DNA Research **18**, 513–522, (2011) doi:10.1093/dnares/dsr036
Advance Access Publication on 15 November 2011

Analysis of the Asian Seabass Transcriptome Based on Expressed Sequence Tags

Jun Hong Xia, Xiao Ping He, Zhi Yi Bai, Grace Lin, and Gen Hua Yue*

Molecular Population Genetics Group, Temasek Life Sciences Laboratory, National University of Singapore, 1 Research Link, Singapore 117604, Republic of Singapore

*To whom correspondence should be addressed. Tel. +65-68727405. Fax. +65-68727007.
Email: genhua@tll.org.sg

Edited by Mikita Suyama
(Received 19 April 2011; accepted 24 September 2011)

Abstract

Analysis of transcriptomes is of great importance in genomic studies. Asian seabass is an important fish species. A number of genomic tools in it were developed, while large expressed sequence tag (EST) data are lacking. We sequenced ESTs from nine normalized cDNA libraries and obtained 11 431 high-quality ESTs. We retrieved 8524 ESTs from dbEST database and analyzed all 19 975 ESTs using bioinformatics tools. After clustering, we obtained 8837 unique sequences (2838 contigs and 5999 singletons). The average contig length was 574 bp. Annotation of these unique sequences revealed that 48.9% of them showed significant homology to RNA sequences in GenBank. Functional classification of the unique ESTs identified a broad range of genes involved in different functions. We identified 6114 putative single-nucleotide polymorphisms and 634 microsatellites in ESTs. We discovered different temporal and spatial expression patterns of some immune-related genes in the Asian seabass after challenging with a pathogen *Vibrio harveyi*. The unique EST sequences are being used in developing a cDNA microarray to examine global gene expression and will also facilitate future whole-genome sequence assembly and annotation of Asian seabass and comparative genomics.

Key words: Asian seabass; EST; function; expression

1. Introduction

Transcriptome reflects the subset of genes from the genome that are functionally active in a selected tissue and species of interest.[1,2] Since transcriptome analysis is of growing importance, research in the field has been a key area of biological inquiry for past decades and has made great progress.[2] An expressed sequence tag (EST) is a short subsequence of a transcribed cDNA sequence.[3] ESTs can be used to identify gene transcripts and are instrumental in gene discovery.[4] Cloning and sequencing of ESTs are also an effective approach for the recovery of full-length cDNA, discovery of novel genes and development of molecular markers.[5–8] Large-scale EST data represent a snapshot of the transcriptome of an organism. Since ESTs are important genomic resources, their numbers in public databases are rapidly increasing.

Fish involving >30 000 species[9] offer unique systems for studies of transcriptomics, genomics and evolutionary biology. So far, some assembled genome sequences of fish species (e.g. fugu, medaka, stickleback, tetraodon, zebrafish and tilapia) are available in the public databases, and more genome sequencing projects are ongoing. Only in a few fish species (e.g. zebrafish,[10] salmon,[11,12] trout,[13] catfish,[14] gilthead seabream,[15] European seabass,[16] Atlantic halibut,[17] cod[18]), transcriptomes were characterized based on large EST data. A few comparative transcriptomic studies were carried out among fish species (e.g. catfish[19] and carp[20]).

These studies have dramatically expanded the biological data available across these species and provided a starting point for detailed studies on functional genomics.

The Asian seabass (*Lates calcarifer* Bloch, 1790), also known as Barramundi, is widely distributed in the Indo-West Pacific region.[21] This fish is of large commercial importance and extensively used for aquaculture.[21] Currently, a number of genomic tools have been developed in Asian seabass. These include a large number of microsatellites,[21–23] a genetic linkage map,[24,25] a bacterial artificial chromosome (BAC) library,[26] a physical map based on BAC fingerprinting,[27] some ESTs related to immune responses,[28] microRNAs[29] and quantitative trait loci for growth[30] and an adaptive trait.[31] So far, the EST resource of the Asian seabass presented in the public NCBI dbEST database has grown to over 8000 ESTs (5637 ESTs from the brain[32] and 2887 immune-related ESTs from the spleen[28]). However, to date, the large-scale EST data for the Asian seabass are lacking.

In order to expand our knowledge of the transcriptome, facilitate future whole-genome sequence assembly and annotation in the Asian seabass and enable a comparison of transcriptomes among fish species, we sequenced ESTs from nine normalized cDNA libraries of the Asian seabass and obtained 11 431 high-quality ESTs. We analyzed 19 975 ESTs including 8524 ESTs from dbEST database using bioinformatic tools and examined the expression patterns of immune-related genes after challenge with a pathogen *Vibrio harveyi*.

2. Materials and methods

2.1. Fish and sampling

To construct full-length cDNA libraries, five individuals at the age of 90 days post-hatch (dph) with an average body weight of 48.7 ± 21.6 g were used. The fish were kept under the standard hatchery conditions.[33] Tissue samples of the gill, brain, muscle, liver, heart, spleen, eye, intestine and kidney of the five individuals were collected, immersed in Trizol (Invitrogen, Carlsbad, USA) and stored at -80°C until use.

To explore the temporal and spatial expression patterns of the immune-related genes in Asian seabass challenged with a bacterium *V. harveyi* which is a significant pathogen of marine vertebrates and invertebrates,[34] 30 individuals at the age of 90 dph with an average body weigh of 35.0 ± 2.68 g were transferred to two tanks holding 200 l of seawater at the fish facility of the Temasek Life Sciences Laboratory. For 15 fishes in the test tank, each fish was injected intraperitoneally with 0.1 ml of *V. harveyi* ($\sim e^{10}$ copy/ml) in phosphate-buffered saline as described in Xia *et al.*[29] In the control tank, each of the 15 fishes received an intraperitoneal injection of 0.1 ml of phosphate-buffered saline. Three fishes from each tank were sacrificed at 1, 3, 6, 12 and 24 h post-injection (hpi). The spleen and liver were taken for each fish from each tank and kept in Trizol reagent at -80°C until use.

