

## Research Article

# Genome Sequencing of *Hericium coralloides* by a Combination of PacBio RS II and Next-Generation Sequencing Platforms

Caixia Zhang <sup>1</sup>, Lijun Xu <sup>2</sup>, Jian Li,<sup>3</sup> Jiansong Chen,<sup>4</sup> and Manjun Yang <sup>1,4</sup>

<sup>1</sup>Tibet Vocational Technical College, Lhasa, Xizang, 850000, China

<sup>2</sup>Tibet University of Tibetan Medicine, Lhasa, Xizang, 850000, China

<sup>3</sup>China Institute of Veterinary Drug Control, Beijing 100081, China

<sup>4</sup>School of Life Sciences, Instrumental Analysis and Research Center, Sun Yat-sen University, Guangzhou 510006, China

Correspondence should be addressed to Manjun Yang; manjunyang@126.com

Caixia Zhang and Lijun Xu contributed equally to this work.

Received 27 August 2021; Revised 7 December 2021; Accepted 9 January 2022; Published 31 January 2022

Academic Editor: Ernesto Picardi

Copyright © 2022 Caixia Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The fruiting bodies or mycelia of *Hericium coralloides* (*H. coralloides*) contain many physiologically active compounds that are used to treat various diseases, including cardiovascular disorders and cancers. However, the genome of *H. coralloides* has not been sequenced, which hinders further investigations into aspects, such as bioactivity or evolutionary events. The present study is aimed at (i) performing *de novo* sequencing of the assembled genome; (ii) mapping the reads from PE400 DNA into the assembled genome; (iii) identifying the full length of all the repeated sequences; and (iv) annotating protein-coding genes using GO, eggNOG, and KEGG databases. The assembled genome comprised 5,59,05,675 bp, including 307 contigs. The mapping rate of reads obtained from PE400 DNA in the assembled genome was 92.46%. We identified 2,525 repeated sequences of 14,23,274 bp length. We predicted ncRNAs of 48,895 bp and 11,736 genes encoding proteins that were annotated in the GO, eggNOG, and KEGG databases. We are the first to sequence the entire *H. coralloides* genome (NCBI; Assembly: ASM367540v1), which will serve as a reference for studying the evolutionary diversification of edible and medicinal mushrooms and facilitate the application of bioactivity in *H. coralloides*.

## 1. Introduction

Wild edible mushrooms are extensively consumed owing to their unique and delicate flavors, abundant polysaccharides, proteins, fibers, and amino acids, and low lipid content that is good for low-calorie diets [1–3]. Except for the nutritional characteristics, mushrooms also contain rich bioactive compounds with medicinal properties, such with antimicrobial, antioxidant, lipid-lowering, and antitumor activity [4–6].

*Hericium coralloides* (1794) is an edible and medicinal mushroom species. The members of genus *Hericium* produce fleshy, whitish basidiomata, and their fruiting bodies or mycelia have been widely applied in traditional Chinese medicine (TCM) to treat various diseases, including cardiovascular disorders, cancer [7, 8], and gastric ulcers *in vivo* [9–13]. Therefore, we aimed to detect *H. coralloides* bioactiv-

ity by sequencing its genome. As a result of its potential activity, investigation of its genome may be necessary.

The recent advent of next-generation sequencing (NGS) technologies has facilitated genome sequencing and *de novo* assembly and increased the efficiency of genetic studies that have become a relative routine for genome analysis [14–16]. We previously spliced the *H. coralloides* genome by single-molecule real-time (SMRT) sequencing using individual polymerase molecules and the PacBio RS sequencing platform [17–19].

Here, we isolated mycelia from the fruiting bodies of *H. coralloides*, samples of which were collected from Lulang town, and enriched them by liquid fermentation. We assessed the overall profiles of the genes of *H. coralloides* and annotated their functions and obtained information about nonprotein coding RNAs (ncRNAs). We then investigated the underlying

TABLE 1: Statistics for the final assemblies of the *H. coralloides* genome.

Property	Contig
Total sequence Num.	307
Total sequence length (bp)	55,905,675
Min. sequence length (bp)	4,325
Max. sequence length (bp)	2,271,665
N50	441,150
GC (%)	53.84
N Num.	0
N rate (%)	0

TABLE 2: Prediction of genes in the assembled *H. coralloides* genome.

Property	Value
Total gene length (bp)	25,264,974
Total gene Num.	11,736
Average gene length (bp)	2,152
Gene percentage of genome (%)	45.19
Total exon length (bp)	20,204,958
Total exon Num.	73,583
Average exon length (bp)	274
Average exons per gene	6.2
Exon percentage of genome %	36.14
Total CDS length (bp)	16,343,018
Average CDS length (bp)	1,392
CDS percentage of genome %	29.23
Average intron length (bp)	76.6

genetic basis of *H. coralloides*, by using the PacBio RS II, Illumina MiSeq, and Illumina NextSeq500 platforms in combination to sequence the *H. coralloides* genome and assemble it *de novo*. The results of the present study deepen the understanding of the *H. coralloides* genome and provide a basis for future studies of the genetic mechanisms that underlie the biological activities and potential medicinal value of this species. The genome dataset provides invaluable information about *H. coralloides* genes and their potential functional metabolites.

## 2. Materials and Methods

**2.1. Sample Collection and Culture.** Fruiting bodies of *H. coralloides* collected from Lulang town in Nyingchi City, Tibet Autonomous Region, China (29°56'10.4" N, 94°47'2.65" E) on August 12, 2014, were morphologically identified and further characterized by sequencing the DNA of the internal transcribed spacer (ITS) region. Then, the mycelium of *H. coralloides* was isolated and purified from the fruiting body named *H. coralloides* tvtc0002 using PDA medium (potato dextrose agar medium, (w/v) 1% potato extract powder, 2% dextrose, and 1.5% agar, pH 6.5) and incubating at 25°C for 14 days. We then enlarged the mycelia by liquid fermentation in enriched PD medium (w/v; 1% potato extract

TABLE 3: Result of 17-mer analysis.

Property	Value
k-mer	17
k-mer Num.	55,483,329
k-mer peak (×)	148
Low frequency k-mer Num. (≤2)	1,340,810
Avg. read length (bp)	220
Total read Num.	31,921,174
Heterozygosity (%)	0.847
Repetitive 17-mer fraction	0.026
Genome size (bp)	43,990,397

k-mer: sequence of k bases; k-mer Num.: number of k-mer; k-mer peak (×): peak value of k-mer; low frequency k-mer Num. (≤2): k-mer of low-frequency; Avg. read length (bp): average length of reads; total read Num.: total number of reads; repetitive 17-mer fraction: proportion of repeated sequence.

powder, 2% dextrose, 0.1% MgSO<sub>4</sub>, 1% peptone, 0.5% beef extract powder, and 0.01% vitamin B<sub>1</sub>; pH 6.5) at 25°C and 200 rpm for 14 days.

**2.2. DNA and RNA Isolation.** The concentration, quality, and integrity of genomic DNA extracted using the cetyltrimethyl ammonium bromide (CTAB) method with minor modifications were determined using a Qubit Fluorometer (Invitrogen, USA) and a NanoDrop spectrophotometer (Thermo Scientific, USA). Sequencing libraries were generated using the TruSeq DNA Sample Preparation (Illumina, USA) and Template Prep Kits (Pacific Biosciences, USA).

