BMC Research Notes

DATA NOTE Open Access

De novo assembly of transcriptome dataset from leaves of *Dryobalanops aromatica* (Syn. *Dryobalanops sumatrensis*) seedlings grown in two contrasting potting media

Iskandar Zulkarnaen Siregar[1*], Fifi Gus Dwiyanti[1], Ulfah Juniarti Siregar[1] and Deden Derajat Matra[2]

Abstract

Objectives: Efforts to restore tropical peat swamp forests in Indonesia face huge challenges of potential failures due to socio-economic factors and ecological dynamics attributed to lack of knowledge on the adaptive mechanisms of potential tree species such as Kapur (*Dryobalanops aromatica* C.F.Gaertn Syn. *Dryobalanops sumatrensis* J.F. Gmelin A.J.G.H Kostermans). This species is a multi-purpose tree that, commonly grows in mineral soils, but also in peat swamp as previously reported, which raised a fundamental question regarding the molecular mechanism of this adaptation. Therefore, a dataset was created aiming to detect candidates of adaptive genes in *D. aromatica* seedlings, cultivated in two contrasting potting media, namely mineral soil and peat media, based on RNA Sequencing Transcriptome Analysis.

Data description: The RNA transcriptome data of *D. aromatica*'s seedlings derived from young leaves of three one-year-old seedlings, raised in each dry mineral soil media and peat media, were generated by using Illumina HiSeq 4000 platform in NovogenAIT, Singapore. The acquired data, as the first transcriptome dataset for *D. aromatica,* is of a great importance in understanding molecular mechanism and responses of the involved genes of *D. aromatica* to the contrasting, growing potting media conditions that could also be useful to generate molecular markers.

Keywords: Adaptive genetic variation, *Dryobalanops aromatica*, *Dryonalanops sumatrensis*, Peat swamp, Transcriptome

Objective

The past genetic research on *Dryobalanops aromatica* focused on pattern of genetic variation and population structure in North-eastern Borneo, Sumatera, and the Malay Peninsula using nuclear microsatellite markers [1]. The investigated ecosystem types for all populations were from mineral soil forest types, in which *D. aromatica* could be found abundantly on deep, humid, yellow, sandy soils with a propensity for ridges [2]. However, it was recently discovered that this species also grows in peat swamp forest, as found in Singkil Wildlife Reserve (Suaka Margasatwa Singkil), Aceh, Sumatera. According to this finding, the former investigation was then concentrated on how to understand life-history characteristics such as comparing shoot cuttings ability of *D. aromatica* in peat and coco peat media [3]. In addition, due to lack of in-depth investigation of adaptive genetic variation of this species grown in mineral soil and peat media, an experiment was carried out through RNA sequencing (RNA-Seq) transcriptome analysis. Studies on adaptive genetic analysis using RNA-Seq in tropical forest trees

*Correspondence: siregar@apps.ipb.ac.id
[1] Department of Silviculture, Faculty of Forestry and Environment, IPB University (Bogor Agricultural University), Bogor, Indonesia
Full list of author information is available at the end of the article

have previously been reported, such as research on *Shorea balangeran* adaptation grown in mineral and peat potting media [4] and gall-rust infected and uninfected trees of *Falcataria moluccana* [5]. Considering potential application of transcriptome analysis on forest trees, similar research was also conducted on *D. aromatica*. Objective of the research was to detect candidates of adaptive genes in *D. aromatica* seedlings, grown in two contrasting potting media, namely mineral soil and peat media. The findings were expected to provide more accurate information on molecular adaptive mechanism for practical use to support rehabilitation and conservation of degraded peat swamp forests in Indonesia. Results of the study are presented in Table 1.

