Biochemical and Biophysical Research Communications 431 (2013) 111–115

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

De novo characterization of *Lentinula edodes* C_{91-3} transcriptome by deep Solexa sequencing

Mintao Zhong, Ben Liu, Xiaoli Wang, Lei Liu, Yongzhi Lun, Xingyun Li, Anhong Ning, Jing Cao, Min Huang *

Department of Medical Microbiology, Dalian Medical University, 9 Western Section, Lvshun South Road, Lvshunkou District, Dalian 116044, China

ARTICLE INFO

Article history:
Received 22 November 2012
Available online 22 December 2012

Keywords:
Solexa
Next-generation sequencing
Lentinula edodes
Antitumor
Transcriptome
Functional genomics

ABSTRACT

Lentinula edodes, has been utilized as food, as well as, in popular medicine, moreover, its extract isolated from its mycelium and fruiting body have shown several therapeutic properties. Yet little is understood about its genes involved in these properties, and the absence of *L. edodes* genomes has been a barrier to the development of functional genomics research. However, high throughput sequencing technologies are now being widely applied to non-model species. To facilitate research on *L. edodes*, we leveraged Solexa sequencing technology in de novo assembly of *L. edodes* C_{91-3} transcriptome. In a single run, we produced more than 57 million sequencing reads. These reads were assembled into 28,923 unigene sequences (mean size = 689 bp) including 18,120 unigenes with coding sequence (CDS). Based on similarity search with known proteins, assembled unigene sequences were annotated with gene descriptions, gene ontology (GO) and clusters of orthologous group (COG) terms. Our data provides the first comprehensive sequence resource available for functional genomics studies in *L. edodes*, and demonstrates the utility of Illumina/Solexa sequencing for de novo transcriptome characterization and gene discovery in a non-model mushroom.

1. Introduction

Lentinula edodes, commonly known as the Shiitake mushroom, is the second most popular and widely cultivated edible mushroom in the world. As the main medicinal fungi, its extract has demonstrated high immunopotentiating and antimetastatic activities, antibacterial, antitumor activity, antifungal, and antidiabetic activities [1–4]. Numerous studies showed Lentinan could inhibit tumor cells growth, improve patient's symptoms, and reduce adverse reactions. It is now officially used for clinical medicine as adjuvant chemotherapy drugs [5–9].

However, antitumor activity of *L. edodes* was rarely reported at protein level. In our previous studies, the protein components of *L. edodes* C_{91-3} mycelia had significant antitumor effects on inducing apoptosis in vivo and in vitro, especially a direct anti-tumor effect in vitro [10,11]. Because the strain was studied beginning in March 1991, it was named as "C_{91-3}" "C" represents "China".

Despite its global importance, genomic sequence resources available for *L. edodes* are scarce, especially for *L. edodes* C_{91-3}. Currently (April 23th, 2012), there are about 26,541 EST and 1021 nucleotide sequences available on NCBI for *L. edodes*, but most of them have no functional annotation. Hence these existing data on NCBI, obtained by traditional Sanger sequencing method, are insufficient to discover the functional genes for *L. edodes*.

Over the past several years, the next generation sequencing (NGS) technology has emerged as a cutting edge approach for high-throughput sequence determination and this has dramatically improved the efficiency and speed of gene discovery [12,13].

With the development of cost efficient and massively parallel high-throughput sequencing technologies, genome-scale studies in non-model organisms are being actively pursued for gene discovery, expression profiling, and studies in functional, comparative, and evolutionary genomics in taxa where little or no previous genomic information exists [14–16]. Despite its obvious potential, next generation sequencing methods have not yet been applied to *L. edodes* research.

This study was conceived to develop an extensive expressed gene sequence resource in *L. edodes* C_{91-3} by deep Illumina/Solexa sequencing for evolutionary and functional genomics. The first comprehensive transcriptome characterization for *L. edodes* C_{91-3} was presented, including an assessment of transcriptome coverage, gene sequences and functional annotation by bioinformatics analysis. In this study, over four billion bases of high-quality DNA sequence were generated with Illumina/Solexa technology, which demonstrated the suitability of short-read sequencing for de novo assembly and annotation of genes expressed in a eukaryote without the prior genome information. In a single run, we got 28,923 unigene sequences (mean size = 689 bp) including 18,120 unigenes

* Corresponding author. Fax: +86 411 86110007.
E-mail address: huangminchao@163.com (M. Huang).

