Genomics 97 (2011) 364–371

Contents lists available at ScienceDirect

Genomics

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

Transcriptome profiling and digital gene expression by deep-sequencing in normal/regenerative tissues of planarian *Dugesia japonica*

Yun-Fei Qin [a,1], Hui-Min Fang [a,1], Qing-Nan Tian [a], Zhen-Xia Bao [a], Ping Lu [a], Jin-Mei Zhao [a], Jia Mai [a], Zhao-Yu Zhu [a], Lin-Lin Shu [a], Li Zhao [b], San-Jun Chen [a], Feng Liang [a], Yi-Zhe Zhang [a], Shou-Tao Zhang [a,c,*]

[a] Department of Bioengineering, Zhengzhou University, Zhengzhou, Henan 450001, China
[b] China National Center for Biotechnology Development, Beijing, 100036, China
[c] Mobile Station of Basic Medical Sciences, Zhengzhou University, Zhengzhou, Henan 450001, China

ARTICLE INFO

Article history:
Received 15 January 2011
Accepted 3 February 2011
Available online 17 February 2011

Keywords:
Digital gene expression
Dugesia japonica
Gene Ontology
Planarian
Regeneration
Transcriptome

ABSTRACT

Planarians exhibit an extraordinary ability to regenerate lost body parts which is attributed to an abundance of pluripotent somatic stem cells called neoblasts. In this article, we report a transcriptome sequence of a Planaria subspecies *Dugesia japonica* derived by high-throughput sequencing. In addition, we researched transcriptome changes during different periods of regeneration by using a tag-based digital gene expression (DGE) system. Consequently, 11,913,548 transcriptome sequencing reads were obtained. Finally, these reads were eventually assembled into 37,218 unique unigenes. These assembled unigenes were annotated with various methods. Transcriptome changes during planarian regeneration were investigated by using a tag-based DGE system. We obtained a sequencing depth of more than 3.5 million tags per sample and identified a large number of differentially expressed genes at various stages of regeneration. The results provide a fairly comprehensive molecular biology background to the research on planarian development, particularly with regard to its regeneration progress.

1. Introduction

Regeneration has been a rather enigmatic topic in modern developmental biology. Partially, this is because the most spectacular regeneration occurs in organisms such as hydra or planarians that are not standard laboratory model organisms as yet. Therefore, the investigation of regeneration by using modern molecular genetics techniques has been limited [1–3], and regeneration is believed to utilize the same molecular mechanisms involved in embryonic development.

Planarians are best known for their ability to regenerate complete animals from tiny fragments of their bodies. In an experiment, T. H. Morgan discovered that a segment corresponding to 1/279th of a planarian could successfully regenerate into a new worm [4]. This size (~10,000 cells) is typically accepted as the smallest fragment that can regroup into a new planarian. The type of regeneration undertaken by planarian is called epimorphic regeneration. The strong regenerative ability of planarians is empowered by a system of pluripotent stem cells, called neoblasts [4]. Following injury, neoblasts migrate to the wound region, divide, and their progeny eventually differentiate to replace missing structures. Neoblasts also function in homeostasis by serving to replace cells lost during the normal cellular turnover. Neoblasts of planarian organisms are equivalent to stem cells; this makes these remarkable creatures an excellent model system for the study of stem cell biology.

In recent times, research studies on the planarian organisms have received extensive attention. Several of the regeneration-related genes of planarian organisms have been identified. In particular, Fernandéz-Taboada et al. reported on the characterization of the first planarian protein of the LSm protein superfamily, Smed-SmB, which is expressed in stem cells and neurons in the planarian organism *Schmidtea mediterranea* [5]. In addition, Wnt and other signaling pathways, depicting conservative signaling flows, were identified and studied in planarian organisms [6,7]. This elucidated on the mechanism of planarian regeneration. In general, researchers cloned certain planarian nucleotide sequences by sequencing and comparison with known sequences, annotated hypothetical functions, and then utilized available tools (e.g., RNAi and genetic screens, microarrays, in situ hybridizations) to verify their functions. This method is very traditional, but there are difficulties associated with the identification of a large number of genes in a single test, particularly with regard to specific genes. In addition, in research on planaria, most laboratories conduct research on the *Schmidtea mediterranea*, a type of planarian. However, studies involving other planarians have rarely been reported. The main reason for this could be that the background of other planarian organisms in molecular biology is not very clear.

* Corresponding author at: Department of Bioengineering, Zhengzhou University, Zhengzhou, Henan 450001, China. Fax: +86 371 67783235.
E-mail address: zhangst@zzu.edu.cn (S.-T. Zhang).
[1] These authors contributed equally to this work.

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.02.002

Dugesia japonica is a common planarian organism in East Asia; it also comprises the most extensive planarian species in China. It belongs to a planarian family Dugesiidae, the same family as *Schmidtea mediterranea*. Both organisms exhibit strong regeneration abilities. To better understand this organism, we employed the planarian *Dugesia japonica* as a model and utilized next-generation sequencing technology to isolate a large number of planarian genes and differential expression of genes during stages of regeneration.

In previous decades, next-generation sequencing 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 [8,9]. These new technologies also have the potential to overcome some of the aforementioned limitations. Illumina sequencing of transcriptomes for organisms with completed genomes confirmed that the relatively short reads thus produced can be effectively assembled and employed for gene discovery and comparison of gene expression profiles. Transcriptome-seq technology has been applied to human, yeast, mouse and Arabidopsis models, thus opening up the entire transcriptional landscape of gene activity and AS in a high-throughput and quantitative manner [10–18]. Further, high-quality DNA sequence with Illumina technology demonstrated the suitability of short-read sequencing for de novo assembly and annotation of genes expressed in a eukaryote without prior genomic information [19].

