Marine Genomics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Marine Genomics

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

Genomics/technical resources

Transcriptome profiling of larvae of the marine medaka *Oryzias melastigma* by Illumina RNA-seq

Bo-Mi Kim [a], Beom-Soon Choi [b], Hui-Soo Kim [a], Jae-Sung Rhee [c], Doris W.T. Au [d], Rudolf S.S. Wu [e], Ik-Young Choi [b,*], Jae-Seong Lee [a,*]

[a] Department of Biological Science, College of Science, Sungkyunkwan University, Suwon 440-746, South Korea
[b] National Instrumentation Center for Environmental Management, College of Agriculture and Life Sciences, Seoul National University, Seoul 151-921, South Korea
[c] Department of Marine Science, College of Natural Sciences, Incheon National University, Incheon 406-772, South Korea
[d] State Key Laboratory on Marine Pollution, City University of Hong Kong, Kowloon, Hong Kong, China
[e] School of Biological Sciences, University of Hong Kong, Hong Kong SAR, China

ARTICLE INFO

Article history:
Received 6 June 2015
Received in revised form 27 June 2015
Accepted 15 July 2015
Available online xxxx

Keywords:
Transcriptome
RNA-seq
Marine medaka
Oryzias melastigma
Larvae
Ecotoxicogenomics

ABSTRACT

We sequenced the whole transcriptome of the 24 h-old larval stage of the marine medaka *Oryzias melastigma* using Illumina RNA-seq. *De novo* assembly of 64,914,324 raw reads was performed using Trinity, resulting in 144,953 contigs. TransDecoder found 58,246 candidate coding contigs with homology to other species based on BLAST analysis. Functional gene annotation was performed by GO, KEGG pathway, and COG analyses. We determined an expressed gene catalog for *O. melastigma* for gene information-based environmental genomic and ecotoxicogenomic research. This information will serve as a resource for elucidating the molecular mechanisms underlying the response of *O. melastigma* to environmental stresses and chemicals.

1. Introduction

The Japanese medaka, *Oryzias latipes*, and the zebrafish, *Danio rerio*, are freshwater aquatic model animals with well-characterized, sequenced genomes (Kasahara et al., 2007; Howe et al., 2013). Although they have proven suitable model animals for diverse research areas, they are less suitable for marine environmental science studies. The marine medaka *Oryzias melastigma* has a wide distribution along the coastal region of Northeast Asia. This species has attracted attention as a model for ecotoxicology and environmental research because of its daily spawning, small size (3–4 cm), transparency of embryos, sexual dimorphism, ease of maintenance and breeding under laboratory conditions, and availability of two- or multi-generation tests with a short generation time (≈3 months).

Transcriptome and genomic information for the marine medaka is accumulating due to the development of new technologies for next-generation sequencing (NGS) and bioinformatics (Rhee et al., 2013; Lai et al., 2015). Transcriptome information of *O. melastigma* larvae will be useful for understanding the molecular responses of early stage marine medaka to environmental fluctuations. In this manuscript, we analyzed transcriptome information from the 24 h-old larval stage of the marine medaka *O. melastigma* to further develop this species as a resource for marine environmental genomic and ecotoxicogenomic studies.

2. Data description

2.1. Fish

All animal handling and experimental procedures were approved by the Animal Welfare Ethical Committee and Animal Experimental Ethics Committee of Sungkyunkwan University (Suwon, South Korea). Marine medaka *O. melastigma* individuals were kindly provided by Dr. Doris W.T. Au (City University of Hong Kong, Hong Kong SAR, China) and were maintained at the aquarium facility of the Department of Biological Science, Sungkyunkwan University (Suwon, South Korea). Briefly, fish were maintained at the aquarium facility of the Department of Biological Science, Sungkyunkwan University (Suwon, South Korea) under the controlled conditions of 26 °C, 12 L:12D in artificial seawater (TetraMarine Salt Pro, Tetra™, Cincinnati, OH, USA; 5.71 ± 0.19 mgO_2/L, 12 psu). Fish were maintained in glass aquaria (60 L capacity) and each aquarium accommodated up to 30 adult fish (both sexes). Fish were fed *Artemia salina* (<24 h after hatching) once a day until satiation.

* Corresponding authors.
E-mail addresses: choii@snu.ac.kr (I.-Y. Choi), jslee2@skku.edu (J.-S. Lee).

http://dx.doi.org/10.1016/j.margen.2015.07.013
1874-7787/

Table 1
Summary of the annotated transcriptome.

Raw data information		
Raw data	Read no.	78,932,138
	Read length (bp)	7,972,145,938
Raw sequences after QC	Read no.	65,306,322
	Read length (bp)	6,434,455,005
Microbial removed data	Read no.	64,914,324
	Read length (bp)	6,396,375,373
De novo assembly		
Trinity (ver. 2.0.6)	Contig no.	144,953
	Contig length (bp)	142,695,688
	Length distribution (bp)	224 to 21,061
	Average length (bp)	984
	Median length (bp)	457
	N50 (bp)	1938
TransDecoder	Contig no.	58,246
	Contig length (bp)	122,453,786
	Largest length (bp)	21,061
	Average length (bp)	895
	N50 (bp)	2961
	GC Ratio (%)	47.12

