High quality genome sequence reveals the 12 pseudo-chromosomes of *Ganoderma boninense*

Condro Utomo^{a#}, Zulfikar Achmad Tanjung^b, Redi Aditama^b, Rika Fithri Nurani Buana^c, Antonius Dony Madu Pratomo^d, Reno Tryono^e, Tony Liwang^f

^aDepartment of Biotechnology, PT SMART Tbk, Bogor, Indonesia

^bSection of Bioinformatics, PT SMART Tbk, Bogor, Indonesia

^cSection of Microbiome Technology, PT SMART Tbk, Bogor, Indonesia

^dSection of Clonal Technology, PT SMART Tbk, Bogor, Indonesia

^eSection of Genetic Engineering, PT SMART Tbk, Bogor, Indonesia

^fDivision of Plant Production and Biotechnology, PT SMART Tbk, Bogor, Indonesia

#Address correspondence to Condro Utomo, condro.utomo@sinarmas-agri.com

Author's contribution:

C.U. and T.L.: conceive the study, lead and supervise the project

Z.A.T., R.A.: perform genome assembly and analysis

R.F.N.B. and A.D.M.P.: perform fungal cultivation

R.T.: conceive the study and wrote the manuscript

All authors approved the submitted manuscript

Abstract

Ganoderma boninense is the dominant fungal pathogen in causing Basal Stem Rot (BSR) disease on oil palm. The whole genome of this fungus was sequenced using PacBio RS II platforms to gain the whole-genome shotgun libraries. These libraries were smoothed with Illumina Hiseq 4000 paired end sequencing to polish the genome assembly. Subsequently, a combination of Dovetail Chicago and HiC libraries with the HiRise assembly pipeline revealed the 12 pseudo-chromosomes of *G. boninense*. This is the first report of chromosomal-scale genome assembly for the most important fungal pathogen on oil palm.

Keywords:

Ganoderma boninense, chromosomal-assembly, oil palm.

Introduction

The basidiomycete fungus of *Ganoderma boninense* Pat. is the causal agent of basal stem rot (BSR) disease on oil palm¹. The disease is reported as a major economic importance on oil palm plantations in Indonesia, Malaysia, and Papua New Guinea, by reducing yield up to 50-80%^{2,3}. Considering its economic impact, a draft genome sequence was assembled to complement the lack of genomic information of this fungus using Pacific Biosciences (PacBio) and Illumina platforms for the Indonesian strain (G3)⁴. However, the chromosome level assembly for this fungus remains uncovered.

Current sequencing technologies and assembly algorithms change the genome sequencing landscape. Chromosomal assembly can be done through single-molecule sequencing and chromatin conformation capture (3C) data analysis⁵⁻⁷. PacBio and Illumina sequencing lay the fundamental layer of sequence scaffolds contiguity and resolve repetitive DNA regions^{8,9}. The Chicago technique improves scaffolding of de novo PacBio and Illumina sequencing through chromatin reconstructing usage as a substrate to obtain proximity ligation libraries *in vitro*¹⁰. Additionally, the HiC which is an adaptation of 3C approach, is used to detect chromatin interaction in the nucleus through *in vivo*

formaldehyde-crosslink and physically form 3D architecture of the genome¹¹. The HiRise algorithm assembly pipeline in both Chicago and HiC analysis anchors, orders, and orients the assembly sequences within each chromosome. In this study, we reported the first chromosome genome assembly of *G. boninense* using PacBio long-reads, Illumina short-reads, Chicago sequencing and HiC technology.

Methods

Sample collection

A *G. boninense* G3 strain was isolated from an oil palm tree with severe symptoms of BSR disease in North Sumatera Province, Indonesia. Freshly pure revived mycelia were grown in 100 ml yeast malt broth in dark at 28°C for 14 days. Mycelia were harvested on a layer of Whatman paper no. 1 and air dried for 15 minutes. Half of the sample was proceeded for genomic DNA isolation using GenElute plant genomic DNA miniprep kit (Sigma-Aldrich Co., St. Louis, MO, USA) according to the manufacturer's instructions for PacBio and Illumina platforms sequencing. Another part of the sample, mycelia were freeze-dried and shipped for Chicago and HiC platforms sequencing.

