BMC Research Notes

DATA NOTE

Open Access

# Generating long-read sequences using Oxford Nanopore Technology from *Diospyros celebica* genomic DNA

Iskandar Zulkarnaen Siregar[1,2*], Fifi Gus Dwiyanti[1,2], Rahadian Pratama[2,3], Deden Derajat Matra[2,4] and Muhammad Majiidu[2]

## Abstract

**Objectives:** Development of sequencing technology has opened up vast opportunities for tree genomic research in the tropics. One of the aforesaid technologies named ONT (Oxford Nanopore Technology) has attracted researchers in undertaking testings and experiments due to its affordability and accessibility. To the best of our knowledge, there has been no published reports on the use of ONT for genomic analysis of Indonesian tree species. This progress is promising for further improvement in order to acquire more genomic data for research purposes. Therefore, the present study was carried out to determine the effectiveness of ONT in generating long-read DNA sequences using DNA isolated from leaves and wood cores of Macassar ebony (*Diospyros celebica* Bakh.).

**Data description:** Long-read sequences data of leaves and wood cores of Macassar ebony were generated by using the MinION device and MinKnow v3.6.5 (ONT). The obtained data, as the first long-read sequence dataset for Macassar ebony, is of great importance to conserve the genetic diversity, understanding the molecular mechanism, and sustainable use of plant genetic resources for downstream applications.

**Keywords:** *Diospyros celebica*, Genome, ONT, Scaffolds, Long-read sequences

## Objective

The third-generation sequencing from Oxford Nanopore Technologies (ONT) that is capable of generating long-read sequences was applied to fill the existing technological gaps, particularly with respect to capital cost, use of native DNA/RNA samples, simplicity, portability, ease of use for library preparation, etc. In particular, these technologies provided an on-site analysis that is of a significant advantage considering possible constraints due to existing gaps in the current regulation (e.g. Nagoya Protocol and others) on sample transfer permits either from field to laboratory, both within the country and overseas [1, 2]. The use of ONT on-site required fewer efforts in arranging the administration process for sample transfer, hence these functions enabled to accelerate data generation for various immediate needs, ones of which were urgent decision-making for species identification and conservation and even for on-site forensic investigation. The ONT could also be used in a hybrid system with other sequencing platforms, such as short-read sequencing in order to analyze missing fragments, structural variations, etc. [3]. In the tropics, research on the use of ONT to dissect biodiversity has been still limited due to the new finding scarcity, especially in regards to tree genomic variation analysis. Associated problems such as DNA/RNA yields and quality have been still consistently found depending on species and sample sources led by mainly more complex chemical compounds (such as phenols) and samples' accessibility. In addition, site conditions

*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

might also influence the DNA yields forcing the use of only one general protocol across the samples. Macassar ebony—an endemic and vulnerable species in Sulawesi (Celebes), Indonesia, was utilized in the experiment and designed to determine the utilization efficacy aiming for long-read sequencing using samples from both leaves and small wood cores collected in Celebes [4]. Results of this study are presented in Table 1.

## Data description

Total genomic DNA from 15 individuals of Macassar ebony (*Diospyros celebica* Bakh.) leaves (n = 11) and wood core (n = 4) collected by using Pickering Punch in three provinces in Indonesia, namely Central Sulawesi, West Sulawesi, and South Sulawesi, were extracted using a modified CTAB methods [5] in which the CTAB buffer contained CTAB 10%, Tris HCl, NaCl 5 M, EDTA 0.5 M, PVP 1%, β-Mercaptoethanol, and $dH_2O$. DNA quality was evaluated by electrophoresis using a Gel Doc EZ System (Bio-Rad, USA) and DNA concentration was measured by using NanoPhotometer NP80 (IMPLEN, Germany).

The library preparation of genomic DNA sample was followed the Nanopore Protocol for Native barcoding genomic DNA (with EXP-NBD104 and SQK-LSK109), version NBE_9065_v109_revJ_23May2018. Sequencing was done in two rounds using two flowcells (FLO-MIN106). The list of samples per flowcell, as well as the native barcode (NBD01–NBD12) used in the study, were listed in Data File 1.

