Gene 552 (2014) 265–271

Contents lists available at ScienceDirect

Gene

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

Transcriptome analysis reveals positive selection on the divergent between topmouth culter and zebrafish

Li Ren [a,1], Xing-Jun Tan [a,1], Ya-Feng Xiong [a], Kang Xu [a], Yi Zhou [a], Huan Zhong [a], Yun Liu [a], Yun-Han Hong [a,b,*], Shao-Jun Liu [a,**]

[a] Key Laboratory of Protein Chemistry and Fish Developmental Biology of Education Ministry of China, College of Life Sciences, Hunan Normal University, Changsha 410081, China
[b] Department of Biological Sciences, National University of Singapore, Singapore 117543, Singapore

ARTICLE INFO

Article history:
Received 17 June 2014
Received in revised form 29 August 2014
Accepted 25 September 2014
Available online 28 September 2014

Keywords:
Liver
Natural selection
Repeat sequence

ABSTRACT

The topmouth culter (*Erythroculter ilishaeformis*) is a predatory cyprinid fish that distributes widely in the East Asia. Here we report the liver transcriptome in this organism as a model of predatory fish. Sequencing of 5 Gb raw reads led to 27,741 unigenes and produced 11,131 annotatable genes. A total of 7093 (63.7%) genes were found to have putative functions by gene ontology analysis. Importantly, a blast search revealed 4033 culter genes that were orthologous to the zebrafish. Extracted from 38 candidate positive selection genes, 4 genes exhibit strong positive selection based on the ratio of nonsynonymous (Ka) to synonymous substitutions (Ks). In addition, the four genes also indicated the strong positive selection by comparing them between blunt snout bream (*Megalobrama amblycephala*) and zebrafish. These genes were involved in activator of gene expression, metabolic processes and development. The transcriptome variation may be reflective of natural selection in the early life history of Cyprinidae. Based on Ks ratios, date of the separation between topmouth culter and zebrafish is approximately 64 million years ago. We conclude that natural selection acts in diversifying the genomes between topmouth culter and zebrafish.

1. Introduction

Up to date, compared with zebrafish (*Danio rerio*) as an omnivorous and small-sized fish, topmouth culter belonging to the genera of *Culter* fishes, is a predatory fish distributed in the East Asia region and is different to several economically cultivated carp, including common carp (Ji et al., 2012), crucian carp (Liao et al., 2013) and grass carp (Chen et al., 2012). It shows rapid growth rates which could be up to 10 kg. With regard to artificial breeding of fish, topmouth culter shows a great culture performance as a cultured fresh water fish (Zhang et al., 1999; Chen et al., 2003; Zhao et al., 2009).

The Cyprinidae are the largest fish family of freshwater fishes, which are an extremely diverse group of fishes that show a remarkable level of diversity affecting their morphology, ecology, behavior and genomes, as well as multiple other facets of their biology. As to tectonic movements and climatic changes, the divergence of genera *Culter* and *Danio* occurred in the early life history of Cyprinidae based on the phylogenetic analysis (Bermingham and Avise, 1986; Dominguez-Dominguez et al., 2008; He et al., 2008). Although traditionally recognized subfamily groupings of Cyprinidae were identified by monophyletic analysis (Chen et al., 1984), the phylogenetic analysis based on S7 ribosomal protein-coding gene helps us in understanding the divergence events between the genera *Culter* and *Danio* clearly (He et al., 2008). The lack of information in numerous genes referring to different types of evolution limited us to understand the change of the genetic structure that is shaped by aquatic environment (Bermingham and Avise, 1986).