2.2. Construction of normalized cDNA libraries and EST sequencing

Eight normalized cDNA libraries for the gill, brain, muscle, liver, heart, spleen, kidney and the mixture of tissues (gill, brain, muscle, liver, heart, spleen, eye, intestine and kidney) were constructed as described.[35] Briefly, total RNA from each tissue was isolated using the Trizol kit (Invitrogen). Purification of mRNA from total RNA was carried out using Oligotex mRNA Midi Kit (Qiagen, Valencia, USA). The mRNA from the same tissue of five fishes was pooled with equal quantity for the construction of tissue-specific cDNA libraries; and the resulting mRNA mixtures from different tissues of five fishes were also pooled with equal quantity for the construction of a full-length cDNA library of all seven tissues. cDNAs were synthesized, normalized and cloned into a Gateway™ pCMV•SPORT6 *Not*I/*Sal*I Cut vector (Invitrogen) and transformed into *Escherichia coli* strain XL-1 (Stratagene, CA, USA) as described.[35] In addition, a normalized cDNA library of all seven tissues (gill, brain, muscle, liver, heart, spleen and kidney) enriched microsatellites constructed previously[35] was also used to sequence ESTs (see details about the cDNA libraries in Supplementary Table S1).

For each library, 5000 randomly picked clones were stored in the LB medium with 25% glycerol in -80°C for later sequencing. Some clones from each of the nine libraries were sequenced from the 5′ end by using the M13/pUCR universal primer (5′-AGC GGA TAA CAA TTT CAC ACA GG-3′) and BigDye chemicals on an ABI 3730xl Genetic Analyzer (Applied Biosystems, Foster city, CA, USA). The EST sequences were deposited in the dbEST database of GenBank: JG732280– JG743071.

2.3. EST preprocessing and cluster analysis

EST sequences from the nine cDNA libraries constructed by our laboratory and sequences of ESTs[28,32] of Asian seabass from GenBank were included in the following analysis. Base calling from chromatogram traces and trimming of vector and adaptor sequences and low-quality regions from EST sequences were performed by using commercial software Sequencher 4.9 (Gene Codes, Ann Arbor, USA).

Trimming of the sequences were first performed from two ends until the last 25 bases containing less than five ambiguities, and then that of the vectors and adaptors were performed under the parameters that minimum overlap and approximate match percentage to consider as contamination were set to 8 and 99%, respectively. Other parameters were used following the original software defaults. High-quality ESTs ($\geq$ 50 bp) were clustered by using software NGen (DNASTAR, Madison, USA). The clustering parameter settings were automatically determined based on the Sanger sequencing read technology and De novo Transcriptome Assembly and used a conservative philosophy that aligned portions of ESTs must share 95% sequence identity in at most 20 bp of overhanging sequence.

2.4. *Functional annotation of EST sequences*

A total of 21 databases were retrieved from Ensembl (http://www.ensembl.org/info/data/ftp/index.html) and NCBI databases (http://mirrors.vbi.vt.edu/mirrors/ftp.ncbi.nih.gov/blast/db/), respectively, and were used to perform BLAST analyses for EST annotation (Supplementary Tables S2–S5). Singletons and consensus sequences of each contig referred to as unique sequences were compared against the 11 protein databases using BLASTx and the 10 cDNA databases using the BLASTn algorithm with an *E*-value threshold of E^{-5} and a minimum alignment length of 22 bp. Batch blast of the unique sequences was performed using the local BLAST tools that available in ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/.

2.5. *Gene ontology analysis by Blast2GO, GO Slimmer and KAAS*

The Blast2GO annotation tool[36] was used to assign most probable gene ontology (GO) terms to the unique sequences that annotated against the SwissProt database (Supplementary Table S5) with the *E*-value set to $1 \times E^{-5}$ and other parameters following the defaults. The significant unigene information (*E*-value cutoff was $\leq 1 \times E^{-5}$ and a minimum alignment length of 22 bp) of the query set of unique sequences subsequently were mapped into several subcategories of the three level 1 categories of 'cellular component', 'molecular function' and 'biological process', respectively, with program GO Slimmer (http://amigo.geneontology.org/cgi-bin/amigo/slimmer?) and by searching against all available species databases and evidence codes (AmiGO version: 1.8). In case the unigenes have matched in multiple databases, only one record was kept in the summary (Supplementary Table S6). The KO (KEGG orthology) assignments and the KEGG pathway reconstruction were performed in KAAS (Automatic Annotation Server Ver. 1.6a; http://www.genome.jp/tools/kaas/) with the default parameters (Supplementary Table S7).

2.6. *Gene expression analysis using quantitative real-time PCR*

Examination of gene expressions was conducted using quantitative real-time PCR as described in Xia *et al.*[29] Briefly, 1 μg aliquot of the DNase-treated total RNA was reverse transcribed to cDNA by M-MLV reverse transcriptase (Promega, Madison, USA) with poly-dT as RT primer following the manufacturer's protocol. For analysis of gene expression patterns, 10 times dilution of the resulting single-strand cDNA were assayed as DNA template by real-time PCR using primers (Supplementary Table S8) for 26 immune-related genes and *EF1A* gene as control. PCRs were performed with the iQ SYBR Green Supermix (Bio-Rad, Hercules, CA, USA) as described by the manufacturer in an iQ^{TM}5 Real Time PCR Detection Systems (Bio-Rad). PCRs were performed in triplicates.

For analysis of the changes of gene expression, the values of triplicate real-time PCRs were normalized to *EF1A* gene expression, calculated by the $\Delta\Delta C_t$ method. The normalized values for the test fishes challenged with the bacteria *V. harveyi* were compared with the control level at the respective time and tissue. The resulting expression ratio was then converted to natural logs and analyzed with Cluster 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv).

2.7. *Identification of repetitive elements and single-nucleotide polymorphisms*

Identification and annotation of transposable elements (TEs) were carried out with the program CENSOR in Repbase[37] by blasting against zebrafish conserved repeats. The microsatellite-containing ESTs (with $\geq$8 repeat number) were detected using the program Tandem Repeats Finder[38] with the default parameters (Supplementary Table S9). Putative single-nucleotide polymorphisms (SNPs) were screened by using the program SeqMan (DNASTAR). To minimize the potential errors in the detection of SNPs caused by sequencing errors, putative high-quality SNPs were further identified by using more stringent criteria (only clusters with at least four EST sequences and an SNP mutated within at least two ESTs were selected) and no indel data were included in the analysis (Supplementary Table S10).

