow Cell R9.4.1 (FLO-MIN106D) on MinION Mk1B. After sequencing, the raw reads were base called using Guppy 6.1.2 with default parameters. Next, data pre-processing followed <https://github.com/felixgrunberger/microbepore> protocol includes demultiplexing and NanoStat v1.2.1 to assess the reads quality and reads' statistics [13]. Next, full-length reads with remaining SSP (strand-switching primer) and VNP (oligo-dT30VN) primers were identified using pypochopper v2.5.0 (<https://github.com/nanoporetech/pypochopper>). Then, polyA-tails and the remaining SSP adapters were removed using Cutadapt [14]. De novo assembly was then performed on clean full-length reads using the RATTLE program [15] (Table 1).

2.4. Functional annotation

The full-length polished transcripts were BLASTed using BlastX [16] with a cut-off of 10^{-5} using the filtered-UNIPROT database (Magnoliopsida (TaxID: 3398), downloaded on 19 October 2021). The blasted output was performed using Blast2Go software [17]. Open reading frames (ORFs) transcripts were predicted by the TransDecoder (<https://github.com/TransDecoder/TransDecoder>) with default parameters [18]. Microsatellite regions were observed using MISA software [19] (<http://pgrc.ipk-gatersleben.de/misa>) with modified parameters [20].

Ethics statements

Not applicable.

CRediT author statement

Deden Derajat Matra: Conceptualization, Methodology, Writing - Original Draft, Funding acquisition, **M Adrian:** Validation, Investigation, **Karmanah:** Resources, Writing - Review & Editing, **Jakty Kusuma:** Data Curation, Writing - Review & Editing, **Jérôme Duminil:** Conceptualization, Writing - Review & Editing, **Sobir:** Supervision, Funding acquisition, **Roedhy Poerwanto:** Supervision, Conceptualization, Funding acquisition.

Declaration of Competing Interest

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.

Data Availability

RujakBase project - Myristica database for Whole Genome and Transcriptome Studies (Original data) (European Nucleotide Archive (ENA)).

RujakBase project - Myristica database for Whole Genome and Transcriptome Studies (Original data) (European Nucleotide Archive (ENA)).

Acknowledgments

Funding: This research was supported by the Ministry of Education, Culture, Research, and Technology (KEMENDIKBUDRISTEK), the Republic of Indonesia, through the National Competitive of Fundamental Research Scheme (PDKN-Penelitian Dasar Kompetitif Nasional) to Deden Derajat Matra as PI with grant number: 0267/E5/AK.04/2022.

References

- [1] J. Kusuma, N. Scarcelli, M. Couderc, et al., Microsatellite markers development for Indonesian nutmeg (*Myristica fragrans* Houtt.) and transferability to other Myristicaceae spp, *Mol. Biol. Rep.* 47 (2020) 4835–4840, doi:[10.1007/s11033-020-05535-y](https://doi.org/10.1007/s11033-020-05535-y).
- [2] European Bioinformatics Institute. 2022. <https://identifiers.org/insdc.sra:ERX6798613>. Accessed November 3, 2022.
- [3] D.D. Matra, Reads quality statistics of *Myristica fatua*, Figshare (2022) [10.6084/m9.figshare.21526272](https://doi.org/10.6084/m9.figshare.21526272).
- [4] D.D. Matra, Transcriptome assembly transcripts of *Myristica fatua*, Figshare (2022) [10.6084/m9.figshare.21524457](https://doi.org/10.6084/m9.figshare.21524457).
- [5] D.D. Matra, Functional annotation from protein sequence database of Uniprot, *Myristica fatua*, Figshare (2022) [10.6084/m9.figshare.21524646](https://doi.org/10.6084/m9.figshare.21524646).
- [6] D.D. Matra, E-value distribution of *Myristica fatua*, Figshare (2022) [10.6084/m9.figshare.21557838](https://doi.org/10.6084/m9.figshare.21557838).
- [7] D.D. Matra, Similarity distribution of *Myristica fatua*, Figshare (2022) [10.6084/m9.figshare.21557841](https://doi.org/10.6084/m9.figshare.21557841).
- [8] D.D. Matra, Species distribution of *Myristica fatua*, Figshare (2022) [10.6084/m9.figshare.21557844](https://doi.org/10.6084/m9.figshare.21557844).
- [9] D.D. Matra, Gene Ontology results of *Myristica fatua*, Figshare (2022) [10.6084/m9.figshare.21524634](https://doi.org/10.6084/m9.figshare.21524634).
- [10] D.D. Matra, KEGG results of *Myristica fatua*, Figshare (2022) [10.6084/m9.figshare.21557931](https://doi.org/10.6084/m9.figshare.21557931).
- [11] Matra D.D. Open Reading Frames (ORFs) prediction of *Myristica fatua*. Figshare. 2022 [10.6084/m9.figshare.21558750](https://doi.org/10.6084/m9.figshare.21558750).
- [12] D.D. Matra, Results of microsatellite region finding of *Myristica fatua*, Figshare (2022) [10.6084/m9.figshare.21524607](https://doi.org/10.6084/m9.figshare.21524607).
- [13] W. De Coster, S. D'Hert, D.T. Schultz, M. Cruts, C Van Broeckhoven, NanoPack: visualizing and processing long-read sequencing data, *Bioinformatics* 34 (15) (2018) 2666–2669, doi:[10.1093/bioinformatics/bty149](https://doi.org/10.1093/bioinformatics/bty149).
- [14] Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, [S.I.]17, 1, 10–12. doi: [10.14806/ej.17.1.200](https://doi.org/10.14806/ej.17.1.200).
- [15] I. de la Rubia, A. Srivastava, W. Xue, J.A. Indi, S. Carbonell-Sala, J. Lagarde, M.M. Albà, E. Eyra, RATTLE: reference-free reconstruction and quantification of transcriptomes from Nanopore long-read sequencing, *Genome Biol* 23 (2022) 153, doi:[10.1186/s13059-022-02715-w](https://doi.org/10.1186/s13059-022-02715-w).
- [16] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, *J. Mol. Biol.* 215 (1990) 403–410.
- [17] A. Conesa, S. Götz, J.M. García-Gómez, J. Terol, M. Talón, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics* 21 (18) (2005) 3674–3676, doi:[10.1093/bioinformatics/bti610](https://doi.org/10.1093/bioinformatics/bti610).
- [18] F. Indriani, U.J. Siregar, D.D. Matra, I.Z. Siregar, De novo transcriptome datasets of *Shorea balangerana* leaves and basal stem in waterlogged and dry soil, *Data Brief* 28 (2019) 104998, doi:[10.1016/j.dib.2019.104998](https://doi.org/10.1016/j.dib.2019.104998).
- [19] T. Thiel, W. Michalek, R. Varshney, A. Graner, Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.), *Theor. Appl. Genet.* 106 (2003) 411–422, doi:[10.1007/s00122-002-1031-0](https://doi.org/10.1007/s00122-002-1031-0).
- [20] D.D. Matra, M.A.N. Fathoni, M. Majiidi, et al., The genetic variation and relationship among the natural hybrids of *Mangifera casturi* Kosterm, *Sci. Rep.* 11 (2021) 19766, doi:[10.1038/s41598-021-99381-y](https://doi.org/10.1038/s41598-021-99381-y).