GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Screening potential SSR markers of the anadromous fish *Coilia nasus* by de novo transcriptome analysis using Illumina sequencing

D.-A. Fang[1,2], Y.-F. Zhou[2], J.-R. Duan[2], M.-Y. Zhang[2], D.-P. Xu[2], K. Liu[2], P. Xu[1,2]* and Q. Wei[3]*

[1]Wuxi Fisheries College, Nanjing Agricultural University, Wuxi, China
[2]Freshwater Fisheries Research Center,
Chinese Academy of Fishery Sciences, Wuxi, China
[3]Key Lab of Freshwater Biodiversity Conservation Ministry of Agriculture,
Yangtze River Fisheries Research Institute,
Chinese Academy of Fishery Sciences, Wuhan, China

*These authors contributed equally to this study.
Corresponding authors: P. Xu / Q. Wei
E-mail: xup@ffrc.cn / weiqw@yf.iac.cn

Genet. Mol. Res. 14 (4): 14181-14188 (2015)
Received January 6, 2015
Accepted June 17, 2015
Published November 13, 2015
DOI http://dx.doi.org/10.4238/2015.November.13.1

ABSTRACT. RNA-Seq technology has been widely applied to transcriptomics, genomics, molecular marker development, and functional gene studies. In the genome, microsatellites are simple sequence repeats (SSR) with a high degree of polymorphism that are used as DNA markers in many molecular genetic studies. Using traditional methods such as magnetic bead enrichment, only a few microsatellite markers have been isolated. *Coilia nasus* is an anadromous, small-to-moderately sized fish species that is famous as an important fishery resource. Here, we have identified a large number of microsatellites from the fish brains by using Illumina sequencing. About 20 million Illumina reads were assembled into 148,845 unigenes. A total of 13,038 SSR motifs

were identified via analysis of 3,958,293,117 (3.96 Gb) nucleotides to produce a comprehensive transcript dataset for the *C. nasus* brain, including mono-, di-, tri-, tetra-, and penta-repeat motifs. The most abundant type of repeat motif was di-nucleotide (42.97%), followed by mono-nucleotide (38.86%), tri-nucleotide (16.21%), tetra-nucleotide (1.83%), and penta-nucleotide (0.05%) repeat units, which is similar to the results obtained in studies in other species. These data provide a base of sequence information to improve molecular-assisted markers to study *C. nasus* genetic diversity.

Key words: *Coilia nasus*; *De novo* transcriptome; SSR markers; Illumina sequencing

INTRODUCTION

Coilia nasus is a moderately sized Clupeiforme fish in the family Engraulidae (Jiang et al., 2012). It is famous for its importance as an anadromous fishery resource, nutritive value, and delicacy. *C. nasus* is also called the Japanese grenadier anchovy and is widely distributed in the Yellow Sea, East Sea, and Ariake Bay (Li et al., 2010; Jiang et al., 2012; Liu et al., 2014). As an anadromous species, it swims several kilometers up rivers, spawns in fresh water, and then the spherical eggs float down and hatch near the river mouth (Li et al., 2010; Liu et al., 2014). However, adult *C. nasus* spend most of their lives in marine environments. Every year, these fish migrate from the sea to the middle lower reaches of the Yangtze River and its affiliated lakes in China. *C. nasus* reaches sexual maturity at 2-3 years of age and spawns from April to October, breeding once every year. Mature fish migrate upriver and spawn in the lower and middle reaches of the Yangtze River and other adjacent rivers in China. *C. nasus* may also spawn in lakes adjacent to the Yangtze River, including Poyang and Taihu Lakes, where anadromous migrations have ceased and the fish permanently reside (Jiang et al., 2012; Liu et al., 2014).

However, excessive fishing and changes in aquatic ecology have almost caused extinction of the species in the middle reaches of the Yangtze River (Li et al., 2007; Jiang et al., 2012). In recent years, a number of research projects have been conducted by using artificial breeding and larval rearing techniques (Xu et al., 2011). As a result, because of overfishing and water environmental deterioration, the threat to the *C. nasus* resource has been alleviated (Yang et al., 2011).

Prior genetic studies of this anadromous species were limited to traditional genetic diversity researches (Zhong et al., 2007; Yang et al., 2011). Next-generation sequencing (NGS) technologies facilitate generation of a large number of sequences and have been used recently to isolate simple-sequence repeat (SSR) markers in studies of non-model animals, plants, and fish (Parchman et al., 2010). Transcriptome sequencing can yield a subset of genes from the genome that are functionally active in a selected tissue and species of interest (Venter et al., 2001; Martinez-Alcantara et al., 2009; Metzker, 2010; Schatz et al., 2010).

