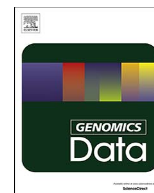
Genomics Data 14 (2017) 126–131

Contents lists available at ScienceDirect

Genomics Data

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



Transcriptome analysis of the Chinese giant salamander (*Andrias davidianus*) using RNA-sequencing

Yong Huang*, Jian Li Xiong, Xiao Chan Gao, Xi Hong Sun

College of Animal Science and Technology, Henan University of Science and Technology, Luoyang 471023, China

ARTICLE INFO

Keywords:
Andrias davidianus
Transcriptome
Unigenes
Gene annotation

ABSTRACT

The Chinese giant salamander (*Andrias davidianus*) is an economically important animal on academic value. However, the genomic information of this species has been less studied. In our study, the transcripts of *A. davidianus* were obtained by RNA-seq to conduct a transcriptomic analysis. In total 132,912 unigenes were generated with an average length of 690 bp and N50 of 1263 bp by de novo assembly using Trinity software. Using a sequence similarity search against the nine public databases (CDD, KOG, NR, NT, PFAM, Swiss-prot, TrEMBL, GO and KEGG databases), a total of 24,049, 18,406, 36,711, 15,858, 20,500, 27,515, 36,705, 28,879 and 10,958 unigenes were annotated in databases, respectively. Of these, 6323 unigenes were annotated in all database and 39,672 unigenes were annotated in at least one database. Blasted with KEGG pathway, 10,958 unigenes were annotated, and it was divided into 343 categories according to different pathways. In addition, we also identified 29,790 SSRs. This study provided a valuable resource for understanding transcriptomic information of *A. davidianus* and laid a foundation for further research on functional gene cloning, genomics, genetic diversity analysis and molecular marker exploitation in *A. davidianus*.

1. Introduction

The Chinese giant salamander (*A. davidianus*) is the largest extant amphibian in the world [1,2]. Now, it is classified as an endangered species by the International Union for Conservation of Nature and Nature Resources, and is the class II state major protection species in China. In the evolution history of vertebrate, *A. davidianus* occupies a seat at the phylogenetic and species evolution process which is representing a transitional form that links the aquatic animals to terrestrial organisms because it has existed for > 350 million years [3]. Therefore, this species has an important value in scientific research. In history, *A. davidianus* was widely distributed in central and southern China. However, in the past few decades, *A. davidianus* population has declined sharply due to deterioration of habitat, environmental pollution, climate change, infectious diseases, commercial trade and infrastructure development for human settlement [4,5].

Owing to little genomic data was available previously for *A. davidianus*, it has hindered the understanding of the molecular mechanisms associated with growth, reproduction, immunization and sex determination of *A. davidianus*. In recent years, RNA sequencing technologies has been accepted as a powerful approach for large-scale transcriptome profiling for studying non-model species [6,7], which has improved the efficiency and speed of gene discovery. Compared to the whole genome sequencing, RNA sequencing technologies provide a cost-effective approach to produce transcriptome sequences and molecule markers [8–10]. For example, a few of amphibians were undertaken a large-scale analysis of transcriptome sequenced by RNA sequencing technologies [11,12]. Currently, transcriptome analysis reports of *A. davidianus* were only focused on skin, spleen, kidney, liver, intestines and gonad tissues [13,14]. Therefore, further enriching the transcriptome analysis of *A. davidianus* has significant scientific value.

In this study, we are the first to characterize complete transcriptome of *A. davidianus* through the analysis of large-scale transcript sequences generated from a pooled mixed tissues including the spleen, liver, muscle, kidney, skin, testis, gut and heart by using the Illumina Hiseq 2500 high-throughput sequencing platform. These analyses identified a substantial number of unigenes which significantly improve our understanding on the genome prints of *A. davidianus*. Our results provide a global view of the transcriptome and pave the way for further functional characterization of *A. davidianus*.

2. Materials and methods

2.1. Ethics statement and sample collection

The healthy second generation of the farmed male *A. davidianus*

* Corresponding author.
E-mail address: huangyong1979111@126.com (Y. Huang).

http://dx.doi.org/10.1016/j.gdata.2017.10.005
Received 21 July 2017; Received in revised form 21 October 2017; Accepted 21 October 2017
Available online 28 October 2017
2213-5960/ © 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/BY/4.0/).

(three years old) was obtained from Luoyang Huani Bio-Tech Co., Ltd. (Luoyang City, Henan, China). This study has also been reviewed and approved by the Ethics Committee of Henan University of Science and Technology according to the Regulations for the Administration of Affairs Concerning Experimental Animals (Ministry of Science and Technology, China; revised in June 2004). Subsequently, *A. davidianus* were anesthetized and sacrificed by decapitation. The organ samples including the spleen, liver, muscle, kidney, skin, testis, gut and heart were pooled and frozen at − 80 °C until RNA extraction.

2.2. RNA extraction and Illumina sequencing

Total RNA was extracted using TRIzol Reagent (Invitrogen, CA, USA) according to the manufacturers' directions, and then treated with RNase-free DNaseI. The extracted RNA content, integrity and purity were checked by a 2100 Bioanalyzer (Agilent Technologies). The pooled sample, 10 mg of total RNA was used for cDNA library construction following the protocol supplied with the Truseq™ RNA sample prep Kit. Briefly, Poly (A) mRNA was isolated using oligodT 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 *A. davidianus* was constructed and sequenced on the Illumina HiSeq 2500 platform (Sangon Biotech Co., Ltd., Shanghai, China). All raw data for *A. davidianus* obtained in this study were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) with the accession numbers SRP099564.

