Mar Biotechnol (2011) 13:608–621
DOI 10.1007/s10126-010-9286-y

ORIGINAL ARTICLE

A Review of the Major Penaeid Shrimp EST Studies and the Construction of a Shrimp Transcriptome Database Based on the ESTs from Four Penaeid Shrimp

Jiann-Horng Leu · Shu-Hwa Chen · Yu-Bin Wang · Yen-Chen Chen · Sheng-Yao Su · Chung-Yen Lin · Jan-Ming Ho · Chu-Fang Lo

Received: 24 February 2010 / Accepted: 10 March 2010 / Published online: 17 April 2010

Abstract By economic value, shrimp is currently the most important seafood commodity worldwide, and these animals are often the subject of scientific research in shrimp farming countries. High throughput methods, such as expressed sequence tags (ESTs), were originally developed to study human genomics, but they are now available for studying other important organisms, including shrimp. ESTs are short sequences generated by sequencing randomly selected cDNA clones from a cDNA library. This is currently the most efficient and powerful method for providing transcriptomic data for organisms with an uncharacterized genome. This review will summarize the sixteen major shrimp EST studies that have been conducted to date. In addition, we analyzed the EST data downloaded from NCBI dbEST for the four major penaeid shrimp species and constructed a database to host all of these EST data as well as our own analysis results. This database provides the shrimp aquaculture research community with an outline of the shrimp transcriptome as well as a tool for shrimp gene identification.

Keywords Penaeid shrimp · cDNA libraries · EST · Transcriptome · Gene expression

Jiann-Horng Leu and Shu-Hwa Chen contributed equally to this work.

J.-H. Leu
Center for Marine Bioenviroment and Biotechnology,
National Taiwan Ocean University,
No. 2, Pei-Ning Road,
Keelung 20224 Taiwan, Republic of China

S.-H. Chen · Y.-B. Wang · Y.-C. Chen · S.-Y. Su · C.-Y. Lin (✉) · J.-M. Ho (✉)
Institute of Information Science, Academia Sinica,
No. 128, Sec. 2, Academia Road,
Taipei 11529 Taiwan, Republic of China
e-mail: cylin@iis.sinica.edu.tw
e-mail: hoho@iis.sinica.edu.tw

C.-Y. Lin
Division of Biostatistics and Bioinformatics,
National Health Research Institutes,
No. 35, Keyan Road,
Zhunan, Miaoli County 35053 Taiwan, Republic of China

C.-Y. Lin
Institute of Fisheries Science, National Taiwan University,
No. 1, Sec. 4, Roosevelt Road,
Taipei 10617 Taiwan, Republic of China

C.-F. Lo (✉)
Institute of Zoology, National Taiwan University,
No.1, Sec. 4, Roosevelt Road,
Taipei 10617 Taiwan, Republic of China
e-mail: gracelow@ntu.edu.tw

Introduction

In the past few decades, shrimp has gradually become the most important commodity in the international seafood trade. According to Gillett (2008), the total amount of global shrimp production now reaches about 6 million tonnes annually, with global trade value at US $10 billion, or 16% of the global seafood trade, far exceeding any other seafood item. Nearly half of this shrimp production is from farms, and shrimp culture has attained great economic and social importance in shrimp farming countries. To improve shrimp aquaculture production, the study of shrimp biology, particular shrimp physiology, immunology, and genetics has become increasingly important. The availability of shrimp transcriptome data will not only facilitate this

research at molecular levels, but will also help in the identification of shrimp genes. Currently, the most efficient and cost-effective method to retrieve transcriptomic data from an organism is by analysis of the expressed sequence tags (ESTs) of a cDNA library.

ESTs are short cDNA sequences (200–800 nucleotide bases in length) that are generated by single-pass 5′ or 3′ end sequencing of clones randomly selected from cDNA libraries (Adams et al. 1991). Therefore, in a sense, an EST provides a partial description of the transcribed regions or activities of a genome in a tissue or an organism under specific experimental conditions. For organisms with limited genomic information, like the shrimp, ESTs provide a rich resource for gene identification. ESTs also aid in complementation of genome annotation, gene structure identification, detection of alternative transcripts, characterization of single nucleotide polymorphism (SNP) and facilitating the analysis of proteomes (Jongeneel 2000; Dong et al. 2005; Rudd 2003).