Divergent time events are often estimated by information of the genome, mitochondrial DNA, and microRNA sequences. Sequencing and comparative analysis of whole-genome sequences of teleost fishes, such as fugu, tetraodon, and medaka (Aparicio et al., 2002; Jaillon et al., 2004; Kasahara et al., 2007), have enabled calculation of the rate of evolution in fish lineage. In recent years, along with the rapid development of next generation sequencing (NGS) technology, the distribution of synonymous substitution (Ks value) based on coding sequences has been used to calculate the divergent event in species (Thorne and Kishino, 2002), especially for the divergence of fish, which comprise the largest amount of vertebrates (Elmer et al., 2010; Wang et al., 2012). This calculation is an effective method for

Abbreviations: bp, base pair; COG, Clusters of Orthologous Groups of proteins; ESTs, expressed sequence tags; gata4, GATA-binding protein 4; GO, gene ontology; GY, Goldman and Wang method; Ka, nonsynonymous substitution rate; KEGG, Kyoto Encyclopedia of Genes and Genomes; Ks, synonymous substitution rate; MA, model averaging; MS, model selection; MYN, modified Yang and Nielsen method; nceh1, neutral cholesterol ester hydrolase 1; NG, Nei–Gojobori method; NGS, next generation sequencing; NR, nonredundant nucleotide database; NT, nucleotide(s); ORFs, open reading frames; phf13, PHD finger protein 13; tbl3, transducin (beta)-like 3; YN, Yang and Nielsen method.

* Correspondence to: Y. Hong, Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543, Singapore.
** Corresponding author.
E-mail addresses: dbshyh@nus.edu.sg (Y.-H. Hong), lsj@hunnu.edu.cn (S.-J. Liu).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.gene.2014.09.053
0378-1119/

determining the divergence time. The whole-genome sequence of topmouth culter is not available yet. Therefore, transcriptome analysis will help us to determine the molecular genetics of topmouth culter. Meanwhile, comparison with principal Cyprinidae fish models (zebrafish) (Volff, 2005) will help understand the mechanisms driving differential evolution.

NGS has been regarded as a fast and effective method to investigate not only species diversity but also genetic marker (Rokas and Abbot, 2009; Metzker, 2010). The combination of NGS and microsatellite markers has opened new horizons for simple sequence repeat markers (Blanca et al., 2011; Dutta et al., 2011). The study of morphological variations among seven populations of topmouth culter showed that the species could be divided into three patterns (Wang et al., 2007). However, we still lack detailed molecular marker research on this species.

In the current study, we compared the genes between topmouth culter and zebrafish: 1) to discover new molecular markers in coding sequences in topmouth culter; 2) to identify the genes under positive selection for biological function; and 3) to assess the relative age of separation between topmouth culter and zebrafish.

2. Materials and methods

2.1. Animal materials

The topmouth culter was obtained from the Engineering Center of Polyploidy Fish Breeding of the National Education Ministry located in Hunan Normal University, China. Topmouth culter was kept in a tank (23 °C) for 2 days before being killed. After anesthetizing the fish with 2-phenoxyethanol, liver tissue was carefully excised. Samples were stored in RNALater (Ambion) at −80 °C. The total RNA was extracted from the liver tissue of topmouth culter. After RNALater was removed, the samples were homogenized using a pestle and mortar. RNA isolation was performed according to the standard Trizol protocol (Invitrogen). After using agarose gel electrophoresis to detecting, the purity and concentration were used for assessing the quality of the RNA by NanoDrop (NanoDrop Technologies, Wilmington, DE, USA; http://www.nanodrop.com).

2.2. The obtaining and processing quality control of raw Illumina data

Poly(A) mRNA isolation was performed using beads with oligo(dT) after total RNA collection from eukaryotes. Fragmentation buffer was added for interrupting mRNA to short fragments. After taking these short fragments as templates, first-strand cDNA was synthesized by random hexamer primer. Second-strand cDNA was then synthesized using buffer, dNTPs, RNaseH, and DNA polymerase I. Short fragments were purified with the QiaQuick PCR extraction kit (Qiagen) and resolved with elution buffer. These fragments were run with agarose gel electrophoresis after adding sequencing adapters. PCR amplification templates of the suitable fragment were selected as PCR amplification templates. During the quality control steps, the Agilent 2100 Bioanalyzer and ABI StepOnePlus™ Real-Time PCR System were used to qualify and quantify the sample library. Finally, the library was sequenced using Illumina HiSeq™ 2000. The blunt snout bream liver transcriptome data was downloaded from the NCBI's Sequence Read Archive (SRA) (Accession ID SRX685580).

