MYCOBIOLOGY
2022, VOL. 50, NO. 4, 254–257
https://doi.org/10.1080/12298093.2022.2109874

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Draft Genome Sequence of the Reference Strain of the Korean Medicinal Mushroom *Wolfiporia cocos* KMCC03342

Bogun Kim[a], Byoungnam Min[a]*, Jae-Gu Han[b], Hongjae Park[a]*, Seungwoo Baek[a], Subin Jeong[a] and In-Geol Choi[a]

[a]Department of Biotechnology, College of Life Sciences and Biotechnology, Korea University, Seoul, Republic of Korea; [b]Mushroom Research Division, National Institute of Horticultural and Herbal Science, RDA, Eumseong, Chungbuk, Republic of Korea

ABSTRACT

Wolfiporia cocos is a wood-decay brown rot fungus belonging to the family Polyporaceae. While the fungus grows, the sclerotium body of the strain, dubbed Bokryeong in Korean, is formed around the roots of conifer trees. The dried sclerotium has been widely used as a key component of many medicinal recipes in East Asia. *Wolfiporia cocos* strain KMCC03342 is the reference strain registered and maintained by the Korea Seed and Variety Service for commercial uses. Here, we present the first draft genome sequence of *W. cocos* KMCC03342 using a hybrid assembly technique combining both short- and long-read sequences. The genome has a total length of 55.5 Mb comprised of 343 contigs with N50 of 332 kb and 95.8% BUSCO completeness. The GC ratio was 52.2%. We predicted 14,296 protein-coding gene models based on *ab initio* gene prediction and evidence-based annotation procedure using RNAseq data. The annotated genome was predicted to have 19 terpene biosynthesis gene clusters, which was the same number as the previously sequenced *W. cocos* strain MD-104 genome but higher than Chinese *W. cocos* strains. The genome sequence and the predicted gene clusters allow us to study biosynthetic pathways for the active ingredients of *W. cocos*.

ARTICLE HISTORY
Received 17 February 2022
Revised 11 July 2022
Accepted 1 August 2022

KEYWORDS
Wolfiporia cocos; secondary metabolite biosynthesis gene cluster; whole genome sequence

1. Introduction

Wolfiporia cocos is a medicinal basidiomycete fungus decaying wood and has a subterranean growth habit in association with pine trees [1]. The fungus is known to develop a hard endurable underground sclerotium body during its life cycle [2]. The sclerotium of *W. cocos* has been widely used as a key component of traditional medicine in East Asia because of its pharmacological properties including diuretic and sedative effects [2]. Various polysaccharides and triterpenoids are thought to be the major bioactive component of *W. cocos* but their biosynthetic pathways are not fully understood yet [3]. Among many active components isolated from *W. cocos*, pachymic acid is one of the well-known triterpenoids that display antitumor and anti-inflammatory activities [4]. Comprehensive genomic analysis of *W. cocos* is required to understand the genetic basis of various biosynthetic pathways, which can guide scientists to breed commercial strains and use *W. cocos* as a potent medicine to treat a variety of human diseases. The genome sequence of the *W. cocos* strain MD-104 isolated in Florida, United States, had been first revealed by the U.S. Department of Energy Joint Genome Institute (JGI) as a part of the 1000 Fungal Genomes Project [5]. Several *W. cocos* strains sampled in China have been reported but only two of those Chinese *W. cocos* strain genome sequences were publicly available [6,7]. *W. cocos* strain KMCC03342 (Cultivar name: Bokryeong1ho) is the reference dikaryotic strain of the *W. cocos* in South Korea and is maintained by the Korea Seed and Variety Service. Here, we report the high-quality genome sequence of the reference strain, *W. cocos* KMCC03342, for the scientific community.

2. Methods and materials

2.1. DNA/RNA extraction and sequencing

Total DNA was extracted with a modified protocol based on DNeasy® Plant Mini Kit (Qiagen, Hilden, Germany), described in the previous fungal genome project [8]. RNA was extracted with the Qiagen RNeasy® Mini Kit (Qiagen) following the manufacturer's protocol. The short read sequencing library for DNA and RNA sequencing was prepared with

CONTACT In-Geol Choi igchoi@korea.ac.kr
*Current address: BM, U.S. Department of Energy Joint Genome Institute, Berkeley, California, USA; HP, Institute of Hydrobiology, Biology Center CAS, Ceske Budejovice, Czech Republic

Illumina® DNA Prep kit (Illumina, CA, USA) and NEBNext® Ultra™ II RNA Library Prep Kit (New England Biolabs, USA), respectively. Sequencing was carried out on the Illumina MiSeq platform (Illumina) using Illumina MiSeq reagent kit V3 (300 bp paired-end). The long-read sequencing library was prepared using Oxford Nanopore Ligation Sequencing Kit (Oxford Nanopore, Oxford, UK). Sequencing was carried out on a MinION sequencing device (Oxford Nanopore) equipped with a MinION flow cell (R9.4.1) (Oxford Nanopore). PacBio single-molecule real-time (SMRT) sequencing was performed by Macrogen (Seoul, South Korea) on four SMRT cells using the PacBio RS II system.