Data description

Dryobalanops aromatica's seedlings, collected from Lae Kombih Forest Park, Aceh, Sumatera and transported to greenhouse of Department of Silviculture, IPB University, Bogor, were treated under two contrasting types of potting (diameter 10 cm) fine media, i.e., mineral soil (n = 3 seedlings) and peat (n = 3 seedlings) with regular watering. Peat media was classified as fibric peat, which has pH of 4.0 and 135.32% water content, whereas mineral soil media is classified as clay loam soil which has pH of 5.0 and 32.09% water content. Total RNA from young leaves collected from three one-year-old seedlings cultivated in each mineral soil media and peat media were extracted by using Plant Total RNA Mini Kit (Geneaid Biotech Ltd), following manufacturer's instructions. The integrity and quantity of extracted-RNA were measured by using NanoDrop ND-1000 spectrophotometer and Agilent 2100 Bioanalyzer.

The RNA sequencing was undertaken using Illumina HiSeq 4000 (Novogene-AIT, Singapore) that produced pre-processing reads, which afterwards became subjects to discard the library adaptors and low-quality reads below $Q < 30$ (data set 1). The clean reads were de novo assembled by Trinity 2.3.2 [6], and the redundant transcripts were removed using CAP3, cd-hit-est, and corset 1.08, respectively [7–9]. Sequencing the yielded 221 million reads produced total 114,268 contigs. The contigs

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Transcriptome assembly contigs	Fasta file (.fasta)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5 [19]
Data file 2	Summary for alignment of clean reads to reference transcriptome	Document file (.docx)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5
Data file 3	Functional annotation from non-redundant nucleotide NCBI	BLAST output in txt/-outfmt 6 option (.txt)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5
Data file 4	Functional annotation from non-redundant protein NCBI	BLAST output in XML/-outfmt 5 option (.xml)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5
Data file 5	Functional annotation from protein sequence database of SwissProt	BLAST output in XML/-outfmt 5 option (.xml)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5
Data file 6	Functional annotation from protein sequence database of TrEMBL	BLAST output in txt/-outfmt 6 option (.txt)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5
Data file 7	Statistics related to contig length distribution and the Blast results: e-value distribution, contig similarity distribution, top-hit species distribution	PNG files in compressed file (.zip)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5
Data file 8	Functional annotation from Complete TREP nucleotide database	BLAST output in txt/-outfmt 6 option (.txt)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5
Data file 9	Functional annotation from Hypothetical TREP protein database	BLAST output in txt/-outfmt 6 option (.txt)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5
Data file 10	Assessing transcriptome assembly and annotation completeness with single-copy orthologs by BUSCO	BUSCO output in txt (.txt)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5
Data file 11	Gene Ontology and KEGG analysis	Blast2GO file (.b2g)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5
Data file 12	Open Reading Frames (ORFs) prediction	Fasta File (.fasta)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5
Data file 13	Results of microsatellite region finding	Misa file (.txt)	Figshare https://doi.org/10.6084/m9.figshare.12326177.v5
Data set 1	Raw RNA-seq. reads and assembled contigs	Fastq files (.fastq)	DNA Data Base of Japan (https://identifiers.org/insdc.sra:DRP005979 [20]

ranged from 201 to 50,886 base pairs with N50 of 1970 bp (data file 1). To assess the quality of transcriptome reference, clean reads were mapped to reference using Bowtie2 [10] (Data file 2).

The functional annotation of contigs was performed using BLAST + 2.7.1 program against the NCBI nr (data file 3), NCBI nt (data file 4) (downloaded by 6th October 2018 and subjected to Euphyllophyta) and SwissProt (data file 5) and TrEMBL (data file 6) (downloaded by 3rd January 2020) databases with an E-value cutoff of 10^{-5} [11, 12]. Statistics of transcriptome reference were analyzed using Blast2GO 5.2 [13] that produced statistics of length distribution and Blast results with NCBI nr as follows: e-value distribution, contig similarity distribution and top-hit species distribution (data file 7). Functional analysis showed that 80,507 (70.45%) indicated significant matches with NCBI nr as well as 59,353 (51,94%) in the SwissProt database. The transposon sequence analysis was analyzed using BLAST program with TREP database [14] (data file 8, data file 9). Transcriptome reference was assessed using Busco v.3.2 [15] under Maser platform [16] (data file 10). The SwissProt-annotated contigs were used to analyze GO and KEGG pathways using Blast2GO 5.2 (data file 11).