Since the publication of the first shrimp EST analysis in 1999 by Lehnert et al., various related reports have been released. In most of these studies, the ESTs came from four penaeid species: *Fenneropenaeus chinensis, Litopenaeus vannamei, Marsupenaeus japonicus* and *Penaeus monodon*, which reflects the economic importance of these four shrimp. *P. monodon* and *L. vannamei* alone probably account for approximately 85% of global cultured shrimp production (Rosenberry 2006). Most of the penaeid shrimp ESTs in the published reports and NCBI dbEST are derived from these two species. The shrimp EST studies reviewed here were conducted to identify genes of interest, such as immune-related or reproduction-related genes, to investigate gene expression differences after challenge with a pathogen, or to establish EST databases (Table 1).

In the first part of this review, the findings of 16 published shrimp EST papers will be briefly summarized. In the second part, based on shrimp data downloaded from NCBI dbEST, we will present the results of our analysis of the ESTs from four penaeid shrimp. No such analysis has been done before, and this is the first attempt at a description of the penaeid shrimp transcriptome. All the analysis results are accessible at our database: http://sysbio.iis.sinica.edu.tw/page/.

The Earliest Shrimp EST Study

The first penaeid shrimp EST paper was published in 1999 by Lehnert et al. (1999). In this pioneer study, three cDNA libraries were constructed from the cephalothorax, eyestalks and pleopods of *P. monodon*. This was a small-scale EST study. A total of 172 clones were selected for sequencing, of which 88, 56 and 32 clones were from the cephalothorax, eyestalk and pleopod libraries, respectively. Among the three libraries, the pleopod ESTs contained a high propor-

Table 1 EST libraries discussed in this review

Species	Tissue source: EST no.	Reference
Fenneropenaeus chinensis	Cephalothorax: 10446	Xiang et al. 2008
Fenneropenaeus chinensis	Hemocyte: 2371	Dong and Xiang 2007
Litopenaeus vannamei	Hepatopancreas: 516 Hemocyte: 504	Gross et al. 2001
Litopenaeus vannamei	Multiple tissues, 40 libraries: 13,656	O'Leary et al. 2006
Litopenaeus vannamei	Gill, WSSV-infected: 872	Clavero-Salas et al. 2007
Litopenaeus setiferus	Hepatopancreas: 480 Hemocyte: 545	Gross et al. 2001
Marsupenaeus japonicus	Hemocyte: 635 Hemocyte, WSSV-infected: 370	Rojtinnakorn et al. 2002
Marsupenaeus japonicus	Eyestalk: 1988	Yamano and Unuma 2006
Penaeus monodon	Cephalothorax: 88 Eyestalks: 56 Pleopods: 32	Lehnert et al. 1999
Penaeus monodon	Hemocyte: 615	Supungul et al. 2002
Penaeus monodon	Hemocyte, *V. harveyi*-challenged: 446	Supungul et al. 2004
Penaeus monodon	Multiple tissues, 15 libraries: 10,100	Tassanakajon et al. 2006
Penaeu monodon	Postlarvae, normal: 7636 Postlarvae, WSSV-infected: 8345	Leu et al. 2007
Penaeus monodon	Lymphoid organ: 408 Lymphoid organ, *V. harveyi*-challenged: 625	Pongsomboon et al 2008.
Penaeus monodon	Ovary: 1051	Preechaphol et al. 2007
Penaeus monodon	Testis: 896	Leelatanawit et al. 2009

tion of mitochondrial sequences (19 of 32 clones), whereas the cephalothorax and eyestalk cDNA libraries contained only 8.0% and 1.8% mitochondrial sequences, respectively. These differences might be related to the function of the respective tissue, as the pleopod is a locomotive organ that needs a lot of mitochondria in its cells to supply energy. Lehnert et al. explained that the presence of mitochondrial sequences in *P. monodon* cDNA libraries might be due to the high A+T content in the mitochondiral genome, which could have caused the mitochondrial RNAs to be bound by the oligo-dT column used for purifying poly(A)-containing mRNA. When using an EST approach, the presence of a high proportion of mitochondrial sequences in a cDNA library hampers the rate of novel gene discovery; consequently, in the Lehnert et al. study, only a limited number of clones from the pleopod cDNA library was sequenced.