In this analysis, an annotated brain transcriptome library was constructed by *de novo* assembly of hundreds of millions of short raw DNA reads generated from NGS without prior genomic sequence information (Cox et al., 2010). Development of a large number of sequence-based genetic markers is an essential step for linkage map construction and screening potential SSR markers of the anadromous fish *C. nasus* for further genetic diversity study.

MATERIAL AND METHODS

Animals

Sixty healthy fish samples of *C. nasus* from six geographical populations were collected from major regional habitats in Yangtze River (Figure 1) during the anadromous period (from March to July). Six populations of fish were collected from the following regions: Anqing, Dangtu, Zhenjiang, Jingjiang, Nantong, and Chongming. All fishes were transferred to the laboratory in dry ice boxes. The brain of each fish was then surgically removed, immediately frozen in liquid nitrogen, and stored at -80°C until use. Brain tissues of five different individuals were selected from different regions for RNA extraction and then they were all pooled as one sample for the transcriptome library construction. All fish experimental procedures were performed in accordance with the Regulations for the Administration of Affairs Concerning Experimental Animals and were approved and authorized by the State Council of China.

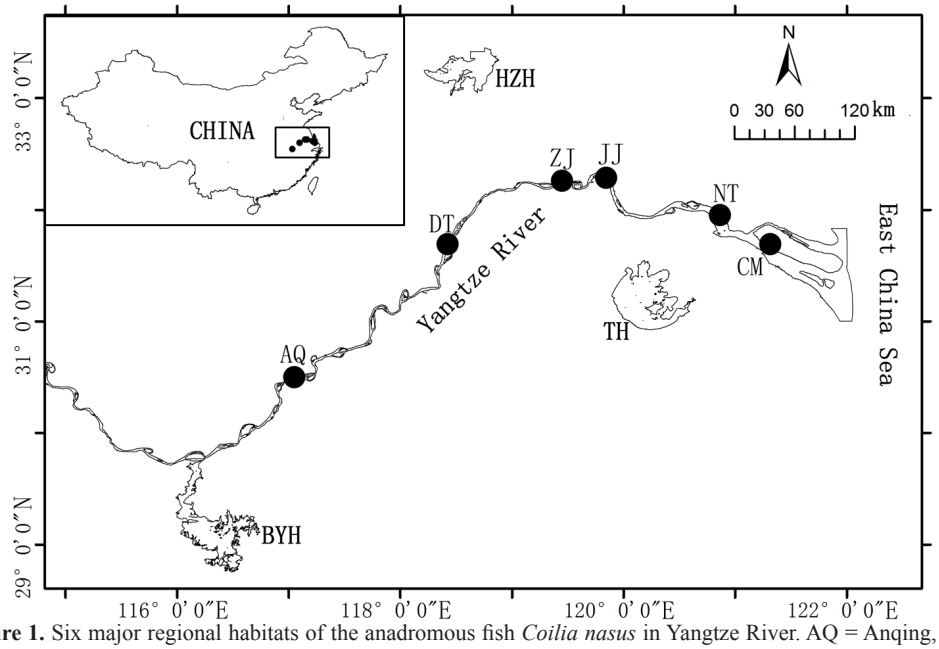

Figure 1. Six major regional habitats of the anadromous fish *Coilia nasus* in Yangtze River. AQ = Anqing, DT = Dangtu; ZJ = Zhenjiang; JJ = Jingjiang; NT = Nantong; CM = Chongming; HZH = Hongzehu; TH = Taihu; BYH = Boyanghu.

RNA extraction and cDNA library preparation

Total RNA was extracted using Trizol Lysis Reagent and then purified using an RNeasy Kit (Qiagen, Beijing, China) following manufacturer instructions. The RNA integrity, with a score of 7.8, and quantity were estimated by spectrophotometry (absorbance at 260 nm) and agarose gel electrophoresis, respectively. Equal amounts of total RNA that was purified from each brain were pooled, and mRNA was isolated using the Oligotex mRNA Kit (Invitrogen, Beijing, China) according to the manufacturer's protocol.

Isolated mRNA fragments were used as templates to synthesize the first-strand cDNA with a cDNA Library Construction Kits (Roche, Beijing). The paired-end library was synthesized using the Genomic Sample Prep Kit (Illumina, Beijing, China). Short fragments were purified using a QIAquick PCR Extraction Kit (QIAGEN, Beijing, China) and resolved with ethidium bromide buffer for end reparation and addition of poly (A). Subsequently, the short fragments were connected with sequencing adapters. After agarose gel electrophoresis, suitable fragments were selected for PCR amplification as templates. A mixed cDNA sample that represented different anadromous stages of brain was prepared and sequenced using the Illumina HiSeq™ 2000 and Solexa sequencing technology.

