Journal of FISH BIOLOGY
Published for the Fisheries Society of the British Isles

Journal of Fish Biology (2010) **76**, 1190–1204
doi:10.1111/j.1095-8649.2010.02592.x, available online at www.interscience.wiley.com

Generation and analysis of 10 000 ESTs from the half-smooth tongue sole *Cynoglossus semilaevis* and identification of microsatellite and SNP markers

Z. Sha*, S. Wang†, Z. Zhuang*, Q. Wang*, Q. Wang*‡, P. Li†, H. Ding*, N. Wang*, Z. Liu† and S. Chen*§

**Key Laboratory for Sustainable Utilization of Marine Fisheries Resources, Ministry of Agriculture, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, 106 Nanjing Road, Qingdao, Shandong 266071, China, †The Fish Molecular Genetics and Biotechnology Laboratory, Department of Fisheries and Allied Aquacultures and Program of Cell and Molecular Biosciences, Aquatic Genomics Unit, 203 Swingle Hall, Auburn University, Auburn, AL 36849, U.S.A. and ‡College of Fisheries, Qingdao Agricultural University, Chengyang, Qingdao 266109, China*

(Received 5 October 2009, Accepted 13 January 2010)

Three normalized cDNA libraries were constructed, two of which were constructed from reproductive tissues ovary and testis, and the other one from pooled immune tissues including head kidney, intestine, liver and spleen. A total of 10 542 clones were sequenced generating 10 128 expressed sequence tags (ESTs). Cluster analysis indicated a total of 5808 unique sequences including 1712 contigs and 4096 singletons. A total of 4249 (73%) of the unique ESTs had significant hits to the non-redundant protein database, 2253 of which were annotated using Gene Ontology (GO) terms. A total of 311 microsatellites (with 246 having sufficient flanking sequences for primer design) and 6294 putative SNPs were identified. These genome resources provide the material basis for future microarray development, marker validation and genetic linkage and QTL analysis.

Key words: expressed sequence tag; half-smooth tongue sole; microsatellite; SNP.

INTRODUCTION

Half-smooth tongue sole *Cynoglossus semilaevis* Günther is becoming one of the most promising species for mariculture. Two aspects, in particular, have afforded much attention to this species. The first is its strong sex bimorphism and the second is the serious disease problems associated with intensive aquaculture. The female *C. semilaevis* grow two to four times faster than males and therefore all female populations are desirable for production. Not only do the females grow faster and larger than males, their meat quality is also better. It is known that sex control in this species is operated by a chromosomal control mechanism with the female being digametic (ZW) (Zhuang *et al*., 2006), but its sex control genes are not known at

§Author to whom correspondence should be addressed. Tel.: +86 532 85844606; fax: +86 532 85811514; email: chensl@ysfri.ac.cn

present. Efforts have been made to develop sex-linked molecular markers for early identification of sex. For instance, seven amplified fragment-length polymorphism (AFLP) markers have been identified to be linked with the sex chromosome (Chen *et al*., 2008). Such dominant markers, however, have very limited use for routine applications. Projects involving the development of polymorphic co-dominant markers such as microsatellites have been initiated (Liao *et al*., 2007). Recently, artificial gynogenesis has been achieved with this species (Chen *et al*., 2009). In spite of the economic significance and the strong research interest, genome resources of this species are still very limited.

Development of expressed sequence tags (EST) is one of the efficient ways to develop genomic resources because it allows not only rapid gene discovery and identification but also identification of molecular markers such as microsatellites and single nucleotide polymorphisms (SNP). EST resources have been developed in a large number of fish species (Zeng & Gong, 2002; Li *et al*., 2007; Haidle *et al*., 2008; von Schalburg *et al*., 2008*b*; Wynne *et al*., 2008; Yazawa *et al*., 2008; Bai *et al*., 2009) and applied for functional and comparative genome analysis (Liu *et al*., 2008; Peatman *et al*., 2008; Sarropoulou *et al*., 2008; Kueuktas, 2009; Liu *et al*., 2009). In order to generate genome resources for the study of sex control genes and eventually to identify co-dominant markers tightly linked with sex control genes and immune-related genes, here the generation of EST resources from *C. semilaevis* is reported. The objectives were to generate ESTs from the reproductive organs gonad (testis and ovary) and the immune-related organ of head kidney, intestine, liver and spleen. In order to include genes potentially induced after infection, the libraries were made after the fish were challenged with *Vibrio anguillarum*, the causative agents of vibrosis disease. A total of 10 128 ESTs were sequenced and deposited in the GenBank database with consecutive numbers (GH229188–GH239315). These EST resources should serve as basis for microarray development (Schalburg *et al*., 2008*a*). From this EST resource, 311 microsatellites were identified from 281 unique sequences, and 6294 putative SNPs were identified. These markers should be useful for genetic linkage analysis for the study of traits and for population genetic studies (Pujolar *et al*., 2009).

MATERIALS AND METHODS

TISSUE SOURCE AND RNA ISOLATION

Specimens of *C. semilaevis* were collected from Laizhou Mingbo Aquatic Ltd (www.mbaquatic.com). For the collection of immune function-related tissues, fish were challenged with *Vibrio anguillarum*. The bacterium was incubated to mid-logarithmic stage at 28° C in medium 2216E, collected by centrifugation and re-suspended to *c*. $4{\cdot}6 \times 10^9$ colony forming units (CFU) ml^{-1} in phosphate-buffered saline (PBS). The challenge experiment was performed as described (te Kronnie *et al*., 1999) with an injection of 0·5 ml of bacterial suspension. A total of 12 fish were killed using a lethal dose of MS-222 anaesthetic (300 ppm) 48 h after infection. Equal amounts of tissues (50 mg) from liver, head kidney, spleen and intestine of each fish were dissected and pooled for extraction of RNA. For the ovary library, ovary tissue (100 mg) from one 6 month-old fish was used. All the dissected tissues were flash-frozen in liquid nitrogen and stored at −80° C until RNA extraction. Total RNA was prepared from each sample using TRIzol reagent (Invitrogen; www.invitrogen.com) according to manufacturer's recommendations. Total RNA quality and quantity were checked by agarose gel electrophoresis containing formaldehyde and using a spectrophotometer.