2.2. Illumina sequencing

Thirty 24 h-old larvae of *O. melastigma* were homogenized in three volumes of TRIZOL® reagent (Invitrogen, Paisley, Scotland) with a tissue grinder. Total RNA was extracted according to the manufacturer's instructions. DNA was digested using DNase I (Sigma, St. Louis, MO, USA). Total RNA was quantified by absorption of light at A260 using a spectrophotometer, and quality was checked by analyzing A230/260 and A260/280 ratios (QIAxpert®, Qiagen, Hilden, Germany). Synthesis of a paired-end library and sequencing with Illumina HiSeq™ 2000 (Illumina, San Diego, CA, USA) were performed at the National Instrumentation Center for Environmental Management (NICEM, Seoul National University, Seoul, South Korea). Briefly, adaptor-ligated fragments were separated by size on an agarose gel, and the desired range of cDNA fragments (200 ± 25 bp) was excised from the gel. Fragments were selected and purified and subsequently PCR-amplified to create the final cDNA library template. Image deconvolution, base calling, and quality value calculations were performed using the Illumina GA pipeline (ver. 1.6) according to the manufacturer's instructions.

2.3. Transcriptome assembly

Sequencing and assembly results are summarized in Table 1. Illumina HiSeq™ 2000 produced 78,932,138 reads representing a total of 7,972,145,938 nucleotides (Table 1). Raw reads were cleaned by filtering out adaptor-only reads, trimming adaptor sequences and empty nucleotides ('N' at the end of reads), and removing low quality sequences (reads containing more than 50% bases with a Q-value ≤ 20). *De novo* assembly of the clean reads was performed to generate non-redundant unigenes. Large contigs filtered by quality controls were constructed using the *de novo* assembler Trinity (ver. 2.0.6; Grabherr et al., 2011). Using the sequence reads that passed the quality control, we assembled 144,953 contigs with a size range of 224 to 21,061 bp. TransDecoder (http://transdecoder.sourceforge.net/) was used to identify candidate coding regions from the assembled transcripts and/or contigs; the candidate coding regions were subjected to BLAST analysis against the NCBI non-redundant (nr) protein database. Among the assembled contigs, TransDecoder found 58,246 contigs that contained candidate coding regions (Table 1). Average read size, N50 value, and GC ratio were 895 bp, 2961 bp, and 47%, respectively.

2.4. Unigene annotation and classification

BLAST analysis found that 51,148 unigenes (88%) had positive matches (e-value < 1e-06) to homologous genes of other species (Supplementary file S1). Most genes were related to those of the Japanese medaka *O. latipes* (34,295 genes; 67%) (Fig. 1). Of 51,148 unigenes, over 98% had a highest BLAST hit with a fish species unigene (Supplementary file S1). Gene Ontology (GO) and KEGG pathway analyses of contigs were performed using Blast2GO (Conesa et al., 2005). All GO analysis results were analyzed at the second level, and GO terms related to the top domains are described in Supplementary file S2. The molecular function class was the most highly represented (53%), followed by biological process (31%) and cellular component (16%) (Supplementary Fig. 1). Analysis of KEGG pathways showed that most annotated sequences were involved in metabolic pathways with 118 predicted pathways (Supplementary file S3). Functional annotation using the Clusters of Orthologous Groups (COGs) database (Tatusov et al., 2000) resulted in classification of 21,736 unigenes into 25 categories (Fig. 2, Supplementary file S4). The vast majority of unigenes were involved in 'general function prediction' (5695 genes), followed by 'DNA

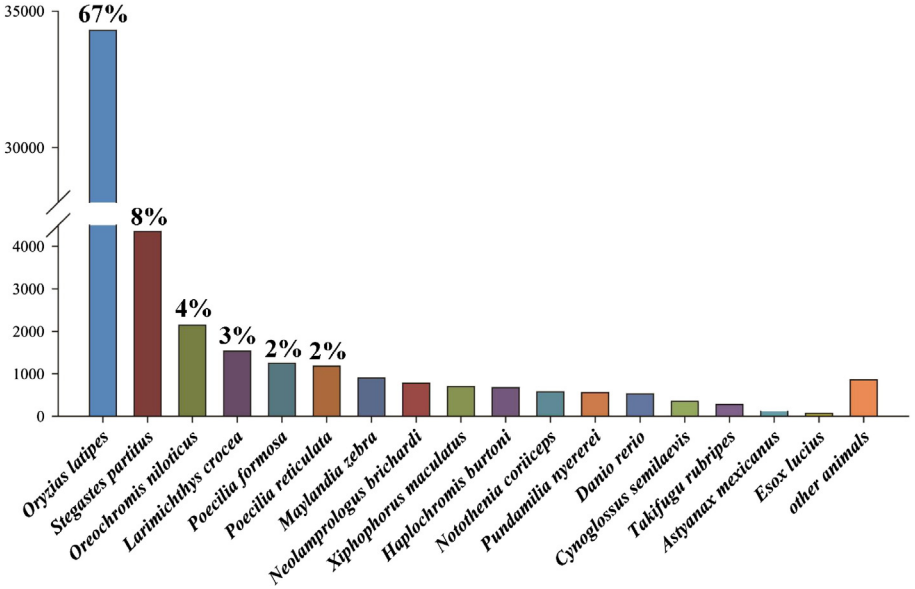

Fig. 1. Distribution of top BLAST hit species (A) and their composition (B).