0006-291X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2012.12.065

with coding sequence. Furthermore, based on similarity search with known proteins, the unigene sequences were annotated with gene function. Therefore, the assembled, annotated transcriptome sequences and gene functional annotation provide an invaluable resource for functional genomics studies in *L. edodes*.

2. Materials and methods

2.1. L. edodes $C_{91\text{-}3}$ strain and culture conditions

L. edodes $C_{91\text{-}3}$ strain were obtained from the Department of Microbiology in Dalian Medical University. It was cultured in the potato culture medium containing 1% vitamin B1, 2.0% agar, 0.15% $MgSO_4$, 0.3% K_2HPO_4, and 2.0% glucose having pH 6.0.

2.2. cDNA library construction and sequencing

For Illumina/Solexa sequencing, the total RNA of every sample was extracted using TaKaRa RNAiso™ Plus from *L. edodes* $C_{91\text{-}3}$ mycelia which cultured for 10 days, and then treated with TaKaRa RNase-free DNase I for 45 min according to the manufacturer's protocols. Beads with oligo(dT) were used to isolate poly(A) mRNA after total RNA was collected from eukaryote. Fragmentation buffer was added for interrupting mRNA to short fragments. The mRNA was fragmented into small pieces using divalent cations at elevated temperature. The cleaved mRNA fragments were converted to double-stranded cDNA using SuperScript II, RNaseH, and DNA Pol I, primed by random primers. Short cDNA fragments were purified with QiaQuick PCR extraction kit and resolved with EB buffer for end reparation and adding poly(A). After that, the cDNA short fragments were connected with sequencing adapters. For PCR amplification, suitable fragments were selected as templates based on the result of agarose gel electrophoresis. Finally, the cDNA library products were sequenced using the 1G Illumina Genome Analyzer.

2.3. Sequence assembly

All the Solexa reads, stored in fastq format, were filtered to remove poly(A/T), low quality sequences and empty reads using the SeqClean program. Resulting sequences and quality files were assembled using SOAPdenovo with default parameters [17]. SOAPdenovo firstly combined reads with certain length of overlap to form longer fragments without N, which were called contigs. Then the reads are mapped back to contigs; with paired-end reads, it is able to detect contigs from the same transcript as well as the distances between these contigs. The contigs were connected using N to represent unknown sequences between each two contigs, and then scaffolds were made. Paired-end reads were used again for gap filling of scaffolds to get sequences with least Ns and cannot be extended on either end. Such sequences were defined as unigenes. When multiple samples were sequenced from the same species, unigenes from each sample's assembly should be taken into further process of sequence splicing and redundancy removing with TGICL software to acquire non-redundant unigenes as long as possible [18]. Finally, blastx alignment (e-value < 0.00001) was performed between unigenes and protein databases like nr, Swiss-Prot, KEGG, and COG. The best aligning results were used to decide sequence direction of unigenes. When some unigene happened to be unaligned to none of the above databases, ESTScan software would be introduced to predict its coding regions as well as to decide its sequence direction [19].

2.4. Functional annotation and KEGG pathway analysis

The annotation of unigenes was based on sequence homology using BLASTX software. The unigene sequences were searched against the Swiss-Prot database, the Nr database, the KEGG database, the COG database and the Nt database (e-value < 1×10^{-5}). The unique sequences were assigned to special biochemical pathways according to the KEGG standards using BLASTX [20]. The terms of GO classification were assigned to all well-annotated sequences by performing Blast2GO program [21,22]. To reduce the redundancy, each sequence that had BLAST hit in the Nr database was given a unigene ID according to the best homologue they were aligned to.

2.5. CDS prediction

Unigenes were firstly aligned by blastx (e-value < 0.00001) to protein databases in the priority order of nr, Swiss-Prot, KEGG and COG. Unigenes aligned to databases with higher priority would not enter the next circle. The alignments ended when all circles were finished. Proteins with highest ranks in blast results were taken to decide the coding region sequences of unigenes. Unigenes that cannot be aligned to any database were scanned by ESTScan for getting nucleotide sequence (5′–3′) and the coding regions [19].

