Gene xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Gene

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

Illumina-based *de novo* transcriptome sequencing and analysis of *Amanita exitialis* basidiocarps

Peng Li [a,b,1], Wang-qiu Deng [b,1], Tai-hui Li [a,b,*], Bin Song [b], Ya-heng Shen [b]

[a] School of Bioscience & Bioengineering, South China University of Technology, Guangzhou 510006, China
[b] State Key Laboratory of Applied Microbiology, South China (The Ministry–Province Joint Development), Guangdong Institute of Microbiology, Guangzhou 510070, China

ARTICLE INFO

Article history:
Accepted 4 September 2013
Available online xxxx

Keywords:
Toxin
Gene family
Next-generation sequencing
Gene polymorphism

ABSTRACT

Amanita exitialis is a lethal mushroom that was first discovered in Guangdong Province, China. The high content of amanitin in its basidiocarps makes it lethal to humans. To comprehensively characterize the *A. exitialis* transcriptome and analyze the *Amanita* toxins as well as their related gene family, transcriptome sequencing of *A. exitialis* was performed using Illumina HiSeq 2000 technology. A total of 25,563,688 clean reads were collected and assembled into 62,137 cDNA contigs with an average length of 481 bp and N50 length of 788 bp. A total of 27,826 proteins and 39,661 unigenes were identified among the assembled contigs. All of the unigenes were classified into 166 functional categories for understanding the gene functions and regulation pathways. The genes contributing to toxic peptide biosynthesis were analyzed. From this set, eleven gene sequences encoding the toxins or related cyclic peptides were discovered in the transcriptome. Three of these sequences matched the peptide toxins α-amanitin, β-amanitin, and phallacidin, while others matched amanexitide and seven matched unknown peptides. All of the genes encoding peptide toxins were confirmed by polymerase chain reaction (PCR) in *A. exitialis*, and the phylogenetic relationships among these proprotein sequences were discussed. The gene polymorphism and degeneracy of the toxin encoding sequences were found and analyzed. This study provides the first primary transcriptome of *A. exitialis*, which provided comprehensive gene expression information on the lethal amanitas at the transcriptional level, and could lay a strong foundation for functional genomics studies in those fungi.

1. Introduction

Amanita mushrooms are responsible for approximately 90% of the mushroom poisoning fatalities (Bresinsky and Besl, 1990). The lethal *Amanita* species contain various peptide toxins including amatoxins, phallotoxins and virotoxins, which are bicyclic octapeptides, bicyclic heptapeptides, and monocyclic heptapeptides respectively (Faulstich et al., 1980; Wieland, 1986). Although the peptide toxins of *Amanita* are highly toxic, they have many applications in molecular biology, medicine, and pharmacy. Amatoxins have been extensively used as agents to inhibit RNA polymerase II and/or protein synthesis in biological research (Bushnell et al., 2002; Kroncke et al., 1986; Letschert et al., 2006), while phallotoxins exert their function by stabilizing F-actin (Bamburg, 1999; Lengsfeld et al., 1974). However, like most other ectomycorrhizal Basidiomycetes, the lethal amanitas grow slowly and cannot form basidiocarps in culture, and only wild basidiocarps produce high concentrations of the toxins (Zhang et al., 2005). Great efforts have been made to synthesize the amanitins, but no satisfactory results have been achieved to date (Wieland and Faulstich, 1991). The production of *Amanita* peptide toxins remains a big challenge to both academia and industry.

Although 6929 expressed sequence tags (ESTs) have been developed on several important *Amanita* species according to the summary in NCBI (http://www.ncbi.nlm.nih.gov/nucest/?term=amanita), gene families encoding the major toxins of lethal amanitas were only reported in three *Amanita* species (*Amanita bisporigera*, *Amanita phalloides*, and *Amanita ocreata*) (Hallen et al., 2007) and one *Galerina* species (*Galerina marginata*) (Luo et al., 2012). Studies have demonstrated that α-amanitin and phallacidin were synthesized on ribosomes; the proproteins of α-amanitin and phallacidin were composed of 35 and 34 amino acids respectively, and the prolyl oligopeptidase (POP) was predicted to specifically cleave these proproteins (Luo et al., 2010). Studies have also shown that different lethal *Amanita* species contain diverse peptide toxins and related peptides (Chen et al., 2003; Hallen et al., 2007). For example, thirteen new, related sequence encoding

Abbreviations: AeBA, *A. exitialis* basidiocarp; AMA, amanitin; bp, base pair; CDS, coding sequences; COG, Clusters of Orthologous Groups; dNTP, deoxyribonucleotide triphosphate; EST, expressed sequence tag; E-value, Expect-value; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; NCBI, National Center for Biotechnology Information; Nr, non-redundant; ORF, open reading frame; PCR, polymerase chain reaction; PHA, phallacidin; POD, phalloidin; POP, prolyl oligopeptidase; RPMK, reads per kb exon model per million uniquely mapping reads; TGICL, TGI Clustering tools; UN-PRO, unknown proproteins; USDA, United States Department of Agriculture.

* Corresponding author at: Room 404, Building 59, NO.100, Xianlie Road, 510070 Guangzhou, Guangdong Province, China. Tel./fax: +86 20 87137619.
E-mail address: mycolab@263.net (T. Li).

[1] The authors contributed equally to this work.

0378-1119/$ – see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.gene.2013.09.014

peptides have been found in *A. bisporigera* (Hallen et al., 2007). Therefore, the peptide toxins and related encoding genes of other lethal *Amanita* species should be studied.

Amanita exitialis Zhu L. Yang & T. H. Li (Fig. 1), one of the most poisonous mushrooms worldwide, was discovered in Guangdong Province, China (Yang and Li, 2001), where it is the most common cause of mushroom poisoning (Deng et al., 2011; Yang, 2005). Recent studies have shown that *A. exitialis* produces numerous cyclic peptide toxins and related peptides including α-amanitin, β-amanitin, amaninamide, phallacidin, phallisin, phallacin, phallisacin, phalloin, desoxoviroidin, and amanexitide (Deng et al., 2011; Xue et al., 2011). High-performance liquid chromatography results indicated that the pileus contains the highest amount and richest toxins at the vigorous stage (Hu et al., 2012). However, gene family encoding the toxins or related peptides of *A. exitialis* has not yet been reported.

Transcriptome sequencing is an efficient approach for obtaining microbial functional genomics information. In recent years, next-generation sequencing techniques such as Solexa/Illumina (Illumina), 454 (Roche), and SOLID (ABI) platforms have emerged as the useful tool for transcriptome analysis. These tools are widely used in the detection of gene expression, discovery of novel transcripts, and identification of differentially expressed genes (Garber et al., 2011; Gibbons et al., 2009; MacLean et al., 2009; Pop and Salzberg, 2008; Trombetti et al., 2007). The Illumina produces orders of magnitude more sequences with higher coverage and lower costs than other sequencing technologies. Compared with other *de novo* transcriptome assemblers, the Trinity recovers more full-length transcripts across a broad range of expression levels with sensitivity similar to those of methods that rely on genome alignments (Grabherr et al., 2011).

In this study, the transcriptome of *A. exitialis* was sequenced using Illumina HiSeq 2000 technology and the *de novo* assembly of the full-length transcripts was performed using the Trinity method. The transcriptome of *A. exitialis* will be characterized and the toxin gene family and related new genes will be systematically explored.