The concentration, quality, and integrity of the total RNA isolated were determined using the TRIzol reagent (Invitrogen Life Technologies) and a NanoDrop spectrophotometer (Thermo Scientific). Three micrograms of the total RNA was used as input for the RNA sample preparation. Sequencing libraries were generated using the TruSeq RNA Sample Preparation Kit (Illumina, San Diego, CA, USA). Briefly, the mRNA was purified from total RNA using poly T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations at elevated temperatures in an Illumina proprietary fragmentation buffer. First-strand cDNA was synthesized using random oligonucleotides and SuperScript II; then, second-strand cDNA was synthesized using DNA Polymerase I and RNase H. Remaining overhangs were converted into blunt ends using exonuclease/polymerase, after which they were removed. The 3' ends of the DNA fragments were adenylated and then ligated with PE adapter oligonucleotides to prepare for hybridization. To select 200 bp cDNA fragments, the library fragments were purified using the AMPure XP system (Beckman Coulter, Beverly, CA, USA). Thereafter, the DNA fragments with adaptor molecules ligated to both ends were selectively enriched using the Illumina PCR Primer Cocktail in a 15-cycle PCR reaction. The products were purified using the AMPure XP system and quantified using Agilent High Sensitivity DNA assay and a Bioanalyzer 2100 system (Agilent). The library was sequenced on a HiSeq platform (Illumina) at Shanghai Personal Biotechnology Cp. Ltd.

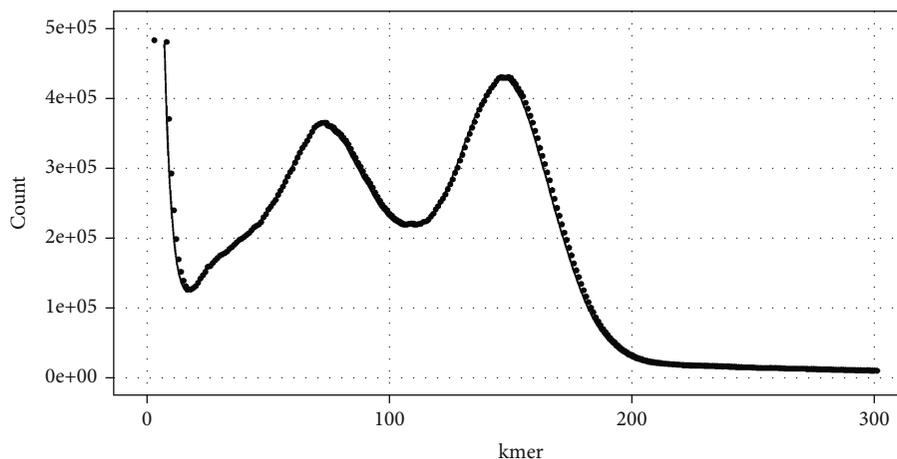


FIGURE 1: Distribution of 17-mers in *H. coralloides* genome. X-axis is 17-mer deep (X); Y-axis is the number of sequencing reads at that depth.

TABLE 4: Evaluation of integrity and continuity of genome assembly.

Property	Number	Percent (%)
Complete BUSCOs	283	97.6
Complete and single-copy BUSCOs	238	82.1
Complete and duplicated BUSCOs	45	15.5
Fragmented BUSCOs	1	0.3
Missing BUSCOs	6	2.1
Total BUSCO groups searched	290	100

Total BUSCO groups searched: the single-copy direct line homologous gene database used by BUSCO is fungi\_odb9, which provides 290 single-copy genes from 85 fungi.

**2.3. PacBio 20 K DNA Library Construction.** Samples (20  $\mu$ g) of DNA ( $OD_{260}/OD_{280} \approx 1.8$ ) were sheared using a Covaris® g-TUBE® device (Covaris, USA), diluted to 200–300 ng/ $\mu$ L in elution buffer, and centrifuged at 5,500 rpm (2029 g) for 2 min on a MiniSpin Plus (Eppendorf). We constructed SMRTbell libraries according to the Procedure & Checklist-20kb Template Preparation using the BluePippin™ Size Selection protocol ([http://files.pacb.com/Training/IntroductiontoSMRTbellTemplatePreparation/story\\_content/external\\_files/Introduction%20to%20SMRTbell%E2%84%A2%20Template%20Preparation.pdf](http://files.pacb.com/Training/IntroductiontoSMRTbellTemplatePreparation/story_content/external_files/Introduction%20to%20SMRTbell%E2%84%A2%20Template%20Preparation.pdf)). Briefly, the library was run on a BluePippin system (Sage Science, MA, USA) to select SMRTbell templates > 10 kb. Sequencing primers were annealed to the hairpins of the templates and bound with P5 sequencing polymerase and MagBeads (Pacific Biosciences, CA, USA). The libraries were sequenced on a PacBio RS II platform at Shanghai Personal Biotechnology Cp. Ltd.

**2.4. PE400 DNA Library Construction.** The sequencing libraries were generated using the Nextera XT DNA Library Prep Kit (Illumina Inc.) as described by the manufacturer. Briefly, unfragmented gDNA was cleaved, tagged with Illumina adapters, and amplified using a limited-cycle PCR program with a unique combination of i7 and i5 index primers.

TABLE 5: Statistical result of repeated sequences.

Elements	Number of elements	Length (bp)	Percentage of genome
Interspersed repeats	2,353	1,389,268	2.49
Retroelements	1,827	1,303,927	2.33
DNA transposons	466	71,198	0.13
Unclassified	60	14,143	0.03
Satellites	33	9,162	0.02
Simple repeats	121	20,741	0.04
Low complexity	18	4,103	0.01
Summary	2,525	1,423,274	3.00

The PCR products were purified using AMPure XP beads (Beckman Coulter Inc.) to remove short library fragments and then normalized with Nextera XT Library Normalization Beads (Illumina Inc.). The libraries were then sequenced on a MiSeq platform at Shanghai Personal Biotechnology Cp. Ltd.

**2.5. Data Quality Control.** The raw reads were filtered based on quality using FastQC with default parameters (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>), and the Q20, Q30, and GC contents were determined. For the raw data from NGS, 3'-adapter contaminant and tag sequence were removed (AdapterRemoval, version 2) [20]. Furthermore, we used NextClip (version 1.3.1) to remove the tag sequence of the reads in the Nextera Long Mate Pair library [21]. After quality correction of all the reads using Quake (version 0.3) and setting the k-mer to 17 [22], the reads that were  $\leq 50$  bp were removed.

**2.6. Survey Analysis.** Counting the number of occurrences of every k-mer in a DNA sequence is a central subproblem in many applications, including genome assembly, error correction of sequencing reads, fast multiple sequence

