Whole genome sequence of *Auricularia heimuer* (Basidiomycota, Fungi), the third most important cultivated mushroom worldwide

Yuan Yuan\(^{a,b}\), Fang Wu\(^b\), Jing Si\(^b\), Yi-Fan Zhao\(^c\), Yu-Cheng Dai\(^{a,b,*}\)

\(^{a}\) Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Beijing Forestry University, Beijing 100083, China
\(^{b}\) Institute of Microbiology, Beijing Forestry University, Beijing 100083, China
\(^{c}\) Beijing Genomics Institute, Shenzhen 518083, China

**Abstract**

*Heimuer, Auricularia heimuer,* is one of the most famous traditional Chinese foods and medicines, and it is the third most important cultivated mushroom worldwide. The aim of this study is to develop genomic resources for *A. heimuer* to furnish tools that can be used to study its secondary metabolite production capability, wood degradation ability and biosynthesis of polysaccharides. The genome was obtained from single spore mycelia of the strain Dai 13782 by using combined high-throughput Illumina HiSeq 4000 system with the PacBio RSII long-read sequencing platform. Functional annotation was accomplished by blasting protein sequences with different public available databases to obtain their corresponding annotations. It is 49.76 Mb in size with a N50 scaffold size of 1,350,668 bp and encodes 16,244 putative predicted genes. This is the first genome-scale assembly and annotation for *A. heimuer*, which is the third sequenced species in *Auricularia*.

**1. Introduction**

*Auricularia* is a very important genus of wood-decaying fungi, and several species are widely cultivated as edible and medicinal mushrooms in China and some other Asian countries. Among them, Heimuer, previously identified as *Auricularia auricula* (L.) Underw or *A. auricula-judae* (Bull.) Quél. but currently as *A. heimuer* F. Wu, B.K. Cui & Y.C. Dai [1], is the most common and popular species of the genus, and is the third most important cultivated mushrooms worldwide (Fig. 1; [2]). It has good taste and a wide spectrum of health-associated properties, including antitumor, cholesterol-lowering, anticoagulant, antioxidant, immunomodulatory, anti-inflammatory, and antimicrobial activities [3]. The earliest report about its medicinal value appeared in the famous Chinese medicinal monograph *Compendium of Materia Medica* [3]. The cultivation of Heimuer could be traced back to 1300 years ago [6].

Fruiting bodies of *A. heimuer* were collected by Yu-Cheng Dai on Quercus in Ning’an County, Heilongjiang Province, northeastern China. The specimen (Dai 13782) was deposited in the herbarium of the Institute of Microbiology, Beijing Forestry University (BJFC), with the accession number BJFC017513. A single spore strain was obtained from the specimen following the methods of Choi et al. [17]. Genomic DNA was extracted from the single spore mycelia by using Cetyltrimethyl Ammonium Bromide (CTAB) method [18].

We sequenced the genome of *A. heimuer* strain Dai 13782 using a whole-genome shotgun strategy by high-throughput Illumina HiSeq...
In order to ensure the accuracy of the follow-up analyses, reads of low complexity, low quality with adapter and duplication contaminations were removed. Several steps were performed.

For filtering the raw data of 350 bp library: Read 1 and read 2 were filtered out based on adapter contaminations (15 bp overlap between adapter and reads as default, parameter setting at 15 bp) were removed; reads with a certain proportion of Ns’ base or low complexity reads (10% as default, parameter setting at 10 bp) were removed; adapter contaminations (15 bp overlap between adapter and reads as default, parameter setting at 15 bp) were removed; duplication contaminations were removed.

For filtering the raw data of 10 kb library: Read 1 and read 2 were cut by 1–100 bp; reads with a certain proportion of low quality (Read Quality ≤ 20) bases (40% as default, parameter setting at 40 bp) were removed; reads with a certain proportion of Ns’ base or low complexity reads (10% as default, parameter setting at 5 bp) were removed; adapter contaminations (15 bp overlap between adapter and reads as default, parameter setting at 15 bp) were removed; duplication contaminations were removed.

The above processes are applied to filter read 1 and read 2 based on conventional BGI pipelines. After that, 10%–20% of the data is eliminated generally (For small insert size reads). Because large insert size reads have high duplication, the data eliminated is more but there is no certain proportion.

Reads of PacBio RSII were filtered using custom BGI pipelines with following methods: polymerase reads with length < 100 bp were filtered out; polymerase reads with quality score < 0.8 were filtered out; adapter sequences were filtered out and then subreads were produced; subreads with length < 1000 bp were filtered out; subreads with quality score < 0.8 were filtered out for further analysis.