2. Materials and methods

2.1. Sample preparation and RNA extraction

The fresh basidiocarps of *A. exitialis* (AeBA) were collected at the Baiyun Mountain in Guangzhou City, Guangdong Province, China, in March 2012. One clean pileus was selected and frozen at −80 °C until RNA extraction. No specific permits were required for the described field studies, and the localities where the samples came from are not protected in any way.

Total RNA was extracted using TRIzol reagent (Invitrogen, USA) following the manufacturer's protocol, and treated with RNase-free DNase I. Integrity and size distributions were checked using Agilent 2100 with an RNA integrity number (RIN: 8.0) and GE ImageQuant 350.

Fig. 1. The basidiocarps of *A. exitialis*.

2.2. cDNA library construction and Illumina sequencing

The extracted RNA samples were used for the cDNA synthesis. Poly (A) mRNA was isolated using oligo-dT beads (Qiagen). All mRNA was broken into short fragments (200 nt) by adding fragmentation buffer. First-strand cDNA was generated using random hexamer-primed reverse transcription, followed by the synthesis of the second-strand cDNA using RNase H and DNA polymerase I. The cDNA fragments were purified using a QIAquick PCR extraction kit. These purified fragments were then washed with EB buffer for end reparation poly (A) addition and ligated to sequencing adapters. Following agarose gel electrophoresis and extraction of cDNA from gels, the cDNA fragments (200 bp ± 25 bp) were purified and enriched by PCR to construct the final cDNA library. The cDNA library was sequenced on the Illumina sequencing platform (Illumina HiSeq™ 2000) using the single-end paired-end technology in a single run, by Beijing Genomics Institute (BGI)-Shenzhen, Shenzhen, China. The original images process to sequences, base-calling and quality value calculation were performed by the Illumina GA Pipeline (version 1.6), in which 90 bp paired-end reads were obtained.

2.3. De novo transcriptome assembly and analysis

Prior to assembly and mapping, reads with adapters, ambiguous bases > 10%, and low quality in which the percentage of low quality bases (base quality ≤ 20) is >40% were removed. Transcriptome data was *de novo* assembled using the Trinity assembly program (Grabherr et al., 2011), at the parameters of "–kmer method jellyfish – min contig length 100 – jaccard clip". Then unigenes from four libraries were further spliced and assembled to obtain non-redundant unigenes by TGICL with the minimum overlap length of 100 bp (Pertea et al., 2003), and this was used for further analysis in this study. Functional annotations of unigenes include protein sequence similarity, KEGG pathway analysis, and Clusters of Orthologous Groups (COG) and Gene Ontology (GO) database analysis. AeBA-Unigene sequences against protein databases (NR, SwissProt, KEGG, and COG) using BLASTX (E-value < 10^{-5}) were searched. Protein function information can be predicted from annotation of the most similar protein in those databases. Unigene sequences that have hits in a former database will not go to the next round for searching against a later database. The BLAST results information was used to extract coding sequences (CDS) from unigene sequences and translate them into peptide sequences. The BLAST results information is also used to train ESTScan (Iseli et al., 1999). CDS of unigenes that have no hit in BLAST were predicted using ESTScan and then translated into peptide sequences.

With NR annotation, the Blast2GO program (Conesa et al., 2005) is used to obtain the GO annotation of AeBA-Unigene. After GO annotation was obtained for every AeBA-Unigene, WEGO software (Ye et al., 2006) was used to perform GO functional classification for all unigenes and to understand the species' functional gene distribution from the macro level. The COG annotation was performed using the BLASTX algorithm (E-value < 10^{-5}) against the COG database to predict and classify possible functions. To reconstruct the metabolic pathways involved in *A. exitialis*, annotated sequences were mapped to the KEGG database (Ogata et al., 1999) using the Blast2GO platform.

2.4. Searching the Amanita toxin related genes in the transcriptome

The characterization of the "MSDIN" family of *Amanita* toxins reported from *A. bisporigera* was consulted (Hallen et al., 2007). After a search of the CDS database of the *A. exitialis* transcriptome is performed, the two queries must be satisfied. First, the upstream conserved consensus sequence MSDINATRLP (MSDIN, R, and P are invariant) and the

downstream conserved consensus sequence CVGDDV (the first D is invariant) can be used as queries; second, all of the putative toxin regions start immediately downstream of the invariant Pro residue and end after an invariant Pro residue.

2.5. Validation of gene family encoding the major toxins

PCR was performed to validate the *Amanita* toxins and toxin related genes searched among the transcriptome data. After extraction of the total RNA and genome DNA, the total RNA was reversed using QuantScript RT Kit (Tiangen Biotech Co., Ltd., Beijing, China) and the genome DNA was amplified using the following degenerate primers: forward (5′ATGTCNGAYATYAAYGCNACNCG3′) (Hallen et al., 2007) and reverse (5′CCAAGCCTRAYAWRGTCMACAAC3′). The cycling conditions were set as follows: initial denaturation at 94 °C for 4 min, followed by 33 cycles of denaturation at 94 °C for 30 s, annealing at 51 °C for 30 s, extension at 72 °C for 30 s, and a final extension at 72 °C for 7 min. The PCR products were quantified by gel electrophoresis on a 1% agarose gel and then purified using Sangong's purification kit (Sangong, China). All of the purified PCR products were recovered, ligated to the pMD18-T vector (Takara), and then transformed by the DH5α competent cells, and then ten clones were sequenced by Invitrogen Biotechnology Co., Ltd. The carrier sequences were removed by the online software VecScreen (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html) using DNAMAN6.0 to predict the amino acid sequences.

2.6. Phylogenetic analyses of proprotein sequences

The dataset was analyzed. Six proprotein sequences from *A. bisporigera* (EU196140; EU196143), *A. ocreata* (EU196158), *A. phalloides* (FN555142), *Amanita verna* (FN555143), and *Amanita virosa* (FN555144) were download from the GenBank, and twenty proprotein sequences of *A. exitialis* are shown in Table 2 (accession nos.: KF387476–KF387495). All of these sequences were aligned using MACSE (Ranwez et al., 2011). Phylogenetic analysis was performed using MEGA version 5.0 Beta (Tamura et al., 2011). The phylogenetic tree was constructed using the Maximum Likelihood method and bootstrap values were calculated from 10,000 replicates. The Jones–Taylor–Thornton model was selected, gamma distributed among sites was selected, and all positions with gaps/missing data were treated as partial deletion (site coverage cutoff 90%). Branches corresponding to partitions reproduced in less than 50% of the bootstrap replicates were collapsed.

3. Results

3.1. Assembly of A. exitialis transcriptome

The pileus of *A. exitialis* at the vigorous stage was prepared and sequenced. After the removal of the ambiguous nucleotides, low-quality sequences (quality scores < 20), and contaminated microbial sequences, a total of 25,563,688 clean reads with an average length of 90 bp each, comprising 2,300,731,920 nucleotides were obtained. The Q20 and GC percentages were 93.89% and 51.58% respectively. All high-quality reads were assembled *de novo* using the Trinity program (Grabherr et al., 2011), and it produced 62,137 cDNA contigs with an average length of 481 bp and N50 length of 788 bp. A total of 39,661 unigenes with an average length of 662 bp and N50 length of 862 bp were obtained (Table 1). Among them, 27,848 unigenes were annotated. The number and length distribution of the assembled contigs and unigenes were listed (Fig. 2). Above the 90% coverage rate cutoff, there were 24,186 unigenes. These unigenes had an average depth of more than 91 and a size of 200–4739 bp with no gaps. The longest 10% of the unigenes were 1325–4819 bp long.