To predict ORFs, the contigs were analyzed using TransDecoder 5.5.0 [17] (data file 12). A total of 84,175 contigs was identified as ORFs with 5′prime partial of 13,430 (15,95%), 3′prime partial of 8574 (10,19%) and complete ORFs type of 57,306 (68,08%). Contigs containing microsatellite were extracted by using the MISA program [18], with minimum repeats such as: 10 for one base, 6 for two bases, and 5 for 3, 4, 5, and 6 bases; and the interruptions between sites of microsatellite were 100 bases. The microsatellite motifs containing contigs were summed up to 39,025 (data file 13).

Limitations

The seedlings were not collected directly from the field due to the lack of natural regeneration and remarkably lengthy distance. Rather, seedlings were treated in two types of potting media (i.e. mineral and peat) grown in the green house with regular maintenance. Furthermore, RNA extraction samples were obtained from the leaves, only leaving other plant parts to be analyzed for better comparisons due to already established RNA extraction methods for the leaves. The extraction was also carried out solely once during sampling point in order to meet the sufficient replicates.

Abbreviations

RNA-Seq: RNA sequencing transcriptome analysis; nr: Non-redundant protein; nt: Nucleotide sequences; TREP: The TRansposable Elements Platform; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; BP: Biological processes; MF: Molecular function; CC: Cellular component; ORFs: Open reading frames.

Acknowledgements

The authors thank Mr. Abdul Rahman Ali from Subulussalam, Aceh, for providing *Dryobalanops aromatica* seedlings and Mr. Riya Kamba from Singkil-Aceh, for assisting in seedlings transfer from Aceh to Bogor. High appreciation goes to Ms. Ridha Wati for contributing in research material preparation in greenhouse of Department of Silviculture, Faculty of Forestry and Environment, IPB University.

Authors' contributions

IZS designed experiments and managed the study. FGD performed experimental treatments and managed RNA extraction and RNA sequencing analysis. DDM performed, analyzed and interpreted the RNA-sequencing data. DDM and IZS fabricated the first draft of manuscript, whilst FGD and UJS made major contributions to the writing. All authors reviewed and discussed the contents of the manuscript. All authors read and approved the final manuscript.

Funding

This research was supported by the Basic Research 2019–2020 (International Collaboration and International Publication Scheme No: 3/E1/KP.PTNBH/2019 and No: 1/AMD/E1.KP.PTNBH/2020) titled "Adaptive Genetic Divergence on *Dryobalanops aromatica* Growing in Sumatran Tropical Peat Swamp Forests" awarded by Ministry of Research and Technology/National Agency for Research and Innovation (RISTEK-BRIN) of the Republic of Indonesia. Funding is used to cover research design, laboratory expenses, sample preparation, sequencing, data collection, analysis of results, manuscript writing, and article processing charge.

Availability of data and materials

The data described in this Data note can be freely and openly accessed on figshare (https://doi.org/10.6084/m9.figshare.12326177.v5) and DNA Data Base of Japan (https://identifiers.org/insdc.sra:DRP005979). Please see Table 1 and references list [19, 20] for details and links to the data.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Silviculture, Faculty of Forestry and Environment, IPB University (Bogor Agricultural University), Bogor, Indonesia. [2] Department of Agronomy and Horticulture, Faculty of Agriculture, IPB University (Bogor Agricultural University), Bogor, Indonesia.