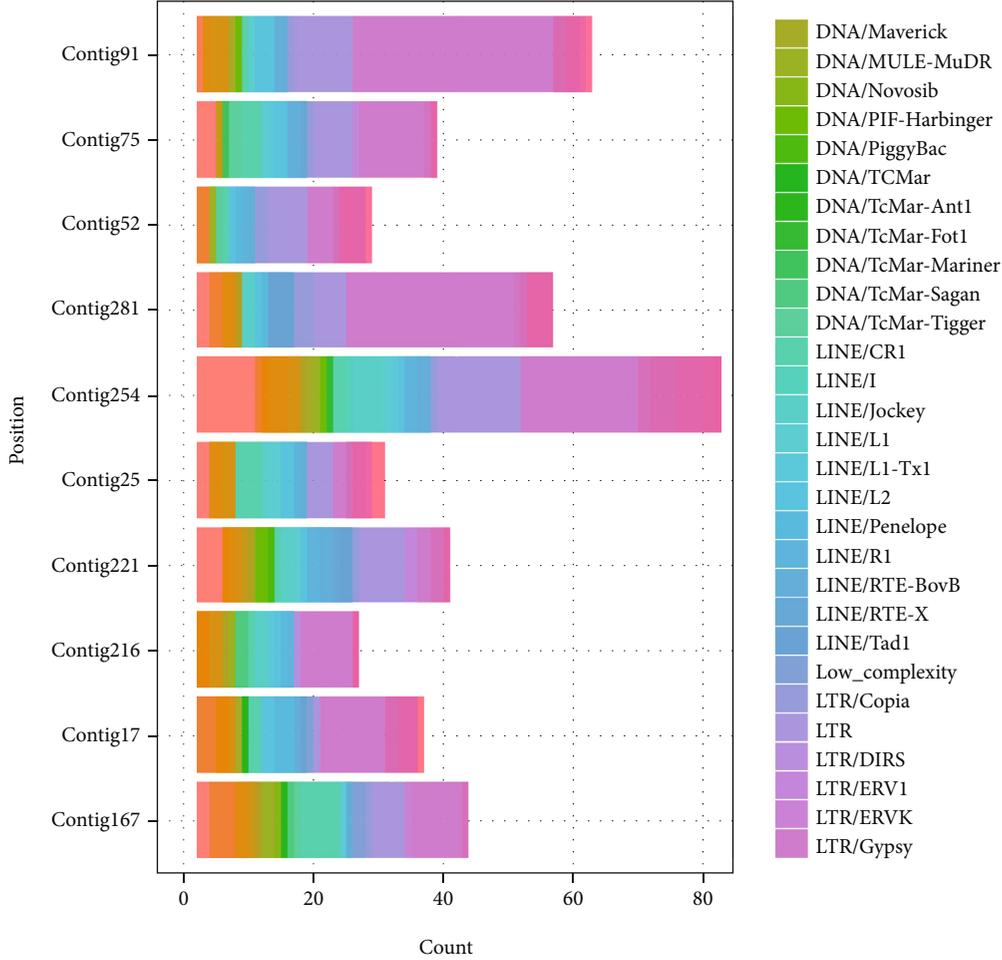


FIGURE 2: Composition of repeated sequences in each chromosome in the genome. Title is the class and family of repetitive DNA shown as DNA TIRs (terminal inverted repeats), LINEs (long interspersed nuclear elements), and LTRs (long terminal repeat). The family of TIRs of *H. coralloides* are shown as Maverick (virus-like DNA transposon), MULE-MuDR (Mutator-like transposable elements-MuDR transposon), Novosib (Novosib transposon), PIF-Harbinger (PIF-Harbinger transposon), PiggyBac (PiggyBac transposon), and TcMar (TcMar transposon).

TABLE 6: Results of noncoding RNA (ncRNA) prediction.

Type	Copy	Avg. length (bp)	Total length (bp)	% of genome
rRNA	9	1,923	17,309	0.030961
tRNA	270	88	23,845	0.042652
snoRNA	6	99	594	0.001062
CD-box	6	99	594	0.001062
HACA-box	-	-	-	-
scaRNA	-	-	-	-
snRNA	21	130	2733	0.004888
Other ncRNA	32	137	4414	0.007895
Summary	338	-	48,895	0.087458

ncRNA type: type of ncRNA; copy: copy number of ncRNA; Avg. length (bp): average length of ncRNA; total length (bp): total length of ncRNA; % of genome: percentage of ncRNA in genome.

alignment, and repeat detection [21]. We estimated genome size and heterogeneity using a survey analysis based on the k-mer counting algorithm [23]. We estimated the size of the genome ( $G$ ) as follows:

$$G = \frac{(N \times (L - K + 1) - B)}{D}, \quad (1)$$

where  $N$  is the number of reads,  $L$  is the average length of the reads,  $K$  is k-mer,  $B$  is the k-mer for low frequency, and  $D$  is the peak value in the k-mer distribution diagram.

2.7. *Genome Assembly and Analysis.* The Falcon software (<https://github.com/PacificBiosciences/FALCON-integrate>) [24] assembles long reads and is thus suitable for genome assembly in diploid organisms. We used Falcon to assemble reads obtained from third-generation single molecular sequencing and constructed a contig with default parameters. The assembly results were then corrected based on NGS data using the Pilon software [25]. Finally, GapCloser

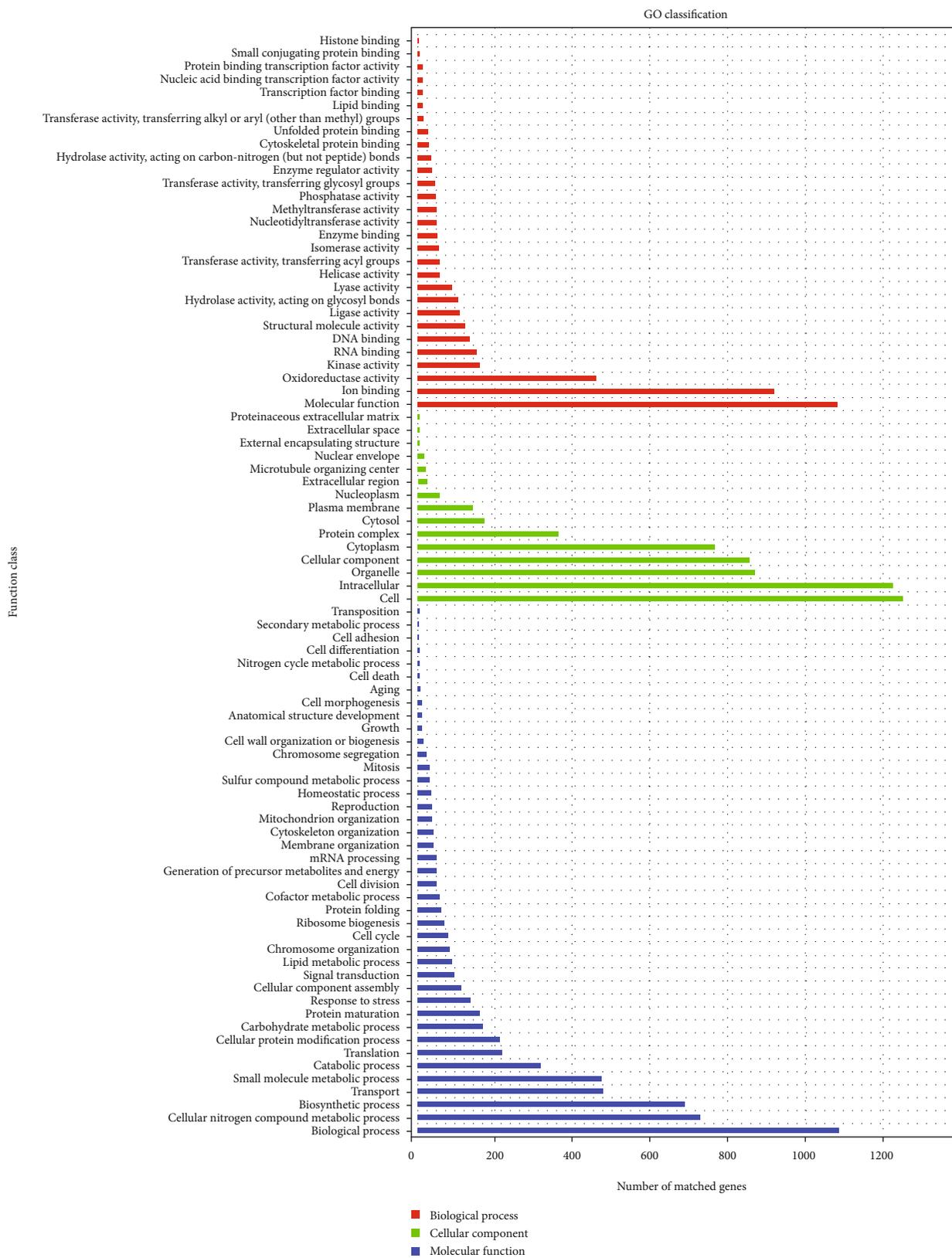


FIGURE 3: Gene Ontology (GO) functional annotation of genome. Blue, biological process; red, molecular function; green, cellular component. X-axis, number of matched genes; Y-axis (right), function class.

