Data Article

Draft genome assembly dataset of the Basidiomycete pathogenic fungus, *Ganoderma boninense*

Suhaila Sulaiman a,*, Nur Qistina Othman a, Joon Sheong Tan b, Yang Ping Lee a

a FGV R&D Sdn. Bhd., FGV Innovation Centre (Biotechnology), PT 23417 Lengkuk Teknologi, 71760, Bandar Enstek, Negeri Sembilan, Malaysia
b PT. Tunggal Yunus Estate, Oil Palm Research Station- Topaz, Jl. Soekarno Hatta No.7, 8, 9, 10, Pekanbaru, Riau, 28125, Indonesia

**A R T I C L E   I N F O**

Article history:
Received 13 December 2019
Received in revised form 30 December 2019
Accepted 15 January 2020
Available online 23 January 2020

Keywords:
*Ganoderma boninense*
Genome sequencing
Pathogenic
Basal stem rot

**A B S T R A C T**

*Ganoderma boninense* is a soil-borne Basidiomycete pathogenic fungus that eminent as the key causal of devastating disease in oil palm, named basal stem rot. Being a threat to sustainable palm oil production, it is essential to comprehend the fundamental view of this fungus. However, there is gap of information due to its limited number of genome sequence that is available for this pathogenic fungus. This implies the hitches in performing biological research to unravel the mechanism underlying the pathogen attack in oil palm. Therefore, here we report a dataset of draft genome of *G. boninense* that was sequenced using Illumina Hiseq 2000. The raw reads were deposited into NCBI database (SRX7136614 and SRX7136615) and can be accessed via Bioproject accession number PRJNAS03786.

© 2020 The Authors. 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: suhaila.s@fgvholdings.com (S. Sulaiman).

https://doi.org/10.1016/j.dib.2020.105167
2352-3409/© 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1. Data description

This data consist of raw reads of the cultured *G. boninense* genome that were sequenced via Illumina Hiseq 2000 technology [1]. The data sets were named as s1_1.fastq, s1_2.fastq, s8_1.fastq and s8_2.fastq, whereby this involved paired-end reads sequencing in two lanes, denoted by s1* and s8* file names. The data reported here covers the pre-processing of raw reads, assembly data statistics and similarity search. Table 1 shows pre-processing statistics of the genome reads, consisting of raw reads and cleaned reads, which the latter indicates reads with high quality. Table 2 summarizes the main assembly statistics of the assembled draft genome. Fig. 1 shows assessment of draft genome completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO) software while using fungi dataset of Basidiomycota odb9 a reference. Fig. 2 shows the distribution of similarity search of assembled draft genome against Swiss-Prot database which delineated into different levels of similarity in the sense of E-value parameter.

2. Experimental design, materials, and methods

2.1. Genome sequencing

Genomic DNA (gDNA) was isolated from the fruiting body of *G. boninense*. A total of 5 μg of DNA was used to prepare a 400 bp paired-end sequencing library using an Illumina paired-end DNA sample preparation kit. The quality of the library was assessed by Q-PCR before continuing to cluster
Table 1
Pre-processing statistics of the genome reads. Clean reads refer to high quality reads with at least Phred quality value of Q20 and longer than 30 bp.

| Sample Name | Total Raw Reads | Total Raw Reads Bases | Total Clean Reads | Clean Reads (%) |
|-------------|-----------------|-----------------------|-------------------|-----------------|
| s_1_1.fastq | 81,292,176      | 8,210,509,776         | 76,710,474        | 94.36           |
| s_1_2.fastq | 81,292,176      | 8,210,509,776         | 76,116,018        | 93.63           |
| s_8_1.fastq | 95,001,316      | 9,595,132,916         | 88,377,542        | 93.03           |
| s_8_2.fastq | 95,001,316      | 9,595,132,916         | 87,615,553        | 92.23           |
| TOTAL       | 352,586,984     | 35,611,285,384        | 328,819,587       | 93.31 (average) |

Table 2
Assembly statistics of draft genome.

| Attributes               | Value         |
|--------------------------|---------------|
| Number of contigs        | 2,040         |
| Total residues (bp)      | 66,570,000    |
| Average length (bp)      | 32,634        |
| N50 contig (bp)          | 239,351       |
| L50 contig (bp)          | 78            |
| Largest contig (bp)      | 1,452,011     |
| Smallest contig (bp)     | 197           |

Fig. 1. Assessment of draft genome completeness using BUSCO software. Fungi dataset of Basidiomycota odb9 that consist of 1,335 total BUSCO groups was used a reference.

Fig. 2. Distribution of similarity search of assembled draft genome against Swiss-Prot database. About 74.31% of the assembled sequence were similar to the manually curated protein database in Swiss-Prot database.
Sequencing was performed using two lanes of Illumina HiSeq 2000 paired-end flow cell using 202 cycles to produce 2 × 100 bp paired-end reads.

2.2. Quality assessment and reads pre-processing

Prior to bioinformatics analysis, the quality of raw reads were assessed using FASTQC [2]. The raw reads were pre-processed using Perl-coded computer scripts to trim low quality bases and filter short reads to obtain high quality reads, which refer to reads with Phred quality value of Q20 and longer than 30 bp [3]. The improved quality of cleaned reads were confirmed using FASTQC [2]. Table 1 shows the pre-processing statistics of the genome reads.

2.3. De novo genome draft assembly

The high quality reads of Illumina were assembled using de novo approach by Trinity tools [4,5]. Assembly statistics for both approaches is shown in Table 2. The completeness of de novo assembled draft genome was evaluated using BUSCO [6] on a local workstation. Fungi dataset of Basidiomycota odb9 was used as its single-copy orthologs database and the result is shown in Fig. 1. The assembled sequence was searched against Swiss-Prot database [7] using Blastx program [8] which was downloaded locally. The similarity search shows about 74.31% of the assembled sequence were similar to the manually curated protein database (Fig. 2).

CRediT author statement

Suhaila Sulaiman: Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation. Nur Qistina Othman: Methodology, Validation, Data curation, Writing- Original draft preparation, Resources. Joon Sheong Tan: Conceptualization, Methodology, Supervision, Writing- Reviewing and Editing. Yang Ping Lee: Conceptualization, Supervision, Writing- Reviewing and Editing.

Acknowledgments

This work was supported by FGV Agri Services Sdn. Bhd. We are also grateful to Malaysian Genomics Resource Centre Berhad (MGRC) for the sequencing service of genome data.

Conflict of 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2020.105167.

References

[1] B.U. Experience, U. Output, HiSeq™ 2000 Sequencing System, 2000.
[2] S. Andrews, FastQC A Quality Control Tool for High Throughput Sequence Data, 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed 9 September 2019).
[3] C. Phred, Q. Scores, Quality Scores for Next-Generation Sequencing, 2000, pp. 1–2.
[4] M. Baker, De novo genome assembly: what every biologist should know. Nat. Methods 9 (2012) 333–337.
[5] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacon, A. Niranke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (2011) 644–652, https://doi.org/10.1038/nbt.1883.
[6] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212, https://doi.org/10.1093/bioinformatics/btv351.

[7] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal, A.J. Bridge, S. Poux, L. Bougueleret, I. Xenarios, UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view, Methods Mol. Biol. 1374 (2016) 23–54, https://doi.org/10.1007/978-1-4939-3167-5_2.

[8] 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, https://doi.org/10.1016/S0022-2836(05)80360-2.