A total of 27,826 CDS were identified from the dataset with a sequence length of 113–4521 bp. Most of those identified CDS by BLAST and EST-scan were <2.0 kb and <1.0 kb, respectively, and no gap was observed. Most of the identified protein sequences contained <500 amino acids and had no gaps.

Table 1
Output statistics of *Amanita exitialis* transcriptome sequencing and assembly.

	Sequences (nt)	All numbers	Mean length (bp)	N50 (bp)
Total clean reads	2,300,731,920	25,563,688	90	
Total contigs	29,887,897	62,137	481	788
Total unigenes	26,255,582	39,661	662	862
GC percentage				51.58%
N percentage				0.00%
Q20 percentage				93.89%

3.2. Functional annotation

Functional information, protein sequence similarity, KEGG pathway, COG, and GO information were provided from the unigene annotations. With an E-value cutoff of 1e-10, a total of 21,466 unigenes had significant hits, corresponding to 20,682 unique protein accessions in the NR protein database. The GO analyses were conducted and plotted on those proteins (Fig. 3). Briefly, the genes involved in the cellular and

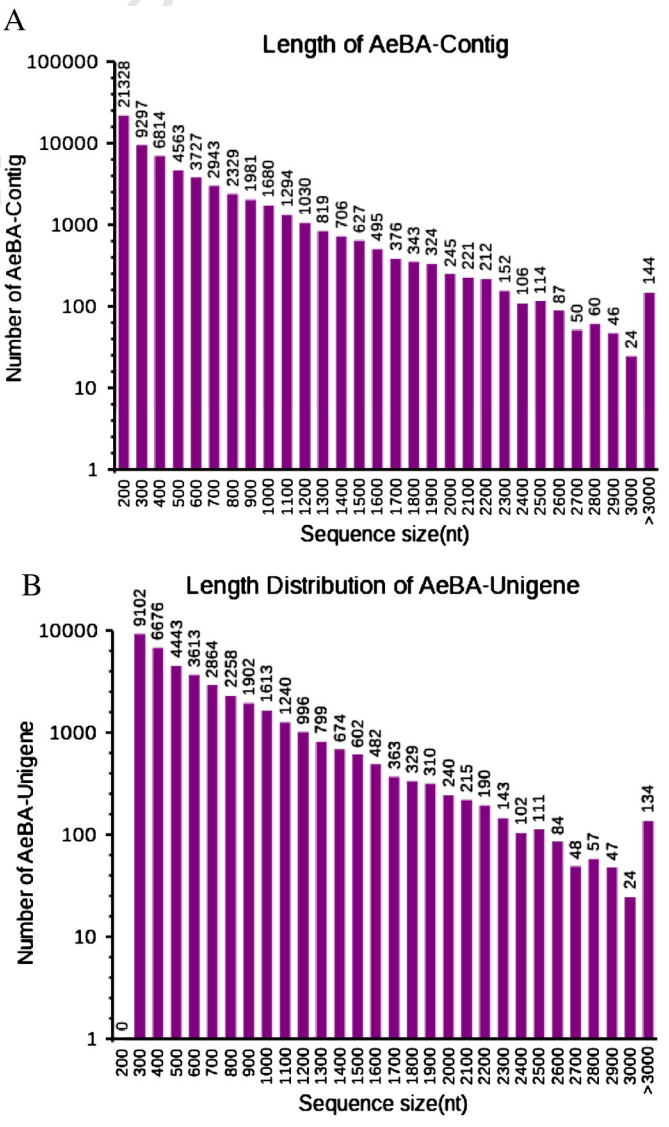

Fig. 2. Overview of the *A. exitialis* transcriptome assembly. (A) Length distribution of the contigs. (B) Length distribution of the unigenes.

Please cite this article as: Li, P., et al., Illumina-based de novo transcriptome sequencing and analysis of Amanita exitialis basidiocarps, Gene (2013), http://dx.doi.org/10.1016/j.gene.2013.09.014

metabolic processes (GO: 0004553) were highly represented in the category of biological processes. Catalytic activity (GO: 0044237) was the most represented GO term, followed by binding (GO: 0019941) in the category of molecular functions. Cells and cell part (GO: 0003723) were the most represented in the categories of cellular components.

A total of 21,420 unigenes were assigned to the COG classifications (Fig. 4). Among the 25 COG categories, general function prediction (3186; 14.87%) represented the largest group, followed by translation, ribosomal structure and biogenesis (1909; 8.91%), replication, recombination and repair (1583; 7.39%), transcription (1571; 7.33%), carbohydrate transport and metabolism (1534; 7.16%), posttranslational modification, protein turnover, and chaperones (1410; 6.58%), and amino acid transport and metabolism (1295; 6.05%). Two additional categories were listed, including function unknown (853; 3.98%) and secondary metabolite biosynthesis, transport, and catabolism (697; 3.25%).

Pathway analyses were also performed on all assembled unigenes as an alternative approach to functional categorization and annotation. Briefly, of these sequences with KEGG annotation, the numbers of unigenes involved in the metabolic pathways, biosynthesis of secondary metabolites, and microbial metabolism in the diverse environments were 4681 (28.74%), 1884 (11.57%), and 1726 (10.6%) respectively. A total of 10,022 unigenes were classified into metabolism processing, including 874 unigenes in the majority groups of purine metabolism, 793 unigenes in the group of starch and sucrose metabolism, 538 unigenes in the group of pyrimidine metabolism, 516 unigenes in the group of amino sugar and nucleotide sugar metabolism, and 2736 unigenes in the group of all amino acid metabolism. Genetic information processing accounted for 1848 unigenes. The main sub-groups were protein processing in the endoplasmic reticulum (549 unigenes) and mRNA surveillance pathway (293 unigenes), nucleotide excision repair (238 unigenes), mismatch repair (222 unigenes), base excision repair (216 unigenes), DNA replication (212 unigenes), and protein export (118 unigenes). Biosynthesis processing accounted for 3712 unigenes. The main sub-groups were biosynthesis of secondary metabolites (1884 unigenes); aminoacyl-tRNA biosynthesis (285 unigenes); valine, leucine and isoleucine biosynthesis (151 unigenes); N-glycan biosynthesis (126 unigenes); pantothenate and CoA biosynthesis (125 unigenes); phenylalanine, tyrosine and tryptophan biosynthesis (120 unigenes); biosynthesis of unsaturated fatty acids (113 unigenes); and ubiquinone and other terpenoid–quinone biosynthesis (103 unigenes). Meanwhile, 4550 unigenes were involved in degradation processing. The main sub-groups were limonene and pinene degradation (508 unigenes), RNA degradation (458 unigenes), aminobenzoate degradation (440 unigenes), naphthalene degradation (432 unigenes) and polycyclic aromatic hydrocarbon degradation (410 unigenes) [Supporting Information (SI 1) file].

3.3. Analysis of genes encoding the toxins and related peptides in the *A. exitialis* transcriptome

All of the toxin-related genes in the *A. exitialis* transcriptome were searched. Eleven related toxin unigene sequences were discovered and their amino acid sequences were predicted (Table 2A). One sequence (IWGIGCNP) matched the amino acid sequence of α-amanitin, one matched β-amanitin (IWGIGCDP), one matched phallacidin (AWLVDCP), and one matched to amanexitide (VFSLPVFFP). The seven unknown sequences had not been reported before.