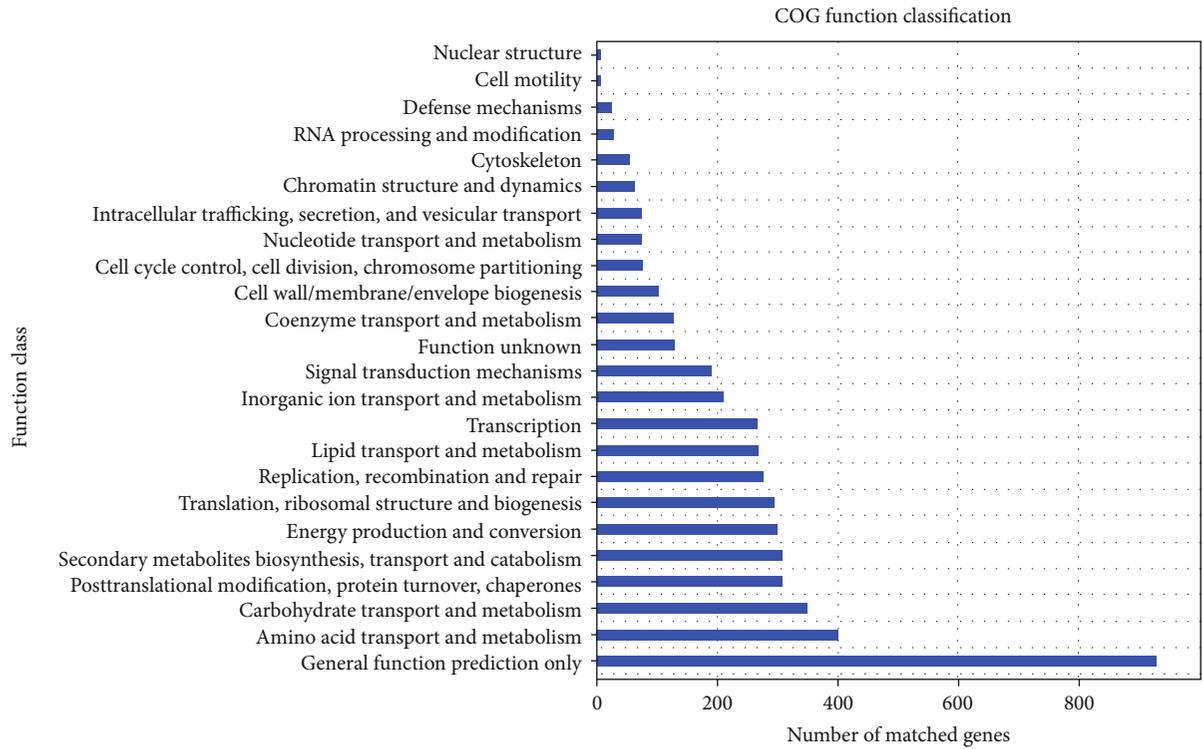


FIGURE 4: Clusters of orthologous groups (COG) of protein annotations of genome. X-axis, COG categories; Y-axis, number of matched genes. Total number of genes enriched in KEGG pathways was 1,311.

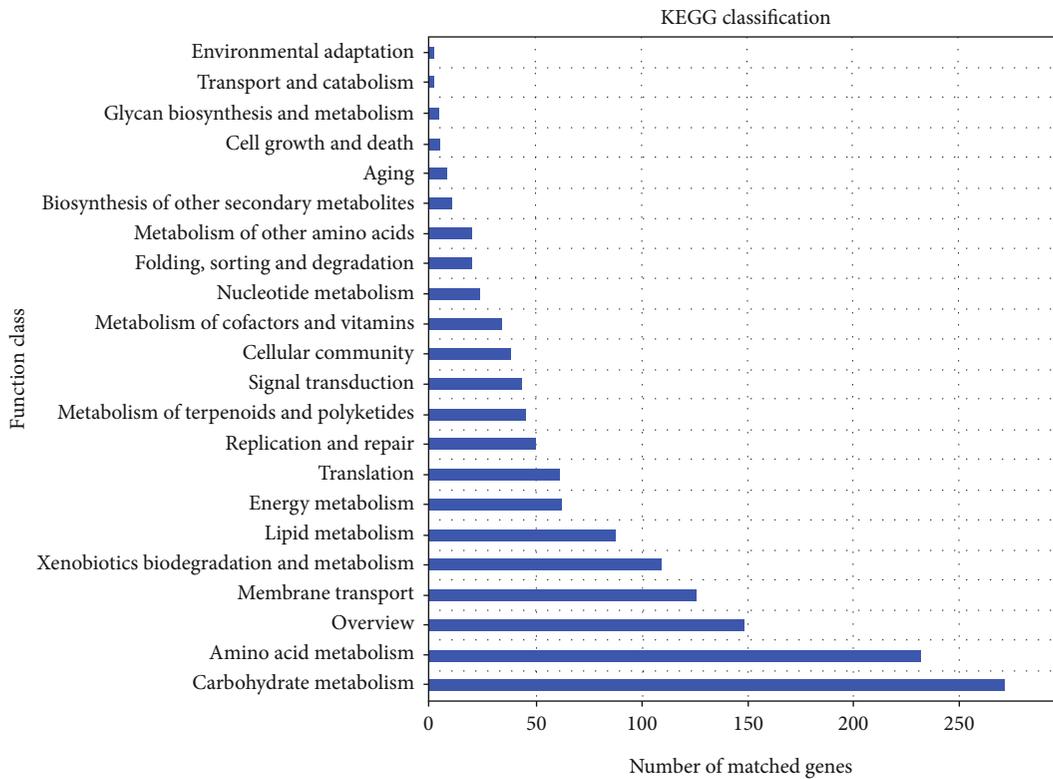


FIGURE 5: Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation of genome. X-axis, number of matched genes; Y-axis (right), function class.