PacBio/Illumina sequencing and de novo assembly

Single-molecule sequencing was performed using PacBio RS II with the latest P6-C4 chemistry systems and Illumina HiSeq 4000 system according to each manufacturer's instruction. The PacBio reads was assembled using WTDBG2 with default parameters¹². Subsequently, two rounds of Racon was used for consensus calling for this assembly¹³. At the final assembly step, two rounds of PacBio-Racon assemblies were polished by basecall correction through Pilon software with default parameters¹⁴. Pilon uses Illumina reads to perform base corrections and derive an accurate consensus sequence.

Chicago library preparation and sequencing

A Chicago library was prepared as described previously¹⁰. Briefly, ~500ng of high molecular weight of genomic DNA (gDNA) (mean fragment length = 59) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq X to produce 165 million 2x150 bp paired end reads, which provided 1,887.11 x physical coverage of the genome (1-100 kb pairs).

Dovetail HiC library preparation and sequencing

A Dovetail HiC library was prepared in a similar manner as described previously¹⁵. Briefly, for each library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq X

to produce 194 million 2x150 bp paired end reads, which provided 41,650.61 x physical coverage of the genome (10-10,000 kb pairs).

Scaffolding the assembly with HiRise

The input *de novo* assembly, shotgun reads, Chicago library reads, and Dovetail HiC library reads were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies¹⁰. An iterative analysis was conducted. First, Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (<http://snap.cs.berkeley.edu>). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After aligning and scaffolding Chicago data, Dovetail HiC library sequences were aligned and scaffolded following the same method. After scaffolding, shotgun sequences were used to close gaps between contigs.

Results and Discussion

The single-molecule genome sequencing of the G3 strain was performed using PacBio RS II platforms combined with Illumina Hiseq 4000 paired end sequencing technology to obtain the whole-genome shotgun libraries. For PacBio RSII platform, a 20 kb library was built and sequenced. Simultaneously, paired-end reads were generated by sequencing of initial libraries of 300 bp using Illumina HiSeq4000 system.

The 3C analysis was conducted using sequencing-based Chicago and HiC assemblies. Using Chicago approach, Library 1 produced 165 million reads of 2x150 bp and provided 1,887.11x physical coverage of the genome (1-100 kb pairs). In HiC approach, Library 1 produced 194 million read pairs of 2x150 bp and provided 41,650.61x physical coverage of the genome (10-10,000 kb pairs).

Furthermore, the libraries from PacBio and Illumina were assembled using WTDBG2, Racon, and Pilon to produce 55.82 Mb with 592 scaffolds and N50 value of 0.357 Mb. Both Chicago and HiC libraries were assembled with HiRise pipeline. Using PacBio and Illumina as an input data, the Chicago HiRise assembly resulted in L90/N90 of 75 scaffolds. This result was used as an input for HiC HiRise assembly and resulted the L90/N90 output into 12 scaffolds (**Table 1**). The contiguity comparison graph showed the significance improved of assembly (**Figure 1**).

Table 1. Comparison of de novo and proximity-guided genome assembly.

Platform/Library	PacBio + Illumina	Chicago	HiC
Assembler + Polisher	WTDBG2 + Racon + Pilon	HiRise	HiRise
Longest scaffold (bp)	1,682,192	3,150,284	6,651,874
Number of scaffolds	592	264	112
Number of scaffolds > 1kb	592	264	112
Assembled genome size (Mb)	55.82	55.86	55.87
N50 (Mb)	0.357	0.993	4.126
L90	257	75	12

The improvement of scaffolds assembly occurred from the role of the usage of biotin-labeled nucleotide in HiC which enables selective purification of chimeric DNA ligation junctions¹¹. HiC data provides spanning even whole chromosomes to improve scaffold contiguity of assemblies chromosome-length scaffolds for genomes¹⁶.

Figure 1. A comparison of the contiguity of the input assembly and the final HiRise scaffolds. Each curve shows the fraction of the total length of the assembly present in scaffolds of a given length or smaller. The fraction of the assembly is indicated on the Y-axis and the scaffold length in basepairs is given on the X-axis. The two dashed lines mark the N50 and N90 lengths of each assembly. Scaffolds less than 1 kb are excluded.

In addition, Dovetail Genomics' HiRise pipeline generated a HiC linkage density histogram plot (**Figure 2**). The plot compares the positions of read pair sequences (the pair of end sequences from each and every sequenced DNA fragment obtained by chromatin cross-linking) versus the positions of each individual DNA sequence within the genome assembly^{10,17}. The alignment produces a diagonal of lines from lower left to upper right in the plot that represent each of the 12 pseudo-chromosomes. Dots (sequences) within boxes at the last column are probably un-scaffold DNA sequence.