2.8. Verification of potential SNPs from ESTs

In order to verify the potential SNPs indentified in ESTs, six DNA samples of wild seabass from Southeast Asia were used. All unique transcripts containing putative high-quality SNPs and enough flanking regions were selected from the ESTs for primer design using the program PrimerSelect (DNASTAR, Wilmington, DE). PCR products were sequenced directly in both directions with forward or reverse primers (Supplementary Table S11). SNP genotype was analyzed by using software Sequencher 4.9 (Gene Codes).

3. Results and discussions

3.1. Generation and clustering of ESTs

Identification and characterization of a transcriptome for the Asian seabass can make a significant contribution to future research of gene functions in the community. At the start of the transcriptome project, the dbEST database contained 5637 Asian seabass ESTs (gi:169659332–169664968), originated from a brain cDNA library[32] and two subtracted spleen cDNA libraries (2887 ESTs; gi: GT219120–GT222006).[28] In order to enrich the existing transcriptome data sets and analyze genes that expressed in different organs of the Asian seabass, we sequenced ESTs from nine normalized cDNA libraries from various tissues of the Asian seabass (Supplementary Table S1). A total of 11 451 high-quality novel EST sequences were obtained from these libraries after removing vector and low-quality sequences. The detailed information of the EST resources is presented in Supplementary Table S1.

To create a large global collection of the Asian seabass ESTs, the published 8524 ESTs[28,32] were retrieved from the dbEST database and included in the following clustering and annotation. After combining these data, a total of 19 975 high-quality ESTs were obtained and used to identify EST clusters representing redundant transcripts under high stringency (95% similarity; Supplementary Table S10). The clusters were composed of 8837 unique sequences (putative transcripts) including 2838 contigs and 5999 singletons. The number of ESTs in a contig ranged from 2 to 289. On average, each contig contained 4.9 EST sequences. The most abundant transcript containing 289 ESTs shows high similarity to the parvalbumin beta-1 mRNA sequence. Only 58 contigs consisted of more than 20 ESTs. The length distribution of high-quality EST sequences in the clusters ranged from 50 to 2098 bp with an average size of 425 bp. A total of 4406 (50%) unique sequences were >500 bp, 2653 sequences were between 300 and 500 bp and 121 sequences were <100 bp (Fig. 1). The average contig length was 574 bp, being comparable with that of other studies, such as cod[18] and halibut.[39] The unique sequences are being used in the development of a cDNA microarray to examine global gene expression in Asian seabass.

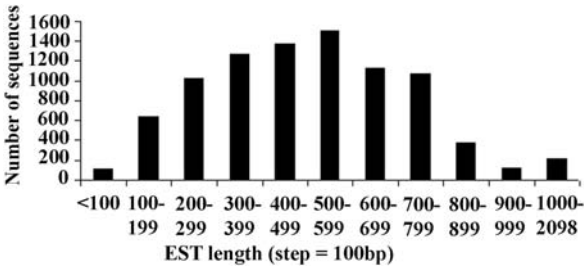

Figure 1. Length distribution of the 8837 unique transcriptome clusters of Asian seabass.

Recently, the rapid development of novel massively parallel cDNA sequencing, or RNA-Seq, has allowed many advances in the characterization and quantification of transcriptomes in fish, such as salmonids,[40] whitefish,[41] catfish[42] and Japanese seabass.[43] Compared with traditional Sanger sequencing, RNA-Seq platforms allow the entire transcriptome to be surveyed in a very high-throughput, cost-effective and quantitative manner.[7,44,45] However, there are still substantial challenges in using RNA-Seq for the assembly and annotation of transcriptomes from organisms lacking a genome reference. For example, the assembly of short RNASeq reads is difficult. Without cloning, it is difficult to know *a priori* which reads came from which transcripts; and many less highly or less broadly expressed transcripts are only weakly or incompletely supported by current RNA-Seq.[46] Since the Sanger sequencing can determine the sequence of a full-length gene from cDNA libraries, the utilization of methods that combine RNA-Seq with the Sanger sequencing-based strategies will collectively resolve several technical challenges of RNA-Seq. The combination of the two sequencing methods is expected to improve the annotation of ESTs and facilitating studies on transcriptome.

3.2. Annotation of the 8837 putative transcripts of the Asian seabass

To compare our data with the other existing data, we downloaded 18 species-specific data sets from Ensembl database and three data sets (refseq_rna, refseq_protein and SwissProt) from NCBI database. Ten of the cDNA or protein data sets originated from five fish species, and eight of the data sets originated from human, chicken, lizard and frog. By local blast search of 8837 unique sequences from seabass against these data sets, 2198 (25% of the unique

sequences; lizard) to 4100 (46%; Stickleback) unique sequences were annotated by BLASTn and 2929 (33%; Lizard) to 3471 (39%; Stickleback) had significant matches by BLASTx. Both results showed that the transcriptome of the Asian seabass was more similar to that of the Stickleback (Supplementary Table S2) than to other fish species. In addition, 2364 (BLASTn) to 3002 (BLASTx) of the seabass unique sequences showed high conservation in at least seven species; and 1516 (BLASTn) to 349 (BLASTx) of the unique sequences had positive hits with three or less species.

Compared with these species-specific databases, the two reference sequence databases (RefSeq) and SwissProt database could provide a more comprehensive, integrated, non-redundant and well-annotated set of sequences. The annotation of the unique 8837 putative transcripts (Supplementary Tables S3–S5) was performed on the basis of the best match data that found after local blast searches against RefSeq_RNA database by BLASTn and RefSeq_Protein and SwissProt protein databases by BLASTx. Of the annotated unique sequences in the Asian seabass, ~17 (Refseq_Protein) to 24% (Refseq_RNA) had significant homology with *E*-values between $1 \times E^{-5}$ and $1 \times E^{-15}$, being weakly similar to the counterparts in the databases. A smaller sets of genes (5.5% for SwissProt to 22% for Refseq_rna) that had a very good BLAST hit ($<1 \times E^{-105}$) were considered as highly significant homologs. The remaining was considered moderately similar (*E*-values between $1 \times E^{-15}$ and $1 \times E^{-105}$; Fig. 2).