2.3. Assembly of the Illumina contigs

After raw reads were produced from sequencing machines, some read adaptors and unknown nucleotides larger than 5% were removed, including low quality reads in which the percentage of low quality bases (base quality ≤ 10) was more than 20%. Transcriptome de novo assembly was carried out with a short read assembling program (Trinity) (Grabherr et al., 2011), including three independent software modules called Inchworm, Chrysalis, and Butterfly. Trinity partitioned the sequence data into many individual de Bruijn graphs, and then processed each graph independently to extract full-length splicing isoforms and to tease apart transcripts derived from paralogous genes. Firstly, Module Inchworm was used to assemble the RNA-seq data into unique sequences of transcripts. Secondly, Module Inchworm was used to cluster and construct complete de Bruijn graphs for each cluster. Finally, Module Butterfly processed the individual graphs in parallel, tracing the paths that reads and pairs of reads took within the graph. This ultimately reported full-length transcripts for alternatively spliced isoforms, and teased apart transcripts that corresponded to paralogous genes. Finally, the resulting sequences of Trinity underwent the further process of sequence splicing, and redundancy was removed using sequence clustering software to acquire non-redundant contigs as long as possible.

2.4. Gene annotation

BLASTX alignment (e-values < $1e^{-5}$) of the above-mentioned contigs was performed using NCBI nonredundant nucleotide database (NR), NCBI-nucleotide (NT), and Swiss-Prot as the reference databases. The best alignment results were used to decide the sequence direction of contigs. The associated gene name and gene ontology (GO) term accession number were obtained from BLASTX alignment (e-values < $1e^{-6}$) with zebrafish in Ensembl BioMart (Flicek et al., 2013). WeGo software was used for analysis of the GO annotations (Ye et al., 2006).

2.5. Predicting and extracting open reading frames (ORFs)

Contigs were aligned by BLASTX (e-value < $1e^{-5}$) to protein databases of the NCBI-NR, Swiss-Prot, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Clusters of Orthologous Groups of proteins (COG) for extracting the ORFs using a Perl script. We first aligned contigs to NR, the unaligned contigs were aligned to Swiss-Prot, KEGG, and finally COG, step by step. Contigs that could not be aligned to the above-mentioned databases were scanned by the software package ESTScan (Iseli et al., 1999), which produced nucleotide sequences with direction (5′–3′) and amino sequences of the predicted coding region.

2.6. Identification of SSRs

The various types of adaptors were removed by data processing software in a sequencing system. Therefore, the transcriptome data could be used for identifying SSRs using MIcroSAtellite (http://pgrc.ipk-gatersleben.de/misa/mis-a.html). Contigs with a length less than 150 bp were removed. The remaining sequences were obtained to design primers. The microsatellite loci were identified by Perl5 script. The different microsatellites were classified as dimers, trimers, tetramers, pentamers, and hexamers. After SSRs containing expressed sequence tags (ESTs) were identified, flanking primers were designed using primer3 software with a Perl script. Finally, primers which were aligned to more than one contig or SSRs never found in the primer were removed.

2.7. Identification of Ka and Ks in orthologs

To identify putative orthologs between topmouth culter and zebrafish, the sequences from topmouth culter and zebrafish were aligned with the reciprocal BLAST (BLASTN) hit with an e-value of $1e^{-20}$ (Blanc and Wolfe, 2004). Two sequences were defined as orthologs if each of them was the best hit of the other and if the sequences were aligned over 300 bp. Using the available zebrafish protein database in Ensembl, the sequences from topmouth culter and zebrafish were aligned with the reciprocal BLAST (BLASTX) if the aligned regions were >100 amino acids and a hit with the expected e-value was < $1e^{-15}$ (Blanc and Wolfe, 2004). If no same best match was found in the two