Journal of Bioscience and Bioengineering
VOL. 121 No. 4, 380–384, 2016

www.elsevier.com/locate/jbiosc

De novo transcriptome analysis of *Inonotus baumii* by RNA-seq

Li Zou, Tingting Sun,* Danlei Li, Yun Tan, Guoquan Zhang, Feng Wang, and Jian Zhang

College of Forestry, Northeast Forestry University, Hexing Road 26, Xiangfang District, Harbin 150040, Heilongjiang, China

Received 3 February 2015; accepted 5 September 2015
Available online 19 October 2015

Inonotus baumii*, a basidiomycete white rot fungus, has been widely used as traditional herbal medicine in China, Korea, Japan and other Asian countries for many years. Its extract is of great medicinal importance and plays a valuable role in the immune response and disease resistance. However, limited genetic resources for *I. baumii* have hindered exploration of this species. In order to gain a molecular understanding of this fungus, Illumina high-throughput technology was used to sequence and analyze the transcriptome of *I. baumii*, and 280,691 contigs, 43,890 scaffolds and 30,051 unigenes were obtained. Additionally, based on similarity search with known proteins, unigenes were annotated with gene descriptions, gene ontology (GO), clusters of orthologous group (COG), and database of protein families (Pfam) terms. According to the annotation of unigenes, a total of 12 candidate genes involved in the triterpenoid biosynthesis pathway and 21 putative FOLymes (fungal oxidative lignin enzymes) and 176 CAZymes (carbohydrate-active enzymes) were obtained using homology-based BlastX. Moreover, for better understanding of the transcripts function, the BlastX algorithm was used to search for homologous sequences against the Yeast genome. This is the first study on transcriptome analyses of *I. baumii*, which provided a dataset for functional gene mining and laid a basis for further functional genomics studies of *I. baumii*.

[**Key words:** *Inonotus baumii*; Transcriptome; Triterpenoid biosynthesis; Lignin degradation; Carbohydrate-active enzymes; Next-generation sequencing]

Inonotus baumii, a white-rot fungus (used to be identified as *Phellinus baumii*), has been used as a traditional herb in China and some other Asian countries for centuries (1). Its extract has been demonstrated to be effective on a diversity of diseases. Furthermore, this fungus has the potential to be used as lignocellulose degrading enzymes in recent years (2).

The main bioactive components of *I. baumii* and other fungi belonging to genera *Inonotus* and *Phellinus* include polysaccharides, triterpenoids, flavonoids and phenolic compounds (3–6). Among these, polysaccharides and triterpenoids have received considerable attention due to their important value. To date, the polysaccharides have been reported for anti-cancer, anti-diabetes, anti-inflammation, antioxidation, and so on (7–9). As for the triterpenoids, play an important role in antitumor, anti-inflammation and immune regulation (10–12). Moreover, a novel extracellular laccase from this fungus have been demonstrated with antiproliferative activity (13).

In recent years, the next-generation sequencing technology, such as Roche 454, Solexa/Illumina and ABI-SOLiD, have greatly improved the efficiency of genes discovery and have been widely used in many areas, including gene expression detecting, novel transcripts discovering, and differentially expressed genes testing (14–16). At present, several mushroom transcriptomes have been sequenced using the next-generation sequencing technology, including *Lentinula edode*, *Auricularia polytricha*, *Agrocybe aegerita*, *Ganoderma lucidum*, *Cordyceps militar*, *Amanita exitialis*, and so on (17–22).

Although *I. baumii* has huge important properties, the publicly available genetic data for *I. baumii* is so scarce. To date, there are only about 66 nucleotide sequences available on NCBI for *I. baumii*, and most of them are used for phylogenetic analysis (23,24). The current genomic sequence resources are too insufficient to reveal the pharmacological mechanisms of *I. baumii* at a molecular level.

This study was carried out using Illumina HiSeq 2000 technology to determine the transcriptomes of *I. baumii*. By *de novo* assembly, a total of 30,051 unigenes were obtained with 20,634 annotated, and by comprehensive analysis, these results provide an invaluable resource for identifying the synthesis of pharmaceutical components and mining lignocelluloses degrading enzymes in *I. baumii*.