A total of 4323 unique sequences (48.9%) had significant annotations that presented in Refseq_RNA database using BLASTn; and 3108 sequences (35.2%) to 3320 sequences (37.6%) had significant matches to the SwissProt and Refseq_Protein protein databases using BLASTx (Supplementary Tables S2, S3 and S5). The low annotation ratio (35.2–48.9%) of the unique sequences with significant BLAST hits is similar to the level that reported in the Catfish EST project (37 and sim 50%),[14,19] European seabass *Dicentrarchus labrax* (~ 41%)[16] and the salmon EST project (45%).[11] The low significant annotation success ratio in fish species might reflect the poor annotation rates of the fish genes in contrast to those of mammalian species. For example, currently EST entries in NCBI database for bony fishes are 5 142 023 hits, which are only 26.8% of the mammals EST entries (19 183 053; http://www.ncbi.nlm.nih.gov/nucest; 27 August 2010). Alternatively, the number of unigenes (8837) that were estimated from the number of unique transcripts is probably an overestimation of isolated genes in this study. The unclustered ESTs from a single transcript might result from alternate splicing, sequence polymorphisms, sequencing errors and non-overlapping ESTs.[47]

A substantial proportion of the seabass unique sequences lack significantly functional annotations or protein identities. These sequences might correspond to proteins that have not yet been identified in related organisms to date or they were only the 5′ untranslated regions of known genes due to incomplete genomic information. In zebrafish, a total of 30 796 genes had already been discovered (http://zfin.org/zf_info/zfin_stats.html; 13 September 2010). Since the gene numbers in fish transcriptomes are usually similar, the gene number (<8837) that were found for the Asian seabass in the study is much less than expected. Future additions of ESTs to the clusters will undoubtedly improve the rate of gene discovery in the Asian seabass. The information about the significant annotation of the query set of unique sequences is presented in Supplementary Tables S3 and S5.

ESTs from multiple normalized cDNA libraries will significantly increase the gene discovery,[47] providing a more broadly applicable data set for functional genomic applications. In this study, 11 cDNA libraries from various tissues of the Asian seabass were analyzed. The annotated unique sequences were further subdivided according to their library of origin (Supplementary Table S3). Of the ESTs, 9780 (49%) have significant annotations by BLASTx search against Refseq_Protein database and 12 394 (62%) have significant annotation by BLASTn search against Refseq_RNA database. Further analysis revealed a few genes were more highly presented in specific tissues and thus were candidate genes for further studies (Supplementary Table S3). For example, in the brain, several abundant transcripts were involved in transport processes, such as ATP synthase F0 subunit 6 and cytochrome *c* oxidase subunit II; and in the kidney and liver, some genes with high abundance related to immune system processes were identified, such as cathepsin D, IgG heavy chain, thioredoxin-like 1 and NF-kappa-B-activating protein. A list of top eight most abundant unique genes in each tissue is presented in Supplementary Table S4.

3.3. *Functional classification of the Asian seabass genes*

In order to achieve a functional annotation of unique transcripts, we took the advantage of GO database using the automated annotation tool Blast2GO[48] and GO Slimmer in AmiGO database, and Kyoto Encyclopedia of Genes and Genomes (KEGG) database. These analyses should provide the base information for further research of gene functions in the Asian seabass.

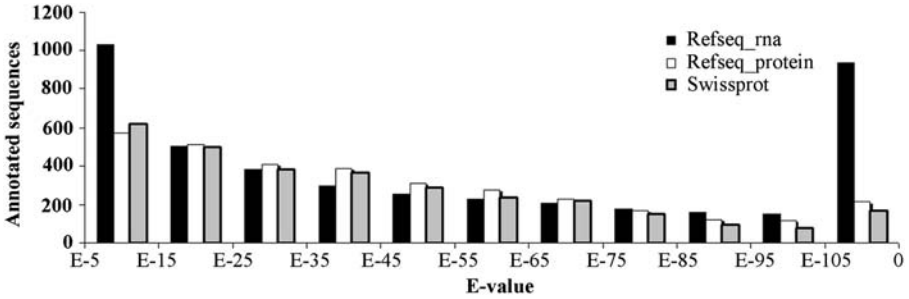

Figure 2. The distribution of annotated seabass unique sequences in various *E*-value intervals showing conservation with the homologs in the NCBI database.

SwissProt protein database can provide a high level of annotations, a minimal level of redundancy and high level of integration with other databases. Therefore, functional classification of the seabass unique sequences based on the annotation results of SwissProt databases by BLASTx program was performed. The annotation information for 3108 unique sequences (containing 2489 unigenes) was first functionally classified by using Blast2GO software. We obtained 44 338 GO terms for 3042 unique sequences (97.9%) using this method and the remaining 66 of the unique sequences (2.1%) were unmapped (Supplementary Table S5). Therefore, the majority of the seabass genes could be classified bioinformatically.

By using the GO Slimmer program under the three level 1 categories (cellular components, molecular functions and biological processes), the 2489 unigenes (Supplementary Table S5) were then classified into 23 of 24 subcategories (second level GO terms) in the category of biological process, 13 of 14 subcategories in cellular component and 12 of 16 subcategories in molecular function. Functional classifications of the Asian seabass unigenes for each of the three main GO categories are given in Supplementary Table S6.

Genes encoding for proteins associated with cellular process (1161; 46.6% of the 2489 unigenes), metabolic process (854; 34.3%) and biology regulation (674; 27.1%) in the category of biological processes were the three largest annotated subcategories, indicative of the high metabolic characteristics in juvenile seabass. Two following groups were found to encode products related to the regulation of biological process (636; 25.6%) and response to stimulus (518; 20.8%). We found 140 (5.6%) genes acting in immune system process, e.g. chemokine (C-C motif) ligand 13 (*ccl13*). Figure 3 and Supplementary Table S6 show the distributions of unigenes involved in the immune system process and response to stimulus (under the third level GO terms) according to the GO consortium.