Brain transcriptome assembly

Transcriptome *de novo* assembly was carried out with the Short Read Assembling Program *de novo* (Dohm et al., 2008). All subsequent analyses were based on clean reads. The reads of certain lengths of overlap with no uncalled bases (N) were combined in contigs to form longer fragments (Cock et al., 2010). Contigs were then connected using N to represent the unknown sequence between each pair of contigs to form scaffolds (Simpson et al., 2009). Paired-end reads were used for gap filling of scaffolds to obtain sequences with the smallest number of N's. These sequences were defined as unigenes. In the final step, BLASTx alignments between unigenes and sequences in protein databases, including the National Center for Biotechnology Information (NCBI) non-redundant (nr) database, Swiss-Prot, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Clusters of Orthologous Groups (COG) were performed to identify the direction of unigene sequences (Tatusov et al., 2000; Kanehisa et al., 2004). If results of different databases conflicted, a priority order of alignments from the nr, Swiss-Prot, KEGG, and COG databases was followed to decide the sequence direction. When a unigene did not align with any sequences in the above databases, the software program EST Scan was used to define the sequence direction (Iseli et al., 1999). For unigenes with determined sequence directions, we identified their sequences from the 5' end to 3' end; for those with undetermined directions, we determined their sequence based on the assembly software.

Homology searches and functional unigene annotation

In the functional annotation, unigene sequences were first aligned using BLASTx to the nr, Swiss-Prot, KEGG, and COG protein databases, retrieving proteins with the highest sequence similarity with the given unigenes, along with their functional protein annotations. Homology searches were carried out by query of the NCBI nr protein database using the BLASTx algorithm (Altschul et al., 1997; Conesa et al., 2005). After nr annotation, we used the Blast2GO program to obtain Gene Ontology (GO) annotations and perform GO functional classification of all unigenes in order to understand the distribution of gene functions (Conesa et al., 2005). Using Enzyme Commission numbers, biochemical pathway information was collected by downloading relevant maps from the KEGG database (http://www.genome.jp/kegg/) (Kanehisa et al., 2004). After obtaining the KEGG pathway annotations, unigenes were aligned to the COG database to predict and classify potential functions based on known orthologous gene products. Every protein in COG is assumed to have evolved from an ancestor protein, and the entire database is built on coding proteins with complete genomes as well as systematic evolutionary relationships of bacteria, algae, and eukaryotic organisms (Tatusov et al., 2000).

SSR detection and validation

SSRs were searched for in the assembled sequences using MIcroSAtellite identification tool (MISA, Version 1.0) (http://pgrc.ipk-gatersleben.de/misa) using the following parameters: dinucleotide to hexanucleotide motifs with a minimum of five repetitions and with at least 50-nt flanking sequence. Mononucleotide repeats were ignored because distinguishing genuine repeats from polyadenylated products and single nucleotide sequencing errors is difficult.

Data deposition

De novo assembly sequence data from *C. nasus* were deposited in the NCBI Sequence Read Archive (SRA) (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi). The full dataset is also available from Di-an Fang on request (fangdyan@gmail.com).

RESULTS AND DISCUSSION

Illumina sequencing and assembly of the transcriptome

In total, 19,596,949 raw reads were produced in a typical run using the Illumina HiSeq 2000 platform. The reads averaged 202 bases in length and contained 39.6 Mb of total sequence data (Table 1).

Table 1. Summary of the transcriptome and results of the microsatellite search.

Summary of the transcriptome	Number of sequences	Mean length (bp)
Name		
Total reads	19,596,949	202
Total nucleotides (nt)	3,958,293,117	-
Total number of contigs	2,264,149	100
Total number of transcripts	220,602	938
Total number of unigenes	148,845	580
Results of microsatellite search		
Name	Number	
Total number of sequences examined	18,099	
Total size of sequences examined (bp)	38,775,423	
Total number of identified SSRs	13,038	
Number of SSR-containing sequences	8,190	
Number of sequences containing more than 1 SSR	3,019	
Number of SSRs present in compound formation	1,214	