2.2. Genome assembly and gene prediction

The initial assembly was assembled using the FALCON assembler (v0.4.0) with default options [9]. Draft *de novo* assembly was assembled using Canu assembler (v2.0) with default options [10]. Duplicated contigs from the draft genome were removed using the purge_dups (v1.2.5) program with default options [11]. Adapter sequences of short reads were removed using TrimGalore (v0.6.7) [12] with the '–paired' option. Errors in the draft genome sequence were corrected with Racon (v1.4.11) [13] and Pilon (v1.24) [14] with default options. The mitochondrial genome sequence was removed from the assembly by BLAST+ (v.2.12.0+) [15] alignment of *W. cocos* strain BL16 mitochondrial genome sequence (GenBank accession: NC_050681.1) to the *W. cocos* strain KMCC03342 assembly. Genome completeness analyses were performed using BUSCO (v5.2.2) [16] with the OrthoDB fungi v10 (fungi_odb10) database.

Gene prediction was performed with FunGAP (v1.1.0) [17] using *Laccaria bicolor* for the AUGUSTUS species model and 20,875,982 reads from RNA-seq results as evidence for the gene models. Transposable element-related genes were removed with the detect_te_genes.py script from FunGAP.

2.3. Functional annotation

Functional annotation of predicted protein-coding genes was carried out with InterProScan (v.5.51-85) [18] for protein domain annotation. Secondary metabolite biosynthesis gene cluster analysis was performed by antiSMASH (v6.0.1) [19] with a 'strict' strictness option.

2.4. Genome tree building using single copy ortholog concatenation

A total of 34 fungi genomes of the order Polyporales were retrieved from the NCBI database for comparative analysis. The species tree was built using FastTree (v2.1) [20] from the single copy ortholog genes identified by OrthoFinder (v2.5.4) using diamond for sequence alignment [21]. Mafft (v7.490) [22] and ClipKIT (v1.3.0) [23] were used to align multiple sequences to extract the conserved sequences with '-m gappy' for ClipKIT parameters.

3. Results and discussion

A total of 5.5 billion bases from 430,844 reads with an average read length of 12,702 bases were retrieved from long-read sequencing by the PacBio platform. The initial assembly assembled with PacBio reads only was comprised of 442 contigs and had a total length of 46.4 Mb but BUSCO revealed the genome completeness of 90.3%. To improve the quality of the reference genome, we added more sequencing data and employed a hybrid assembly technique using both short- and long-reads obtained from Illumina and Oxford Nanopore sequencing platforms, respectively. First, we obtained a total of 1.1 billion bases from 156,279 reads by sequencing with the Oxford Nanopore MinION platform. Reads from PacBio and Oxford Nanopore sequencing were combined for *de novo* assembly. Overlapping contigs from the diploid *W. cocos* KMCC03342 genome assembly was purged to a single contig. The assembly was polished with a total of 2.2 billion bases from 3,665,972 reads obtained from the Illumina MiSeq platform. The final polished assembly resulted in 343 contigs with the longest contig length of 1,489,262 bp and an N50 value of 332,393 bp. We found that genome completeness was also increased after polishing with short reads from 90.3 to 95.8% by the BUSCO analysis. The total length of the *W. cocos* KMCC03342 genome was 55,457,880 bp and the GC ratio was 52.2%. When compared to JGI *W. cocos* MD-104 genome assembly, the assembly of *W. cocos* KMCC03342 was considerably improved, showing a larger contig N50 value (332,393 bp) than that of the MD-104 (109,659 bp) with a smaller number (343) of contigs than that of the MD-104 (2,228) (Table 1). The genome of *W. cocos* KMCC03342 was missing 24 BUSCOs (3.1%) while the JGI MD-104 assembly

Table 1. Summary of the genome assembly and gene prediction of *Wolfiporia cocos* KMCC03342 in comparison to *W. cocos* MD-104 (JGI).