3. Results

3.1. Output statistics of sequencing, assembly quality and CDS prediction

The output of sequenced data is an important indicator of the contract. According to the contract, clean reads in each sample must contain a total base number of no less 1G. After cleaning and quality checks, 43 million of 75 bp reads were obtained from sequencing one plate, G + C content was 48.31%, and Q20 value was 95.39% more than 80%. For sequence assembly using SOAPdenovo software 106,495 contigs were obtained and the mean contig size was 236 bp. Using paired-end joining and gap-filling, the contigs were further assembled into 55,322 scaffolds with a mean size of 424 bp including 4906 scaffolds larger than 1000 bp. The scaffolds were further assembled into 28,923 unigenes with a mean size of 689 bp, including 25,344 unigenes without gap (Table 1). All the sequences had randomly covered the full-length of unigene from 5′ to 3′ end (Fig. 1). To demonstrate the accuracy of sequencing data, 17 unigenes were randomly selected for RT-PCR amplification. The identity of all 17 PCR products were confirmed by Sanger sequencing (Table 2). Comparing to original sequence of 17 unigenes, sequencing results of 9 PCR products showed 100% similarity, 5 PCR products showed 99% similarity, 2 PCR products showed 98% similarity, and 1 PCR

Table 1
Output statistics of sequencing.

Total number of reads	57,998,508
Total base pairs (bp)	4,349,888,100
Average read length (bp)	75
Total number of contigs	106,495
Mean length of contigs (bp)	236
Total number of scaffolds	55,322
Mean length of scaffolds (bp)	424
Total number of Unigenes	28,923
Mean length of Unigenes (bp)	689
Q20 percentage	95.39%
G + C percentage	48.31%

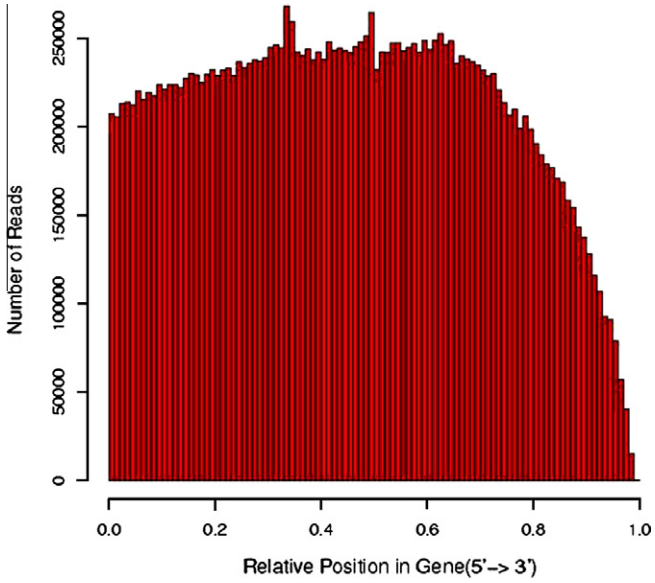

Fig. 1. Randomness on reads from *Lentinula edodes* $C_{91\text{-}3}$ sample (all the reads sequencing from *Lentinula edodes* $C_{91\text{-}3}$ had randomly covered the full-length of unigene from 5′ to 3′)

Table 2
The identity between the original sequence of Unigenes and PCR products.

Gene ID	The identity (%)
Unigene1245	94
Unigene24718	100
Unigene28447	100
Unigene10627	99
Unigene20035	100
Unigene21859	98
Unigene23050	100
Unigene4347	100
Unigene14872	99
Unigene24886	99
Unigene2562	99
Unigene3375	99
Unigene8290	100
Unigene19804	98
Unigene24277	100
Unigene6339	100
Unigene13853	100

products showed 94% similarity. Hence, average accuracy rate of Solexa sequencing was 99.18%.

Through the blast on the protein databases in the priority order of nr, Swiss-Prot, KEGG and COG, 14,075 unigenes with CDS were observed. 4045 unigenes with CDS were gained using the ESTscan prediction. Fig. 2 indicates the proportion distribution of Unigene with CDS matches in nr, Swiss-Prot, KEGG and COG databases and through ESTScan prediction. Obviously, a 51% of match efficiency was observed for sequences in nr databases, 24% in Swiss-Prot databases, 21% in ESTScan prediction, 3% in COG databases and 1% in KEGG databases.