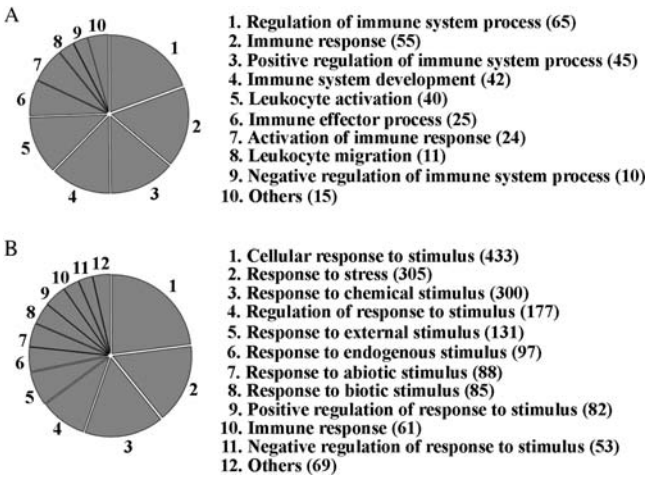

Figure 3. The distributions of seabass unigenes involved in the subcategories of immune system processes (A) and response to stimulus (B). The number in parenthesis shows the number of unigenes that classified into the subcategory.

Regarding the cellular component, a large proportion of unigenes was classified into cell (1136; 45.6%), cell part (1136; 45.6%) and organelle (818; 32.9%) followed by organelle part (460; 18.5%) and macromolecular complex (377; 15.1%). Binding (956; 38.4%) and catalytic activity (517; 20.8%) were the most dominant subcategories in the molecular function category followed by transporter activity (106; 4.3%). Nineteen genes involved in antioxidant activity such as superoxide dismutase 1 (*sod1*) and muconate cycloisomerase (*catb*). The remaining genes encode products involved in many other diverse biological activities.

KO of the putative transcripts (8837) were identified through BLAST searching against the KEGG database.[49] KEGG categories were found for 1381 seabass unique sequences (15.6%) with 1192 KO term, and 735 unique genes (8.3%) fell into 210 pathways (Supplementary Table S7). A total of 7463 (84.5%) seabass unique sequences had no assigned function. According to the electronic annotation, 56.8% of all the pathways (210 of 370) on the molecular

interaction and reaction networks in the database have been captured in this study. This result was similar to the halibut EST analysis,[39] in which 7.5% sequences were found for KEGG categories and fell into 185 KEGG pathways (50%).

The three largest pathway categories were Metabolic pathways (224), ribosome (77) and oxidative phosphorylation (69), suggesting that the fish was undergoing tremendous metabolic activities. Large amounts of the annotated genes were also found associated with diseases and immune processes, e.g. Huntington's disease (67), Alzheimer's disease (63) and complement and coagulation cascades (22). Alternatively, many EST clusters were presented in 25 signaling pathways, e.g. MAPK signaling pathway (32), Chemokine signaling pathway (31) and Jak-STAT signaling pathway (16). A number of important regulatory molecules were identified in these pathways, such as proto-oncogene protein (*c-fos*). The KEGG pathway information of the annotated genes is summarized in Supplementary Table S7.

3.4. *Expression patterns of immune-related genes in the Asian seabass transcriptome after challenging with a bacterium V. harveyi*

Vibrio infection of fish can cause significant mortality in mariculture.[50] Previous studies showed that the transcriptome profiles of fish challenged with *V. harveyi* were considerably altered in some fish species, such as Japanese seabass,[43] turbot[51] and rainbow trout.[52] In Asian seabass, elevated antibody activities in sera were found in all treatment groups[53]; and the miRNAs level was highly changed in acute inflammatory immune responses with protection against *Vibrio* infection.[29] These studies suggested that *V. harveyi* can cause immune responses in aquatic organisms. However, so far little is known about temporal and spatial expression patterns of immune-related genes in fishes after challenge with *V. harveyi*.

The Toll-like receptor (TLR) family is a family of receptors involved in microbial recognition by the immune system and the activation of signaling pathways that result in immune responses against microbial infections.[54,55] Of the annotated genes in this study, 14 genes (Supplementary Table S7) were classified into the TLR signaling pathway through BLAST search against the KEGG database. To study the expression changes after challenge with *Vibrio*, effective qRT–PCR primers for 13 of these genes were developed. Temporal expression of these genes at the spleen and liver were analyzed by qRT–PCR using RNA isolated from seabass challenged with *V. harveyi* at five different times of points. In addition,

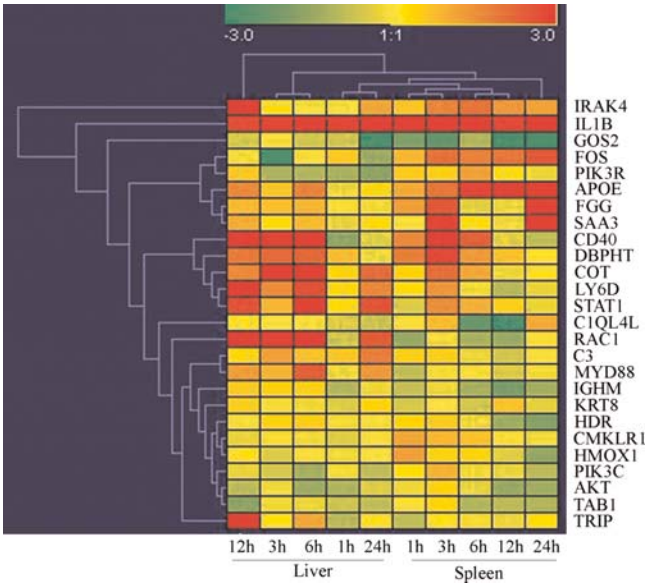

Figure 4. Temporal and spatial expression patterns of 26 immune-related genes in the spleen and liver revealed by qRT–PCR, Real-time PCR was used to examine gene expressions in the liver and spleen of Asian seabass challenged with *V. harveyi*. Samples were collected at 1, 3, 6, 12 and 24 h post-injection. Relative gene expression changes are shown on a natural logarithmic scale, and the values represent the mRNA expression changes after challenged with *V. harveyi* relative to a control samples at each point of sampling time and tissue. Red color represents increasing level of the gene expression and blue color indicates decreasing of the gene expression after challenging with *V. harveyi* at the respective time and tissue.

another 13 immune-related genes showing important roles in host inflammatory responses[28,29] were also analyzed in this study (Fig. 4 and Supplementary Table S8).