MATERIALS AND METHODS

The fungal strain The DL101 strain was collected from *Syringa reticulata* at Liangshui Nature Reserve, Lesser Xing'an Mountains in Yichun city, Heilongjiang Province, China, in July 2009. This strain was identified as *I. baumii* according to ITS sequence alignment (GenBank accession number KP974834).

Sample preparation and RNA extraction The mycelium of *I. baumii* was maintained on potato dextrose agar plate and kept in the dark at 25°C for 10 days. Then inoculated into the liquid potato dextrose (PD) media and cultured using an shaking incubator at 120 rpm and 25°C. The *I. baumii* mycelium were harvested after 8 days of growth in the liquid PD media, and frozen at −80°C until RNA extraction.

Total RNA was extracted using TRIzol reagent according to the manufacturers' directions, then treated with RNase-free DNaseI. The extracted RNA content, integrity and purity were checked by a BioPhotometer D30.

* Corresponding author. Tel./fax: +86 451 82190384.
E-mail address: stt_fighting@126.com (T. Sun).

1389-1723/$ – see front matter
http://dx.doi.org/10.1016/j.jbiosc.2015.09.004

TABLE 1. Overview of the sequencing and assembly.

	Transcripts
Total number of reads	27259264
Total nucleotides (nt)	2453333760
Q20%	92.88%
GC percentage	51.32%
Step-wise assembly	
Contig	
Total number of contigs	280691
Length of all contigs (nt)	29989910
Average sequence size of contigs (nt)	164
Contigs N50 (nt)	187
Scaffold	
Total number of scaffolds	43890
Length of all scaffolds (nt)	18693492
Average sequence size of scaffolds (nt)	416
Scaffolds N50 (nt)	716
Unigene	
Total number of unigenes	30051
Length of all unigenes (nt)	16854970
Average sequence size of unigenes (nt)	561
Unigenes N50 (nt)	831

cDNA library construction and Illumina sequencing The qualified RNA samples were used for the cDNA synthesis. Poly(A) mRNA was isolated using oligo-dT beads. All mRNA was broken into short fragments (200 nt) by adding fragmentation buffer. First-strand cDNA was generated using random primers and reverse transcriptase, then the second-strand cDNA was obtained with RNase H and DNA polymerase I. The cDNA fragments were purified using a QIAquick PCR extraction kit and washed with EB buffer for end reparation poly(A) addition. After that, the cDNA fragments were ligated to sequencing adapters. PCR amplification was then performed by selecting suitable fragments as templates. Finally, the cDNA library of *I. baumii* was constructed and sequenced on the Illumina sequencing platform (Illumina HiSeq 2000).

***De novo* assembly** Low-quality reads, such as adaptor sequences or with unknown nucleotides >10%, were filtered. Basing on the clean reads, transcriptome *de novo* assembly was carried out using short reads assembling program-SOAPdenovo (25). Briefly, reads with the overlap were combined to generate contigs. With paired-end reads, contigs from the same transcript were connected by SOAPdenovo using N to represent unknown sequences, then scaffolds were obtained. Next, paired-end reads were used again for gap filling of scaffolds to get unigenes which have least Ns. Finally, blastx alignment (E-value $<10^{-5}$) is carried out between unigenes and protein databases like Nr, Swiss-Prot, KEGG and COG, then the sequence direction of unigenes are decided by the best aligning results. When a unigene happens to be aligned to none of the above databases, a software named ESTScan (26) would be performed to decide its sequence direction.

Functional annotation and analysis of transcriptome For the functional annotation, the BLASTx algorithm was used to search for homologous sequences against the NCBI Nr database (E-value $<10^{-5}$). With Nr annotation, Blast2GO program (27) was used to obtain the GO annotation of unigenes of *I. baumii*. WEGO software (28) was used to perform GO functional classification for all unigenes. For Pfam domain annotation, the predicted protein sequences were searched against the Pfam-A database (v27.0) using HMMER v3.0 at gathering threshold (29). KEGG database was used to analyze the metabolic pathways of cell, genetic information processing, cellular processes, organismal systems and and so on (30). Unigenes were submitted to the KEGG Automatic Annotation Server (KAAS) (31) to obtain the pathway annotation.