Figure 2. Dovetail Genomic's HiC linkage density histogram. In this figure, the x and y axes give the mapping positions of the first and second read in the read pair respectively, grouped into bins. The color of each square gives the number of read pairs within that bin. White vertical and black horizontal lines have been added to show the borders between scaffolds. Scaffolds less than 1 Mb are excluded.

Overall, the single-molecule sequencing combined with chromatin cross-linking deep sequencing resulted in 12 scaffolds. The density histogram plot is able to guide the anchoring of scaffolds to pseudo-chromosomes^{18,19}. It is thus corresponding to the pseudo-chromosome number in *G. boninense*. This number is resembling two other close-related *Ganoderma* species that the chromosomal assembly has been generated earlier i.e. *G. lucidum* (13 chromosomes) and *G. sinense* (12 chromosomes)^{20,21}. In term of gene number, *G. boninense* is the highest with 21,074 coding sequences followed with *G. lucidum* with 16,113, and *G. sinense* with 15,688 genes. Gene density of each pseudo-chromosome in *G. boninense* is scattered evenly in all pseudo-chromosomes (**Figure 3**). This study showed that the final HiC assembly provides a robust, fast, and valid data for generating de novo assemblies with chromosome-length scaffolds. However, these assemblies still contain 504 gaps (Table 2).

Figure 3. Gene density distribution in all pseudo-chromosomes of *G. boninense*. This graph excludes 100 scaffolds with total of 4.3 Mb length that unmapped which size ranges between 1.1 Mb and 2.6 Kb.

Table 2. Assembly statistics for HiC libraries and HiRise assembly of *G. boninense*.

Characteristics	Number
Number of pseudo-chromosome	12
Number of scaffold	112
Longest scaffold (Mb)	6.5
N50 scaffold (Mb)	304.34
L50 (scaffolds)	6
N90 (Mb)	3.032
L90 (scaffolds)	12
Length of genome assembly (Mb)	55.87
GC contents (%)	56.2
Number of protein-coding genes	21,074
GC content of protein-coding genes (%)	59.2
Average gene length (bp)	1.265
Average number of exons per gene	7
Average exon size (bp)	178
Average intron size (bp)	97
Average size of intergenic regions (bp)	1.250
Number of gaps	504
Non-gap bases (bp)	55.821.405

Conclusion

This study demonstrates a chromosomal-scale genome assembly of *G. boninense* through combination of single-molecule sequencing (PacBio long reads and Illumina short reads) and 3C data analysis (Chicago and HiC with HiRise pipeline). These combined technologies harbored 55.87 MB length of genome assembly within 12 pseudo-chromosomes and an additional 4.3 Mb un-scaffold sequences.

Acknowledgement

This work was supported by the management of PT SMART Tbk. We thank Roberdi, Marcelinus Rocky Hatorangan, and Victor Aprilyanto for proof-read this manuscript, and Sanju Rianintika and Hani Feorani for their technical supports.