PacBio data were assembled using the SMRT Analysis v2.3.0 workflow into contigs. GATK 3.4.0, SOAPsnp and SOAPindel were used to correct single bases from the assembly results. The parameters of GATK were: -cluster 2 -window 5 -stand_call_conf 50 -stand_emit_conf 20 -dovc 200 MQ0 ≥ 4. The parameters of SOAPsnp and SOAPindel were the default. After that, SPSPACE Basic v2.0 [19] was used to construct the scaffold by using HiSeq long-insert-size data (default parameters). For the PacBio data, Pbsuite 14.9.9 was used to fill gaps by default parameters.

BUSCO (Benchmarking Universal Single-Copy Orthologs) software was used to assess the completeness of genome assembly and annotation with single-copy orthologs [20]. BUSCO v2.0 was run on the scaffolded genome assembly (using “–m genome”). The lineage dataset of BUSCO was fungi_odb9 (Creation date: 2016-10-21, number of species: 85, number of BUSCOs: 290). In addition, assembly of A. heimuer was compared with A. auricula-judae, A. subglabra and E. glandulosa.

2.2. Genome annotation methods

Protein-encoding genes were predicted applying the ab initio gene predictors SNAP [21], Augustus [22], and GeneMark-ES [23]. The resulting gene sets were integrated to obtain the most comprehensive and non-redundant reference genes.

Transposon sequence analysis of A. heimuer was carried out for the assembled gene sequences with the transposon Repbase database.

4000 sequencing and the PacBio RSII long-read sequencing platform at BGI-Shenzhen, China. Paired-end reads were generated by sequencing of insert libraries of 350 bp and 10 kb using Illumina HiSeq4000 system. For PacBio RSII platform, a 20 kb library was built and sequenced.

In order to ensure the accuracy of the follow-up analyses, reads of low complexity, low quality with adapter and duplication contamination were removed from the raw data. Several steps were performed.

For filtering the raw data of 350 bp library: Read 1 and read 2 were cut by 1–100 bp; reads with a certain proportion of low quality (Read Quality ≤ 20) bases (40% as default, parameter setting at 40 bp) were removed; reads with a certain proportion of Ns’ base or low complexity reads (10% as default, parameter setting at 10 bp) were removed; adapter contaminations (15 bp overlap between adapter and reads as default, parameter setting at 15 bp) were removed; duplication contaminations were removed.

For filtering the raw data of 10 kb library: Read 1 and read 2 were cut by 1–49 bp; reads with a certain proportion of low quality (Read Quality ≤ 20) bases (40% as default, parameter setting at 20 bp) were removed; reads with a certain proportion of Ns’ base or low complexity reads (10% as default, parameter setting at 5 bp) were removed; adapter contaminations (15 bp overlap between adapter and reads as default, parameter setting at 15 bp) were removed; duplication contaminations were removed.

The above processes are applied to filter read 1 and read 2 based on conventional BGI pipelines. After that, 10%–20% of the data is eliminated generally (For small insert size reads). Because large insert size reads have high duplication, the data eliminated is more but there is no certain proportion.

Reads of PacBio RSII were filtered using custom BGI pipelines with following methods: polymerase reads with length < 100 bp were filtered out; polymerase reads with quality score < 0.8 were filtered out; adapter sequences were filtered out and then subreads were produced; subreads with length < 1000 bp were filtered out; subreads with quality score < 0.8 were filtered out for further analysis.

PacBio data were assembled using the SMRT Analysis v2.3.0 workflow into contigs. GATK 3.4.0, SOAPsnp and SOAPindel were used to correct single bases from the assembly results. The parameters of GATK were: -cluster 2 -window 5 -stand_call_conf 50 -stand_emit_conf 40 -dovc 200 MQ0 ≥ 4. The parameters of SOAPsnp and SOAPindel were the default. After that, SPSPACE Basic v2.0 [19] was used to construct the scaffold by using HiSeq long-insert-size data (default parameters). For the PacBio data, Pbsuite 14.9.9 was used to fill gaps by default parameters.