The cDNA sequences of the α-amanitin gene (*α-AMA*), β-amanitin (*β-AMA*) gene, and phallacidin (*PHA*) gene of the *A. exitialis* transcriptome were also analyzed (Fig. 5). The DNA sequences of *α-AMA*, *β-AMA*, *PHA*, and their translation products were aligned and found to be similar in overall size and sequence. A total of 29 of 30 (97%) of the upstream regions between *α-AMA* and *β-AMA* were common, while 49 of 54 (90.7%) downstream regions were shared, and 20 of 24 (83.3%) of the toxin regions were shared (Fig. 5A). The upstream regions of *α-AMA* and *PHA* with the same DNA sequences, whereas 46 of 54 (85.2%) of the downstream sequences were shared, and only 12 of 24 (50%) of the toxin regions in common (Fig. 5B). Therefore, a variable toxin region flanked by the conserved upstream and downstream regions composed the proproteins of α-amanitin, β-amanitin, and phallacidin.

3.4. Confirmation of the major toxin encoding genes

The authenticity and reliability of the toxin-encoding genes found in the *A. exitialis* transcriptome were verified. By amplifying the genome DNA and cDNA of some basidiocarps, the products with approximate 280 bp and 200 bp were obtained (SI 2), respectively, the results of cloning sequencing revealed nine related complete or almost complete peptide sequences (Table 2B). Three of translations from the three different basidiocarps matched the amino acid sequence of α-amanitin (IWGIGCNP); one matched to β-amanitin (IWGIGCDP), one matched to phallacidin (AWLVDCP), and one matched to amanexitide (VFSLPVFFP).

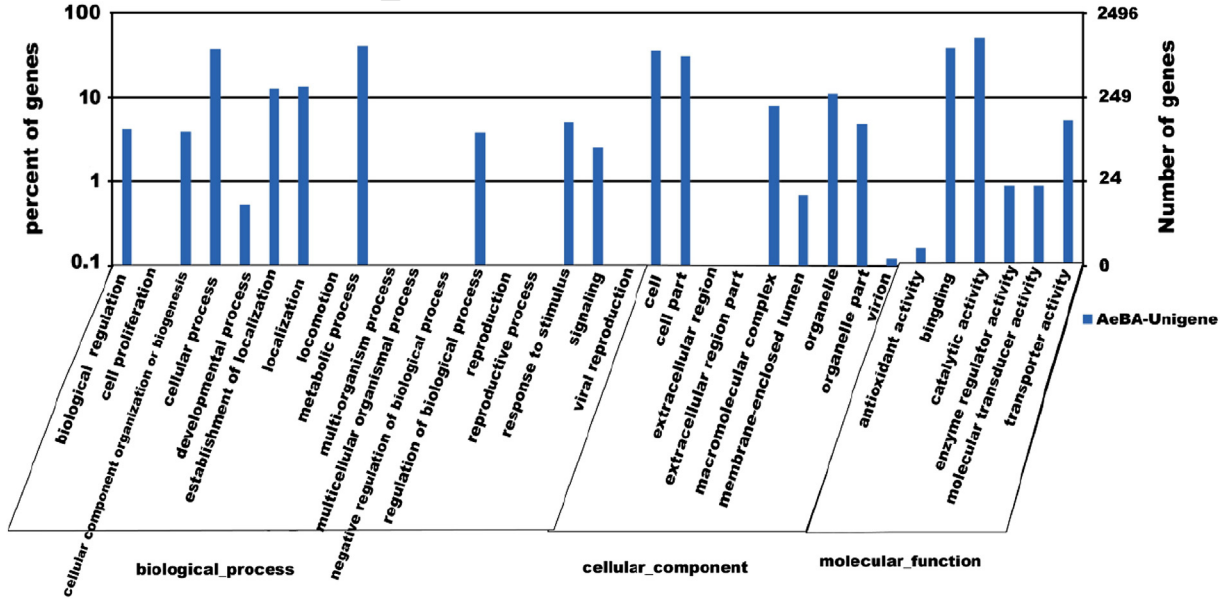

Fig. 3. Gene Ontology (GO) analysis and functional classification of the *A. exitialis* transcriptome.

Please cite this article as: Li, P., et al., Illumina-based de novo transcriptome sequencing and analysis of Amanita exitialis basidiocarps, Gene (2013), http://dx.doi.org/10.1016/j.gene.2013.09.014

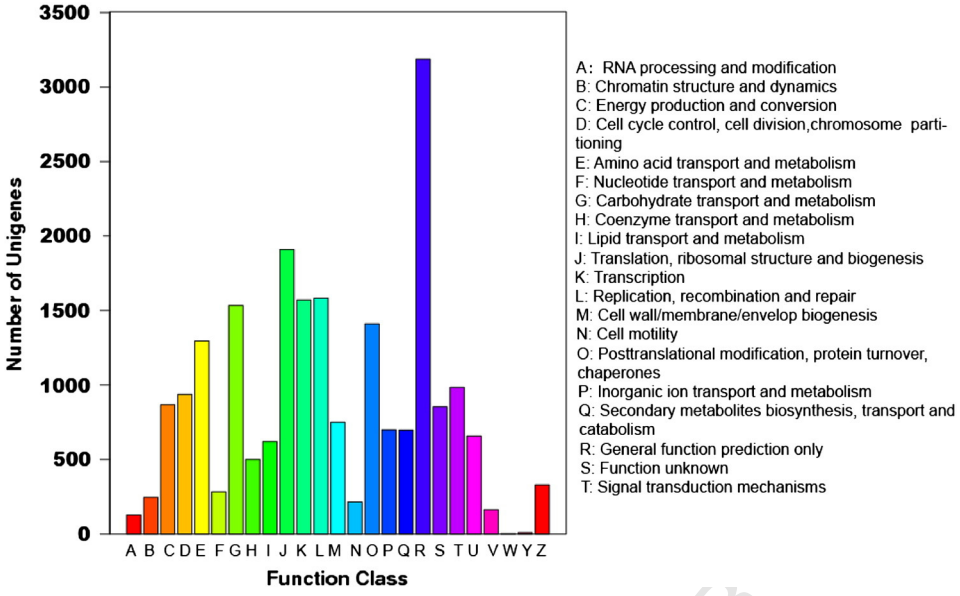

Fig. 4. Clusters of Orthologous Groups (COG) function classification of the *A. exitialis* transcriptome.

Therefore, nearly all of the toxin genes from the *A. exitialis* transcriptome were confirmed by PCR.

3.5. Polymorphism analysis of the major Amanita toxin genes

In this study, when the cDNA was amplified by the degenerate primers from the multiple *A. exitialis* individuals, different gene sequences encoding the same toxins were obtained. For example, there were three sequences of *α-AMA* and *PHA* in *A. exitialis* (Figs. 6A1, A3). Among the three *α-AMA* sequences of *A. exitialis*, there are six polymorphic sites with a difference rate of 5.56% (6/108). Among the three *PHA* sequences of *A. exitialis*, there are also six different bases with a difference rate 5.56% (6/108).