The expression profiles in the liver at 3 and 6 hpi and at 1 and 24 hpi were clustered together, respectively. The expression profiles for the spleen at 1 hpi were clustered with 3 hpi, and the profiles at 6 hpi were more similar to 12 hpi. Clustering of the expression patterns for the 26 genes revealed the increasing expression of *IL1b* was ubiquitous at five sampling stages and in two organs, but *GOS2* gene showed strongly decreasing expression in most of the sampling time points, especially in the spleen. The expression profiles for gene *FOS*, *PIK3R*, *APOE*, *FGG* and *SAA3* were clustered together. These genes show generally increasing expression at each time point in the spleen; especially gene *FOS* and *APOE* showed highly increasing expression during experiment in the spleen, but *FGG* and *SAA3* show highly increasing expression at 3 and 24 hpi and slightly increasing expression at 6 and 12 hpi. In addition, *CD40*, *DBPHT*, *COT*, *LY6D* and *STAT1* showed similarly increasing expression in the liver at 3–24 hpi, and in the spleen at 3 hpi. Interestingly, *RAC1* show increasing expression in the liver at 3–24 hpi, but highly

decreasing expression in the spleen at 1, 6 and 12 hpi. Our study demonstrates that temporal dynamics of immune-related gene expression in response to *Vibrio* challenge in the spleen and liver and suggested that some genes may respond differentially to pathogen challenge at different times and tissues. Future works on these genes to clarify their functions during pathogen challenge will help us to understand the innate and early immune response of fish. We are currently developing an oligo-cDNA microarray using the 8837 unique EST sequences identified in this study to study global gene expressions. The application of the microarray is expected to identify more genes responding to pathogens in Asian seabass.

3.5. Identification of DNA markers in ESTs

Microsatellites and SNPs in ESTs/genes are useful as DNA markers because they represent transcribed genes and can be used as anchor markers for comparative mapping, evolutionary and association studies.[56] Large-scale EST sequences can provide an enormous resource for marker development.[14,42] In the Asian seabass, we identified 634 microsatellite loci in all EST sequences (Supplementary Table S9). A total of 70 different repeat motifs (2–10 bases) were found. Of which, (GT/AC)*n* (351 loci; 55.4%) was the largest group and followed by (CT/AG)*n* (110 loci; 17.4%) in the di-nucleotide microsatellite group. In halibut, 46.5% of the microsatellites detected in the cDNA libraries were di-nucleotide repeats.[39] In our previous studies,[24,25] 150 primer pairs were designed for these EST-derived loci and were tested on the three parents of the mapping panel containing two full-sib families. Finally, 55 of these loci were polymorphic and were mapped in the linkage map of Asian seabass.[24,25] Therefore, we could expect that that 36% of EST microsatellites could be mapped in the linkage map of Asian seabass.[25]

By using the program SeqMan, 6144 putative SNPs were detected in the 2838 contigs of the current EST clusters for the Asian seabass (Supplementary Table S10). To minimize the potential errors in the detection of SNPs caused by sequencing errors, 193 putative high-quality SNPs consisting of 137 transitions and 56 transversions (Supplementary Table S11) were further identified by using more stringent criteria. We further validated the potential SNPs by genotyping six wild samples captured from Southeast Asia. Fifty-one (26.4%) of these SNP candidates showed polymorphism in the samples (Supplementary Table S12). The rate of polymorphic SNPs was probably an underestimate due to the fact that only six individuals were used and primers were developed only for 90 unique sequences where enough flanking sequences were available for designing primers for the validation. Using more individuals, the polymorphic rate of potential SNPs should be higher.

3.6. ESTs containing repetitive elements

TEs are usually smaller than 15 kb and can be classified into DNA transposons (type 1) and retrotransposons (LTR and non-LTR retrotransposons; type 2).[57] Repbase (http://www.girinst.org/repbase/index.html), a reference database of eukaryotic repetitive DNA, is the most widely used database of TEs.[58] By using the software tool Censor,[37] we analyzed the repeats of the Asian seabass 8837 unique sequences in the Repbase database. Besides microsatellites (Supplementary Table S9), additional 461 repeat fragments (TEs) were found in the 8837 unique sequences (Supplementary Table S13). Of which, DNA transposons (208; 45.1% of the repeats) were the most abundant type of TEs, followed by LTR retrotransposon (127; 27.5%) and non-LTR retrotransposon (84; 18.2%). Fragments containing Gypsy (85 hits), hAT (44 hits) and CR1 (37 hits) were the three largest groups in respective categories of the most abundant TEs in the species.

TEs in the Asian seabass represent around 0.5 % (0.04/8.5 Mb) of the total length of clustered sequences. These results are broadly consistent with the observed fractions of repetitive DNA in the genomes of fish species, such as 0.7% TEs found in Salmonid EST Database.[11] The hierarchy of classes of repeats for the 8837 unique sequences is presented in Supplementary Table S13.

3.7. Conclusion

A transcriptome including 19 975 EST sequences of the Asian seabass was characterized in this study. The clustering of these EST data produced 8837 putative transcripts. The annotation and functional classification of the unigenes identified a broad range of genes involved in different functions, processes and compartments, and a comprehensive list of signaling regulatory molecules working in disease, immune, growth and developments. The EST cluster presented here provides an unprecedented look at the seabass transcriptome. This transcriptome data provide a crucial starting point for further comparison of transcriptomes with other fish species and will enhance the progress of gene discovery and characterization and facilitate future whole-genome sequence assembly and annotation of Asian seabass.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This study is funded by the Ministry of National Development and National Research Foundation of Singapore.