CONSTRUCTION OF NORMALIZED cDNA LIBRARY, PLASMID ISOLATION AND DNA SEQUENCING

In brief, the Creator SMART cDNA Library Construction Kit (Clontech; www.clontech.com) and components from the Trimmer-Direct Kit from Evrogen (www.evrogen.com) were used for the construction of three normalized cDNA libraries, with procedures as provided by the manufacturer. A total of *c*. 7×10^5 primary recombinant clones were obtained from each library. The average insert size was 1000 base pairs (bp). The libraries were amplified and stored as stocks in 25% of glycerol in a -80° C freezer. When needed, clones were plated on Luria–Bertani (LB) agar medium containing chloramphenicol (30 μg ml^{-1}) and grown overnight at 37° C. Colonies were randomly picked, inoculated in 384 well microtitre plate (containing LB medium, 30 μg ml^{-1} chloramphenicol and 10% glycerol), incubated overnight with shaking at 37° C and stored in a -80° C freezer until further use.

Clones were transferred from 384 well plates to 96 well plates containing LB medium with 50 μg ml^{-1} chloramphenicol and grown for 16 h before plasmid isolation. Plasmid DNA was isolated from randomly selected clones using Perfectprep Plasmid 96 Vac, Direct Bind Kit (Eppendorf; www.eppendorf.com). cDNA clones were sequenced from their 5′ end using Big Dye terminator and T7 primer (5′-TAATACGACTCACTATAGGG-3′) on an ABI 3730 Genetic Analyzer (Applied Biosystems; www.appliedbiosystems.com) following the manufacturer's protocol.

EST PROCESSING, CONTIG ASSEMBLY AND ANALYSIS

The chromatogram files were exported to the PHRED program (Ewing & Green, 1998; Ewing *et al*., 1998) for base calling and removal of poor quality sequences. Vector sequences, adapter sequences and poly $(A)^+$ tails were trimmed from the sequences using Vector *NTI Advance*™ 10 (Invitrogen Corporation, 2005). ESTs with at least 100 bp after trimming were then assembled into clusters of contiguous sequences (contigs) using software CAP3 (Huang & Madan, 1999) with a cut-off value of overlapping sequence length of 50 bp and 95% sequence identity.

GENE IDENTIFICATION AND ONTOLOGY ANNOTATION

The unique sequences comprised of all the consensus sequences of the contigs and singletons were used as queries to search against the NCBI non-redundant protein database, zebrafish *Danio rerio* (Hamilton) Refseq and *Tetraodon nigroviridis* Marion de Procé Ensemble databases using BLASTX with a cut-off E-value of e-5. The XML BLASTX results were imported into the programme Blast2GO (Conesa *et al*., 2005) to conduct the gene ontology analysis. The distribution of genes in each of the main ontology categories, *i.e.* biological process, molecular function and cellular component (Ashburner *et al*., 2000), was examined and the percentages of unique sequences in each of the assigned GO terms were computed.

MARKER IDENTIFICATION

The unique sequences were searched for microsatellites using *Msatfinder* (Thurston, 2005). The minimum repeat number used for the search was eight for di-nucleotide microsatellites and five for tri, tetra and penta-nucloetide microsatellites. Microsatellite-containing ESTs possessing 50 bp flanking sequences on both sides were assumed to harbour sufficient flanking sequences for primer design (Rozen & Skaletsky, 2000). Putative SNPs were identified using the AutoSNP programme with default parameters (Barker *et al*., 2003).

RESULTS

cDNA LIBRARY CONSTRUCTION, EST SEQUENCING AND CLUSTER ANALYSIS

Three normalized cDNA libraries were constructed: first was constructed from RNA isolated from ovary, second from RNA isolated from testis and the third from

TABLE I. A summary of cDNA libraries made from gonad and mixed immune tissues of *Cynoglossus semilaevis* and ESTs generated from these libraries

Normalized cDNA library	Tissues	Number of ESTs	Number of unique sequences		Average length (bp)
			Contigs	Singletons	
CSGMMA	Head, kidney, liver, intestine and spleen	3527	590	1707	497
CSGSPA	Testis	3167	484	1944	514
CSGOVB	Ovary	3434	438	1750	508
Total		10 128			504

pooled immune tissues of head kidney, intestine, liver and spleen from 12 fish after challenge with *Vibrio anguillarum*. A total of 10 542 randomly picked cDNA clones were sequenced, resulting in 10 128 EST sequences (96% successful rate) including 3434 ESTs from the ovary library, 3167 ESTs from the testis library and 3527 ESTs from the pooled immune tissue library. The ESTs had an average length of 504 bp (Table I).

Cluster analysis of the 10 128 ESTs indicated the presence of 5808 unique sequences including 1712 contigs and 4096 singletons. The average contig contained 3·5 sequences (Table II), with 858 contigs containing two ESTs, 370 contigs containing three ESTs, 182 contigs containing four ESTs, 85 contigs containing five ESTs, 173 contigs containing six to 10 ESTs, 34 contigs containing 11–20 ESTs and 10 contigs containing >20 ESTs (Table II). More than 87% of the contigs contained five or fewer sequences. The redundancy of ESTs here is similar to that found in other flatfish EST projects (Douglas *et al*., 2007; Cerda *et al*., 2008).