Statistics	*W. cocos* KMCC03342	*W. cocos* MD-104
Total assembly length (bp)	55,457,880	50,483,556
Number of contigs	343	2,228
Largest contig length (bp)	1,489,262	547,220
Contig N50 (bp)	332,393	109,659
Contig L50	38	129
GC content (%)	52.15	49.85
BUSCO completeness (%)	95.8	96.6
Protein coding genes	14,296	12,746

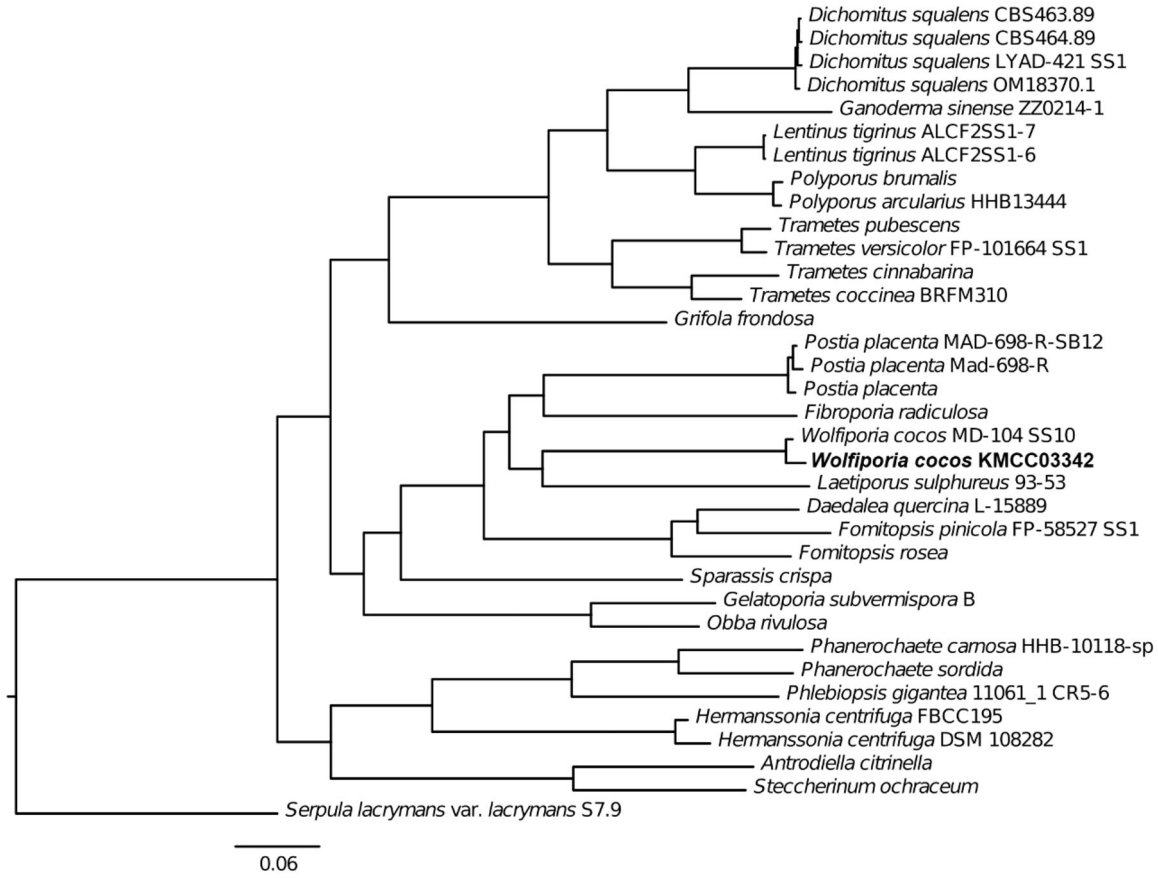

Figure 1. Maximum likelihood (ML) tree generated using single copy ortholog genes from 34 NCBI GenBank Polyporales genomes and *Wolfiporia cocos* KMCC03342. *Serpula lacrymans* var. *lacrymans* S7.9 genome (GenBank: GCA_000218685.1) was used as an outgroup.

was missing 22 BUSCOs (2.9%). The quality of genome assembly was acceptable to proceed with the genome annotation using RNA-seq data. To make reliable gene model predictions based on the transcriptomic data, we additionally conducted RNAseq of *W. cocos* KMCC03342, resulting in a total of 12.5 billion bases from 20,875,982 reads. Using RNAseq as the gene model prediction evidence data, we predicted 14,296 protein-coding genes from the FunGAP annotation pipeline. The genome data (gene models) was used to build the maximum likelihood phylogenetic tree based on single copy ortholog genes, reassuring the taxonomic rank of *W. cocos* KMCC03342 by placing KMCC03342 strain next to *W. cocos* MD-104 among other Polyporales genomes (Figure 1).