In conclusion, all the statistics illustrated a better quality and depth of sequencing produced by Solexa sequencing on *L. edodes* $C_{91\text{-}3}$.

3.2. Functional annotation of predicted proteins

For annotation, distinct unigene sequences were firstly aligned by blastx to protein databases like nr, Swiss-Prot, KEGG and COG (*e*-value < 0.00001), retrieving proteins with the highest sequence similarity with the given unigenes along with their protein

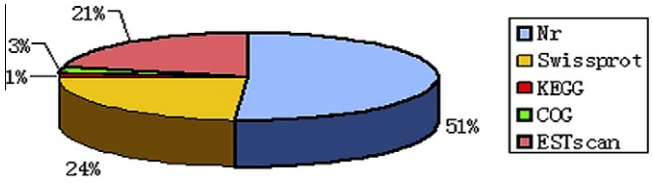

Fig. 2. The distribution of unigene sequences with CDS based on the proteins databases and ESTscan in *Lentinula edodes* $C_{91\text{-}3}$ (a 51% of match efficiency was observed for sequences in nr databases, 24% in Swiss-Prot databases, 21% in ESTScan prediction, 3% in COG databases and 1% in KEGG databases)

functional annotations. Using this approach, 14,075 unigenes (48.7% of all distinct sequences) returned an above cut-off BLAST result. Because of the relatively short length of distinct gene sequences (mean size of 689 bp) and lack of genome information in *L. edodes*, most of the 14,848 assembled sequences could not be matched to known genes (51.3%).

3.3. GO function classification

GO assignments were used to classify the functions of the predicted *L. edodes* $C_{91\text{-}3}$ genes. Based on sequence homology, 2853 Unigenes can be categorized into 38 functional groups (Fig. 3). In each of the three main categories (biological process, cellular component and molecular function) of the GO classification, "metabolic process", "cell part", and "catalytic" terms were dominant respectively; however, we did not find any genes in the clusters of "cell killing", "death", "immune system process", "rhythmic process", "viral reproduction", "extracellular region part", "symplast", "synapse", "synapse part", "virion", "virion part", "auxiliary transport protein", "chemoattractant", "chemorepellent", "metallochaperone" and "proteasome regulator". We also noticed a high-percentage of genes from categories of "cellular process", "organelle" and "binding" and only a few genes from terms of "biological adhesion", "locomotion", "protein tag", "growth" and "nutrient reservoir".

3.4. COG function classification

To further evaluate the completeness of our transcriptome library and the effectiveness of our annotation process, we searched the annotated sequences for the genes involved in COG classifications. In total, 6203 unigenes have a COG classification (Fig. 4), and especially some unigene with many functions can be classified into different COG category. Among the 25 COG categories, the cluster for general function prediction represents the largest group (1809 members) followed by "Carbohydrate transport and metabolism" (1053 members). The following categories: "extracellular structures" (8 members) and "Defense mechanisms" (8 members), represent the smallest groups (Fig. 4).

3.5. KEGG pathway annotation

To identify the biological pathways that are active in *L. edodes* $C_{91\text{-}3}$, we mapped the 14,075 annotated sequences to the reference canonical pathways in Kyoto Encyclopedia of Genes and Genomes. In total, 2264 unigenes were assigned to 214 KEGG pathways, and some unigene which distributed in the different pathway, can also joined into the different links in the same pathway. The pathways with most representation by the unique sequences were purine metabolism (966 members) and pyrimidine metabolism (878 members).

In brief, all these annotations provide a valuable resource for investigating specific processes, functions and pathways associated with the Unigene of *L. edodes* $C_{91\text{-}3}$.