Aligning the three *α-AMA* sequences of *A. exitialis* and the one *α-AMA* sequence of *A. bisporigera*, nine substitutional sites were found with a difference rate 8.33% (9/108), and alignment of the six *α-AMA* sequences of *Amanita* and *Galerina* species revealed 39 substitutional sites, with a difference rate up to 36.11% (39/108) (Fig. 6A1). Among the two *β-AMA* sequences from *A. exitialis* and *A. verna*, seven substitutional bases were found with the difference rate of 6.86% (7/102; six gaps were not calculated) (Fig. 6A2); Among the five *PHA* sequences from *A. exitialis*, *A. bisporigera*, and *A. virosa*, eleven substitutional bases were found, with a difference rate of 10.48% (11/105) (Fig. 6A3).

All of the predicted amino acid sequences of these DNA sequences were also aligned. Among the four α-amanitin sequences of *A. exitialis* and *A. bisporigera*, four different amino acids were found, while among the six α-amanitin sequences of *A. exitialis*, *A. bisporigera*, and *G. marginata*, 16 different amino acids were found (Fig. 6B1). Between the β-amanitin sequences of *A. exitialis* and *A. verna*, there were two different amino acids (Fig. 6B2). Among the five phallacidin sequences of *A. exitialis*, *A. bisporigera*, and *A. verna*, only three different amino acids were found (Fig. 6B3).

Table 2
The predicted amino acid sequences of toxins and related peptides from *Amanita exitialis* basidiocarps.

	Upstream sequence	Toxin or related peptide regions	Downstream sequence	Notes
A	MSDINATRLP	IWGIGCNP	CVGDDVTSVLTRGEALC*	α-AMA1
	MSDINATRLP	IWGIGCDP	CVGDDVTALLTRGEALC*	β-AMA1
	MSDINATRLP	AWLVDCP	CVGDDVNRLLTRGESLC*	PHA1
	MSDINTARLP	VFSLPVFFP	FVSDDCIAVLTRGESLC*	Amanexitide1
	MSDINPTRLP	IFWFIYFP	CVSDVDSTLTRGER*	UN-PRO1
	MSDINATRLP	IIWAPVVP	CISDDNDSTLTRGQR*	UN-PRO2
	MSDINVIRAP	LLILSILP	CVGDDIEVLR RGEGLS	UN-PRO3
	MSDINATRLP	VWIGYSP	CVGDDCIALLTRGEGLC*	UN-PRO4
	MSDINATRLP	LFFPPDFRPP	CVGDADNFTLTRGENLC*	UN-PRO5
	MSDINTTRLP	FVFVASPP	CVGDDIAMVLTRGENLC*	UN-PRO6
	MSDINATRLP	AWLTDCP	CVGDDVNRLLTRGESLC*	UN-PRO7
B	MSDINATRLP	IWGIGCNP	CVGDDVTSVLTRGEA	α-AMA2
	MSDINATRLP	IWGIGCNP	CVGDDVTSVLTRGEALC*	α-AMA3
	MSDINATRLP	IWGIGCNP	CVGDEVAALLTRGEALC*	α-AMA4
	MSDINATRLP	IWGIGCDP	CVGDDVTALLTRGEALC*	β-AMA2
	MSDINATRLP	AWLVDCP	CVGDDVNRLLTRGESLC*	PHA2
	MSDINATRLP	VFSLPVFFP	CVGDDCIALLTRGEGLC*	Amanexitide2
	MSDINATRLP	FVFVASPP	CVGDDIAMVLTRGENLC*	UN-PRO8
	MSDINATRLP	VWIGYSP	FVSDDIQAVLTRGESLC*	UN-PRO9
	MSDINATRLP	IFWFIYFP	CVSDVDSTLTRGER*	UN-PRO10

All the sequences of A were from the transcriptome and all the sequences of B were amplified by polymerase chain reaction. "*" means the stop codons. UN-PRO means the unknown proproteins.

3.6. Sequences comparison and phylogenetic analysis of Amanita peptides

All the predicted protein products of this gene family from *A. exitialis* are characterized by a hypervariable "toxin" region capable of encoding a wide variety of peptides of 7–10 amino acids flanked by conserved

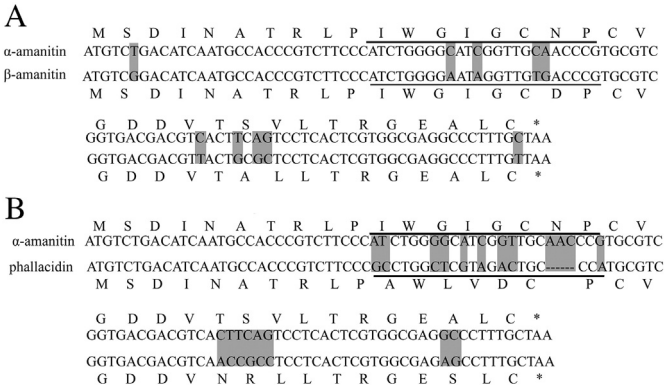

Fig. 5. Alignment of the proproteins' cDNA sequences of the major toxin encoding genes. (A) The proprotein cDNA sequences of the α-amanitin and β-amanitin from *A. exitialis*. (B) The proprotein cDNA sequences of the α-amanitin and phallacidin from *A. exitialis*. The mature toxin sequences are lined. The substitutional bases among the copies are boxed in black.

sequences. The toxin regions start immediately downstream of the invariant Pro residue and end at an invariant Pro residue, which are hypervariable compared with the upstream and downstream sequences. The upstream conserved sequence "MSDINATRLP" and downstream conserved sequence "CVGDDV" (the first D is invariant) become the significant feature of these proteins found in *Amanita* (Table 2), which are consistent with the results of Hallen et al. while the conserved sequences are various in different genera, the upstream sequences of two copies of *G. marginata* (*GmAMA1-1* and *GmAMA1-2*) beginning with "MFDTNA(S)TRLP" and the downstream sequences with "WTAEHVDQTLASGND" (Luo et al., 2012).

A phylogenetic tree was constructed and the phylogenetic relationship among the 26 proproteins' sequences from six *Amanita* species were analyzed (Fig. 7). The alignment comprised 34 characteristics. In the phylogenetic tree, all proprotein sequences from the *Amanita* species were distributed in five clades: amatoxins, phallotoxins, amanexitide, and two unknown peptide clades. In the amatoxins clade, nine amatoxin proteins including α-amanitin and β-amanitin from *A. exitialis*, *A. bisporigera*, *A. phalloides*, and *A. verna* form a cluster with a 60% bootstrap. In the phallotoxins clade, six phallotoxin proproteins including one unknown peptide (UN-PRO7), phallacidin and phalloidin from *A. exitialis*, *A. bisporigera*, *A. ocreata*, and *A. virosa* form a cluster with a 76% bootstrap. In the amanexitide clade, one unknown protein (UN-PRO4) and amanexitide proteins from *A. exitialis* form a cluster with a bootstrap less than 50%. In the unknown peptide I clade, four unknown proteins (UN-PRO1, 2, 9, 10) from *A. exitialis* form a cluster with a low bootstrap. In the unknown peptide II clade, four unknown proteins (UN-PRO3, 5, 6, 8) from *A. exitialis* form a cluster with a low bootstrap.