BUSCO (Benchmarking Universal Single-Copy Orthologs) software was used to assess the completeness of genome assembly and annotation with single-copy orthologs [20]. BUSCO v2.0 was run on the scaffolded genome assembly (using “–m genome”). The lineage dataset of BUSCO was fungi_odb9 (Creation date: 2016-10-21, number of species: 85, number of BUSCOs: 290). In addition, assembly of A. heimuer was compared with A. auricula-judae, A. subglabra and E. glandulosa.
using RepeatMasker software (version 3.3.0) at [http://www.repeatmasker.org/](http://www.repeatmasker.org/) and RepeatProteinMasker software (using the transposon protein library that comes with RepeatMasker). Tandem repeat sequences were searched in the DNA sequence with Tandem Repeat Finder (TRF version 4.04, [24]).

rRNA sequences were predicted by RNAmmer (version 1.2) prediction software with the whole genome sequence [27]. tRNA genes and tRNA secondary structure were identified by tRNAscan-SE software (version 1.23, [28]). Other non-coding RNAs, such as miRNA, sRNA and snRNA, were predicted by Rfam (version 1.23, [28]).

The functional annotations of predicted genes were mainly based on homology to known annotated genes within different databases using BLAST as the main tool. For the purpose of achieving their corresponding annotation, we aligned protein models in SwissProt [30], TrEMBL, the National Center for Biotechnology Information (NCBI) non-redundant (nr) database, InterPro, GO [31,32], NCBI Clusters of Orthologous Groups of proteins (COG) [33,34], and Kyoto Encyclopedia of Genes and Genomes (KEGG) [35,36].

Annotations of carbohydrate-active enzymes (CAZymes) in the A. heimuer genome were actualized with the CAZy database by BLASTP at [http://www.cazy.org/](http://www.cazy.org/). The reference P450 sequences were downloaded from [http://drnelson.uthsc.edu/CytochromeP450.html](http://drnelson.uthsc.edu/CytochromeP450.html). Auricularia heimuer gene models were aligned to search the reference P450 data set using the BLASTP program with a cut off e-value ≤ 1e − 10 [38].

AntiSMASH fungal 4.0.0, a web-based analysis platform, was employed to predict the gene clusters of secondary metabolites [39]. The parameter settings were maintained as the default parameter values. To validate the predicted results, the obtained gene clusters were checked manually. BLAST analysis and gene annotation were carried out within a software platform provided by the JGI genome portal [40]. The BLASTP and TBLASTN algorithms were used to search all hypothetical gene models present in the database.

The sequence data of other taxa used in the present paper were downloaded from the US Department of Energy Joint Genome Institute website at [http://genome.jgi.doe.gov/](http://genome.jgi.doe.gov/) and the National Center for

<table>
<thead>
<tr>
<th>Species</th>
<th>GCA</th>
<th>BUSCO mode</th>
<th>Complete</th>
<th>Fragmented</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. heimuer</td>
<td>002287115.1</td>
<td>Genome</td>
<td>268 (92.4%)</td>
<td>16 (5.5%)</td>
<td>6 (2.1%)</td>
</tr>
<tr>
<td>A. auricula-judae</td>
<td>002092955.1</td>
<td>Genome</td>
<td>280 (96.5%)</td>
<td>4 (1.4%)</td>
<td>6 (2.1%)</td>
</tr>
<tr>
<td>A. subglabra</td>
<td>000265015.1</td>
<td>Genome</td>
<td>274 (94.5%)</td>
<td>8 (2.8%)</td>
<td>8 (2.7%)</td>
</tr>
<tr>
<td>E. glandulosa</td>
<td>001632375.1</td>
<td>Genome</td>
<td>273 (94.1%)</td>
<td>6 (2.1%)</td>
<td>11 (3.8%)</td>
</tr>
</tbody>
</table>

* This is the number of GenBank assembly accession.

* Number of BUSCO proteins (percent of total BUSCOs).
3. Results and discussion

3.1. The data of sequencing and assembly

*Auricularia heimuer* generated 6614 Mb HiSeq-data and 2835 Mb PacBio-data. The details of data generation are listed in Tables 1 and 2, respectively.

3.2. Genome features

The details of gene annotation summary statistics of *Auricularia heimuer* are shown in Table 3.

Currently, three genomes of Auriculariales species (*Auricularia subglabra*, *Auricularia auricula-judae* and *Exidia glandulosa*) are available in the NCBI WGS database. Assembly statistics of *A. heimuer* are compared to other three species in Table 4, showing similar GC content but significant different genome sizes with *A. subglabra* and *E. glandulosa*.

The completeness of genome assembly and annotation with single-copy orthologs test results suggested a well complete annotation set, with 92.4% of the Fungi BUSCOs being present within the RefSeq annotation set, and 5.5% of those fragmented. Details of BUSCO analysis are presented in Table 5.