TABLE II. Assembly and BLAST analysis summary of *Cynoglossus semilaevis* ESTs

Number of putative transcripts	5808
Number of contigs	1712
Number of singletons	4096
Number of significant NR hits	4249 (73·2%)
Number of significant *Danio rerio* protein hits	4082 (70·3%)
Number of significant *Tetraodon nigroviridis* protein hits	3961 (68·2%)
Average number of sequences per contig	3·5
Average contig length (bp)	645
Number of contigs with:	
2 ESTs	858
3 ESTs	370
4 ESTs	182
5 ESTs	85
6–10 ESTs	173
11–20 ESTs	34
>20 ESTs	10

PUTATIVE IDENTITIES OF THE ESTS AND GENE ONTOLOGY ANNOTATION

In order to gain insight into the identities of the sequenced ESTs, the unique sequences were searched against the non-redundant (NR) database, *D. rerio* Refseq database and *T. nigroviridis* ensemble database using BLASTX. Of the 5808 unique sequences, 4249 (73·2%) had significant BLAST hits against the NR database (cut-off E-value of e-5), while the remaining 1559 (26·8%) had no significant similarity to any sequences in the GenBank (Table II). Of the 5808 unique sequences, 4082 (70·3%) had significant hits to the *D. rerio* Refseq protein database, and 3961 (68·2%) had significant hits to the *T. nigroviridis* Ensemble protein database (cut-off E-value of e-5) (Table II). Some of the most abundant ESTs in the three normalized libraries included zona pellucida protein (ZPB), zinc finger protein, proliferating cell nuclear antigen, zona pellucida glycoprotein ZP3, growth arrest and DNA damage inducible protein (GADD45 gamma) (Table III). The vast majority of the most abundant ESTs

TABLE III. *Cynoglossus semilaevis* top 10 most redundant EST contigs (with ≥20 ESTs)

Top BLAST hit accession number	Gene identity of the top BLAST hit	E-value	Percentage identified[a]	EST accession	Tissue of high expression
AF331671	ZPB	2e-89	66	GH234446	Ovary
NP_571014	CTH1 and C3H zinc finger	4e-60	65	GH235857	Ovary
AAT78432	Proliferating cell nuclear antigen	2e-82	98	GH235857	Ovary, immune tissues and testis
XP_685613	Zona pellucida sperm-binding protein 3 precursor	2e-56	34	GH235069	Ovary
NP_001134943	Growth arrest and DNA damage inducible protein gamma	5e-73	80	GH235092	Ovary and immune tissues
BAF98579	60S ribosomal protein L3	6e-148	96	GH231863	Ovary, immune tissues and testis
XP_001341979	Similar to mucin 19	7e-06	43	GH234296	Ovary
CAG07642	Unnamed protein product	1e-27	49	GH235824	Ovary and testis
ACI66137	Histone H2A.Z	9e-65	100	GH235629	Ovary, immune tissues and testis
AAK71522	Moesin/anaplastic lymphoma kinase fusion protein	2e-63	96	GH231122	Immune tissues and testis

[a]Percentage sequence similarity within the aligned region only.

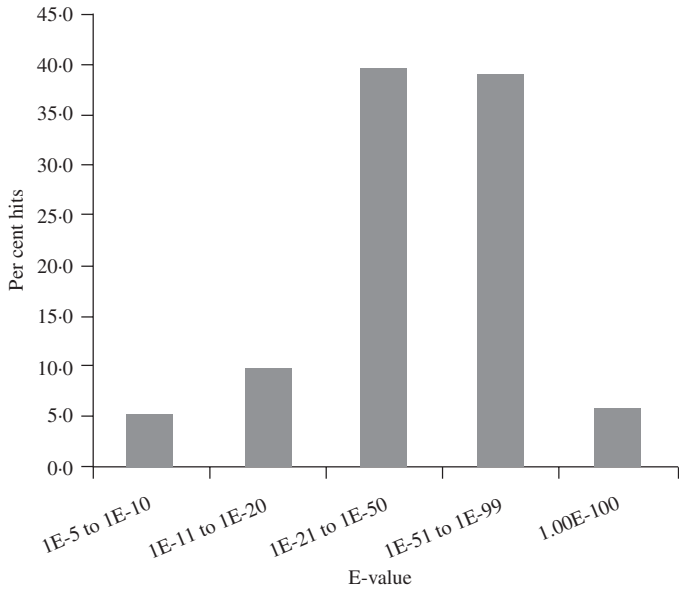

FIG. 1. *Cynoglossus semilaevis* E-value distribution of BLAST searches.

were sequenced from the ovary cDNA library, suggesting its relatively poor normalization. In comparison, the testis library and the pooled immune tissue library appeared to have a greater level of normalization.

Of the unique sequences with hits in the NR protein database, 7% had an E-value of $\leq$1e-100 (Fig. 1) and these are considered to have highly significant homology with known genes (Habermann *et al.,* 2004); 77% had E-values between 1e-20 and 1e-99 and these are considered to have significant homology to known genes (Habermann *et al*., 2004; Coblentz *et al*., 2006), and 16% had E-values between 1e-05 and 1e-19 and these are considered to have low similarities to known genes. Among the ESTs with significant homology or highly significant homology, the *C. semilaevis* ESTs had the highest number of BLASTX hits to the *T. nigroviridis* database (45·8%), followed by the *D. rerio* database (32·8%) (Fig. 2).

Gene ontology (GO) categories were assigned to 2253 unique sequences. Fig. 3 shows the distributions of gene ontology terms (third level GO terms). These annotated gene sequences were involved in 3906 biological process terms, 4177 cellular component terms and 2698 molecular function terms. The largest group of genes of the annotated sequences were involved in cellular physiological process (1225 unique sequences), followed by metabolism (942 unique sequences). Of the biological process subcategory, 161 genes were associated with functions related to immune responses. Of the 1256 unique sequences within the molecular function category, protein binding was the most highly represented GO term (Fig. 3).