The sequencing run of genomic DNA samples was performed using the MinION device and MinKnow v3.6.5. Sequencing was terminated after no more pores actively sequenced the DNA. The high-accuracy base-calling mode was used to base-call the signal in FAST5 files and outputted FASTQ files. Samples were separated according to each barcode, where afterwards the barcodes were set to automatically trimmed from the reads (Data set 1). All samples were combined using *cat* command on Linux Mint terminal and analyzed by using NanoStat v1.2.1 to assess the reads quality and reads' statistics. Meanwhile, distribution plots were generated by using NanoPlot v1.31.0 [6] (Data file 2). We obtained 302 567 reads with 99.5% reads quality > Q7 (nanopore default passed quality). After statistic inspection, all reads quality was filtered through NanoFilt v2.7.1 [6]. Reads with Q-score lower than 7 and less than 500 bp were filtered out, with parameter *–headcrop* and *–tailcrop* of 10 were applied. Reads filtering resulted in 134 220 reads, then subject to correction, trimming and De novo assembly using Canu v2.0 [7] with option of *genome Size = 800 m.* Another De novo long-reads assembler was applied to compare the contig assemblies from plant DNA using SMARTdenovo [8] with minimum read length (*−J*) 2 000. SMARTdenovo utilized corrected reads 'step from Canu correction stage, thus expected to result in better outcome than the Canu assembly's. The contig assemblies were 358 (N50 6.5 kb, GC 39.91%) and 39 (N50 12.7 kb, GC 41.14%) for Canu and SMARTdenovo respectively. The draft assembly then was polished (corrected) against the individual sequencing reads using medaka_consensus v1.0.3 [9] with parameter model for nanopore sequencing (*−m*) *r941_min_high_g330* (Data file 3). The resulting polished assembly statistics was calculated using QUAST v5.0.2

**Table 1 Overview of data files/data sets**

| Label | Name of data file/data set | File types (extension) | Data repository and identifier (DOI or accession number) |
|---|---|---|---|
| Data file 1 | List of ebony samples and Nanopore Native barcodes | Compressed XLSX file (.zip) | https://doi.org/10.6084/m9.figshare.13027991.v1 [13] |
| Data file 2 | Raw reads statistics and plots (quality score, reads length, reads distribution) | Compressed TXT, PNG and HTML files (.zip) | https://doi.org/10.6084/m9.figshare.13028069.v1 [14] |
| Data file 3 | Polished De novo genome contig assemblies of both assembler | Compressed FASTA files (.zip) | https://doi.org/10.6084/m9.figshare.13031195.v1 [15] |
| Data file 4 | Statistics of corrected contig assemblies with *Diospyros celebica* chloroplast | Compressed PDF file (.zip) | https://doi.org/10.6084/m9.figshare.13028177.v1 [16] |
| Data file 5 | Statistics of corrected contig assemblies with *Diospyros lotus* genome | Compressed PDF file (.zip) | https://doi.org/10.6084/m9.figshare.13028180.v1 [17] |
| Data file 6 | Constructed scaffolds from both assembler | Compressed FASTA files (.zip) | https://doi.org/10.6084/m9.figshare.13031702.v1 [18] |
| Data file 7 | Statistics of scaffolds assemblies with *Diospyros celebica* chloroplast | Compressed PDF file (.zip) | https://doi.org/10.6084/m9.figshare.13031693.v1 [19] |
| Data file 8 | GenBank annotation from scaffolds assemblies | Compressed GB files (.zip) | https://doi.org/10.6084/m9.figshare.13031708.v1 [20] |
| Data file 9 | Annotation visualization from scaffolds assemblies | Compressed JPG files (.zip) | https://doi.org/10.6084/m9.figshare.13031714.v1 [21] |
| Data set 1 | Raw genomic DNA reads (trimmed barcode) | FASTQ files (.fastq) | DNA Data Base of Japan (DRP006615) https://identifiers.org/insdc.sra:DR006615 [22] |

[10], with references of *Diospyros celebica* chloroplast (Data file 4) and *Diospyros lotus* genome (Data file 5). This statistic calculation informed how much reference genome fraction covered by the contig assemblies. The polished contigs then were chosen to construct scaffold using LINKS v1.8.7 [11] with default parameter (Data file 6). The resulted scaffolds were 266 (N50 11.3 kb, GC 39.91%) and 33 (N50 17.8 kb, GC 41.14%) for Canu and SMARTdenovo respectively. The longest scaffolds assemblies from both assembler (Canu 141.6 kb, SMARTdenovo 145 kb) were validated with QUAST v5.0.2 with *D. celebica* chloroplast to check the genome fraction that scaffolds covered (Data file 7). These scaffolds assemblies were then annotated by using GeSeq platform for Organellar Genomes [12], resulted in the GenBank annotation (Data file 8) and their visualization (Data file 9).

## Limitations

The long-read sequencing of the Macassar ebony tree equipped with nanopore sequencing was quite challenging. Extraction of genomic DNA shall be optimized to obtain high-quality gDNA without excessive fragmentation. The resulting fragmented DNA required to be removed prior to library preparation as they might occupy nanopores within the flowcells and cause too many short reads across the sequencing outputs. The library preparation shall be optimized as well, for example, the DNA concentration was measured with a spectrophotometer, which could lead to a biased number of the aforementioned concentration. DNA fluorometer was preferred to accurately calculate DNA concentration. The correct DNA concentration loaded into the MinION flowcell would enable the optimal DNA sequencing process and pores occupancy. Achieving higher sequencing throughput is necessary to improve the read accuracy limitation of MinION as been observed in this study.