After removing the contaminated mitochondrial sequences from the ESTs, homology searches against the NCBI non-redundant (nr) protein database showed that 48 of 83 (57.8%) ESTs from the cephalothorax cDNA library, 22 of 55 (40.0%) from the eyestalk library, and 6 of 13 (46.2%) from the pleopod library had significant matches. For the remaining ESTs, three from the cephalothorax library and one from the eyestalk library had matches in GenBank EST databases. Nine cDNA clones contained a microsatellite. Overall, at least 60 genes with database matches were first identified in *P. monodon*, and 49 of these were entirely new to Crustacea. Additionally, 42 ESTs with no matches were detected.

Since many shrimp organs (including stomach, hepatopancreas, gut, gills, gonads, lymphoid organ) are located in the cephalothorax, the ESTs from this library were much more diverse than those from the eyestalk and pleopod libraries. The most abundant transcript in this library was the hemocyanin gene, which reflects the fact that shrimp hemocyanin is a highly expressed protein mainly synthesized in the hepatopancreas (Rainer and Brouwer 1993). Other highly abundant transcripts included ribosomal proteins and polypeptide elongation factors. Muscle-specific transcripts were also highly abundant in this library. Other identified transcripts included trypsin, thrombospondin and opsin. Proteins related to the nervous system were also detected.

The eyestalk of *P. monodon* is both an optical and an endocrine organ, and in this library, in addition to transcripts encoding proteins related to general cellular metabolism and muscle action, there were transcripts specifically involved in vision and the neuroendocrine system, including arrestins, phospholipase C, clathrin and phenylalanine/tryptophan hydroxylase. *P. monodon* contained two different classes of arrestins, as has previously been reported for *Drosophila* (Matsumoto and Yamada 1991) and other insect (Raming et al. 1993). (Since the Lehnert et al. study, a much larger eyestalk EST library has been constructed from *Marsupenaeus japonicus*. The *M. japonicus* library contained 1988 ESTs and emphasized the identification of reproduction-related genes; please see below.)

Identification of Immune-Related Genes Through EST Analysis

Subsequent to the Lehnert et al. report, a second shrimp EST study was conducted using two closely related species of penaeid shrimp, the Pacific white shrimp, *L. vannamei*, and the Atlantic white shrimp, *Litopenaeus setiferus* (Gross et al. 2001). Four different cDNA libraries were constructed from two immune-related tissues (hemocytes and hepatopancreas) from each shrimp species. The Gross et al. study was designed to identify immune-related genes as well as compare the four libraries. A total of 2,045 clones were sequenced, and a relatively high proportion of ESTs (77%) had homologous genes in the NCBI nr database. Contamination by mitochondrial 16S and nuclear 18S rRNAs was noted. A total of 268 ESTs representing 44 immune-related genes were identified. These immune-related ESTs were abundant in both of the hemocyte libraries (27.6% in *L. setiferus* and 21.2% in *L. vannamei*), but were far less common in the hepatopancreas libraries (4.4% in *L. setiferus* and 5.6% in *L. vannamei*). Among the immune-related ESTs, antimicrobial peptides (AMP) were most prominent (64%; 172/268) although they were restricted to the hemocyte libraries. Lectins (6.7%; 18/268) were the largest group of immune ESTs in the hepatopancreas, and they were only identified in the hepatopancreas libraries. Other immune-related ESTs encoded serine protease, protease inhibitors, heat shock proteins, clottable protein and β-1,3 glucan binding protein. Most of these were only present in one tissue, with only two of them (clottable protein and ferritin) being detected in both tissues.