MICROSATELLITE AND SNP MARKERS IDENTIFICATION

A total of 311 microsatellites were found from 281 unique sequences (Table IV). The major types of the identified microsatellites were tri-nucleotide (64%), followed by di-nucleotide (30%), tetra-nucleotide (4·8%); only three penta-nucleotide and one

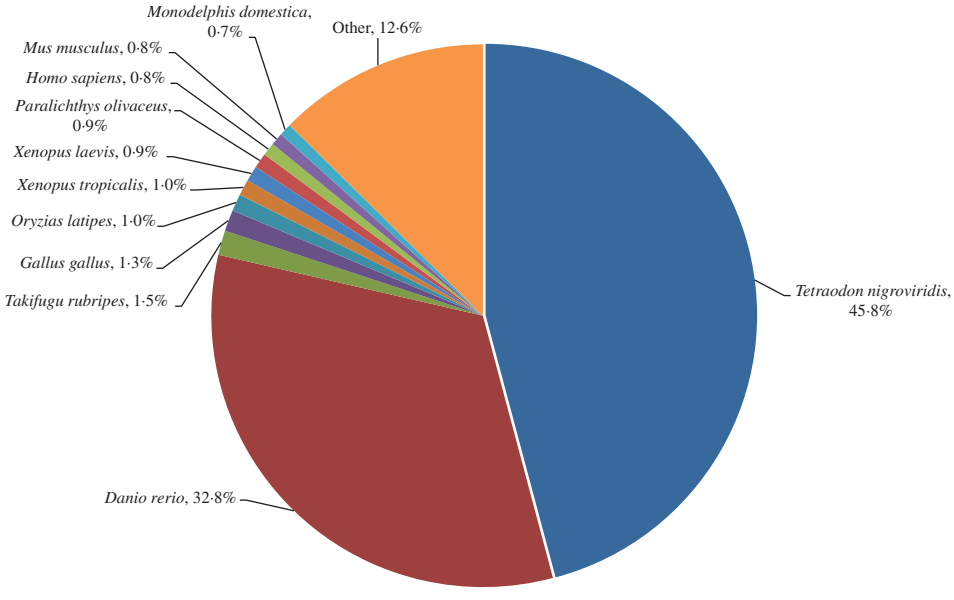

FIG. 2. Percentages of top BLASTX hits based on organism. BLASTX searches were conducted and the top hits were categorized based on organism. For instance, 45·8% of the unique sequences with BLASTX hits had top hits from *Tetraodon nigroviridis*, suggesting that the largest proportion of the *Cynoglossus semilaevis* gene sequences were most similar to those from *T. nigroviridis*, followed by those from *D. rerio*, *etc*. in the GenBank non-redundant protein database.

hexa-nucleotide repeats were found (Table IV). Of the 281 unique sequences containing microsatellites, 246 contained sufficient flanking sequences for primer design.

A total of 6294 putative SNPs were identified from the 1365 contigs. The putative SNPs included 3863 transitions, 1432 transversions and 999 indels (Table V). These SNPs represented a rate of 5·7 SNP per kilobase pairs (Table VI). As shown in Table VI, 2738 (43·5%) putative SNPs were identified from contigs with only two sequences, 1953 (31·0%) putative SNPs were identified from contigs with three sequences; 1316 (20·9%) putative SNPs were identified from contigs with four sequences; and 287 (4·6%) putative SNPs were identified from contigs with five or more sequences (Table VI). A total of 396 SNPs were identified with minor allele sequence frequency, at least twice from contigs containing four or more sequences. These 396 SNPs should have a higher validation rate for SNP genotyping (Wang *et al*., 2008). Along with the identified microsatellites, these SNPs should be useful for genetic linkage mapping and comparative analysis of the sole genome.

DISCUSSION

This is the first major EST project of *C. semilaevis*. A total of 10 128 ESTs representing 5808 unique sequences were generated. With only modest resources for EST sequencing available, three cDNA libraries were chosen to normalize in order to maximize sequencing efficiency. The normalization (Evrogen Trimmer kit) was effective in reducing the number of highly expressed cDNAs (Douglas *et al*., 2007). In this project, the EST sequence redundancy factor of 1·7 suggested that

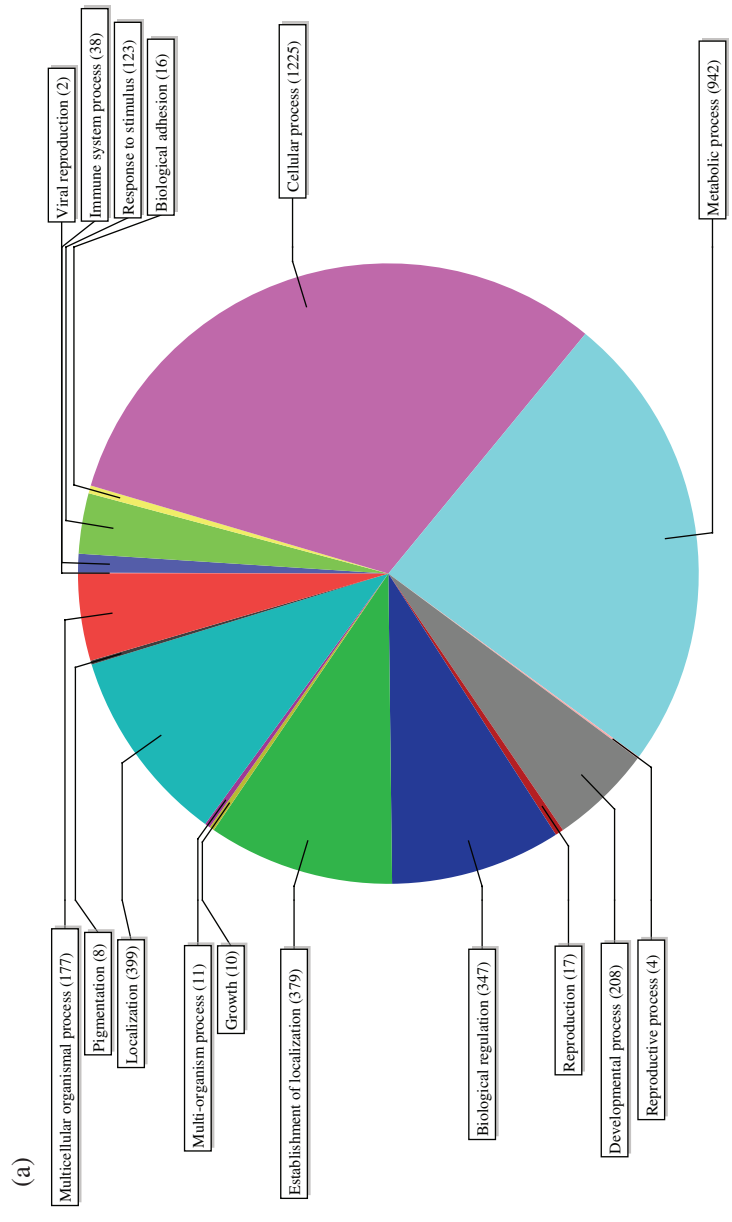

FIG. 3. Distribution of third level GO annotation terms in *Cynoglossus semilaevis*: (a) biological process, (b) molecular function and (c) cellular component.