4. Discussion

4.1. A. exitialis transcriptome data

De novo transcriptome assemblies have been created using next generation sequencing technologies for some important organisms including plants: *Eucalyptus grandis* (Novaes et al., 2008), *Pinus contorta* (Parchman et al., 2010), and *Panax quinquefolius* (Sun et al., 2010); insects: *Melitaea cinxia* (Vera et al., 2008), *Sarcophaga crassipalpis* (Hahn et al., 2009), and *Erynnis propertius* (O'Neil et al., 2010); fishes: *Zoarces viviparous* (Kristiansson et al., 2009), *Coregonus* sp. (Renaut et al., 2010), and *Poecilia reticulate* (Fraser et al., 2011); ten species of birds (Kunstner et al., 2010) and a few fungi: *Ganoderma lucidum* (Yu et al., 2012) and *Schizosaccharomyces pombe* (Bah et al., 2012), however, the transcriptome of lethal amanitas has not been reported.

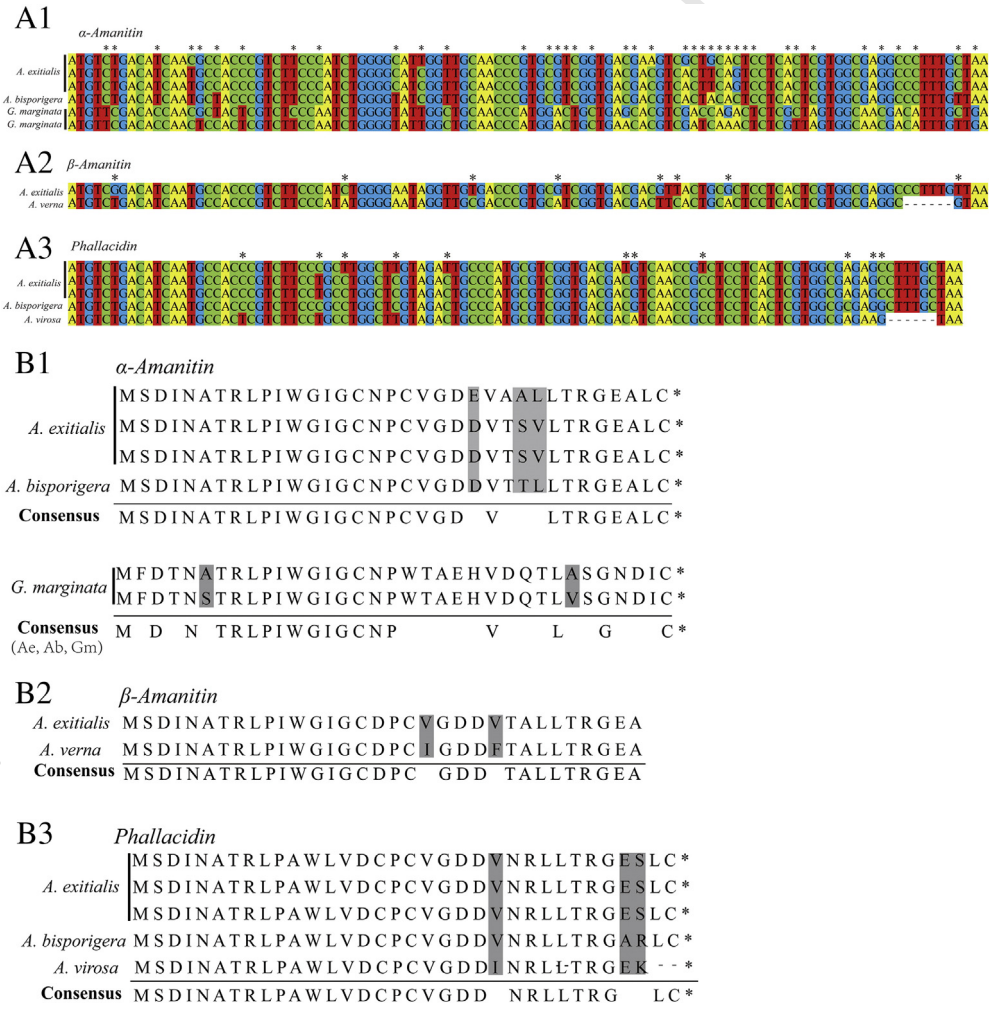

Fig. 6. Alignment of the DNA and proprotein sequences related to *α-amanitin*, *β-amanitin*, and *phallacidin*. (A) Alignment of the encoding sequences of α-amanitin, β-amanitin, and phallacidin; (B) Alignment of the proprotein sequences of *α-amanitin*, *β-amanitin*, and *phallacidin*. Ae means the *A. exitialis*, Ab means the *A. bisporigera*, Gm means the *G. marginata*. Asterisks indicate the polymorphic sites in all the sequences. The different amino acids among the peptide sequences are boxed in black. The three *α-amanitin* sequences and the three *phallacidin* sequences were from the three different *A. exitialis* basidiocarps. The sequences of *A. bisporigera*, *A. phalloides*, and *G. marginata* were from Hallen (2007) and Luo (2012), and the sequences of *A. virosa* and *A. verna* were from the GenBank.

Please cite this article as: Li, P., et al., Illumina-based de novo transcriptome sequencing and analysis of Amanita exitialis basidiocarps, Gene (2013), http://dx.doi.org/10.1016/j.gene.2013.09.014

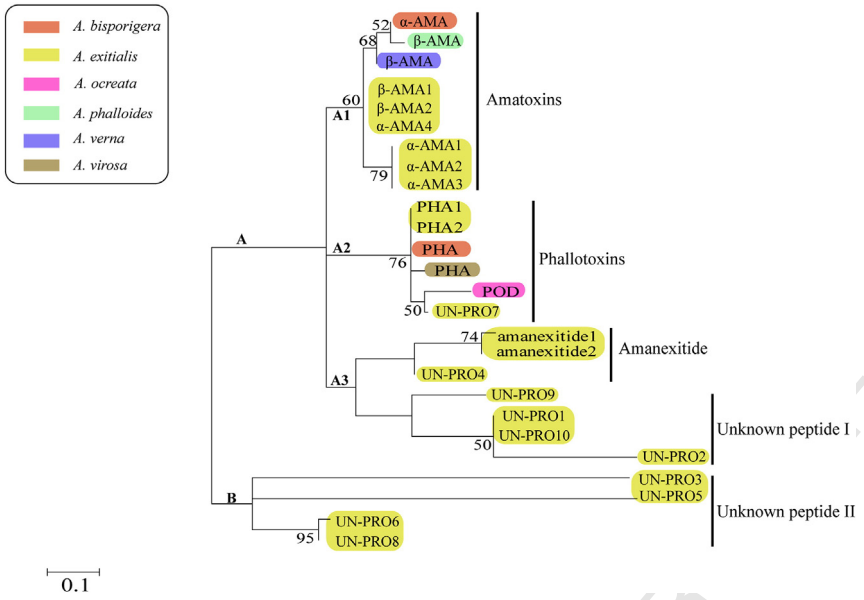

Fig. 7. The phylogenetic tree of *Amanita* peptides' proprotein sequences.

In this study, the transcriptome of *A. exitialis* was first performed using Illumina HiSeq 2000 technology, while the *de novo* assembly of full-length transcripts was assembled using the Trinity method. The average length of contig and unigene lengths in our study were 481 bp and 662 bp, respectively, and the average number of reads per unigene was 66. These findings are comparable to those of other studies using similar technologies (mean, 415 bp) (Bah et al., 2012; Fraser et al., 2011; Hahn et al., 2009; Kristiansson et al., 2009; Kunstner et al., 2010; Novaes et al., 2008; O'Neil et al., 2010; Parchman et al., 2010; Renaut et al., 2010; Sun et al., 2010; Vera et al., 2008; Yu et al., 2012).