In this study, we generated more than 2 billion bases of high-quality DNA sequence with the Illumina technology. In a single run, 37,128 unigenes were identified and assembled by 295,218 distinct contiguous sequences. Further, we compared the gene expression profiles of planarians during different periods of regeneration by using a digital gene expression system. The assembled, annotated transcriptome sequences and gene expression profiles provide an invaluable resource for the identification of the planarian genes involved in regeneration.

2. Materials and methods

2.1. Animals

Planarian organisms utilized in this experiment were asexual specimens of *D. japonica*. The organisms were maintained at 19 °C, cultured in deionized water and fed twice each week with earthworms. Study organisms were starved for one week prior to experimentation.

2.2. RNA isolation and library preparation for transcriptome analysis

Total RNA were isolated by the SV total RNA isolation system (Promega) in accordance with the manufacturer's protocol. RNA integrity was confirmed by the 2100 Bioanalyzer (Agilent Technologies). Beads with Oligo(dT) are utlized to isolate poly(A) mRNA following the collection of total RNA from the organism. A fragmentation buffer was added to break the mRNA into short fragments. With these short fragments as templates, a random hexamer-primer was used to synthesize the first-strand of cDNA. The second-strand of cDNA was synthesized by the buffers dNTPs, RNaseH and DNA polymerase I, respectively. Short fragments are purified with the QiaQuick PCR extraction kit and resolved with an EB buffer for end reparation and by addition of poly(A). Thereafter, short fragments are connected with sequencing adapters. Following agarose gel electrophoresis, suitable fragments were selected as templates for PCR amplification. Finally, the library was sequenced by the Illumina HiSeq™ 2000.

2.3. Analysis of Illumina sequencing results

Transcriptome de novo assembly was carried out with a short reads assembling program—the SOAPdenovo [19]. The SOAPdenovo first combines reads with a certain degree of overlap to form longer fragments without N, and these are called contigs. The reads are then mapped back to contigs; with paired-end reads, it is possible to detect contigs from the same transcript as well as the distances between these contigs. Next, the SOAPdenovo connects the contigs by using N to represent unknown sequences between each two contigs, and then forms Scaffolds. Paired-end reads are used again for gap filling of scaffolds to obtain sequences with least number of Ns and these cannot be extended on either end. Such sequences are defined as Unigenes. When multiple samples from the same species are sequenced, Unigenes from each sample's assembly can be taken up for further processing of sequence splicing, and redundancy can be removed with sequence-clustering software to acquire non-redundant Unigenes. In a final step, blastx alignment (E-value<0.00001) between Unigenes and protein databases such as nr, Swiss-Prot, KEGG and COG was performed; optimal alignment results were utilized to decide the sequence direction of Unigenes. If the results of different databases conflict with each other, a priority order of nr, Swiss-Prot, KEGG and COG should be followed when deciding the sequence direction of Unigenes. When a Unigene is not aligned to any of the aforementioned databases, a software program ESTScan is introduced to predict the coding regions and to decide its sequence direction. For Unigenes with sequence directions, we provide their sequences from the 5′ to 3′ end; for Unigenes without any direction, sequences are provided from assembly software. Unigene sequences are first aligned by blastx to protein databases such as nr, Swiss-Prot, KEGG and COG, and retrieve proteins with the highest sequence similarity to the selected Unigenes as well as their protein functional annotations.

2.4. Digital gene expression library preparation and sequencing

Tag library preparation for the three *D. japonica* samples (Day 0 of regeneration, Day 1 of regeneration, and Day 3 of regeneration) was performed in parallel by using the Illumina gene expression sample preparation kit. An extract of 6 μg of total RNA was obtained and treated with Oligo(dT) magnetic bead adsorption to purify mRNA; Oligo(dT) was then used as a primer to synthesize the first- and second-strand cDNA. The 5′-ends of tags can be generated by two types of Endonuclease: NlaIII or DpnII. Often, the bead-bound cDNA is subsequently digested with restriction enzyme NlaIII, which recognizes and severs the CATG sites. Fragments other than the 3′-cDNA fragments connected to Oligo(dT) beads are washed away, and the Illumina adaptor 1 is ligated to the sticky 5′-end of the digested bead-bound cDNA fragments. The junction of the Illumina adaptor 1 and the CATG site constitutes the recognition site of MmeI, which is a type of Endonuclease with separate recognition and digestion sites; it breaks 17 bp downstream of the CATG site, producing tags with adaptor 1. After removal of 3′ fragments with magnetic beads precipitation, the Illumina adaptor 2 is ligated to the 3′-ends of tags, acquiring tags with different adaptors of both ends to form a tag library. Following 15 cycles of linear PCR amplification, 95 bp fragments are purified by 6% TBE PAGE Gel electrophoresis. After denaturation, single-chain molecules are fixed onto the Illumina Sequencing Chip (flowcell). Each molecule grows into a single-molecule cluster sequencing template through Situ amplification. Then, four types of nucleotides labeled by four colors were added and sequencing was performed by the sequencing by synthesis (SBS) method. Each tunnel generated millions of raw reads with sequencing length of 35 bp.

2.5. Analysis and mapping of digital gene expression (DGE) tags

Sequencing-received raw image data were transformed by base calling into sequence data. Prior to mapping reads to the reference database, all sequences were filtered to remove adaptor sequence, low-quality sequences (tags with unknown sequences 'N'), empty tags (sequence with only adaptor sequences but no tags), low complexity, and