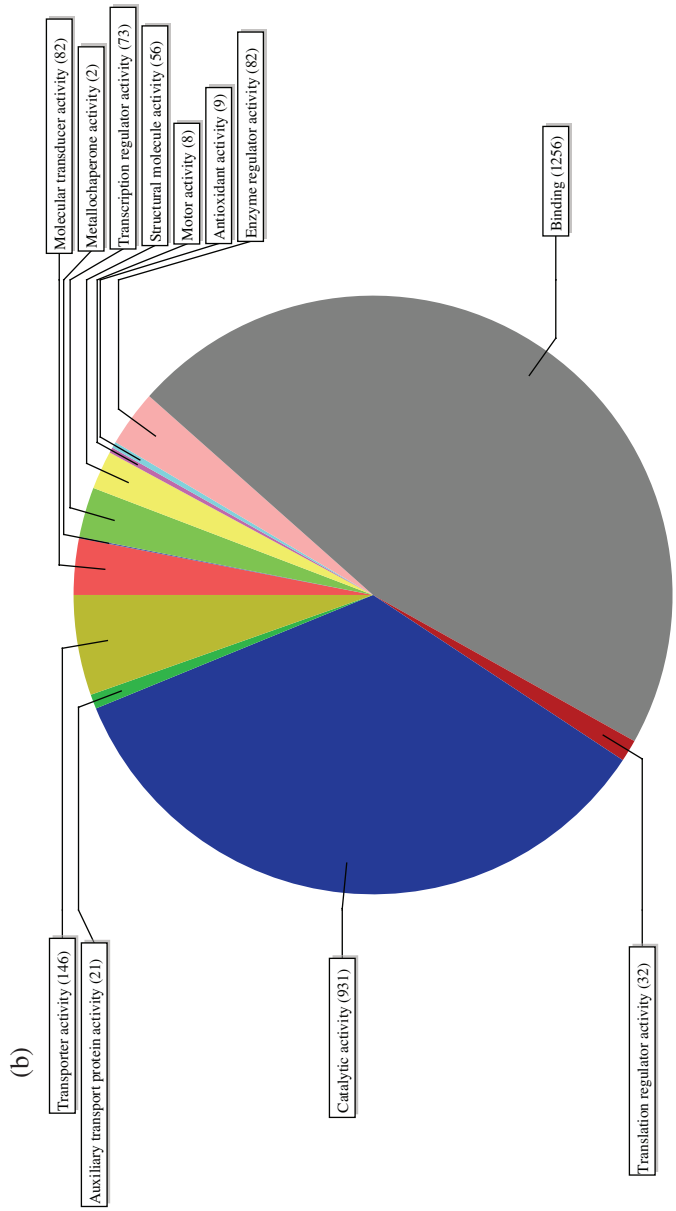

FIG. 3. Continued

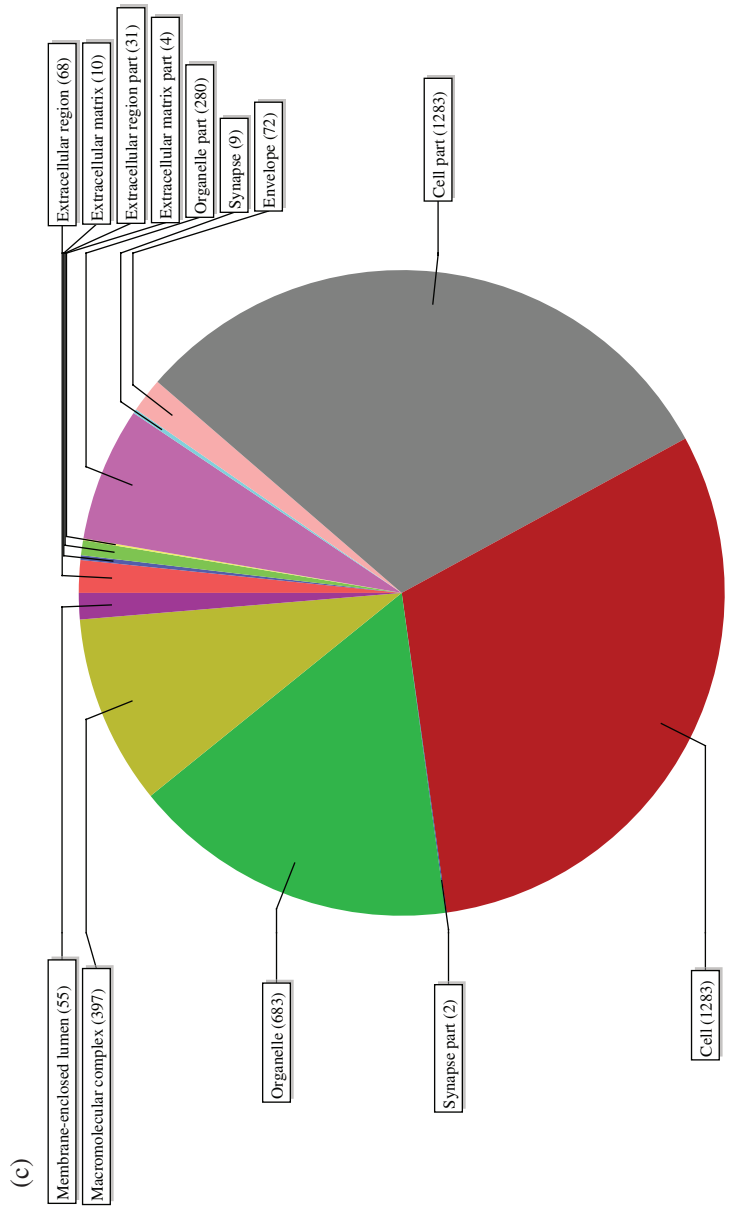

FIG. 3. Continued

TABLE IV. A summary of *Cynoglossus semilaevis* ESTs containing microsatellites

Total number of unique sequences analysed	5808
Number of sequences containing repeats	281
Number of unique sequences containing microsatellites with sufficient flanking sequences for primer design	246
Total number of microsatellites found	311
Di-nucleotide	93
Tri-nucleotide	199
Tetra-nucleotide	15
Penta-nucleotide	3
Hexa-nucleotide	1

