Fish & Shellfish Immunology 51 (2016) 1–8

Contents lists available at ScienceDirect

Fish & Shellfish Immunology

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

Full length article

Transcriptome analysis demonstrate widespread differential expression of long noncoding RNAs involve in *Larimichthys crocea* immune response

Lihua Jiang [a], Wei Liu [a], Aiyi Zhu [a], Jianshe Zhang [a], Jiajian Zhou [b], Changwen Wu [a, *]

[a] National Engineering Research Center of Maricultural Facilities of China, College of Marine Science and Technology, Zhejiang Ocean University, Zhoushan, China
[b] Bsgenomics Co., Ltd., Shenzhen, China

ARTICLE INFO

Article history:
Received 19 October 2015
Received in revised form
19 January 2016
Accepted 3 February 2016
Available online 15 February 2016

Keywords:
RNA-seq
Long noncoding RNAs
Immune response
Larimichthys crocea

ABSTRACT

Long noncoding RNAs (lncRNAs) are a class of transcripts that longer than 200 bp and do not encode proteins. Recent genome-wide studies of vertebrate transcriptomes have annotated lncRNAs that are expressed in various tissues and development stages. The draft genome and several transcriptome sequencing data sets have been collected for the study of protein-coding genes in large yellow croaker (*Larimichthys crocea*), but little is known about the expression and functional roles of lncRNAs in this species. In order to obtain a catalog of lncRNAs for large yellow croaker, several RNA-seq datasets were integrated from various tissues including egg, muscle, liver, and spleen. A total of 48,953 high-confidence transcripts were reconstructed in 38,017 loci, recovering the most of expressed reference transcripts while thousands of novel expressed loci have been identified. The tissue expression profile revealed that most lncRNAs were specifically enriched in different tissues. A stringent set of 210 lncRNAs were identified as being specifically expressed in spleen and potentially involved in immune response. Our study first systematically identify lncRNAs in large yellow croaker, benefiting the future genomic study of this species.

1. Introduction

The development of high throughput sequencing technology during this century has enabled the discovery and functional annotation of transcripts encoded at the genome level [1,2]. Amount of novel RNA transcripts have been found by transcriptome sequencing in eukaryotes [1,3]. Those transcripts that with little protein coding potential were called noncoding RNAs (ncRNAs) [1]. The FANTOM and ENCODE projects annotated thousands of noncoding RNAs including tRNA, rRNA, microRNA, and other noncoding RNA genes [4,5].

lncRNAs have regard as a major type of novel regulatory transcripts that are around 200 nucleotides in length, lncRNAs were found to display spacial expression among different cell types and tissues, which suggesting a specific function. Comparing to small ncRNAs, by contrast, lncRNAs share most characteristics with mRNA, but show less sequence conservation across multiple species [6]. Among the reported lncRNAs, H19 and Xist were some of the earliest ones that were found using traditional gene discovery methods. Then, genomewide epigenetic marker profiles study show that approximately 1600 lncRNAs have been expressed in mouse embryonic stem cell. Thereafter, more than 11,665 lncRNAs were identified in mouse genome by FANTOM consortium. GENCODE project (http://www.gencodegenes.org) focused on the identification and annotation of lncRNAs. They show that over 27,817 lncRNAs (GENCODE v23, http://www.gencodegenes.org/) were identified in human transcriptome, derive from 15,931 lncRNA loci of human genome [5,7]. Some model organisms such as zebrafish have been used for deciphering the precise functional role of the related lncRNAs. The results suggest that most of the lncRNAs have temporal and spatial expression patterns and important roles in different tissues and during embryogenesis in zebrafish [8,9].

The large yellow croaker is a commercial important marine fish in East Asia, With the increasingly development of the croaker culture industry, there have been severe outbreaks of infectious disease caused by bacteria, viruses and *Cryptocaryon irritans*,

* Corresponding author.
E-mail address: wucw2013@126.com (C. Wu).

http://dx.doi.org/10.1016/j.fsi.2016.02.001
1050-4648/

resulting in great economic losses [10]. Analysis of the spleen transcriptome in polyriboinosinic:polyribocytidylic acid [poly(I:C)]-induced large yellow croakers (*Larimichthys crocea*) revealed a number of antivirus-related genes and signaling pathways related to the immune response [11]. The whole-genome sequence of the large yellow croaker demonstrated the existence of a well-developed innate immune system and laid the foundation for genome wide studies in this species [12]. However, few studies have focused on lncRNAs and their biological functions in large yellow croaker. Deciphering the expression pattern of lncRNAs in large yellow croaker would enable a better understanding of how gene regulation in marine fish and provide further improvements in resource protection and mariculture of this species.