Received: 30 May 2020 Accepted: 21 August 2020
Published online: 28 August 2020

References

1. Harada K, Dwiyanti FG, Siregar IZ, Subiakto A, Chong L, Diway B, Lee YF, Ninomiya I, Kamiya K. Genetic variation and genetic structure of two closely related Dipterocarp species, *Dryobalanops aromatica* C.F. Gaertn and *D. beccari* Dyer. Sibbaldia. 2018;16:179–97. https://doi.org/10.23823/Sibbaldia/2018.255.
2. Ashton PS. Dipterocarpaceae. In: Soepadmo E, Saw LG, Chung RCK, editors. Tree flora of Sabah and Sarawak, vol. 5. Malaysia: Forest Research Institute; 2004. p. 388.
3. Siregar IZ, Kustiyarini NF, Wati R, Rachmat HH, Siregar UJ, Dwiyanti FG. Vegetative propagation of *Dryobalanops sumatrensis* and *Dryobalanops oblongifolia* subsp. *oblongifolia* by shoot cuttings. IOP Conf

Ser Earth Environ Sci. 2019;394:012029. https://doi.org/10.1088/1755-1315/394/1/012029.
4. Indriani F, Siregar UJ, Matra DD, Siregar IZ. De novo transcriptome datasets of *Shorea balangeran* leaves and basal stem in waterlogged and dry soil. Data Brief. 2020;28:104998. https://doi.org/10.1016/j.dib.2019.104998.
5. Shabrina H, Siregar UJ, Matra DD, Siregar IZ. The dataset of de novo transcriptome assembly of Falcataria moluccana cambium from gall-rust (*Uromycladium falcatarium*) infected and non-infected tree. Data Brief. 2019;26:104489. https://doi.org/10.1016/j.dib.2019.104489.
6. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52. https://doi.org/10.1038/nbt.1883.
7. Huan X, Madan A. CAP3: a DNA sequence assembly program. Genome Res. 1999;9(9):868–77. https://doi.org/10.1101/gr.9.9.868.
8. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9. https://doi.org/10.1093/bioinformatics/btl158.
9. Davidson MN, Oshlack A. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biol. 2014;15(7):410. https://doi.org/10.1186/s13059-014-0410-6.
10. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.
11. Matra DD, Kozaki T, Ishii K, Poerwanto R, Inoue E. Comparative transcriptome analysis of translucent flesh disorder in mangosteen (*Garcinia mangostana* L.) fruits in response to different water regimes. PLoS ONE. 2019;14(7):e021997. https://doi.org/10.1371/journal.pone.0219976.
12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
13. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization, and analysis in functional genomics research. Bioinformatics. 2005;21(18):3674–6. https://doi.org/10.1093/bioinformatics/bti610.
14. Wicker T, Matthews DE, Keller B. TREP: a database for Triticeae repetitive elements. Trends Plant Sci. 2002;7:561–2. https://doi.org/10.1016/S1360-1385(02)02372-5.
15. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2018;35(3):543–8. https://doi.org/10.1093/molbev/msx319.
16. Kinjo S, Monma N, Misu S, Kitamura N, Imoto J, Yoshitake K, Gojobori T, Ikeo K. Maser: one-stop platform for NGS big data from analysis to visualization. Database. 2018;2018:1–12. https://doi.org/10.1093/database/bay027.
17. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494–512. https://doi.org/10.1038/nprot.2013.084.
18. Thiel T, Michalek M, Varshney RK, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet. 2003;106(3):411–22. https://doi.org/10.1007/s00122-002-1031-0.
19. Siregar, IZ, Dwiyanti FG, Siregar UJ, Matra DD. De novo assembly of transcriptome dataset from leaves of *Dryobalanops aromatica* (Syn. *Dryobalanops sumatrensis*) seedlings grown in two contrasting potting media. *Figshare*. 2020. https://doi.org/10.6084/m9.figshare.12326177.v5.
20. DNA Data Bank of Japan; 2020. https://identifiers.org/insdc.sra:DRP005979.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.