References

1. Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A. and Bassett, D.E. 1997, Characterization of the yeast transcriptome, *Cell*, **88**, 243–51.
2. Bernot, A. 2004, *Genome Transcriptome and Proteome Analysis*. John Wiley & Sons Inc.: West Sussex, England.
3. Adams, M.D., Kelley, J.M., Gocayne, J.D., et al. 1991, Complementary DNA sequencing: expressed sequence tags and human genome project, *Science*, **252**, 1651.
4. Kim, J.M., Lee, K.H., Jeon, Y.J., et al. 2006, Identification of genes related to Parkinson's disease using expressed sequence tags, *DNA Res.*, **13**, 275–86.
5. Parkinson, J. 2009, *Expressed Sequence Tags (ESTs): Generation and Analysis*. Human Press: New York.
6. Liu, Z. 2007, *Aquaculture Genome Technologies*. Wiley: IA, USA.
7. Garg, R., Patel, R.K., Tyagi, A.K. and Jain, M. 2011, De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification, *DNA Res.*, **18**, 53–63.
8. Manickavelu, A., Kawaura, K., Oishi, K., et al. 2011, Comparative gene expression analysis of susceptible and resistant near-isogenic lines in common wheat infected by *Puccinia triticina*, *DNA Res.*, **17**, 211–22.
9. Froese, R. and Pauly, D. 2010, *FishBase*. World Wide Web electronic publication. http://www.fishbase.org/search.php.
10. Gong, Z., Yan, T., Liao, J., Lee, S.E., He, J. and Hew, C.L. 1997, Rapid identification and isolation of zebrafish cDNA clones, *Gene*, **201**, 87–98.
11. Rise, M., von Schalburg, K., Brown, G., et al. 2004, Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics, *Genome Res.*, **14**, 478–90.
12. Edvardsen, R.B., Malde, K., Mittelholzer, C., Taranger, G.L. and Nilsen, F. 2010, EST resources and establishment and validation of a 16k cDNA microarray from Atlantic cod (*Gadus morhua*), *Comp. Biochem. Physiol. D Genomics Proteomics*, **6**, 23–30.
13. Govoroun, M., Le Gac, F. and Guiguen, Y. 2006, Generation of a large scale repertoire of expressed sequence tags (ESTs) from normalised rainbow trout cDNA libraries, *BMC Genomics*, **7**, 196.
14. Wang, S., Peatman, E., Abernathy, J., et al. 2010, Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies, *Genome Biol.*, **11**, R8.
15. Sarropoulou, E., Power, D.M., Magoulas, A., Geisler, R. and Kotoulas, G. 2005, Comparative analysis and characterization of expressed sequence tags in gilthead sea bream (*Sparus aurata*) liver and embryos, *Aquaculture*, **243**, 69–81.
16. Sarropoulou, E., Sepulcre, P., Poisa-Beiro, L., et al. 2009, Profiling of infection specific mRNA transcripts of the European seabass *Dicentrarchus labrax*, *BMC Genomics*, **10**, 157.
17. Bai, J., Solberg, C., Fernandes, J.M. and Johnston, I.A. 2007, Profiling of maternal and developmental-stage specific mRNA transcripts in Atlantic halibut *Hippoglossus hippoglossus*, *Gene*, **386**, 202–10.
18. Olsvik, P. A. and Holen, E. 2009, Characterization of an Atlantic cod (*Gadus morhua*) embryonic stem cell cDNA library, *BMC Res. Notes*, **2**, 74.
19. Li, P., Peatman, E., Wang, S., et al. 2007, Towards the ictalurid catfish transcriptome: generation and analysis of 31,215 catfish ESTs, *BMC Genomics*, **8**, 177.
20. Christoffels, A., Bartfai, R., Srinivasan, H., Komen, H. and Orban, L. 2006, Comparative genomics in cyprinids: common carp ESTs help the annotation of the zebrafish genome, *BMC Bioinformatics*, **7**, S2.
21. Yue, G.H., Li, Y., Chao, T.M., Chou, R. and Orban, L. 2002, Novel microsatellites from Asian sea bass (*Lates calcarifer*) and their application to broodstock analysis, *Mar. Biotechnol.*, **4**, 503–11.
22. Zhu, Z.Y., Wang, C.M., Lo, L.C., Feng, F., Lin, G. and Yue, G.H. 2006, Isolation, characterization, and linkage analyses of 74 novel microsatellites in Barramundi (*Lates calcarifer*), *Genome*, **49**, 969–76.
23. Zhu, Z.Y., Wang, C.M., Lo, L.C., et al. 2010, A standard panel of microsatellites for Asian seabass (*Lates calcarifer*), *Anim. Genet.*, **41**, 208–12.
24. Wang, C.M., Zhu, Z.Y., Lo, L.C., et al. 2007, A microsatellite linkage map of Barramundi, *Lates calcarifer*, *Genetics*, **175**, 907–15.
25. Wang, C.M., Bai, Z.Y., He, X.P., et al. 2011, A high-resolution linkage map for comparative genome analysis and QTL fine mapping in Asian seabass, *Lates calcarifer*, *BMC Genomics*, **12**, 174.
26. Wang, C.M., Lo, L.C., Feng, F., et al. 2008, Construction of a BAC library and mapping BAC clones to the linkage map of Barramundi, *Lates calcarifer*, *BMC Genomics*, **9**, 139.
27. Xia, J.H., Feng, F., Lin, G., Wang, C.M. and Yue, G.H. 2010, A first generation BAC-based physical map of the Asian seabass (*Lates calcarifer*), *PLoS One*, **5**, e11974.
28. Xia, J.H. and Yue, G.H. 2010, Identification and analysis of immune-related transcriptome in Asian seabass *Lates calcarifer*, *BMC Genomics*, **11**, 356.
29. Xia, J.H., He, X.P., Bai, Z.Y. and Yue, G.H. 2011, Identification and characterization of 63 microRNAs in the Asian seabass, *Lates calcarifer*, *PLoS One*, **6**, e17537.
30. Wang, C.M., Lo, L.C., Zhu, Z.Y. and Yue, G.H. 2006, A genome scan for quantitative trait loci affecting growth-related traits in an F1 family of Asian seabass (*Lates calcarifer*), *BMC Genomics*, **7**, 274.
31. Wang, C.M., Lo, L.C., Zhu, Z.Y., et al. 2011, Mapping QTL for an adaptive trait: the length of caudal fin in *Lates calcarifer*, *Mar. Biotechnol.*, **13**, 74–82.
32. Tan, S.L., Mohd-Adnan, A., Mohd-Yusof, N.Y., Forstner, M.R. and Wan, K.L. 2008, Identification and analysis of a prepro-chicken gonadotropin releasing hormone II (preprocGnRH-II) precursor in the Asian seabass, *Lates calcarifer*, based on an EST-based assessment of its brain transcriptome, *Gene*, **411**, 77–86.
33. Wang, C.M., Lo, L.C., Zhu, Z.Y., et al. 2008, Estimating reproductive success of brooders and heritability of growth traits in Asian sea bass (*Lates calcarifer*) using microsatellites, *Aquac. Res.*, **39**, 1612–9.