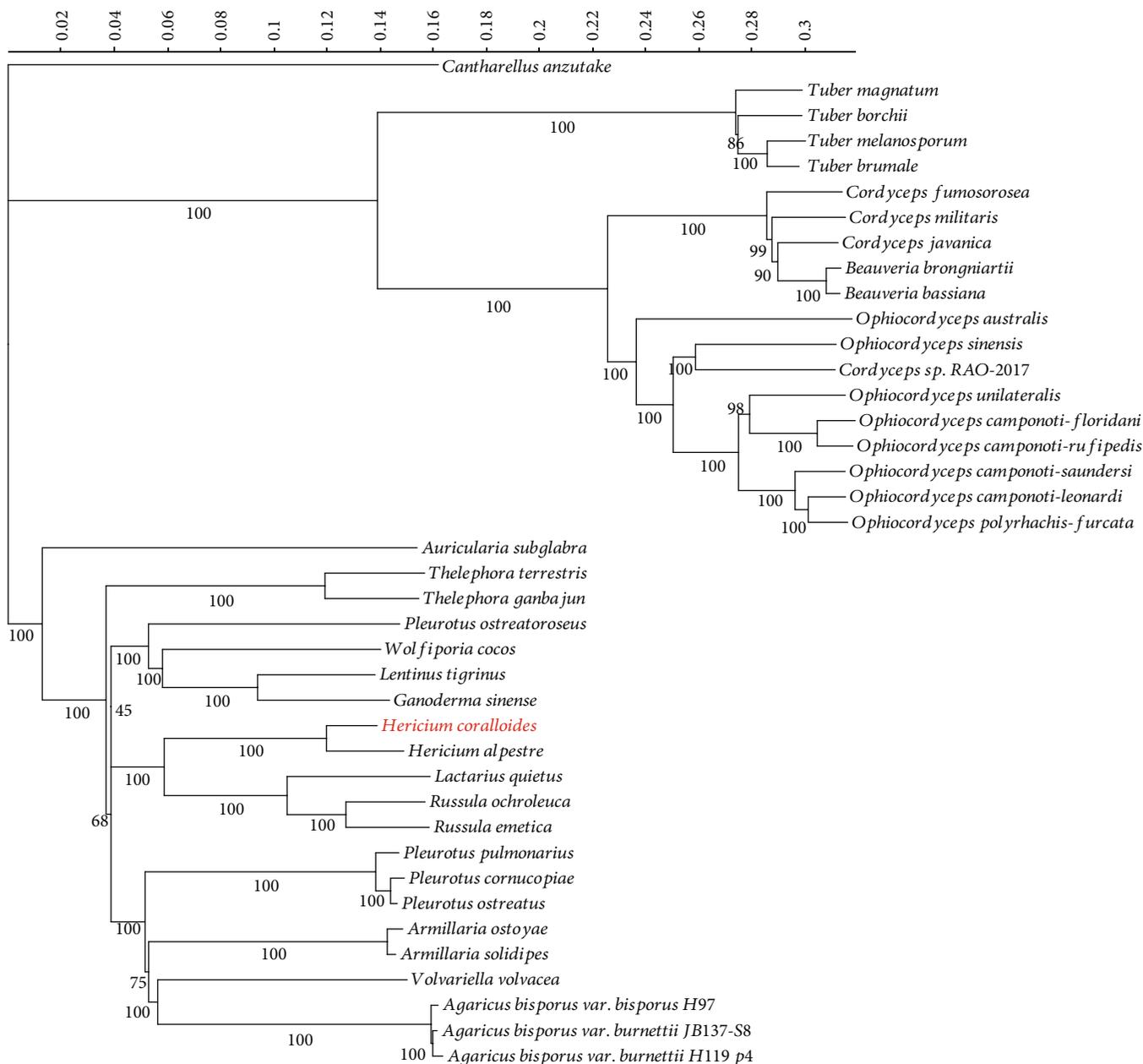


FIGURE 6: The phylogenetic tree based on single-copy ortholog genes between *H. coralloides* and 39 other fungal species.

(<http://soap.genomics.org.cn/soapdenovo.html>) [26] was used at default parameters to fill the gaps within the scaffolds. We mainly optimized the parameter values of Falcon, including -B, -t, -e, and -h, to improve genome assembly.

**2.8. Evaluation of Integrity and Continuity of Genome Assembly.** An accurate set of genes is needed to learn about species-specific properties, train gene-finding programs, and validate automatic predictions. However, many new genome projects lack comprehensive experimental data to derive a reliable initial set of genes. The amino acid sequences of a specific set of proteins are highly conserved over a wide range of eukaryotes. Therefore, comparing the assembled sequences with these proteins can determine the integrity of their sequences, which can then be used to indi-

rectly evaluate the integrity and continuity of the assembled genome. In this study, we used Benchmarking Universal Single-Copy Orthologs (BUSCOs, <http://busco.ezlab.org>, v3.0.2) [27] to ensure the reliability of the assembled genome. BUSCO defined 290 conserved protein sequences and assumed that the sequences were present in all fungi. In the genome of *H. collaroides*, 283 conserved protein sequences were detected by BUSCO, accounting for 95.16% of the total conserved protein sequences. Among these, 238 genes were complete and single-copy BUSCOs, accounting for 82.1% of the total conserved protein sequences.

**2.9. Sequence Alignment Analysis.** High-quality, filtered data were aligned to the genome obtained from assembly using Burrows-Wheeler Aligner (BWA) (v. 0.7.12-r1039) at

default parameters [28]. Duplicates were removed using MarkDuplicates in the Picard package. High-quality sequencing and assembly were second and third generations, respectively. The mapping sequence reached 92.46%, indicating that the results of the third-generation assembly could be analyzed.

**2.10. Repeated Sequence Analysis.** Repeated sequences are patterns of nucleic acids that occur in multiple copies throughout the genome. A significant fraction of genomic DNA is highly repetitive in many organisms [29]. Repeated sequences of nucleic acids occur in multiple copies throughout the genome and have been recognized as potential sources of genetic variation and regulation. We analyzed such sequences by homologous annotation using RepeatMasker v. 4.0.5 [30] based on the Repbase database [31] and by *de novo* annotation using RepeatModeler v. 1.0.4 (<http://repeatmasker.org/RepeatModeler.html>) based on the output files of RECON v.1.0.8 (<http://selab.janelia.org/recon.html>) and RepeatScout v. 1.0.5, <http://repeatscout.bioprospects.org/>).

**2.11. Prediction of ncRNAs.** Most genomes are apparently transcribed into ncRNAs [32] that are involved in many cellular processes, such as translation, RNA splicing, DNA replication, and gene regulation. Abundant and functionally important types of ncRNAs include transfer (tRNAs) and ribosomal RNAs (rRNAs) as well as small RNAs, such as microRNAs, siRNAs, snoRNAs, snRNAs, exRNAs, and the long ncRNAs. We predicted tRNA and rRNA using tRNAscan-SE v. 1.3.1 [33] and RNAmmer v.1.2 [34], respectively. Other types of ncRNAs were predicted by comparison with Rfam [35].

**2.12. Prediction of Genes Encoding Proteins.** We calculated the accuracy of gene prediction in eukaryotic genomic sequences as follows: (1) gene model was predicted *de novo* using Augustus v. 3.03 (<http://augustus.gobics.de/submission>) [36], glimmerHMM v.3.0.1 [37], and SNAP v. 2006-07-28 [38]; (2) homology was predicted using Exonerate v.2.2.0 (<http://www.ebi.ac.uk/about/vertebrate-genomics/software/>); (3) transcriptome was assembled *de novo* from RNA-Seq data using Trinity (v. r20140717) [39] and aligned using Program to Assemble Spliced Alignments (PASA) (v. r20140417) [40]; (4) the results of *de novo*, homologous, and RNA-Seq transcriptome predictions were integrated using EvidenceModeler v. r2012-06-25 with default parameters [40]. Boundaries of the predicted gene models were finally improved using PASA [41].

**2.13. Functional Annotation.** Based on the similarity of amino acid sequences in the protein domain, carbohydrate-active enzymes (CAZymes) are categorized as glycoside hydrolases (GHs), glycosyl transferases (GTs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), carbohydrate-binding modules (CBMs), and auxiliary activities (AAs). We used HMMER (version 3.0) to predict CAZyme genes in the genome.

Genes encoding proteins were annotated using Gene Ontology (GO; <http://www.geneontology.org>), evolutionary

genealogy of genes: Non-supervised Orthologous Groups (eggNOG; <http://eggnoг.embl.de>), the Kyoto Encyclopedia of Genes and Genomes (KEGG; <http://www.genome.ad.jp/kegg/>), and Swiss-Prot (<http://www.expasy.org/sprot/>). Genes were also annotated by GO using BLAST2GO (<http://www.blast2go.org/>) [42] and eggNOG using BLASTP (BLAST version 2.2.28+; <http://blast.ncbi.nlm.nih.gov/cgi>); KEGG orthology and pathways were annotated using the KEGG Automatic Annotation Server (KAAS; <http://www.genome.jp/kegg/kaas/>) [43].

**2.14. Phylogeny Construction.** The protein sequences of another 39 fungal species with available genome sequences were downloaded from the NCBI GenBank database (Table S1). Orthologous genes of *H. coralloides* and other fungal species were obtained using OrthoFinder v. 2.5.4 with the parameters -M dendroblast, -S blast, -A mafft, -T fasttree, and -l 1.5. Core single-copy orthologs were selected for subsequent phylogenetic analyses. The amino acid sequences of single-copy orthogroups were aligned using Muscle version 3.8.31; then, the best-aligned conserved blocks were extracted using Gblocks v. 0.91b at default parameters. A phylogenetic tree was constructed from the concatenated alignment by the neighbor-joining method using MEGAX v.10.2.6 with the p-distance model and 1,000 bootstraps. *Cantharellus anzutake* was set as the outgroup.

### 3. Results

We used the PacBio RS II, Illumina MiSeq, and Illumina NextSeq500 sequencing platforms in combination to sequence and assemble the genome of *H. coralloides de novo*. Table 1 shows 307 contigs of 5,59,05,675 bp length. Table 2 shows 11,736 genes and 73,583 total exons with lengths of 2,52,64,974 and 2,02,04,958 bp, respectively.