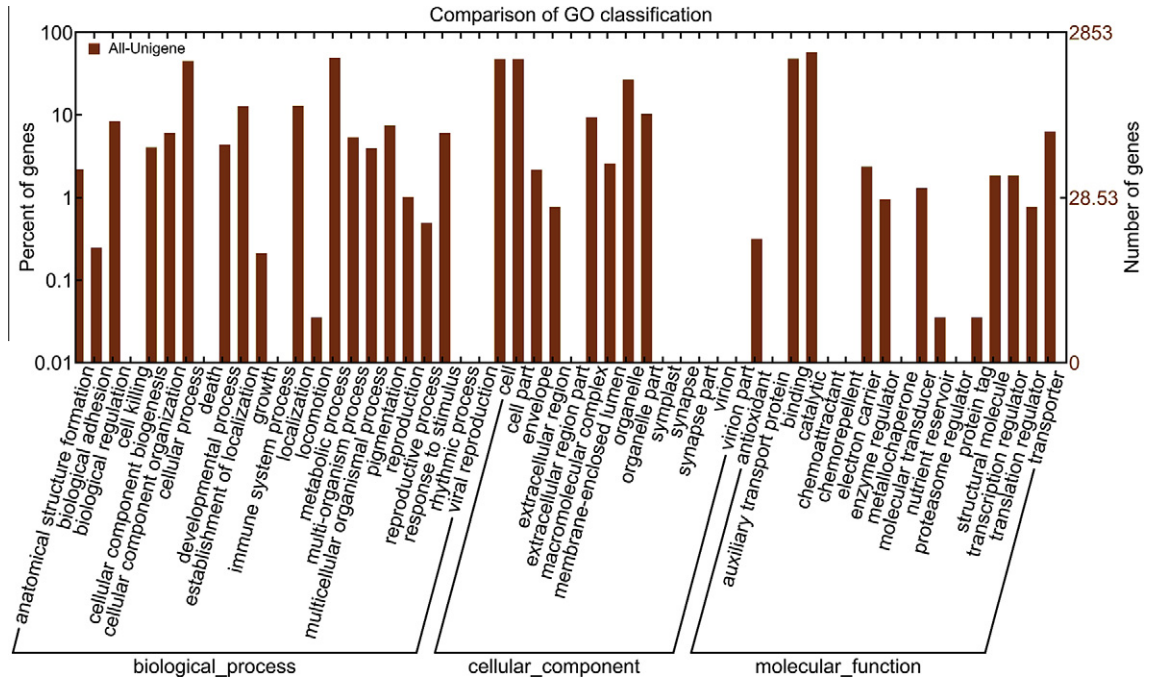

Fig. 3. Histogram presentation of Gene Ontology classification (The results are summarized in three main categories: biological process, cellular component and molecular function. The right *y*-axis indicates the number of genes in a category. The left *y*-axis indicates the percentage of a specific category of genes in that main category)