3.3. Repeat sequences

The total length of repeat sequences was 3,429,083 bp, covering 6.89% of the genomic length. Of the repeat elements, tandem repeat sequences account for 2.32% and transposable elements (TEs) were about 4.57% of the assembled genome. Among the TEs, the proportion of long terminal repeats (LTRs) and non-LTR transposons were 2.96% and 0.99% of the genome, respectively.

3.4. Gene annotation

The annotations of 16,244 protein-encoding genes were supported using publicly available databases. All the protein-encoding genes together had a total sequence length of 28,840,601 bp, including exons areas of 22,015,521 bp, and intron areas of 6,825,080 bp. The protein-encoding region accounted for 44.24% of the genome. On average, each predicted gene contained 5.36 exons. Genes typically contained small exons (average 252.67 bp) and introns (average 98.18 bp), which is similar to other basidiomycetes.

For RNA, 150 tRNAs, 34 snRNAs, and 11 rRNAs were predicted. The
gene density was 5.63 genes/10 kb, and the average size of protein-encoding genes was 1355.30 bp. Among the tRNAs, 12 were possible pseudogenes and the remaining 138 anti-codon tRNAs corresponded to the 20 common amino acid codons. In addition, 79 tRNAs were absence of intron sequences.

A total of 1109 predicted genes encoding hypothetical proteins without apparent homologs to currently available sequences were found in the *A. heimuer* genome. The annotations with NCBI nr, GO, NCBI COG, KEGG, CAZy, SwissProt, PHI, InterPro and TrEMBL protein databases are shown in Tables S2 and S3. These homologous protein genes represent 93.17% of the assembled genome.

By NCBI COG mapping, 1801 proteins (11.08%) were assigned to COG categories (Fig. 2). “General functional prediction only” had the highest number of genes (408); these proteins were not unambiguously assigned to a particular group. This was followed by “Translation, ribosomal structure and biogenesis”, “Posttranslational modification, protein turnover, chaperones”, “Transcription”, “Replication, recombination and repair”, and “Energy production and conversion” as the most gene-rich classes in the COG groupings. These findings were suggestive of the presence of an enriched and varied array of protein metabolism and energy metabolism functions that enable better absorption and transformation of nutrients from substrates.

According to GO database, 7056 predicted proteins that accounted for 43.43% of the entire genome were mainly distributed in five functional entries, “Binding”, “Catalytic activity”, “Cellular process”, “Metabolic process”, and “Single-organism process” (Fig. 3).

To further understand the gene functions in *A. heimuer*, we successfully assigned 5676 (34.94%) putative proteins to their orthologs in the KEGG database. The KEGG function classification is shown in Fig. 4. Similar with the COG annotation, some metabolism and biosynthesis categories in KEGG were highly enriched.

### 3.5. The CAZyme

We mapped our fungal genomes with CAZy to study the presence of CAZymes. A total of 336 genes could be assigned to carbohydrate-active enzyme (CAZymes) families as defined in CAZy database (Table 6). The number of GHs was remarkably larger than GTs, which could be due to the lifestyle, in which its survival is dependent on lignocellulose decomposition. It is concluded that the polysaccharides decomposition is more important than their construction for growth and metabolism of *A. heimuer*. The number of CBMs in *A. heimuer* is more than average, while CEs, GTs, GHs and PLs were less than the basidiomycetes average (Table S4).

### 3.6. The CYPs family

It is worth to note that *A. heimuer* has 72 genes distributed in “Metabolism of xenobiotics by cytochrome P450” and 60 in “Drug metabolism – cytochrome P450” KEGG sub-pathways. *Auricularia heimuer* had a total of 83 CYP genes, which could be
Fig. 5. "Terpenoid backbone biosynthesis" pathway of *Auricularia heimuer*. The red box indicates existing homologous genes of the enzyme, while white box means not. The photo was done by KEGG mapper. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
classified into 21 families according to Nelson’s nomenclature (Table 7) [38]. The CYP64 family was found to have the highest number of genes (26 genes), followed by the CYP2 (10 genes), and CYP4 (9 genes) families (Table 5).

The CYP64 family is involved in “Metabolism – Biosynthesis of other secondary metabolites”, while genes from the CYP73, CYP76 and CYP97 families are necessary in “Metabolism – Metabolism of terpenoids and polyketides”. The CYP2 and CYP4 families play a role in “Lipid metabolism”. However, these predicted CYP genes still need to be validated in further studies.