A total of 6,94,777, 3,33,64,830, and 3,40,63,012 reads were obtained using the PacBio 20 K DNA, PE400\_DNA, and PE400\_RNA libraries, respectively. The ratios (%) of GC for the three libraries were 50.28%, 53.65%, and 55.68%, respectively. The Q20 and Q30 values for the PE400\_DNA and PE400\_RNA libraries were 91.03% and 95.45% and 80.58% and 91.49%, respectively.

Analysis of the sequencing data with 17-mers revealed a genome of length 43.99 Mbp. The extent of heterozygosity was 0.847% (Table 3). In general, the k-mer value is usually set to 17, because the 17<sup>th</sup> iteration of the 4 bases (ATCG; 4<sup>17</sup>) will reach 17G, which is enough to cover the whole genome, whereas k-mer of 15 will only result in 1G, which is insufficient to cover the whole genome. In general, a larger k-mer value is associated with a higher error ratio. We avoided palindromic sequences in k-mer analysis by setting odd k-mer values. Therefore, we set k-mer to 17, and Figure 1 shows a distribution map.

After genome assembly, 307 contigs were obtained, and the genome was 5,59,05,675 bp long. The minimum and maximum lengths of sequences were 4,325 and 22,71,665 bp, respectively, and the ratio of GC was 53.84%. The assembly results of the contig and scaffold were

evaluated using Falcon, Pilon, and GapCloser, and the new genome data can be found online at NCBI (Assembly: ASM367540v1).

BUSCO was used to define 290 conserved protein sequences and assumed that they were common to 85 fungal species. We determined that 283 conserved protein sequences accounted for 97.6% of the total number of conserved protein sequences. Among these, 238 were complete and single-copy BUSCOs, accounting for 82.1% of the total. These results indicated good integrity and continuity of the genome assembly (Table 4). After sequence alignment, 2,95,13,883 reads were mapped to the genome at a rate of 92.46% and an average sequencing depth of 114.6. The coverage  $\geq 4$ ,  $\geq 10$ , and  $\geq 20$  were 96.7%, 95.93%, and 94.35%.

Table 5 shows the results of the repeated sequence analysis. We identified 2,525 repeated sequences, including 2,353 interspersed repeats, 33 satellites, 121 simple repeats, and 18 low complexities. The interspersed repeats included 1,827 retroelements, 466 DNA transposons, and 60 unclassified repeats (Figure 2).

The copy numbers of ncRNAs, rRNAs, tRNAs, snoRNAs, snRNAs, and other ncRNAs were 9, 270, 6, 21, and 32, respectively. Table 6 shows the average and total lengths of ncRNAs, as well as the proportion of ncRNA accounting for the genome. In total, 11,736 genes encoding proteins were predicted, with a total length of 2,52,64,974 bp. The average length of the genes encoding proteins was 2,152 bp.

We assigned 539 *H. coralloides* genes to CAZyme families, as defined in the CAZy database. The results of CAZyme analysis predicted that 152 genes had auxiliary activities, 19 were carbohydrate-binding modules, 83 were CEs, 210 were GHs, and 74 were GTs. The 3,675 annotated genes in the GO database were associated with biological processes (BPs), cellular components (CCs), and molecular function (MF) terms. In detail, 41 BP aspects were annotated, including biological, cellular nitrogen compound metabolic, and biological processes. A total of 15 CC aspects were identified, including cell, intracellular, and organelle. Additionally, 29 MF terms were annotated, including molecular function, ion binding, and oxidoreductase activity (Figure 3).

The comparison of gene sets with the evolutionary eggNOG database using BLASTP (BLAST v. 2.2.28+; <http://blast.ncbi.nlm.nih.gov/Blast.cgi>) resulted in 3,602 annotated genes and clusters of orthologous groups (COG) of proteins. Figure 4 shows that 7.92%, 3.39%, and 2.96% of genes encoding proteins, respectively, were classified into function R (general function prediction only), function E (amino acid transport and metabolism), and function G (carbohydrate transport and metabolism). The pathways associated with metabolism, namely, carbohydrate and amino acid metabolism as well as xenobiotic biodegradation and metabolism, were associated with environmental information processing, such as membrane transport and signal transduction (Figure 5).

We identified 169 single-copy orthogroups from *H. coralloides* and other fungal species and used corresponding amino acid sequences to construct a phylogenetic tree (Figure 6) in which *H. coralloides* was clustered with *Heri-*

*cium alpestre*, which is another species of the same genus. *Russula* and *Lactarius* were the most closely related genera to *Hericium*, followed by *Pleurotus*, *Wolfiporia*, *Lentinus*, and *Ganoderma*.

#### 4. Discussion

In this study, the *H. coralloides* genome was sequenced and assembled *de novo* for the first time using PacBio RS II, Illumina MiSeq, and Illumina NextSeq500 platforms. The assembled genome was 5,59,05,675 bp in length and included 307 contigs. The mapping rate of reads obtained from PE400 DNA in the assembled genome was 92.46%. We identified 2,525 repeated sequences of 14,23,274 bp. We also predicted 48,895 bp ncRNAs and 25,264,974 genes encoding proteins that were annotated in the GO, eggNOG, and KEGG databases.

A significant fraction of genomic DNA is highly repetitive in many organisms. A growing body of literature suggests that such sequences are vital to the genome [44]. The major categories of repeated sequences are terminal, tandem, and interspersed repeats. We identified 2,525 repeated sequences, most (2,353) of which were interspersed. All eukaryotic genomes have interspersed repeats, which are distributed throughout the genome and are not adjacent to each other. The repeated sequences vary depending on the organism and other factors [45]. Interspersed repeats comprise an isolating mechanism that enables new genes to evolve without interference from the progenitor gene. Therefore, repetitive sequence analysis might help to understand the evolution of *H. coralloides*.

The word “gene” has been synonymous for several decades with a genome encoding mRNAs that are translated into proteins. However, recent genome-wide studies have revealed thousands of regulatory ncRNAs that produce a functional RNA product instead of a translated protein [46]. The range and importance of such genes have only recently become apparent, with known ncRNAs playing a wide range of intracellular structural, regulatory, and catalytic roles [35, 47]. Our findings were similar to those of another study on the *Cordyceps guangdongensis* genome, which contains 314 ncRNAs. We predicted 338 ncRNAs among which, 270 were tRNAs that are key components of the translational machinery connecting the genetic code with the amino acid sequences of proteins. They comprise up to 15% of the total cellular RNA and are among the most abundant cellular transcripts [48]. Furthermore, tRNAs regulate numerous cellular and metabolic processes in eukaryotes and prokaryotes. The prediction of ncRNA in *H. coralloides* might contribute to further exploration of its cellular and metabolic processes.

CAZymes build and break down complex carbohydrates and glycoconjugates for many biological roles [49]. A saprophytic lifestyle is closely associated with CAZymes in fungal genomes [50, 51]. *H. coralloides* is a species of wood-rotting fungi, suggesting that it has ligninolytic, cellulolytic, hemicellulolytic, and pectinolytic properties that could be further investigated based on the annotations of CAZymes for some industrial applications. The annotation of CAZymes of *H.*

*coralloides* led to the identification of increased levels of GHs and AAs and decreased levels of CBMs, which might provide a deeper understanding of the ecological roles and carbohydrate metabolic mechanisms of *H. coralloides*. In addition, *H. coralloides* laccase has various substrates, the most sensitive of which is 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulfonic acid) [52]. Furthermore, *H. coralloides* produces an extracellular laccase, which catalyzes epigallocatechin gallate (ETFGg) synthesis from epigallocatechin gallate (EGCg) with gallic acid [11]. To the best of our knowledge, ETFGg has physiological functions. Therefore, more genes encoding useful enzymes might be discovered based on whole-genome sequencing.