TABLE V. Identification of putative SNPs from the ESTs of *Cynoglossus semilaevis*

Total sequences analysed	6032
Number of contigs	1712
Total SNPs detected	6294
Transitions	3863
Transversions	1432
Indels	999
SNP frequency	0·57/100 bp

TABLE VI. Putative SNP distribution in contigs with various numbers of ESTs of *Cynoglossus semilaevis*

Number of sequences in each contig	Number of contigs with SNPs	Number of total SNPs	Total consensus length (bp)	SNP frequency (per 100 bp)
2	718	2738	524 037	0·52
3	354	1953	245 943	0·79
4	179	1316	120 103	1·09
5	34	66	56 968	0·12
6–10	59	141	120 227	0·12
11–20	15	49	26 411	0·18
>20	6	31	6059	0·51
Total	1365	6294	1 099 778	0·57

the libraries were reasonably well-normalized. This result was similar to those with other flatfishes [(*i.e.* Atlantic halibut *Hippoglossus hippoglossus* (L.) and Senegalese sole *Cynoglossus senegalensis* (Kanp)] and medaka *Oryzias latipes* (Temminck & Schlegel). For instance, 12 675 EST sequences resulted in 7738 unique sequences with a redundancy factor of 1·6 from 13 cDNA libraries in *H. hippoglossus* (Douglas *et al*., 2007), and 10 185 ESTs resulted in 5208 unique sequences with a redundancy factor of 1·9 from *S. senegalensis* (Cerda *et al*., 2008). In *O. latipes*, sequencing of 7040 ESTs allowed assembly of 3641 clusters from normalized testis cDNA library with the redundancy factor of 1·9 (Lo *et al.,* 2008). The redundancy rate of the libraries mentioned in this study is significantly lower than that for non-normalized turbot *Psetta maxima* (L.) cDNA library and *D. rerio* cDNA library (Picoult-Newberg *et al*., 1999; Syvanen, 2001). Cluster analysis indicated that no

large contigs were from mitochondrial genes. The 10 largest contigs contained >20 ESTs; five of the 10 largest contigs of ESTs were from the ovary library. Some of these genes such as *ZPB*, *CTH* 1 (*C3H* zinc finger) and zona pellucida sperm-binding protein are known to be highly expressed in the ovary (He *et al*., 2003) and also found to be overrepresented in other fish ovary cDNA libraries (Lo *et al*., 2008; Wang *et al*., 2008; Liao *et al*., 2009) (Table III). In contrast, very few highly redundant ESTs were identified from the testis library. A total of 2428 unique sequences were identified from the 3167 clones from the testis library (Table I), suggesting the level of normalization of the testis library was good, similar to the rainbow trout *Oncorhynchus mykiss* (Walbaum) testis cDNA library (Pardo *et al*., 2008).

The GO terms assignment to ESTs generated from the three libraries indicated that these libraries represent a broad diversity of transcripts, indicative of the characteristics of ovary, testis and immune tissues. A number of ESTs were restricted to only a single tissue library, and as such, they may be good tissue-specific markers for *in situ* hybridization and aid in tracking the appearance of different tissues during development (Lo *et al*., 2003). Some ESTs were identified to be specifically expressed in the ovary library such as zona pellucida protein, vitelline envelope protein and aquaporin. Similarly, some specific ESTs were identified only from the testis library such as periostin.

A large number of immune-related genes were discovered from the pooled immune tissue library such as the complement components, interferon, MHC I and II components, toll-link-receptors (TLR), lectins, defence proteins, cytokines, chemokines, microglobin, glutathione-*S*-transferase, heat-shock proteins and tumor-necrosis factor (TNF) receptor. Further characterization of these genes should provide information to enhance the understanding of the immune system of flatfish and provide molecular tools for further study of disease resistance.

Polymorphic marker development from *C. semilaevis* is necessary for the construction of a high-density genetic map. EST studies can also provide resources for identification of microsatellite and SNP markers. This work allowed the identification of 311 microsatellites from 281 unique sequences, of which 246 had sufficient flanking sequences for primer design. Interestingly, the largest number of microsatellites were tri-nucleotide (Table IV), and the results reveal consistent differences between coding and non-coding regions, in terms of both the quantity of repetitive DNA and the types present. In non-coding regions, all types of microsatellite (mono, di, tri, tetra, penta and hexa-nucleotide repeats) were found. In coding regions, however, a greater number of tri and hexa-nucleotide repeats were found, consistent with the stronger constraints within coding regions (Conner & Hughes, 2003). With tri and hexa-nucleotide repeats, any addition or reduction in repeat numbers would not cause frameshift mutations, while addition or reduction of a repeat unit with di-nucleotide repeats, for instance, would lead to frameshift mutations. As the microsatellites were identified from ESTs, it is reasonable to expect that tri-nucleotide repeat types can be overrepresented. Whether this expectation holds true in general within this species needs additional genome sequencing surveys such as bacterial artifical chromosome (BAC) end sequencing (Xu *et al*., 2006).