2.3. Sequence data processing and de novo assembly

The quality of paired-end raw reads in fastq format was assessed using FastQC software (http://www.bioinformatics.babraham.ac.uk). Low-quality reads, such as adaptor sequences or with unknown nucleotides > 10%, were filtered. The clean reads were then combined to form longer fragments. Transcriptome de novo assembly was carried out using the short read assembly program Trinity with default settings to generate transcript. Finally, the redundancy in these transcripts was removed, and contigs were connected to get unique unigenes. All unigenes were predicted using ORF finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html).

2.4. Annotation of unigenes

Functional annotation of the unigenes was performed by search against the nine public databases. All unigenes were compared with sequences in Nr (NCBI non-redundant protein database), Nt (NCBI non-redundant nucleotide sequence database), KOG (EuKaryotic Orthologous Groups), CDD (Conserved Domains Database), PFAM (The Protein Families), GO (Gene Ontology), KEGG (Kyoto Encyclopedia of Genes and Genomes), Swiss-prot and TrEMBL with E-values $\leq 10^{-5}$.

3. Results

3.1. Transcriptome and assembly characterization

The RIN value of total RNA form *A. davidianus* was 8.2 and the 28S/18S ratios also was 1.80. The cDNA library from *A. davidianus* was constructed and sequenced, which generated 102,659,984 raw reads containing 15,398,997,600 bp with an average length of 150 bp. After stringent quality assessment and data filtering, reads with Q20 bases (those with a base quality > 20) were selected as high quality reads for further analysis. Using the clean reads, Trinity produced 158,103 transcripts with an average length of 810 bp and N50 length of 1659 bp. The length of transcripts ranged from 201 to 16,067 bp. Finally, de novo assembly yielded 132,912 unigenes with an average length of 690 bp and N50 length of 1263 bp. Of these unigenes, 42,327 unigenes (31.85%) were > 500 bp and 21,855 unigenes (16.44%) were > 1000 bp (Table 1). As shown in Fig. 1, the sequence length of these unigenes ranged from 200 bp to > 2000 bp. The number of unigenes decreased with increasing length. In addition, the GC content is also one of the important characteristics of the genome base sequence, which can reflect the structure, function and evolutionary information of the gene. The heterogeneity of GC distribution may lead to the functional difference. The average content of GC of *A. davidianus* was 49.85%, and the unigenes with high GC content (> 80%) or too low (< 20%) did not found. The GC content met normal distribution (Fig. 2), indicating that the sequencing quality was perfect.

Table 1
Transcriptome assembly statistics in *A. davidianus*.

Category	Transcripts	Unigenes
Total length (bp)	128,175,999	91,713,308
Sequence no.	158,103	132,912
≥ 500 bp	59,715	42,327
≥ 1000 bp	34,075	21,855
N50	1659	1263
Max length (bp)	16,067	16,067
Min length (bp)	201	201
Average length (bp)	810	690

N50 of Transcripts or unigenes was calculated by ordering all sequences, then adding the lengths from longest to shortest until the summed length exceeded 50% of the total length of all sequence.

3.2. ORF prediction and annotation of unigenes sequences

The ORF sequence was predicted by using ORF prediction software, and 132,416 were predicted to be encoded amino acids, accounting for 99.63% of all unigenes. The remaining 496 unigenes contained no ORFs, indicating they may either be non-coding sequences coming from untranslated regions (UTR) or de novo unigenes that contain < 150 bp of the start or end of an ORF. Using the BLASTx tool, the unigenes were aligned with sequences recorded in the major databases including CDD, KOG, NR, NT, Pfam, Swisse-Prot, TrEMBL, GO and KEGG. A Venn diagram illustrates the distribution of unigenes annotated to four databases (Fig. 3). Among 132,912 unigenes, a total of 24,049 showed significant matches to CDD, 18,406 to KOG, 36,711 to NR, 15,858 to NT, 20,500 to Pfam, 27,515 to Swiss-Prot, 36,705 to TrEMBL, 28,879 to GO and 10,958 to KEGG respectively. Altogether, 6323 (4.76%) unigenes exhibited a significant match with nine major public databases, and 39,672 unigenes showed significant match, at least one hit to these databases (Table 2).

3.3. Functional classification by GO, KOG and KEGG

GO provides an international standardized gene functional classification system of each assembled unigenes by blasting with the Nr database. In this study, a total of 41,553 unigenes were categorized into 62 subcategories under three main ontologies: molecular function, cellular component, and biological process. For biological process, 21,763 (16.37%) were in the cellular process category, 17,266 (12.99%) were in the metabolic process category and 13,952 (10.50%) were in the single-organism process category. For cellular component, (20,759, 15.62%) cell part and cell represented the majority of this category, respectively. Meanwhile, for molecular function, 18,579 (13.98%) binding and 12,120 (9.12%) catalytic activity were highly represented and assigned to this category, whereas only a few genes

Download English Version:

https://daneshyari.com/en/article/8646371

Download Persian Version:

https://daneshyari.com/article/8646371

Daneshyari.com