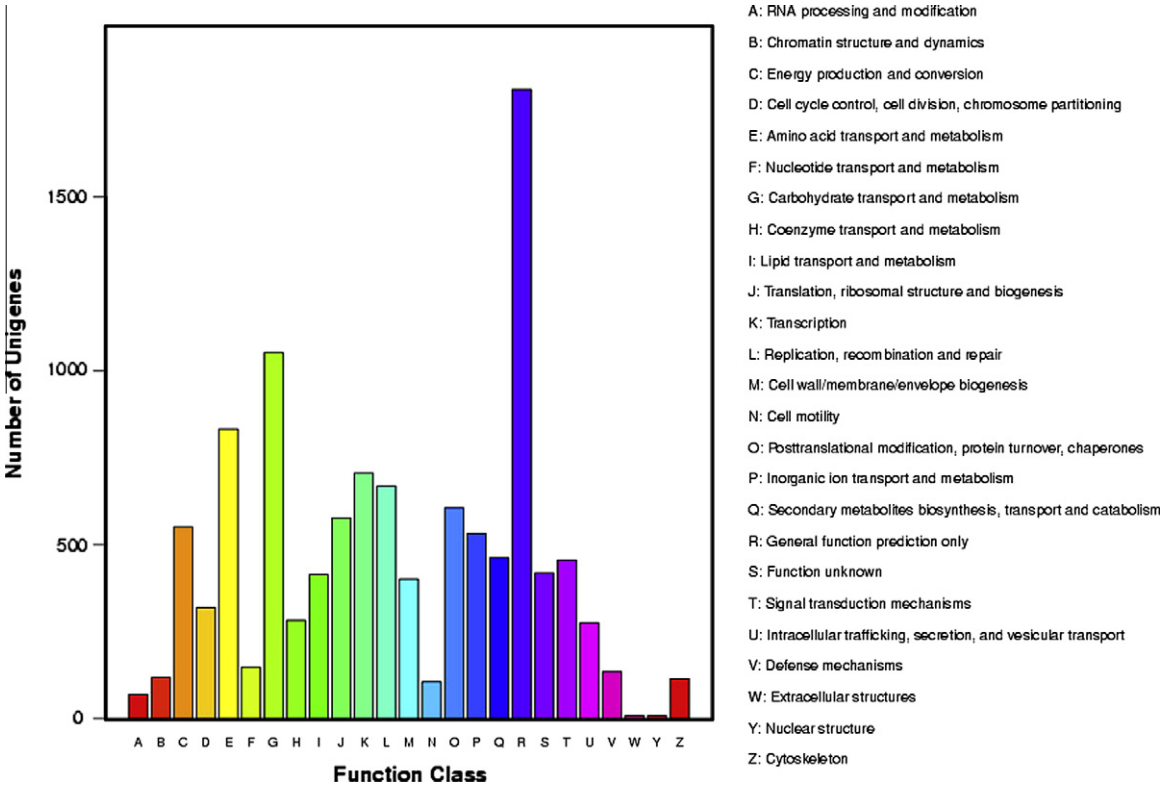

Fig. 4. Histogram presentation of clusters of orthologous groups (COG) classification (6,203 unigene sequences have a COG classification among the 25 categories)

4. Discussion

The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition. Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease. With the development of large-scale genomics, traditional Sanger sequencing technology could not meet the needs of the development [23]. But next-generation sequencing (NGS) overcame current limitations of Sanger sequencing with respect to throughput and costs. Among the

NGS technology, the Roche 454 Genome Sequencer, the Illumina/Solexa Genome Analyzer, and the ABI SOLiD System sequencing platforms have become widely available over the past several years [24,25].

Because of its low cost and vast amounts of data, Solexa sequencing has become the most widely useful technology among the next generation sequencing technology platform. Using the technology, more than 400 research articles have been published in the world since 2008 [26].

For understanding the medical function at protein level in *L. edodes*, insight into transcriptome is required. In this study, a rapid and cost-effective method were presented for transcriptome using Solexa sequencing technology on *L. edodes* $C_{91\text{-}3}$. 28,923 unigene sequences were got including 18,120 unigenes with coding sequence. These findings provide a substantial contribution to gene sequence resources for *L. edodes* $C_{91\text{-}3}$, and this is the first publication using Illumina/Solexa sequencing technology for *L. edodes* without prior genome annotation. Additionally, we have demonstrated the feasibility of using Illumina sequencing technology for gene expression and have provided new leads for functional studies of genes involved in *L. edodes* $C_{91\text{-}3}$.

In short, *De novo* characterization of *L. edodes* $C_{91\text{-}3}$ transcriptome, including lots of assembled, annotated transcriptome sequences and gene functional annotation, firstly provide an invaluable resource for functional genomics studies in *L. edodes*, especially for protein expression and function research on *L. edodes*.

Acknowledgments

This project is supported by, the National Natural Science Foundation of China (30770018). We gratefully acknowledge the technical support of transcriptome sequencing from BGI (Beijing Genomics Institute, Shenzhen, China).

References

[1] R. Hearst, D. Nelson, G. McCollum, B.C. Millar, Y. Maeda, C.E. Goldsmith, P.J. Rooney, A. Loughrey, J.R. Rao, J.E. Moore, An examination of antibacterial and antifungal properties of constituents of Shiitake (*Lentinula edodes*) and Oyster (*Pleurotus ostreatus*) mushrooms, Complementary Therapies in Clinical Practice 15 (2009) 5–7.

[2] U.R. Kuppusamy, Y.L. Chong, A.A. Mahmood, M. Indran, N. Abdullah, S. Vikineswary, *Lentinula edodes* (Shiitake) mushroom extract protects against hydrogen peroxide induced cytotoxicity in peripheral blood mononuclear cells, Indian Journal of Biochemistry and Biophysics 46 (2009) 161–165.

[3] I. Sarangi, D. Ghosh, S.K. Bhutia, S.K. Mallick, T.K. Maiti, Anti-tumor and immunomodulating effects of *Pleurotus ostreatus* mycelia-derived proteoglycans, International Immunopharmacology 6 (2006) 1287–1297.

[4] S. Unursaikhan, X. Xu, F. Zeng, L. Zhang, Antitumor activities of *O*-sulfonated derivatives of (1→3)-alpha-D-glucan from different *Lentinus edodes*, Bioscience, Biotechnology, and Biochemistry 70 (2006) 38–46.