Genomic sequencing has suggested that most genes that specify core biological functions are shared by all eukaryotes [53]. Rational annotation of proteins that are encoded in the sequenced genome can render genome sequences useful for functional and evolutionary investigations [54]. Therefore, we studied a series of functional annotations for the genome and identified numerous metabolic and organismal system-associated functions and pathways, such as cellular nitrogen compound metabolic process and carbohydrate metabolism. Because *H. coralloides* has an abundance of compounds from which metabolites can be extracted using biochemical and physiological means, the present findings serve as an important platform for further biological investigation into the microbe. The positive regulation of these functions might explain the edible and medicinal properties of *H. coralloides*. We also found that *H. coralloides* is a bioactive repository of natural compounds as the extracts of mycelia and fruiting bodies contained many metabolites with neurotrophic effects [55] and antioxidant activity [56], such as corallocins A-C [57], spirobenzofuran [58], and other new compounds [56]. The edible and medicinal properties of *H. coralloides* are further supported by phylogenetic findings indicating its close phylogenetic proximity to the edible fungi genera, *Russula* and *Lactarius*, which have high nutritional value [59].

In summary, we generated and analyzed a draft genome assembly of *H. coralloides* using a combined SMRT long-read sequencing approach. Our novel genomic data will provide valuable resources for edible and medical mushroom investigation and facilitate further studies on the genetic basis of *H. coralloides*. Our findings might also help further explorations on the bioactivity and evolution of *H. coralloides*.

## 5. Conclusions

The study sequenced the entire genome of *H. coralloides* (NCBI; Assembly: ASM367540v1), which is widely applied in TCM. The assembled genome was 5,59,05,675 bp in length and had 307 contigs. We identified 11,736 genes and 73,583 exons of lengths 2,52,64,974 and 2,02,04,958 bp, respectively. The mapping rate of the reads obtained from PE400 DNA in the assembled genome was 92.46%. We also predicted 48,895 bp ncRNAs and 11,736 genes encoding proteins that were annotated in the GO, eggNOG, and KEGG databases. In the future, we plan to detect the bioac-

tive genes in *H. coralloides* and their related processes using genomic data.

## Data Availability

Genomic data can be obtained from NCBI (<https://www.ncbi.nlm.nih.gov/>) (Assembly: ASM367540v1).

## Conflicts of Interest

The authors declare that there are no conflicts of interest.

## Authors' Contributions

Manjun Yang conceived the idea and supervised the project. Caixia Zhang and Lijun Xu as principal authors were responsible for processing experiments and wrote the manuscript. Caixia Zhang, Lijun Xu, Jian Li, and Jiansong Chen performed the experiments and analyzed the data. Caixia Zhang, Jian Li, and Jiansong Chen participated in the data analysis. All authors have carefully read and approved the manuscript. Caixia Zhang and Lijun Xu contributed equally to this work.

## Acknowledgments

We are sincerely thankful for Mr. Tamdrin Tsering's guidance during sampling and Mr. Peng Gao's help in cultivation of *H. coralloides*. This research was funded by the Young Scholar Innovation Project from Tibet Autonomous Region (QC 2015-76 and QCZ 2016-83), Tibetan Natural Scientific Foundation (2016ZR-15-64 and 2016ZR-ZX-02), and Program for Scientific and Technological Innovation Team Construction in Universities of Tibet Autonomous Region (Tibet Vocational Technical College 2014-2017). The genome sequencing was then performed by Personal Biotechnology Company (Shanghai, China).

## Supplementary Materials

Table S1: the list of fungi species used for phylogenetic construction in this study. (*Supplementary Materials*)

## References

- [1] S. A. Heleno, L. Barros, M. J. Sousa, A. Martins, and I. C. F. R. Ferreira, "Study and characterization of selected nutrients in wild mushrooms from Portugal by gas chromatography and high performance liquid chromatography," *Microchemical Journal*, vol. 93, no. 2, pp. 195–199, 2009.
- [2] P. Kalač, "Chemical composition and nutritional value of European species of wild growing mushrooms: a review," *Food Chemistry*, vol. 113, no. 1, pp. 9–16, 2009.
- [3] P. Kalač, "A review of chemical composition and nutritional value of wild-growing and cultivated mushrooms," *Journal of the Science of Food and Agriculture*, vol. 93, no. 2, pp. 209–218, 2013.
- [4] M. J. Alves, I. C. Ferreira, J. Dias, V. Teixeira, A. Martins, and M. Pintado, "A review on antimicrobial activity of mushroom (*Basidiomycetes*) extracts and isolated compounds," *Planta Medica*, vol. 78, no. 16, pp. 1707–1718, 2012.