Functional annotation of *W. cocos* KMCC03342 revealed that 7,564 gene models contain at least one Pfam domain and 30.7% of gene models were multiple domain proteins (≥2 Pfam domains). The secondary metabolite biosynthesis gene cluster prediction program, antiSMASH [19], identified 27 gene clusters in the strain KMCC03342 and annotated 19 of the predicted clusters as potential terpene biosynthesis gene clusters. The number of predicted terpene biosynthetic gene clusters in the strain KMCC03342 (19) was the same as *W. cocos* strain MD-104 (19) and higher than public Chinese *W. cocos* strains, 2018LT001 and CGMCC5.78 (18 and 15, respectively) [6,7]. In addition, 13 terpene synthase genes were found in the *W. cocos* strain KMCC03342 assembly, while only 11 terpene synthase genes were identified in the *W. cocos* strain MD-104 genome. These observations indicate that the capability of *W. cocos* KMCC03342 for the terpene biosynthesis might be higher than other known *W. cocos* strains.

Draft genome sequence of *W. cocos* KMCC03342 will provide a genetic reference to breed better commercial strains and allow us to study the genes related to pachymic acid biosynthesis and other functional compounds found in *W. cocos*. The genome of *W. cocos* KMCC03342 was deposited in GenBank under the accession number JAKOOS000000000, BioProject number PRJNA801446, and BioSample number SAMN25349909.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was funded by the Cooperative Research Program for the National Agricultural Genome Program,

Rural Development Administration, Republic of Korea (project no. PJ01337602) and a National Research Foundation of Korea (NRF) grant funded by the government of the Republic of Korea (MEST) (grant NRF-2019R1A2C1089704). The authors were supported by Korea University grant.

ORCID

In-Geol Choi http://orcid.org/0000-0001-7403-6274

References

[1] Yang L, Tang J, Chen J-J, et al. Transcriptome analysis of three cultivars of *Poria cocos* reveals genes related to the biosynthesis of polysaccharides. J Asian Nat Prod Res. 2019;21(5):462–475.
[2] Ríos J-L. Chemical constituents and pharmacological properties of *Poria cocos*. Planta Med. 2011;77(7):681–691.
[3] Shu S, Chen B, Zhou M, et al. *De novo* sequencing and transcriptome analysis of *Wolfiporia cocos* to reveal genes related to biosynthesis of triterpenoids. PLOS One. 2013;8(8):e71350.
[4] Cheng S, Swanson K, Eliaz I, et al. Pachymic acid inhibits growth and induces apoptosis of pancreatic cancer *in vitro* and *in vivo* by targeting ER stress. PLoS One. 2015;10(4):e0122270.
[5] Floudas D, Binder M, Riley R, et al. The paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science. 2012;336(6089):1715–1719.
[6] Cao S, Yang Y, Bi G, et al. Genomic and transcriptomic insight of giant sclerotium formation of wood-decay fungi. Front Microbiol. 2021;12:746121.
[7] Luo H, Qian J, Xu Z, et al. The *Wolfiporia cocos* genome and transcriptome shed light on the formation of its edible and medicinal sclerotium. Genomics Proteomics Bioinformatics. 2020;18(4):455–467.
[8] Min B, Yoon H, Park J, et al. Unusual genome expansion and transcription suppression in ectomycorrhizal *Tricholoma matsutake* by insertions of transposable elements. PLOS ONE. 2020;15(1):e0227923.
[9] Chin C-S, Peluso P, Sedlazeck FJ, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13(12):1050–1054.
[10] Koren S, Walenz BP, Berlin K, et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res. 2017;27(5):722–736.
[11] Guan D, McCarthy SA, Wood J, et al. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36(9):2896–2898.
[12] Krueger F, James F, Ewels P, et al. 2021. FelixKrueger/TrimGalore: v0.6.7-doi via Zenodo. Zenodo,
[13] Vaser R, Sović I, Nagarajan N, et al. Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Res. 2017;27(5):737–746.
[14] Walker BJ, Abeel T, Shea T, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLOS One. 2014;9(11):e112963.
[15] Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinf. 2009;10(1):1–9.
[16] Manni M, Berkeley MR, Seppey M, et al. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38(10):4647–4654.
[17] Min B, Grigoriev IV, Choi I-G. FunGAP: fungal genome annotation pipeline using evidence-based gene model evaluation. Bioinformatics. 2017;33(18):2936–2937.
[18] Jones P, Binns D, Chang H-Y, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–1240.
[19] Blin K, Shaw S, Kloosterman AM, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res. 2021;49(W1):W29–W35.
[20] Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLOS One. 2010;5(3):e9490.
[21] Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):1–14.
[22] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–780.
[23] Steenwyk JL, Buida TJ III, Li Y, et al. ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. PLOS Biol. 2020;18(12):e3001007.