34. Austin, B. and Zhang, X. 2006, *Vibrio harveyi*: a significant pathogen of marine vertebrates and invertebrates, *Lett. Appl. Microbioll.*, **43**, 119–24.
35. Yue, G.H., Zhu, Z.Y., Wang, C.M. and Xia, J.H. 2009, A simple and efficient method for isolating polymorphic microsatellites from cDNA, *BMC Genomics*, **10**, 125.
36. Götz, S., García-Gómez, J.M., Terol, J., et al. 2008, High-throughput functional annotation and data mining with the Blast2GO suite, *Nucleic Acids Res.*, **36**, 3420–35.
37. Kohany, O., Gentles, A.J., Hankus, L. and Jurka, J. 2006, Annotation, submission and screening of repetitive elements in Repbase: Repbase submitter and censor, *BMC Bioinformatics*, **7**, 474.
38. Benson, G. 1999, Tandem repeats finder: a program to analyze DNA sequences, *Dev. Comp. Immunol.*, **27**, 573–80.
39. Douglas, S.E., Knickle, L.C., Kimball, J. and Reith, M.E. 2007, Comprehensive EST analysis of Atlantic halibut (*Hippoglossus hippoglossus*), a commercially relevant aquaculture species, *BMC Genomics*, **8**, 144.
40. Seeb, J.E., Pascal, C.E., Grau, E.D., et al. 2011, Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids, *Mol. Ecol. Resour.*, **11**, 335–48.
41. Jeukens, J., Renaut, S., St-Cyr, J., Nolte, A.W. and Bernatchez, L. 2010, The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., Salmonidae) divergence as revealed by next-generation sequencing, *Mol. Ecol.*, **19**, 5389–403.
42. Liu, S., Zhou, Z., Lu, J., et al. 2011, Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array, *BMC Genomics*, **12**, 53.
43. Xiang, L.X., He, D., Dong, W.R., Zhang, Y.W. and Shao, J.Z. 2010, Deep sequencing-based transcriptome profiling analysis of bacteria-challenged *Lateolabrax japonicus* reveals insight into the immune-relevant genes in marine fish, *BMC Genomics*, **11**, 472.
44. Wang, Z., Gerstein, M. and Snyder, M. 2009, RNA-Seq: a revolutionary tool for transcriptomics, *Nat. Rev. Genet.*, **10**, 57–63.
45. Ozsolak, F. and Milos, P.M. 2011, RNA sequencing: advances, challenges and opportunities, *Nat. Rev. Genet.*, **12**, 87–98.
46. Haas, B.J. and Zody, M.C. 2010, Advancing RNA-Seq analysis, *Nat. Biotechnol.*, **28**, 421–3.
47. Marques, M.C., Alonso-Cantabrana, H., Forment, J., et al. 2009, A new set of ESTs and cDNA clones from full-length and normalized libraries for gene discovery and functional characterization in *citrus*, *BMC Genomics*, **10**, 428.
48. Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. 2005, Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics*, **21**, 3674–6.
49. Kanehisa, M. and Goto, S. 2000, KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, **28**, 27–30.
50. Toranzo, A.E., Magariños, B. and Romalde, J.S. 2005, A review of the main bacterial fish diseases in mariculture systems, *Aquaculture*, **246**, 37–61.
51. Wang, C., Zhang, X.H., Jia, A., Chen, J. and Austin, B. 2008, Identification of immune-related genes from kidney and spleen of turbot, *Psetta maxima* (L.), by suppression subtractive hybridization following challenge with *Vibrio harveyi*, *J. Fish Dis.*, **31**, 505–14.
52. Arijo, S., Brunt, J., Chabrillón, M., Díaz-Rosales, P. and Austin, B. 2008, Subcellular components of *Vibrio harveyi* and probiotics induce immune responses in rainbow trout, *Oncorhynchus mykiss* (Walbaum), against *V. harveyi*, *J. Fish Dis.*, **31**, 579–90.
53. Crosbie, P.B. and Nowak, B.F. 2004, Immune responses of barramundi, *Lates calcarifer* (Bloch), after administration of an experimental *Vibrio harveyi* bacterin by intraperitoneal injection, anal intubation and immersion, *J. Fish Dis.*, **27**, 623–32.
54. Takeda, K., Kaisho, T. and Akira, S. 2003, Toll-like receptors, *Annu. Rev. Immunol.*, **21**, 335–76.
55. Barton, G.M. and Medzhitov, R. 2003, Toll-like receptor signaling pathways, *Science*, **300**, 1524.
56. Liu, Z. and Cordes, J. 2004, DNA marker technologies and their applications in aquaculture genetics, *Aquaculture*, **238**, 1–37.
57. Kapitonov, V.V. and Jurka, J. 2008, A universal classification of eukaryotic transposable elements implemented in Repbase, *Nat. Rev. Genetics*, **9**, 411–12.
58. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. 2005, Repbase update, a database of eukaryotic repetitive elements, *Cytogenet. Genome Res.*, **110**, 462–7.