The results here demonstrated that the sequencing and assembly strategy substantially improves the assembly results. The high quality of the obtained *A. exitialis* reference transcriptome will become essential for the annotation and study of other lethal amanitas' genomic resources in future studies. Moreover, all of the unigenes were classified into functional categories for the understanding of the gene functions and regulation pathways. The results presented here will enrich the scientific community's knowledge of the genetic resources for biotoxins and the diversities of *Amanita* toxins. However, since no related genome and transcriptome studies of *Amanita* species have been performed until now, transcriptome data comparison among these species was limited.

4.2. Genes encoding the toxins and related peptides

Although 22 *Amanita* peptide toxins have been reported, only four have been studied: *α-AMA* was reported in *A. bisporigera* (Hallen et al., 2007) and *G. marginata* (Luo et al., 2012), *β-AMA* was reported in *A. phalloides*, *PHA1* was reported in *A. bisporigera*, and the phalloidin (POD) gene was reported in *A. ocreata* (Hallen et al., 2007). In the transcriptome of *A. exitialis*, the toxin gene sequences encoding α-amanitin, β-amanitin, and phallacidin were discovered, and the same toxin-encoding genes were also obtained using reverse transcription PCR, which greatly enriches the peptide toxin gene information. In addition, seven genes encoding unknown novel peptides and amanexitide were first found.

In our study, some toxin gene polymorphisms existed in population, species, and genus. The genes *α-AMA* and *PHA* are polymorphic in the population of *A. exitialis*, with difference rates of 5.56% (6/108). Sequential comparison and analyses of *α-AMA*, *β-AMA*, and *PHA* polymorphism in *Amanita* species revealed difference rates of 8.33% (9/108), 6.86% (7/102), and 10.48% (11/105), respectively. However, in *α-AMA* polymorphism analysis in *Amanita* and *Galerina* species the difference rate was up to 36.11% (39/108). Therefore, there are more toxin gene variations of the different genera than in the different species of a single genus. The alignment results of amino acid sequences also support the above results, especially in the upstream region and the toxin region sequences, though several substitutional bases were found in their DNA sequences, they could encode the same amino acids. These results showed that the degeneracy also existed in these toxin encoding genes (Fig. 6). The gene polymorphism and degeneracy of the *Amanita* toxins demonstrate that these lethal *Amanita* species have evolved a polytropic mechanism of biosynthesis that endow them with the ability to more efficiently biosynthesize multitude cyclic peptides and help them to better adapt to the environment. Q18

4.3. Amanita toxins and related peptides

Our study showed that the lethal amanitas contain a variety of *AMA* and *PHA* related genes similar to the MSDIN family, but not all of the MSDIN members were found in all of the lethal amanitas. In contrast, some species have their own particular peptides: the seven related and predicted amino acid sequences VWIGYSP, FVFVASPP, IFWFIYFP, LLILSILP, LFFPPDFRPP, VFSLPVFFP, and AWLTDCP were only found in *A. exitialis*; the thirteen related peptide sequences GFVPILFP, FYQFPDFKYP, FFQPPEFRPP, LFLPPVRMPP, LFLPPVRLPP, YVVFMSFIPP, CIGFLGIP, LSSPMLLP, ILMLAILP, IPGLIPLGIP, GAYPPVPMP, GMEPPSPMP, and HPFPLGLQP were only found in *A. bisporigera*, and the two sequences FNILPFMLPP and IIGILLPP were only found in the *A. phalloides* (Hallen et al., 2007). As such, the lethal amanitas have a broad capacity to synthesize small cyclic peptides including amatoxins and phallotoxins as well as some unknown cyclic peptides. Our discovery of the toxin-encoding genes could also provide direction for isolating the new cyclic peptides; for example, amanexitide, a kind of cyclic nonapeptide, was recently a new separation of the short peptide from *A. exitialis* (Xue et al., 2011), the first encoding gene obtained.

The amatoxin and phallotoxin gene family is predicted to encode Q19 proproteins of 34–37 amino acids with conserved upstream and downstream sequences flanking a hypervariable region of 7–10 amino acids. What are the phylogenetic relationships of the toxin and the related

Please cite this article as: Li, P., et al., Illumina-based de novo transcriptome sequencing and analysis of Amanita exitialis basidiocarps, Gene (2013), http://dx.doi.org/10.1016/j.gene.2013.09.014

proteins? In the study, all known cyclic peptide toxin proteins from *Amanita* and related unknown proteins from *A. exitialis* were analyzed. All of the known amatoxin proteins formed a cluster, while all of the known phallotoxin proteins formed another cluster. The amanexitide clade is the sister cluster of amatoxins and phallotoxins. From the phylogenetic relationship, the classifications of those unknown proteins from *A. exitialis* are inferred. UN-PRO7 and the phallotoxin are clustered into one big clade, which suggests that UN-PRO7 might be an unknown or new phallotoxin; UN-PRO4 might have the homologous functions as the amanexitide, UN-PRO1, UN-PRO2, UN-PRO9, and UN-PRO10 formed a cluster as the sister clade of amanexitide; the other four unknown proteins (UN-PRO3, 5, 6, 8) are far away from amatoxins, phallotoxins, and amanexitides, and may be the other cyclic peptides. However, their true identifications and functions should be researched in a future study.

POP was considered a key enzyme during toxin biosynthesis (Luo et al., 2009). In our study, 12 POP unigenes were obtained that will benefit subsequent research on the toxin biosynthesis (unpublished). Although the *Amanita* toxin genes and key enzymes involved in toxin metabolism have been studied, expression of the toxin gene remains difficult. The main reasons for this are that: 1) the *Amanita* species producing toxins are mycorrhizal fungi that are difficult to artificially cultivate (Yang, 2005); 2) the toxins are toxic and the choice of the expression system is restricted; and 3) the toxins are small cyclic peptides and formed after complex modification (Luo et al., 2012). Meanwhile, for other ribosomal peptide biosynthetic systems, such as cyclotides or patellamides, which could not serve as precedent and no KEGG pathway related to *Amanita* toxin biosynthesis was found, it is still difficult to address the biosynthetic pathway.

5. Accession numbers

The raw next generation sequencing reads are stored in the European Nucleotide Archive under study ERP002373.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2013.09.014.

Competing interests

The authors have declared that no competing interests exist.

Uncited reference

Thompson et al., 1997

Acknowledgments

This work was funded by grants from the National Natural Science Foundation of China (Project No. 31101592). The mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. The USDA is an equal opportunity provider and employer.