Reads were processed to remove low-quality reads, adapters, and primer sequences. Consequently, 12,994,104 (66.31%) high-quality reads were obtained. The length distribution of raw and trimmed reads obtained are shown in Table 1. These reads had an average GC content of 79.7%. All high-quality reads were deposited in the NCBI SRA database under accession No. SRR1238084. The final resulting assembly consisted of 148,845 unigenes. These unigenes had a mean length of 580 bp (Table 1). Our data indicated that the *C. nasus* unigenes were roughly shorter than the average lengths of unigenes that were previously reported in other species such as *Eriocheir sinensis* (382 bp) (He et al., 2012), *Acropora millepora* (440 bp) (Meyer et al., 2009), *Pinus contorta* (500 bp) (Parchman et al., 2010), *Chlamydomonas* spp (665 bp) (Kim et al., 2013), and *Camellia sinensis* (733 bp) (Wu et al., 2013). These results provide sequence information for future gene cloning and transgenic engineering studies.

Annotation and comparison with related species

The 148,845 assembled unigene sequences were searched for against known sequences in five major public databases, NCBI nr, Swiss-Prot, KEGG, GO, and COG, using a BLASTx algorithm with an E-value ≤ 10^{-5} and protein identity of ≥30%. Of the 148,845 unigenes, 41,085 (27.6%) were found to have significant similarity to the nr protein database; however, because of the lack of genomic information on *C. nasus*, 95,228 unigenes did not match any known proteins, indicating that they may be novel genes or are derived from untranslated regions. We found 23,613 unigenes (15.86%) in the search against the Swiss-Prot protein database. After filtering for duplicate unigenes, we obtained 64,698 (43.47%) unigenes that matched unique genes in the nr and Swiss-Prot databases, with only 47,133 (31.67%) annotated unigenes being shared between three other databases (Table 2). The sequence database of *C. nasus* could be considered as a reference for *C. nasus* molecular biology research.

Table 2. Annotation of unigenes from the transcriptome.

Annotated databases	All sequences	≥ 300 bp	≥ 1000 bp
COG	8,192	7,442	4,756
GO	25,850	21,918	11,279
KEGG	13,091	11,109	5,832
Swiss-Prot	23,613	20,377	10,889
nr	41,085	32,488	14,334
All	53,617	40,074	15,535

Characterization and validation of SSRs

Characterization of SSRs was carried out to enable molecular marker development to study genetic diversity of the *C. nasus* in the future. All sequences were scanned using MISA (see **Table S1** for details). In total, 13,038 potential SSRs were identified in 53,617 sequences, of which 3,019 sequences contained more than one SSR, and 2,251 SSRs were in compound form. The SSR frequency in the *C. nasus* transcriptome was 24.32%, and the distribution density was 594.07 per Mb. The most abundant type of repeat motif was di-nucleotide (42.97%), followed by mono-nucleotide (38.86%), tri-nucleotide (16.21%), tetra-nucleotide (1.83%), and penta-nucleotide (0.05%) repeat units (Table 3), which is similar to the findings of studies in other species.

Table 3. SSR analysis of the annotated ungenes.

Searching item	Number	Frequency (%)
Mono-nucleotide	5,066	38.86
Di-nucleotide	5,603	42.97
Tri-nucleotide	2,114	16.21
Tetra-nucleotide	239	1.83
Penta-nucleotide	6	0.05

There were large differences in the relative abundance of special repeat motifs. As shown in **Tables S1** and **S2**, among the di-nucleotide sequences, the motif GT had the highest frequency, representing 22.02% of the sampled sequences, followed by GT (20.83%). Motifs CG and GC (<0.01% each) were comparatively rare. The most frequent tri-nucleotide was GAG (7.72%), whereas ACT, CGG, ACG, and TCG (all <0.01%) were comparatively scarce

(**Table S2**). The frequency distributions from mono- to penta-nucleotide repeats were calculated and shown in **Table S2**. The bulk of repeat sequences were centralized in the domain that was composed of low copy number, and fewer sequences were seen with increasing copy number. The size of each repeat sequence was determined by the copy number of its repeat unit. The frequencies of SSRs with different numbers of tandem repeats are listed in **Table S2**.

CONCLUSION

C. nasus is a very interesting anadromous fish in China that many scientists have substantial interest in researching. Prior to this study, no transcriptome sequencing information was available for *C. nasus* in any of the public databases. We adopted Illumina sequencing technology to analyze the *C. nasus* transcriptome and characterize the transcriptome by *de novo* sequencing without the presence of a reference genome. As a result, we obtained 148,845 unigenes of excellent sequence quality, with a mean size of 580 bp, and 53,617 unigenes (36.02%) were obtained for annotation information. In addition, we found 13,038 potential SSRs that can be used in future genetic studies. These data will be very useful for future studies of breeding, genetic diversity, and gene excavation regarding this anadromous species.