A large number (6294) of putative SNPs were identified from the clustered sequences of ESTs. These putative SNPs are potentially useful for genetic linkage mapping and for the analysis of quantitative traits of the tongue sole. Validation and polymorphic analysis, however, must be performed before these putative SNPs can

be used. That is because a large proportion of these putative SNPs were identified from contigs with just two or three sequences. In the absence of the quality scores, it was not possible for us to differentiate the true SNPs from sequence errors. In spite of this setback, the putative SNPs identified from this study represent the first large set of SNPs from *C. semilaevis*. These SNPs, with additional ongoing validation and polymorphism testing in specific resource families, should be useful markers for genetic analysis in *C. semilaevis*. In conclusion, this is the first report on a major transcriptional analysis in *C. semilaevis*. In this project, 10 128 ESTs were generated representing 5808 unique sequences. The transcript profile comparison among the three libraries allowed the identification of putative genes generally or specifically related with reproduction and immunity. These ESTs should serve as valuable resources for further analysis of genes involved in reproductive or immune function. The ESTs can serve as material basis for the development of microarrays useful for functional genomics analysis. The microsatellites and putative SNPs identified from the sole ESTs should be useful for genetic linkage mapping and for the analysis of quantitative traits.

This project was supported by a grant from Yellow Sea Fisheries Research Institute, CAFS (2008-CHB-04), Qingdao Scientific foundation (07-2-3-5-jch), National Major Basic Research Program (Grant No. 2010CB126303) Taishan Scholar Project of Shandong Province and Shandong Genetic Improvement Key Project for Agricultural Organism. ESTs were sequenced at the Bejing Genomics Institute (BGI).

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25,** 25–29.

Bai, Z., Yin, Y., Hu, S., Wang, G., Zhang, X. & Li, J. (2009). Identification of genes involved in immune response, microsatellite, and SNP markers from expressed sequence tags generated from hemocytes of freshwater pearl mussel (*Hyriopsis cumingii*). *Marine Biotechnology* **11,** 520–530.

Barker, G., Batley, J., O'Sullivan, H., Edwards, K. J. & Edwards, D. (2003). Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* **19,** 421–422.

Cerda, J., Mercade, J., Lozano, J. J., Manchado, M., Tingaud-Sequeira, A., Astola, A., Infante, C., Halm, S., Vinas, J., Castellana, B., Asensio, E., Canavate, P., Martinez-Rodriguez, G., Piferrer, F., Planas, J. V., Prat, F., Yufera, M., Durany, O., Subirada, F., Rosell, E. & Maes, T. (2008). Genomic resources for a commercial flatfish, the Senegalese sole (*Solea senegalensis*): EST sequencing, oligo microarray design, and development of the Soleamold bioinformatic platform. *BMC Genomics* **9,** 508.

Chen, S. L., Deng, S. P., Ma, H. Y., Tian, Y. S., Xu, J. Y., Yang, J. F., Wang, Q. Y., Ji, X. S., Shao, C. W., Wang, X. L., Wu, P. F., Deng, H. & Zhai, J. M. (2008). Molecular marker-assisted sex control in half-smooth tongue sole (*Cynoglossus semilaevis*). *Aquaculture* **283,** 7–12.

Chen, S. L., Tian, Y. S., Yang, J. F., Shao, C. W., Ji, X. S., Zhai, J. M., Liao, X. L., Zhuang, Z. M., Su, P. Z., Xu, J. Y., Sha, Z. X., Wu, P. F. & Wang, N. (2009). Artificial gynogenesis and sex determination in half-Smooth tongue sole (*Cynoglossus semilaevis*). *Marine Biotechnology* **11,** 243–251.

Coblentz, F. E., Towle, D. W. & Shafer, T. H. (2006). Expressed sequence tags from normalized cDNA libraries prepared from gill and hypodermal tissues of the blue crab, *Callinectes sapidus*. *Comparative Biochemistry and Physiology D* **1,** 200–208.

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M. & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21,** 3674–3676.

Conner, S. J. & Hughes, D. C. (2003). Analysis of fish ZP1/ZPB homologous genes – evidence for both genome duplication and species-specific amplification models of evolution. *Reproduction* **126,** 347–352.

Douglas, S. E., Knickle, L. C., Kimball, J. & Reith, M. E. (2007). Comprehensive EST analysis of Atlantic halibut (*Hippoglossus hippoglossus*), a commercially relevant aquaculture species. *BMC Genomics* **8,** 144.

Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8,** 186–194.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8,** 175–185.

Habermann, B., Bebin, A. G., Herklotz, S., Volkmer, M., Eckelt, K., Pehlke, K., Epperlein, H. H., Schackert, H. K., Wiebe, G. & Tanaka, E. M. (2004). An *Ambystoma mexicanum* EST sequencing project: analysis of 17,352 expressed sequence tags from embryonic and regenerating blastema cDNA libraries. *Genome Biology* **5,** R67.

Haidle, L., Janssen, J. E., Gharbi, K., Moghadam, H. K., Ferguson, M. M. & Danzmann, R. G. (2008). Determination of quantitative trait loci (QTL) for early maturation in rainbow trout (*Oncorhynchus mykiss*). *Marine Biotechnology* **10,** 579–592.

He, C., Chen, L., Simmons, M., Li, P., Kim, S. & Liu, Z. J. (2003). Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. *Animal Genetics* **34,** 445–448.

Huang, X. & Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Research* **9,** 868–877.

Kucuktas, H., Wang, S., Li, P., He, C., Xu, P., Sha, Z., Liu, H., Jiang, Y., Baoprasertkul, P., Somridhivej, B., Wang, Y., Abernathy, J., Guo, X., Liu, L., Muir, W. & Liu, Z. (2009). Construction of genetic linkage maps and comparative genome analysis of catfish using gene-associated markers. *Genetics* **181,** 1649–1660.

Li, P., Peatman, E., Wang, S., Feng, J., He, C., Baoprasertkul, P., Xu, P., Kucuktas, H., Nandi, S., Somridhivej, B., Serapion, J., Simmons, M., Turan, C., Liu, L., Muir, W., Dunham, R., Brady, Y., Grizzle, J. & Liu, Z. J. (2007). Towards the catfish transcriptome: generation and analysis of 31,215 catfish ESTs. *BMC Genomics* **8,** 177.

Liao, X. L., Shao, C. W., Tian, Y. S. & Chen, S. L. (2007). Polymorphic dinucleotide microsatellites in tongue sole (*Cynoglossus semilaevis*). *Molecular Ecology Notes* **7,** 1147–1149.