3.7. Secondary metabolisms

For modern pharmacology and therapeutics fields, deep interest has been generated in medicinal fungi due to their diverse and significant secondary metabolites, which have been proved to possess pharmacological properties, such as immunomodulatory, anti-inflammatory, antiviral, antioxidant, antitumor, and antimicrobial activities [41,42].

The biosynthesis genes of secondary metabolites are often in contiguous gene clusters [43]. We searched for genes that may be involved in biosynthesis of secondary metabolites through genome mining based on the previous studies such as terpene synthases (TS), non-ribosomal peptide synthetases (NRPS), and polyketide synthases (PKS) [44–49]. One NRPS gene clusters were identified in the A. heimuer genome. The details of secondary metabolite gene clusters are listed in Table S5.

The potential PKS genes were aligned to Genebank with blastp and their function information were manual checked. The results demonstrate that A. heimuer has no PKS gene. The reason for this phenomenon still needs further research. One lineage of NPRS gene clusters of A. heimuer may be responsible for intracellular siderophore biosynthesis that encodes a single copy gene for siderophore synthetase (Dai13782.GLEAN.10006893). Ferrichrome siderophores perform key functions in fungal cells, which have main role in sexual spore development and asexual development [50,51]. All known fungal siderophores are synthesized by NRPBs [52].

In fungi, one of the largest groups of bioactive natural products that have been identified is the terpenoids, which is derived from the simple five carbon precursor molecules dimethylallyl diphosphate (DMAPP) and isopentenyl diphosphate (IPP) produced from acetyl-CoA via the mevalonate pathway [47,53]. There are 5 gene clusters encoding TS, which are key enzyme for the biosynthesis of terpene. We checked the A. heimuer genes in the “Terpenoid backbone biosynthesis (map00900)” pathway and identified 15 key enzymes distributed in the MVA pathway (Fig. 5). All the core enzymes involved in the MVA pathway are summarized in Table 8. The enzymes protein-S-isoprenylcysteine-O-methyltransferase, farnesyl diphosphate synthase, and diposphomevalonate decarboxylase are each encoded by two copies of the genes, whereas the remaining 12 enzymes are encoded by single-copy of genes.

The search of potential triterpenoid biosynthesis genes in A. heimuer genome resulted in an ORF (Dai13782.GLEAN.10010970) that encodes a single copy gene for lanosterol synthase (LS) (ERG7; K01852; EC5.4.99.7). LSS is an oxidosqualene cyclase family enzyme that catalyzes the cyclization of the triterpenes squalene or 2,3-oxidosqualene to a protersteral cation and finally to lanosterol, the precursor of all steroids [54]. Lanosterol is a key four-ringed intermediate in cholesterol biosynthesis [55]. The CYP73, CYP76, and CYP97 families of P450 genes have been predicted to participate in “Metabolism – Metabolism of terpenoids and polyketides”.

3.8. Biosynthesis of polysaccharides in A. heimuer

Polysaccharides that are accounted for the pharmaceutical potential of A. heimuer are another kind of extensively bioactive compounds.

Among the polysaccharides, water soluble β-1,3-glucan with side chain β-1,6-glucose has the most active immunomodulating and anti-tumor compounds [56]. Table S6 lists 33 enzymes that may be involved in the biosynthesis of uridine diphosphate glucose (UDP-glucose, precursor of glucans) in A. heimuer. These enzymes include 1,3-beta-glucan synthasem, hexokinase, phosphoglucomutase, GTPase-activating protein, and UTP-glucose-1-phosphate uridylyltransferase.

4. Conclusion

In summary, the exposition of the A. heimuer genome has allowed us to predict gene function and study the biosynthesis of active compounds. A research of the secondary metabolite biosynthesis genes in the A. heimuer genome shows that it is particularly enriched with terpenoid biosynthesis genes. Hence, the future study of bioactive molecules can be concentrated in this class of metabolites. Although not complete, the elucidation of A. heimuer genome in this study provides foundational information for studying the biosynthesis of the pharmacologically active compounds and functional food applications in the future.

4.1. Data availability

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession NEKD00000000. The version described in this paper is version NEKD0100000. The genome raw
sequencing data and the assembly reported in this paper is associated with NCBI BioProject: PRJNA382748 and BioSample: SAMN0671819 within GenBank. The SRA accession number is SRBS589175.

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article and supplemental materials.

Acknowledgments

We thank the US Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/) in collaboration with the user community and the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/genome) for providing most of the sequencing data for fungi used in this study. The research was supported by the National Science Foundation of China (Project No. 31372115).

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jgen.2017.12.013.

References