ACKNOWLEDGMENTS

Research supported by the Fundamental Research Funds from the FFRC (#2013JFBR02), the National Natural Science Foundation of China for Young Scientists (#31302169), the Key Lab of Freshwater Biodiversity Conservation (#LFBC0801), and the Jiangsu Postdoctoral Science Foundation (#1302001B).

Supplementary material

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Cock P, Fields C, Goto N, Heuer M and Rice P (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38: 1767-1771.

Conesa A, Götz S, García-Gómez JM, Terol J et al. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676.

Cox M, Peterson D and Biggs P (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11: 485.

Dohm J, Lottaz C, Borodina T and Himmelbauer H (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36: e105.

He L, Wang Q, Jin X, Wang Y, et al. (2012). Correction: Transcriptome profiling of testis during sexual maturation stages in *Eriocheir sinensis* using illumina sequencing. *PLoS One* 7: 10.1371/annotation/1379c1372ccae1379-1373f1379e-1474c-1379e1348-f1371a1372c1302ccf1322.

Iseli C, Jongeneel CV and Bucher P (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 138-148.

Jiang T, Yang J, Liu H and Shen X-Q (2012). Life history of *Coilia nasus* from the Yellow Sea inferred from otolith Sr: Ca ratios. *Environ. Biol. Fish.* 95: 503-508.

Kanehisa M, Goto S, Kawashima S, Okuno Y, et al. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32: D277-D280.

Kim S, Kim M, Jung M, Lee S, et al. (2013). *De novo* transcriptome analysis of an Arctic microalga, *Chlamydomonas* sp. *Genes Genom.* 35: 215-223.

Li WX, Song R, Wu SG, Zou H, et al. (2010). Seasonal occurrence of helminths in the anadromous fish *Coilia nasus* (Engraulidae): parasite indicators of fish migratory movements. *J. Parasitol.* 97: 192-196.

Li Y, Xie S, Li Z, Gong W, et al. (2007). Gonad development of an anadromous fish *Coilia ectenes* (Engraulidae) in lower reach of Yangtze River, China. *Fish. Sci.* 73: 1224-1230.

Liu D, Li Y, Tang W, Yang J, et al. (2014). Population structure of *Coilia nasus* in the Yangtze River revealed by insertion of short interspersed elements. *Biochem. System.Ecol.* 54: 103-112.

Martinez-Alcantara A, Ballesteros E, Feng C, Rojas M, et al. (2009). PIQA: Pipeline for Illumina G1 genome analyzer data quality assessment. *Bioinformatics* 25: 2438-2439.

Metzker M (2010). Sequencing technologies - The next generation. *Nat. Rev. Genet.* 11: 31-46.

Meyer E, Aglyamova G, Wang S, Buchanan-Carter J, et al. (2009). Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* 10: 219.

Parchman T, Geist K, Grahnen J, Benkman C, et al. (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* 11: 180.

Schatz MC, Delcher AL and Salzberg SL (2010). Assembly of large genomes using second-generation sequencing. *Genom. Res.* 20: 1165-1173.

Simpson JT, Wong K, Jackman SD, Schein JE, et al. (2009). ABySS: A parallel assembler for short read sequence data. *Genom Res.* 19: 1117-1123.

Tatusov RL, Galperin MY, Natale DA and Koonin EV (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28: 33-36.

Venter JC, Adams MD, Myers EW, Li PW, et al. (2001). The Sequence of the Human Genome. *Science* 291: 1304-1351.

Wu H, Chen D, Li J, Yu B, et al. (2013). *De novo* characterization of leaf transcriptome using 454 sequencing and development of EST-SSR markers in tea (*Camellia sinensis*). *Plant Mol. Biol. Report.* 31: 524-538.

Xu G, Xu P, Gu R, Zhang C, et al. (2011). Feeding and growth in pond *Coilia nasus* juveniles. *Chin. J. Ecol.* 9: 2014-2018.

Yang Q, Gao T and Miao Z (2011). Differentiation between populations of Japanese grenadier anchovy (*Coilia nasus*) in Northwestern Pacific based on ISSR markers: Implications for biogeography. *Biochem. System. Ecol.* 39: 286-296.

Zhong L, Guo H, Shen H, Li X, et al. (2007). Preliminary results of Sr:Ca ratios of *Coilia nasus* in otoliths by micro-PIXE. *Nucl. Instrum. Meth. B* 260: 349-352.