Liao, X., Ma, H. Y., Xu, G. B., Shao, C. W., Tian, Y. S., Ji, X. S., Yang, J. F. & Chen, S. L. (2009). Construction of a genetic linkage map and mapping of a female-specific DNA marker in half-smooth tongue sole (*Cynoglossus semilaevis*). *Marine Biotechnology* **11,** 699–709.

Liu, Z. J., Li, R. W. & Waldbieser, G. (2008). Utilization of microarray technology for functional genomics in ictalurid catfish. *Journal of Fish Biology* **72,** 2377–2390.

Liu, H., Jiang, Y., Wang, S., Ninwichian, P., Somridhivej, B., Xu, P., Abernathy, J., Kucuktas, H. & Liu, Z. (2009). Comparative analysis of catfish BAC end sequences with the zebrafish genome. *BMC Genomics* **10,** 592.

Lo, J., Lee, S., Xu, M., Liu, F., Ruan, H., Eun, A., He, Y., Ma, W., Wang, W., Wen, Z. & Peng, J. (2003). 15,000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis. *Genome Research* **13,** 455–466.

Lo, L., Zhang, Z., Hong, N., Peng, J. & Hong, Y. (2008). 3640 unique EST clusters from the medaka testis and their potential use for identifying conserved testicular gene expression in fish and mammals. *PLoS ONE* **3,** e3915.

Pardo, B. G., Fernandez, C., Millan, A., Bouza, C., Vazquez-Lopez, A., Vera, M., Alvarez-Dios, J. A., Calaza, M., Gomez-Tato, A., Vazquez, M., Cabaleiro, S., Magarinos, B., Lemos, M. L., Leiro, J. M. & Martinez, P. (2008). Expressed sequence tags (ESTs) from immune tissues of turbot (*Scophthalmus maximus*) challenged with pathogens. *BMC Veterinary Research* **4,** 37.

Peatman, E., Terhune, J., Baoprasertkul, P., Xu, P., Nandi, S., Wang, S., Somridhivej, B., Kucuktas, H., Li, P., Dunham, R. & Liu, Z. J. (2008). Microarray analysis of gene expression in the blue catfish liver reveals early activation of the MHC class I pathway after infection with *Edwardsiella ictaluri*. *Molecular Immunology* **45,** 553–566.

Picoult-Newberg, L., Ideker, T. E., Pohl, M. G., Taylor, S. L., Donaldson, M. A., Nickerson, D. A. & Boyce-Jacino, M. (1999). Mining SNPs from EST databases. *Genome Research* **9,** 167–174.

Pujolar, J. M., Leo, G. A. D., Ciccotti, E. & Zane, L. (2009). Genetic composition of Atlantic and Mediterranean recruits of European eel *Anguilla anguilla* based on EST-linked microsatellite loci. *Journal of Fish Biology* **74,** 2034–2046.

Rozen, S. & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Molecular Biology* **132,** 365–386.

Sarropoulou, E., Nousdili, D., Magoulas, A. & Kotoulas, G. (2008). Linking the genomes of nonmodel teleosts through comparative genomics. *Marine Biotechnology* **10,** 227–233.

Syvanen, A. C. (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Review Genetics* **2,** 930–942.

te Kronnie, G., Stroband, H., Schipper, H. & Samallo, J. (1999). Zebrafish CTH1, a C3H zinc finger protein, is expressed in ovarian oocytes and embryos. *Development Genes and Evolution* **209,** 443–446.

von Schalburg, K. R., Cooper, G. A., Leong, J., Robb, A., Lieph, R., Rise, M. L., Davidson, W. S. & Koop, B. F. (2008*a*). Expansion of the genomics research on Atlantic salmon *Salmo salar* L. project (GRASP) microarray tools. *Journal of Fish Biology* **72,** 2051–2070.

von Schalburg, K. R., Leong, J., Cooper, G. A., Robb, A., Beetz-Sargent, M. R., Lieph, R., Holt, R. A., Moore, R., Ewart, K. V., Driedzic, W. R., Hallers, B. F., Zhu, B., de Jong, P. J., Davidson, W. S. & Koop, B. F. (2008*b*). Rainbow smelt (*Osmerus mordax*) genomic library and EST resources. *Marine Biotechnology* **10,** 487–491.

Wang, S., Sha, Z., Sonstegard, T. S., Liu, H., Xu, P., Somridhivej, B., Peatman, E., Kucuktas, H. & Liu, Z. (2008). Quality assessment parameters for EST-derived SNPs from catfish. *BMC Genomics* **9,** 450.

Wynne, J. W., O'Sullivan, M. G., Cook, M. T., Stone, G., Nowak, B. F., Lovell, D. R. & Elliott, N. G. (2008). Transcriptome analyses of amoebic gill disease-affected Atlantic salmon (*Salmo salar*) tissues reveal localized host gene suppression. *Marine Biotechnology* **10,** 388–403.

Xu, P., Wang, S., Liu, L., Peatman, E., Somridhivej, B., Thimmapuram, J., Gong, G. & Liu, Z. J. (2006). Channel catfish BAC end sequences for marker development and assessment of syntenic conservation with other fish species. *Animal Genetics* **37,** 321–326.

Yazawa, R., Yasuike, M., Leong, J., von Schalburg, K. R., Cooper, G. A., Beetz-Sargent, M., Robb, A., Davidson, W. S., Jones, S. R. & Koop, B. F. (2008). EST and mitochondrial DNA sequences support a distinct Pacific form of salmon louse, *Lepeophtheirus salmonis*. *Marine Biotechnology* **10,** 741–749.

Zeng, S. & Gong, Z. (2002). Expressed sequence tag analysis of expression profiles of zebrafish testis and ovary. *Gene* **294,** 45–53.

Zhuang, Z. M., Wu, D., Zhang, S. C., Pang, Q. X., Wang, C. L. & Wan, R. J. (2006). G-banding patterns of the chromosomes of tonguefish *Cynoglossus semilaevis Günther*, 1873. *Journal of Applied Ichthyology* **22,** 437–440.

Electronic Reference

Thurston, M. I. & Field, D. (2005). Msatfinder: detection and characterisation of microsatellites. Available at http://www.genomes.ceh.ac.uk/msatfinder