### Abbreviations

ONT: Oxford Nanopore Technology; DNA: Deoxyribonucleic acid; RNA: Ribonucleic acid; CTAB: Cetyl Trimethyl Ammonium Bromide; HCl: Hydrochloric acid; NaCl: Sodium chloride; EDTA: Ethylenediaminetetraacetic acid; PVP: Polyvinylpyrrolidone; gDNA: Genomic deoxyribonucleic acid.

### Acknowledgements

The authors thank (i) South Sulawesi Natural Resources Conservation Agency (BKSDA) for providing research permission and assisting during sample collection in Cani Sirenreng Nature Park, Kalaena Nature Reserve, and Mappu-Gandang Dewata National Park, (ii) South Sulawesi Investment and One-stop Integrated Services (DPMPTSP) for providing research permission in Coppo Village, Belabori Secondary Forest, and Tana Toro Protection Forest (iii) West Sulawesi Investment and One-stop Integrated Services (DPMPTSP) for providing research permit in Batu Ampa Protection Forest, and Sondoang Village, and (iv) Central Sulawesi Investment and One-stop Integrated Services (DPMPTSP) for providing research permission in Wawopada Village and Sausu Village. High appreciation also goes to Molecular Science Laboratory, Advanced Research Laboratory, IPB University, Bogor, Indonesia for its contribution in providing lab facilities.

### Authors' contributions

IZS conceived and designed experiments of the study. MMD performed genomic DNA Extraction. FGD, RHP, DDM, and MMD performed experimental treatments and managed DNA sequencing analysis. RHP analyzed and interpreted the DNA sequencing data. RHP and IZS prepared the first draft of manuscript, while FGD, DDM, and MMD made major contributions in writing and editing the manuscript. All authors reviewed, discussed, and revised the contents of the manuscript. All authors read and approved the final manuscript.

### Funding

This research was supported by the World Resources Institute's (WRI) Forest Program grant (Project Code: 04682, period March 1st, 2019–September 30th, 2020) titled "Timber Tracking Project IPB-WRI: Collection of physical timber reference material and setting up a reference data building pipeline for DNA of commercial timber species (*Diospyros celebica*. Bakh)" awarded by Norwegian International Climate & Forest Initiative (NICFI). Funding is used to cover research design, sample collection, laboratory expenses, sample preparation, sequencing, data collection, and analysis of results.

### Availability of data and materials

The data described in this Data note can be freely and openly accessed on figshare [9, 10] (https://doi.org/10.6084/m9.figshare.13027991.v1, https://doi.org/10.6084/m9.figshare.13028069.v1, https://doi.org/10.6084/m9.figshare.13031195.v1, https://doi.org/10.6084/m9.figshare.13028177.v1, https://doi.org/10.6084/m9.figshare.13028180.v1, https://doi.org/10.6084/m9.figshare.13031702.v1, https://doi.org/10.6084/m9.figshare.13031693.v1, https://doi.org/10.6084/m9.figshare.13031708.v1, https://doi.org/10.6084/m9.figshare.13031714.v1, https://identifiers.org/insdc.sra:DRP006615). Please revert to Table 1 and references list [11–20] for details and links to the data.

### Ethics approval and consent to participate

Biological material samples in forms of dried leaves and wood cores were collected from (i) South Sulawesi following permit approvals from South Sulawesi Natural Resources Conservation Agency/BKSDA (No: SK.151/K.8/BIDTEK/KSA/5/2019 for Cani Sirenreng Nature Park and Kalaena Nature Reserve, and No: SK.295/K.8/BIDTEK/KSA/10/2019 for Mappu-Gandang Dewata National Park), and South Sulawesi Investment and One-stop Integrated Services/DPMPTSP (No: 16163/S.01/PTSP/2019 for Coppo Village, Bellabori Secondary Forest, and Tana Toro Protection Forest), (ii) West Sulawesi following permit approvals from West Sulawesi Investment and One-stop Integrated Services/DPMPTSP No. 000451/76.RP.PTSP.B/X/2019 for Batu Ampa Protection Forest, and Sondoang Village, and (iii) Central Sulawesi following permit approvals from Central Sulawesi Investment and One-stop Integrated Services/DPMPTSP No. 070/433/REK-PL/DPMPTSP/2019 for Wawopada Village and Sausu Village. The herbaria vouchers were identified by Mr Denny and are stored in Forest Research and Development Center and Nature Conservation (BZF)-Forestry and Environmental Research Development and Innovation Agency (FOERDIA) of the Ministry of Environment and Forestry of Republic of Indonesia (KLHK).