References

Bah, A., Wischnewski, H., Shchepachev, V., Azzalin, C.M., 2012. The telomeric transcriptome of *Schizosaccharomyces pombe*. Nucleic Acids Res. 40, 2995–3005.
Bamburg, J.R., 1999. Proteins of the ADF/cofilin family: essential regulators of actin dynamics. Annu. Rev. Cell Dev. Biol. 15, 185–230.
Bresinsky, A., Besl, H., 1990. A Color Atlas of Poisonous Fungi: A Handbook for Pharmacists, Doctors and Biologists. Wolfe, Wurzburg, Germany 295.
Bushnell, D.A., Cramer, P., Kornberg, R.D., 2002. Structural basis of transcription: alpha-amanitin–RNA polymerase II cocrystal at 2.8 A resolution. Proc. Natl. Acad. Sci. U. S. A. 99, 1218–1222.
Chen, Z.H., Hu, J.S., Zhang, Z.G., Zhang, P., Li, D.P., 2003. Determination and analysis of the main amatoxins and phallotoxins in 28 species of *Amanita* from China. Mycosystema 22, 565–573.
Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.
Deng, W.Q., Li, T.H., Xi, P.G., Gan, L.X., Xiao, Z.D., Jiang, Z.D., 2011. Peptide toxin components of *Amanita exitialis* basidiocarps. Mycologia 103, 946–949.
Faulstich, H., Buku, A., Bodenmuller, H., Wieland, T., 1980. Virotoxins actin-binding cyclic-peptides of *Amanita virosa* mushrooms. Biochemistry 19, 3334–3343.
Fraser, B.A., Weadick, C.J., Janowitz, I., Rodd, F.H., Hughes, K.A., 2011. Sequencing and characterization of the guppy (*Poecilia reticulata*) transcriptome. BMC Genomics 12, 202. http://dx.doi.org/10.1186/1471-2164-12-202.
Garber, M., Grabherr, M.G., Guttman, M., Trapnell, C., 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods 8, 469–477.
Gibbons, J.G., Janson, E.M., Hittinger, C.T., Johnston, M., Abbot, P., Rokas, A., 2009. Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. Mol. Biol. Evol. 26, 2731–2744.
Grabherr, M.G., et al., 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29, 644–652.
Hahn, D.A., Ragland, G.J., Shoemaker, D.D., Denlinger, D.L., 2009. Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. BMC Genomics 10, 234. http://dx.doi.org/10.1186/1471-2164-10-234.
Hallen, H.E., Luo, H., Scott-Craig, J.S., Walton, J.D., 2007. Gene family encoding the major toxins of lethal *Amanita* mushrooms. Proc. Natl. Acad. Sci. U. S. A. 104, 19097–19101.
Hu, J.S., Zhang, P., Zeng, J., Chen, Z.H., 2012. Determination of amatoxins in different tissues and development stages of *Amanita exitialis*. J. Sci. Food Agric. 92, 2664–2667.
Iseli, C., Jongeneel, C.V., Bucher, P., 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. 138–148.
Kristiansson, E., Asker, N., Forlin, L., Larsson, D.G.J., 2009. Characterization of the *Zoarces viviparus* liver transcriptome using massively parallel pyrosequencing. BMC Genomics 10, 345. http://dx.doi.org/10.1186/1471-2164-10-345.
Kroncke, K.D., Fricker, G., Meier, P.J., Gerok, W., Wieland, T., Kurz, G., 1986. Alpha-amanitin uptake into hepatocytes — identification of hepatic membrane transport systems used by amatoxins. J. Biol. Chem. 261, 2562–2567.
Kunstner, A., et al., 2010. Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. Mol. Ecol. 19, 266–276.
Lengsfeld, A.M., Low, I., Wieland, T., Dancker, P., Hasselba, W., 1974. Interaction of phalloidin with actin. Proc. Natl. Acad. Sci. U. S. A. 71, 2803–2807.
Letschert, K., Faulstich, H., Keller, D., Keppler, D., 2006. Molecular characterization and inhibition of amanitin uptake into human hepatocytes. Toxicol. Sci. 91, 140–149.
Luo, H., Hallen-Adams, H.E., Walton, J.D., 2009. Processing of the phalloidin proprotein by prolyl oligopeptidase from the mushroom *Conocybe albipes*. J. Biol. Chem. 284, 18070–18077.
Luo, H., Hallen-Adams, H.E., Scott-Craig, J.S., Walton, J.D., 2010. Colocalization of amanitin and a candidate toxin-processing prolyl oligopeptidase in *Amanita* basidiocarps. Eukaryot. Cell 9, 1891–1900.
Luo, H., Hallen-Adams, H.E., Scott-Craig, J.S., Walton, J.D., 2012. Ribosomal biosynthesis of alpha-amanitin in *Galerina marginata*. Fungal Genet. Biol. 49, 123–129.
MacLean, D., Jones, J.D.G., Studholme, D.J., 2009. Application of 'next-generation' sequencing technologies to microbial genetics. Nat. Rev. Microbiol. 7, 287–296.
Novaes, E., et al., 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9, 312. http://dx.doi.org/10.1186/1471-2164-9-312.
Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M., 1999. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 27, 29–34.
O'Neil, S.T., Dzurisin, J.D.K., Carmichael, R.D., Lobo, N.F., Emrich, S.J., Hellmann, J.J., 2010. Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. BMC Genomics 11, 310. http://dx.doi.org/10.1186/1471-2164-11-310.
Parchman, T.L., Geist, K.S., Grahnen, J.A., Benkman, C.W., Buerkle, C.A., 2010. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics 11, 180. http://dx.doi.org/10.1186/1471-2164-11-180.
Pertea, G., et al., 2003. TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19, 651–652.
Pop, M., Salzberg, S.L., 2008. Bioinformatics challenges of new sequencing technology. Trends Genet. 24, 142–149.
Ranwez, V., Harispe, S., Delsuc, F., Douzery, E.J.P., 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One 6.
Renaut, S., Nolte, A.W., Bernatchez, L., 2010. Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). Mol. Ecol. 19, 115–131.
Sun, C., et al., 2010. *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 11, 262. http://dx.doi.org/10.1186/1471-2164-11-262.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.
Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.
Trombetti, G.A., Bonnal, R.J.P., Rizzi, E., De Bellis, G., Milanesi, L., 2007. Data handling strategies for high throughput pyrosequencers. BMC Bioinforma. 8, S22. http://dx.doi.org/10.1186/1471-2105-8-S1-S22.
Vera, J.C., et al., 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol. Ecol. 17, 1636–1647.

Please cite this article as: Li, P., et al., Illumina-based de novo transcriptome sequencing and analysis of Amanita exitialis basidiocarps, Gene (2013), http://dx.doi.org/10.1016/j.gene.2013.09.014

Wieland, T., 1986. Peptides of Poisonous *Amanita* Mushrooms. Springer, New York.
Wieland, T., Faulstich, H., 1991. 50 years of amanitin. Experientia 47, 1186–1193.
Xue, J.H., Wu, P., Chi, Y.L., Xu, L.X., Wei, X.Y., 2011. Cyclopeptides from *Amanita exitialis*. Nat. Prod. Bioprospect. 1, 52–56.
Yang, Z.L., 2005. Amanitaceae. Flora Fungorum Sinicorum, 27. Science Press, Beijing.
Yang, Z.L., Li, T.H., 2001. Notes on three white *Amanitae* of section *Phalloideae* (*Amanitaceae*) from China. Mycotaxon 78, 439–448.
Ye, J., et al., 2006. WEGO: a web tool for plotting GO annotations. Nucleic Acids Res. 34, 293–297.
Yu, G.J., et al., 2012. Deep insight into the *Ganoderma lucidum* by comprehensive analysis of its transcriptome. PLoS One 7, e44031. http://dx.doi.org/10.1371/journal.pone.0044031.
Zhang, P., Chen, Z.H., Hu, J.S., Wei, B.Y., Zhang, Z.G., Hu, W.Q., 2005. Production and characterization of Amanitin toxins from a pure culture of *Amanita exitialis*. FEMS Microbiol. Lett. 252, 223–228.

Please cite this article as: Li, P., et al., Illumina-based de novo transcriptome sequencing and analysis of Amanita exitialis basidiocarps, Gene (2013), http://dx.doi.org/10.1016/j.gene.2013.09.014