References

1. Pilotti CA. Stem rots of oil palm caused by *Ganoderma boninense*: Pathogen biology and epidemiology. *Mycopathologia* **159**, 129–137 (2005). doi: 10.1007/s11046-004-4435-3.
2. Aderungboye FO. Diseases of the oil palm. *Pans* **23**, 305–326 (1977). doi: 10.1080/09670877709412457.
3. Corley RHV, Tinker P. B. *The oil palm*. (John Wiley & Sons, Ltd, 2015). doi: 10.1002/9781118953297.
4. Utomo C, Tanjung ZA, Aditama R, Buana RFN, Pratomo ADM, Tryono R, Liwang T. Draft genome sequence of the phytopathogenic fungus *Ganoderma boninense*, the causal agent of basal stem rot disease on oil palm. *Genome Announc.* **6**, (2018). doi: 10.1007/s11046-004-4439-z.
5. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, Burton JN, Huson HJ, Nystrom JC, Kelley CM, Hutchison JL, Zhou Y, Sun J, Crisa A, de Leon FAP, Schwartz JC, Hammond JA, Waldbieser GC, Schroeder SG, Liu GE, Dunham MJ, Shendure J, Sonstegard TS, Phillippy AM, Van Tassel CP, Smith TPL. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650 (2017). doi: 10.1038/ng.3802.Single-molecule.
6. Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy TD, Young R, Lefevre L, Hume DA, Collins A, Ajmone-marsan P, Smith TPL, Williams JL. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat. Commun.* 1–11 (2019). doi: 10.1038/s41467-018-08260-0.
7. Cai M, Zou Y, Xiao S, Li W, Han Z, Han F, Xiao J, Liu F, Wang Z. Chromosome assembly of *Collichthys lucidus*, a fish of Sciaenidae with a multiple sex chromosome system. *Sci. Data* 1–7 (2019). doi: 10.1038/s41597-019-0139-x.
8. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics. Proteomics Bioinformatics* **13**, 278–289 (2015). doi: 10.1016/j.gpb.2015.08.002.
9. Jiao W-B, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, Willing E-M, Piednoel M, Woetzel S, Madrid-Herrero E, Huettel B, Huemann U, Reinhard R, Koch MA, Swan D, Clavijo B, Coupland G, Schneeberger K. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res.* **27**, 778-786 (2017). doi: 10.1101/gr.213652.116.Freely.
10. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS, Green RE. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**,

- 342–350 (2016). doi:10.1101/gr.193474.115.Freely.
11. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012). doi: 10.1016/j.ymeth.2012.05.001.
12. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *bioRxiv* **Jan**, 1–7 (2019). doi: 10.1101/530972.
13. Vaser R, Sovic I, Nagarajan V, Sikic M. Fast and accurate de novo genome assembly from long uncorrected reads. *Methods* **27**, 737–746 (2017). doi: 10.1101/gr.214270.116.5.
14. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakhtikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, 1–14 (2014). doi: 10.1371/journal.pone.0112963.
15. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009). doi: 10.1126/science.1181369.
16. Ghurye J, Pop M, Koren S, Bickhart D, Chin C. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**, 1–11 (2017). doi:10.1371/journal.pcbi.1006994.
17. Nuss AB, Sharma A, Gulia-nuss M. Chicago and Dovetail Hi-C proximity ligation yield chromosome length scaffolds of *Ixodes scapularis* genome. *bioRxiv*, 1–7 (2018). doi: 10.1101/392126.
18. Xue H, Wang S, Yao J-L, Deng CH, Wang L, Su Y, Zhang H, Zhou H, Sun M, Li X, Yang J. Chromosome level high-density integrated genetic maps improve the *Pyrus bretschneideri* ‘DangshanSuli’ v1.0 genome. *BMC Genomics* **19**, 1–13 (2018). doi: 10.1186/s12864-018-5224-6.
19. Linsmith G, Rombauts S, Montanari S, Deng CH, Celton J-M, Guerif P, Liu C, Lohaus R, Zum JD, Cestaro A, Bassil NV, Bakker LV, Schiljen E, Gardiner SE, Lespinasse Y, Durel C-E, Velasco R, Neale D, Chagne D, Van de Peer Y, Troggio M, Binaco L. Pseudo-chromosome length genome assembly of a double haploid ‘Bartlett’ pear (*Pyrus communis* L.). *bioRxiv* May, 1–24 (2019). doi: 10.1101/651778.
20. Chen S, Xu J, Liu C, Zhu Y, Nelson DR, Zhou S, Li C, Wang L, Guo X, Sun Y, Luo H, Li Y, Song J, Henrissat B, Levasseur A, Qian J, Li J, Luo X, Shi L, He L, Xiang L, Xu X, Niu Y, Li Q, Han MV, Yan H, Zhang J, Chen H, Lv A, Wang Z, Liu M, Schwartz DC, Sun C. Genome sequence of the model medicinal mushroom *Ganoderma lucidum*. *Nat. Commun.* **3**, 1–9 (2012). doi: 10.1038/ncomms1923.
21. Zhu Y, Xu J, Sun C, Zhou S, Xu H, Nelson DR, Qian J, Song J, Luo H, Xiang Li, Li Y, Xu Z, Ji A, Wang L, Lu S, Hayward A, Sun W, Li X, Schwartz DC, Wang Y, Chen S. Chromosome-level genome map provides insights into diverse defense mechanisms in the medicinal fungus *Ganoderma sinense*. *Sci. Rep.* 1–14 (2015). doi:10.1038/srep11087.