### 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] Molecular Science Research Group, Advanced Research Laboratory, IPB University (Bogor Agricultural University), Bogor, Indonesia. [3] Department of Biochemistry, Faculty of Mathematic and Natural Science, IPB University (Bogor Agricultural University), Bogor, Indonesia. [4] Department of Agronomy and Horticulture, Faculty of Agriculture, IPB University (Bogor Agricultural University), Bogor, Indonesia.

Received: 29 October 2020 Accepted: 12 February 2021
Published online: 27 February 2021

**References**

1. Runtuwene LR, Tuda JSB, Mongan AE, Suzuki Y. On-site MinION sequencing. Adv Exp Med Biol. 2019;1129:143–50. https://doi.org/10.1007/978-981-13-6037-4_10.
2. Mardiastuti A. Implementation of access and benefit sharing in Indonesia: review and case studies. JMHT. 2019;25(1):35–43. https://doi.org/10.7226/jtfm.25.1.35.
3. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;1(1):1–16. https://doi.org/10.1186/s13059-020-1935-5.
4. Karlinasari L, Noviyanti N, Purwanto YA, Majiidu M, Dwiyanti FG, Rafi M, Damayanti R, Harnelly E, Siregar IZ. Discrimination and determination of extractive content of ebony (*Diospyros celebica Bakh.*) from celebes island by near-infrared spectroscopy. Forests. 2021;12(1):6. https://doi.org/10.3390/f12010006.
5. Doyle JJ, Doyle JL. Isolation of plant DNA from fresh tissue. Focus. 1990;12:13–5.
6. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34(15):2666–9. https://doi.org/10.1093/bioinformatics/bty149.
7. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36. https://doi.org/10.1101/gr.215087.116.
8. Liu H, Wu S, Li A, Ruan J. SMARTdenovo: *A de novo* assembler using long noisy reads. Preprints. 2020. https://doi.org/10.20944/preprints202009.0207.v1.
9. Oxford Nanopore Technologies. Medaka: Sequence correction provided by ONT Research. 2018. https://github.com/nanoporetech/medaka. Accessed 23 Sep 2020.
10. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5. https://doi.org/10.1093/bioinformatics/btt086.
11. Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJM, Birol I. LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. GigaScience. 2015;4(1):35. https://doi.org/10.1186/s13742-015-0076-3.
12. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. GeSeq—versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45(W1):W6–11. https://doi.org/10.1093/nar/gkx391.
13. Siregar IZ, Pratama R, Dwiyanti FG, Matra DD, Muhammad M. Sample list and barcode. Figshare. 2021. https://doi.org/10.6084/m9.figshare.13027991.v1.
14. Siregar IZ, Pratama R, Dwiyanti FG, Matra DD, Muhammad M. Raw reads statistic and plot. Figshare. 2021. https://doi.org/10.6084/m9.figshare.13028069.v1.
15. Siregar IZ, Pratama R, Dwiyanti FG, Matra DD, Muhammad M. Contig assembly. Figshare. 2021. https://doi.org/10.6084/m9.figshare.13031195.v1.
16. Siregar IZ, Pratama R, Dwiyanti FG, Matra DD, Muhammad M. Contig assembly statistic 1. Figshare. 2021. https://doi.org/10.6084/m9.figshare.13028177.v1.
17. Siregar IZ, Pratama R, Dwiyanti FG, Matra DD, Muhammad M. Contig assembly statistic 2. Figshare. 2021. https://doi.org/10.6084/m9.figshare.13028180.v1.
18. Siregar IZ, Pratama R, Dwiyanti FG, Matra DD, Muhammad M. Scaffold assembly. Figshare. 2021. https://doi.org/10.6084/m9.figshare.13031702.v1.
19. Siregar IZ, Pratama R, Dwiyanti FG, Matra DD, Muhammad M. Scaffold statistic. Figshare. 2021. https://doi.org/10.6084/m9.figshare.13031693.v1.
20. Siregar IZ, Pratama R, Dwiyanti FG, Matra DD, Muhammad M. Genbank annotation. Figshare. 2021. https://doi.org/10.6084/m9.figshare.13031708.v1.
21. Siregar IZ, Pratama R, Dwiyanti FG, Matra DD, Muhammad M. Scaffold annotation visualization. Figshare. 2021. https://doi.org/10.6084/m9.figshare.13031714.v1.
22. DNA Data Bank of Japan. 2020. https://identifiers.org/insdc.sra:DRP006615. Accessed 27 Jan 2021.

## Publisher's Note