In this study, a catalog of lncRNAs in large yellow croaker was generated by analysis of the transcriptome in spleen, muscle, liver, and egg tissues. The expression pattern of lncRNAs was also investigated across different tissues and during the immune response of the large yellow croaker. LncRNAs in large yellow croaker share many characteristics with those in other vertebrate species, such as lower levels of expression and higher tissue specificity compared to protein-coding transcripts. Finally, 10 lncRNAs related to the immune response in large yellow croaker were identified and further validated by qPCR experiments. This is the first report of a comprehensive identification of lncRNAs in large yellow croaker. This work will provide a further understanding of the roles of lncRNAs in the regulation of the immune response in large yellow croaker.

2. Materials and methods

2.1. Data collection and filtering

Spleen transcriptome sequencing data (GenBank SRA accession no. SRX450909, run SRR1145693) was previously derived from a polyriboinosinic:polyribocytidylic acid [poly(I:C)]-induced large yellow croaker [11]. The RNA-seq data from muscle, liver, and egg were also published in a previous large yellow croaker genome study [12]. These four data sets were subjected to an in-house Perl script for filtering sequencing reads based on phred quality score. The data set and the filtered results are shown in Supplementary Tables 1–3.

2.2. Transcriptome assembly

The RNA sequencing reads from the four tissues were aligned independently with the large yellow croaker genome (GenBank accession no. JPYK00000000.1) using TopHat (v2.0.3) software (http://tophat.cbcb.umd.edu/) [13–15]. To gain enough sequencing data for lncRNAs assembly, all alignment results were combined using SAMtools [16]. The mapped reads were then assembled into transcripts by Cufflinks (http://cufflinks.cbcb.umd.edu/) [3,17], guided by a reference annotation of the large yellow croaker genome. The assembled transcripts were filtered to remove sequences less than 200 bp in length prior to quantifying their relative abundance. The relative abundance of transcripts was quantified in each of the four different tissues by Cufflinks [3,17]. The FPKM (Fragments Per Kilobase of exons per Million fragments) value is directly proportional to the relative abundance of a transcript in a given sample.

2.3. Identification of lncRNAs

All transcripts longer than 200 bp were subjected to protein-coding potential evaluation by Coding Potential Calculator (CPC, http://cpc.cbi.pku.edu.cn/), which distinguishes coding and noncoding transcripts with high accuracy. CPC applies sequence-based features to predict the protein-coding potential of transcripts and has been widely used to discover lncRNAs. A transcript with a negative score is considered to be a noncoding transcript [1,18,19]. Further, all noncoding transcripts were aligned to the Rfam database in order to eliminate all noncoding transcripts that had been previously annotated as rRNA, miRNA, or other small noncoding RNA transcripts [20]. All remaining transcripts were identified as lncRNAs in large yellow croaker and used for further functional analysis.

2.4. Classification of lncRNAs

The resulting set of lncRNAs was subdivided into two groups: (1) lncRNAs without any overlap with protein coding gene loci were classified as large intergenic noncoding RNA (lincRNA) and (2) the remaining lncRNAs that overlap with protein coding gene region are classified as overlapping-lncRNA. Most of our analyses are focus on lincRNAs.

2.5. Protein-coding transcript annotation

To investigate lncRNAs function in large yellow croaker by associating nearby protein coding genes function annotation. We firstly aligned all protein coding transcripts to the UniProt protein database [21]. Function annotation of the best-aligned protein entry was assigned to the query protein coding transcript. Biological pathway information was retrieved from KEGG (Kyoto Encyclopedia of Genes and Genomes) database. InterProScan (version 5.0) was used to analysis functional domain of protein coding transcript. Then, a house-keeping Perl script was used to retrieve Gene Ontology (GO) annotation from InterProScan result. These protein-coding gene information was utilized for investigation of lncRNA function.

2.6. Tissue-specific transcript analysis

To determine the tissue-specific expression of lncRNAs in multiple tissues, a tissue specificity score of the expression pattern was calculated using a previously proposed method [add the reference] which results in values between 0 for housekeeping genes and 1 for tissue-restricted genes:

$$S = \frac{\sum_{i=1}^{n}\left(1 - \frac{exp_i}{exp_{max}}\right)}{n-1}$$

Where n is the total number of tissues, exp_i is the expression value in tissue i, and $expr_{max}$ represents the maximum expression level across all tissues. Transcripts with a tissue specificity score higher than 0.5 were classified as putative tissue-restricted lncRNAs. Expression values were calculated using FPKM and log2-transformed FPKM; the results of both of these methods were always in agreement. To compare tissue-specific expression patterns of long noncoding transcripts and protein coding transcripts, tissue specificity score for protein coding transcripts were calculated by the same method as done in lncRNAs [22].

2.7. Analysis of lncRNA neighbor genes

The protein-coding genes within 10 kbps away from the expressed lncRNA were defined as the neighbor genes of the lncRNAs. Genes that overlapping lncRNAs were also defined as the neighbor genes of lncRNAs. The lncRNAs neighbor genes expression levels were compared to the expression of protein-coding genes