In addition, Yeast is a model fungi with good gene function and genome annotation. Therefore, the BLASTx algorithm was also used to search for homologous sequences against the Yeast genome (E-value $<10^{-1}$).

Annotation of CAZymes and FOLymes BLASTx search ((E-value $<10^{-20}$) was used to detect carbohydrate-active enzymes and ligin oxidative enzymes from *I. baumii* transcriptome using known protein sequences as queries. Moreover, unigenes which have a sequence identity more than 60% with these biochemically characterized enzymes were considered as a CAZyme or FOLyme.

RESULTS

Illumina sequencing and *de novo* assembly After filtering for the low-quality sequences, a total of 27,259,264 clean reads, comprising 2,453,333,760 nucleotides were obtained. The Q20 and GC percentages were 92.88% and 51.32%, respectively, and an overview of the sequencing and assembly is given in Table 1. Using the clean reads, Trinity produced 280,691 contigs with an average length of 164 bp and N50 length of 187 bp. The contigs were then joined into 43,890 scaffolds (average length of 416 bp) using paired-end information as well as gap-filling process. Finally, *de novo* assembly yielded 30,051 unigenes with an average length of 561 bp and N50 length of 831 bp. Of these unigenes, 9661 unigenes (32.14%) were >500 bp and 4087 unigenes (13.6%) were >1000 bp.

Functional annotation of the RNA-seq data All the 30,051 assembled unigenes were searched against the Nr, Swiss-Prot, KEGG and COG databases. A total of 20,634 (68.66%) unigenes were annotated. Among them 3244 unigenes (19.79%) were annotated in all four databases. The number of unigenes with significant similarity to sequences in Nr, Swiss-Prot, KEGG and COG databases were 20,518 (68.28%), 9505 (31.63%), 10,352 (34.45%), and 6332 (21.07%), respectively.

In GO classification, the 21,560 matched unigenes were classified into 3 functional categories: biological process, cellular component and molecular function. In biological process, the largest subcategory of the molecular function was metabolic process (12.23%) and the second was cellular process (10.38%). In terms of cellular component, the most represented were cell (11.27%) and cell part (9.72%). According to molecular function, the largest number were catalytic activity (17.51%) and binding (13.61%)

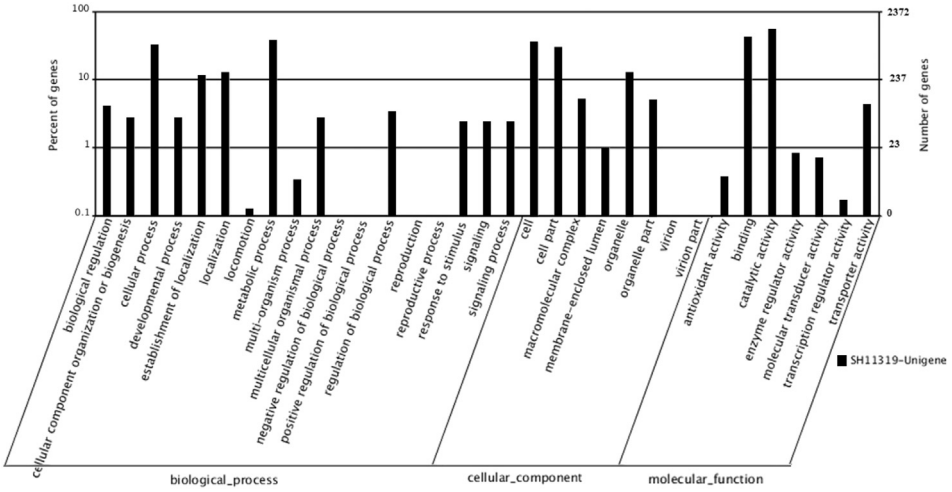

FIG. 1. Histogram presentation of Gene Ontology classification. The results are summarized in three main categories: biological process, cellular component and molecular function.