[5] V.E. Ooi, F. Liu, Immunomodulation and anti-cancer activity of polysaccharide-protein complexes, Current Medicinal Chemistry 7 (2000) 715–729.

[6] P.M. Kidd, The use of mushroom glucans and proteoglycans in cancer treatment, Alternative Medicine Review 5 (2005) 4–27.

[7] R. Zheng, S. Jie, D. Hanchuan, W. Moucheng, Characterization and immunomodulating activities of polysaccharide from *Lentinus edodes*, International Immunopharmacology 5 (2005) 811–820.

[8] N. Fang, Q. Li, S. Yu, J. Zhang, L. He, M.J. Ronis, T.M. Badger, Inhibition of growth and induction of apoptosis in human cancer cell lines by an ethyl acetate fraction from Shiitake mushrooms, Journal of Alternative and Complementary Medicine 12 (2006) 125–132.

[9] C.K. Miyaji, A. Poersch, L.R. Ribeiro, A.F. Eira, I.M. Cólus, Shiitake (*Lentinula edodes* (Berkeley) Pegler) extracts as a modulator of micronuclei induced in HEp-2 cells, Toxicology In Vitro 20 (2006) 1555–1559.

[10] M. Zhong, B. Liu, Y. Liu, X. Wang, X. Li, L. Liu, A. Ning, J. Cao, M. Huang, The antitumor activities of *Lentinus edodes* C91–3 mycelia fermentation protein on S180 (mouse sarcoma cell) in vitro and in vivo, Journal of Medicinal Plant Research 6 (2012) 2488–2492.

[11] B. Liu, M. Zhong, Y. Lun, X. Wang, W. Sun, X. Li, A. Ning, J. Cao, W. Zhang, L. Liu, M. Huang, A novel apoptosis correlated molecule: expression and characterization of protein latcripin-1 from *Lentinula edodes* C91–3, International Journal of Molecular Science 13 (2012) 6246–6265.

[12] S.C. Schuster, Next-generation sequencing transforms today's biology, Nature Methods 5 (2008) 16–18.

[13] W.J. Ansorge, Next-generation DNA sequencing techniques, New Biotechnology 25 (2009) 195–203.

[14] L. Collins, P. Biggs, C. Voelckel, S. Joly, An approach to transcriptome analysis of non-model organisms using short-read sequences, Genome Informatics 21 (2008) 3–14.

[15] T.L. Parchman, K.S. Geist, J.A. Grahnen, C.W. Benkman, C.A. Buerkle, Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery, BMC Genomics 11 (2010) 180.

[16] R. Bettencourt, M. Pinheiro, C. Egas, P. Gomes, M. Afonso, T. Shank, R.S. Santos, High-throughput sequencing and analysis of the gill tissue transcriptome from the deep-sea hydrothermal vent mussel *Bathymodiolus azoricus*, BMC Genomics 11 (2010) 559.

[17] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, J. Wang, De novo assembly of human genomes with massively parallel short read sequencing, Genome Research 20 (2010) 265–272.

[18] G. Pertea, X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, J. Quackenbush, TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets, Bioinformatics 19 (2003) 651–652.

[19] C. Iseli, C.V. Jongeneel, P. Bucher, ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences, in: Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (1999) pp. 138–148.

[20] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, Y. Yamanishi, KEGG for linking genomes to life and the environment, Nucleic Acids Research 36 (2008) D480–484.

[21] A. Conesa, S. Götz, J.M. García-Gómez, J. Terol, M. Talón, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (2005) 3674–3676.

[22] J. Ye, L. Fang, H. Zheng, Y. Zhang, J. Chen, Z. Zhang, J. Wang, S. Li, R. Li, L. Bolund, J. Wang, WEGO: a web tool for plotting GO annotations, Nucleic Acids Research 34 (2006) W293–W297.

[23] Z. Wang, M. Gerstein, M. Snyder, RNA-seq: a revolutionary tool for transcriptomics, Nature Reviews Genetics 10 (2009) 57–63.

[24] S. Audic, J.M. Claverie, The significance of digital gene expression profiles, Genome Research 7 (1997) 986–995.

[25] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nature Methods 5 (2008) 621–628.

[26] B.J. Haas, M.C. Zody, Advancing RNA-Seq analysis, Nature Biotechnology 28 (2010) 421–423.