- [5] I. C. Ferreira, L. Barros, and R. M. Abreu, "Antioxidants in wild mushrooms," *Current Medicinal Chemistry*, vol. 16, no. 12, pp. 1543–1560, 2009.
- [6] I. C. Ferreira, J. A. Vaz, M. H. Vasconcelos, and A. Martins, "Compounds from wild mushrooms with antitumor potential," *Anti-Cancer Agents in Medicinal Chemistry*, vol. 10, no. 5, pp. 424–436, 2010.
- [7] D. A. McCracken and J. L. Dodd, "Molecular structure of starch-type polysaccharides from *Hericium ramosum* and *Hericium coralloides*," *Science*, vol. 174, no. 4007, p. 419, 1971.
- [8] Z. J. Wang, D. H. Luo, and Z. Y. Liang, "Structure of polysaccharides from the fruiting body of *Hericium erinaceus* Pers," *Carbohydrate Polymers*, vol. 57, no. 3, pp. 241–247, 2004.
- [9] J. Chen, X. Zeng, Y. L. Yang et al., "Genomic and transcriptomic analyses reveal differential regulation of diverse terpenoid and polyketides secondary metabolites in *Hericium erinaceus*," *Scientific Reports*, vol. 7, no. 1, p. 10151, 2017.
- [10] S. A. Heleno, L. Barros, A. Martins et al., "Chemical composition, antioxidant activity and bioaccessibility studies in phenolic extracts of two *Hericium* wild edible species," *LWT-Food Science and Technology*, vol. 63, no. 1, pp. 475–481, 2015.
- [11] N. Itoh, S. Takagi, A. Miki, and J. Kurokawa, "Characterization and cloning of laccase gene from *Hericium coralloides* NBRC 7716 suitable for production of epigallocatechin gallate," *Enzyme and Microbial Technology*, vol. 82, pp. 125–132, 2016.
- [12] X. Y. Wang, J. Y. Yin, M. M. Zhao, S. Y. Liu, S. P. Nie, and M. Y. Xie, "Gastroprotective activity of polysaccharide from *Hericium erinaceus* against ethanol-induced gastric mucosal lesion and pylorus ligation-induced gastric ulcer, and its antioxidant activities," *Carbohydrate Polymers*, vol. 186, pp. 100–109, 2018.
- [13] F. Wu, C. Zhou, D. Zhou, S. Ou, X. Zhang, and H. Huang, "Structure characterization of a novel polysaccharide from *Hericium erinaceus* fruiting bodies and its immunomodulatory activities," *Food & Function*, vol. 9, no. 1, pp. 294–306, 2018.
- [14] B. T. M. Dentinger, E. Gaya, H. O'Brien et al., "Tales from the crypt: genome mining from Fungarium specimens improves resolution of the mushroom tree of life," *Biological Journal of the Linnean Society*, vol. 117, no. 1, pp. 11–32, 2016.
- [15] S. Djebali, C. A. Davis, A. Merkel et al., "Landscape of transcription in human cells," *Nature*, vol. 489, no. 7414, pp. 101–108, 2012.
- [16] U. Nagalakshmi, Z. Wang, K. Waern et al., "The transcriptional landscape of the yeast genome defined by RNA sequencing," *Science*, vol. 320, no. 5881, pp. 1344–1349, 2008.
- [17] K. F. Au, V. Sebastiano, P. T. Afshar et al., "Characterization of the human ESC transcriptome by hybrid sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 50, pp. E4821–E4830, 2013.
- [18] J. Eid, A. Fehr, J. Gray et al., "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.
- [19] D. Sharon, H. Tilgner, F. Grubert, and M. Snyder, "A single-molecule long-read survey of the human transcriptome," *Nature Biotechnology*, vol. 31, no. 11, pp. 1009–1014, 2013.
- [20] M. Schubert, S. Lindgreen, and L. Orlando, "AdapterRemoval v2: rapid adapter trimming, identification, and read merging," *BMC Research Notes*, vol. 9, no. 1, p. 88, 2016.
- [21] R. M. Leggett, B. J. Clavijo, L. Clissold, M. D. Clark, and M. Caccamo, "NextClip: an analysis and read preparation tool for nextera long mate pair libraries," *Bioinformatics*, vol. 30, no. 4, pp. 566–568, 2014.
- [22] D. R. Kelley, M. C. Schatz, and S. L. Salzberg, "Quake: quality-aware detection and correction of sequencing errors," *Genome Biology*, vol. 11, no. 11, p. R116, 2010.
- [23] G. Marçais and C. Kingsford, "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers," *Bioinformatics*, vol. 27, no. 6, pp. 764–770, 2011.
- [24] J. Korf, G. Gedman, S. B. Kingan et al., "De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads," *Gigascience*, vol. 6, no. 10, pp. 1–16, 2017.
- [25] B. J. Walker, T. Abeel, T. Shea et al., "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement," *PLoS One*, vol. 9, no. 11, article e112963, 2014.
- [26] M. Boetzer and W. Pirovano, "Toward almost closed genomes with GapFiller," *Genome Biology*, vol. 13, no. 6, p. R56, 2012.
- [27] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs," *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, 2015.
- [28] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.
- [29] A. P. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock, "Repetitive elements may comprise over two-thirds of the human genome," *PLoS Genetics*, vol. 7, no. 12, article e1002384, 2011.
- [30] S. Tempel, "Using and understanding RepeatMasker," *Methods in Molecular Biology*, vol. 859, pp. 29–51, 2012.
- [31] V. V. Kapitonov and J. Jurka, "A universal classification of eukaryotic transposable elements implemented in Repbase," *Nature Reviews Genetics*, vol. 9, no. 5, pp. 411–412, 2008.
- [32] J. S. Mattick and I. V. Makunin, "Non-coding RNA," *Human Molecular Genetics*, vol. 15, supplement 1, pp. R17–R29, 2006.
- [33] T. M. Lowe and S. R. Eddy, "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence," *Nucleic Acids Research*, vol. 25, no. 5, pp. 955–964, 1997.
- [34] K. Lagesen, P. Hallin, E. A. Rødland, H. H. Stærfeldt, T. Rognes, and D. W. Ussery, "RNAmmer: consistent and rapid annotation of ribosomal RNA genes," *Nucleic Acids Research*, vol. 35, no. 9, pp. 3100–3108, 2007.
- [35] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, "Rfam: an RNA family database," *Nucleic Acids Research*, vol. 31, no. 1, pp. 439–441, 2003.
- [36] M. Stanke and B. Morgenstern, "AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints," *Nucleic Acids Research*, vol. 33, no. Web Server, pp. W465–W467, 2005.
- [37] W. H. Majoros, M. Pertea, and S. L. Salzberg, "TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders," *Bioinformatics*, vol. 20, no. 16, pp. 2878–2879, 2004.
- [38] I. Korf, "Gene finding in novel genomes," *BMC Bioinformatics*, vol. 5, no. 1, p. 59, 2004.
- [39] B. J. Haas, A. Papanicolaou, M. Yassour et al., "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis," *Nature Protocols*, vol. 8, no. 8, pp. 1494–1512, 2013.

- [40] B. J. Haas, S. L. Salzberg, W. Zhu et al., "Automated eukaryotic gene structure annotation using EVIDENCEModeler and the program to assemble spliced alignments," *Genome Biology*, vol. 9, no. 1, p. R7, 2008.
- [41] B. J. Haas, A. L. Delcher, S. M. Mount et al., "Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies," *Nucleic Acids Research*, vol. 31, no. 19, pp. 5654–5666, 2003.
- [42] A. Conesa and S. Götz, "Blast2GO: a comprehensive suite for functional analysis in plant genomics," *International Journal of Plant Genomics*, vol. 2008, Article ID 619832, 2008.
- [43] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa, "KAAS: an automatic genome annotation and pathway reconstruction server," *Nucleic Acids Research*, vol. 35, no. Web Server, pp. W182–W185, 2007.
- [44] S. S. Sindi, B. R. Hunt, and J. A. Yorke, "Duplication count distributions in DNA sequences," *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 78, no. 6, article 061912, 2008.
- [45] J. C. Schimenti and C. H. Duncan, "Ruminant globin gene structures suggest an evolutionary role for Alu-type repeats," *Nucleic Acids Research*, vol. 12, no. 3, pp. 1641–1655, 1984.
- [46] J. Gebetsberger and N. Polacek, "Slicing tRNAs to boost functional ncRNA diversity," *RNA Biology*, vol. 10, no. 12, pp. 1798–1806, 2013.
- [47] C. Zhang, W. Deng, W. Yan, and T. Li, "Whole genome sequence of an edible and potential medicinal fungus, *Cordyceps guangdongensis*," *G3-Genes Genomes Genetics*, vol. 8, pp. 1863–1870, 2018.
- [48] Y. Li and H. Zhou, "tRNAs as regulators in gene expression," *Science in China Series C: Life Sciences*, vol. 52, no. 3, pp. 245–252, 2009.
- [49] B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat, "The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics," *Nucleic Acids Research*, vol. 37, no. Database, pp. D233–D238, 2009.
- [50] E. Espagne, O. Lespinet, F. Malagnac et al., "The genome sequence of the model ascomycete fungus *Podospora anserina*," *Genome Biology*, vol. 9, no. 5, p. R77, 2008.
- [51] D. Martinez, J. Challacombe, I. Morgenstern et al., "Genome, transcriptome, and secretome analysis of wood decay fungus *Postia placenta* supports unique mechanisms of lignocellulose conversion," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 6, pp. 1954–1959, 2009.
- [52] Y. J. Zou, H. X. Wang, T. B. Ng, C. Y. Huang, and J. X. Zhang, "Purification and characterization of a novel laccase from the edible mushroom *Hericium coralloides*," *Journal of Microbiology*, vol. 50, no. 1, pp. 72–78, 2012.
- [53] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [54] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, "The COG database: a tool for genome-scale analysis of protein functions and evolution," *Nucleic Acids Research*, vol. 28, no. 1, pp. 33–36, 2000.
- [55] Y. Kimura, M. Nishibe, H. Nakajima et al., "Hericerin, a new pollen growth inhibitor from the mushroom *Hericium erinaceum*," *Agricultural and Biological Chemistry*, vol. 55, pp. 2673–2674, 1991.
- [56] J. Y. Kim, E. E. Woo, I. K. Lee, and B. S. Yun, "New antioxidants from the culture broth of *Hericium coralloides*," *Journal of Antibiotics*, vol. 71, no. 9, pp. 822–825, 2018.
- [57] K. Wittstein, M. Rascher, Z. Rupcic et al., "Coralloins A-C, nerve growth and brain-derived neurotrophic factor inducing metabolites from the mushroom *Hericium coralloides*," *Journal of Natural Products*, vol. 79, no. 9, pp. 2264–2269, 2016.
- [58] P. Kleinwachter, B. Schlegel, H. Dörfelt, and U. Grafe, "Spirobenzofuran, a new bioactive metabolite from *Acremonium sp. HKI 0230*," *Journal of Antibiotics*, vol. 54, no. 6, pp. 526–527, 2001.
- [59] M. Rasalanavho, R. Moodley, and B. S. Jonnalagadda, "Elemental bioaccumulation and nutritional value of five species of wild growing mushrooms from South Africa," *Food Chemistry*, vol. 319, article 126596, 2020.