Comparison between hemoctye and hepatopancreas ESTs showed that these two organs had different gene expression patterns, reflecting their functional difference in shrimp. For example, in addition to the differential abundance of immune-related ESTs, ESTs encoding digestive enzymes and fatty acid binding proteins were only present in the hepatopancreas libraries. Another comparison showed that while the expression patterns of the two hemocyte libraries were quite similar, there was considerable diversity between the expression patterns of the two hepatopancreas libraries. This suggests that gene expression in the hepatopancreas is more easily regulated or influenced. The reasons for such a discrepancy are unknown, but it might be related to the different cellular composition and cell/tissue functions, with the cells in the hepatopancreas being more heterogeneous and diverse in terms of function and composition.

In crustacean, the hemocytes are the major immune effector cells, and they play important roles in both humoral and cellular-mediated immune responses, synthesizing and releasing many immune effector molecules (Sritunyaluksana and Söderhäll 2000; Iwanaga 2002; Smith et al. 2003). Supungul et al (2002) were the first to use an EST approach to analyze gene expression profiles in hemocytes of *P. monodon*. A total of 615 EST clones were obtained, and 51% (315/615) of the ESTs had homologous genes in GenBank, whereas the remaining ESTs (49%) had no matches. The presence of multiple mitochondrial sequences in this EST library was noted. The matched ESTs were classified into several functional categories according to Adams et al. (1991). The most abundant group of ESTs encoded proteins belonging to the category "gene expression and protein synthesis", and these accounted for 17.7% of total ESTs (109/615). Proteins in the category "defense and homeostasis" accounted for 8.9% (55/615) of total ESTs. These defense-related proteins included components of the clotting and prophenoloxidase systems, antioxidative enzymes, AMPs, serine proteinase inhibitors and heat shock proteins (hsp). Three ESTs were found to encode full-length proteins, two of which were AMPs (antilipopolysaccharide factor (ALF), penaeidin), while the third was a heat shock protein (cpn 10). This was the first time that ALF and cpn 10 had been identified in penaeid shrimp. Before this study, ALF had only been identified in horseshoe crabs, and invertebrate cpn10 had not been isolated.

Two years later, Supungul et al. (2004) published a related study in which another 447 EST clones generated from *Vibrio harveyi*-challenged *P. monodon* hemocytes were analyzed together with the above 615 clones from the normal hemocyte library. The 2004 report focused on the identification of AMP homologues, including ALF, crustins and penaeidins. From an analysis of the 1,062 ESTs from both normal and *V. harveyi*-challenged shrimp libraries, 115 clones (10.8%) encoding 30 different immune-related proteins were identified. Transcripts representing AMP were the most abundant class of immune-related genes, accounting for 29.2% and 64.0% of the immune genes discovered in the normal and challenged libraries, respectively. Penaeidins were the major AMP in the normal library (36.84%) followed by crustins and ALF (26.3% each). Conversely, in the challenged library, ALF outnumbered (50.0%) both crustins (25.0%) and penaeidins (18.8%), suggesting that *V. harveyi* infection induces the expression of ALF. Sequence alignment further showed that at least five ALFs and four crustin homologues were present in the libraries, whereas only one penaeidin (penaeidin-3) was identified. A recent study showed that these ALF homologues were encoded by two different genetic loci (Tharntada et al. 2008). ALF, crustin and penaeidin-3 genes were mainly expressed in hemocytes. In *V. harveyi*-challenged shrimp, ALF expression was increased, whereas expression of crustin and penaeidin-3 was decreased (Supungul et al. 2004).

In conclusion, these two studies showed the power of an EST approach to identify immune genes in *P. monodon* hemocytes as well as characterize AMP isoforms. Further, comparative EST analysis between normal and *V. harveyi*-challenged hemocyte libraries revealed the immune gene expression changes in the challenged shrimp. (EST analyses of hemocytes in *F. chinenesis* and in normal and WSSV-infected *M. japonicus* have since been published; please see below)

Identification of Reproduction-Related Genes

Yamano and Unuma (2006) used ESTs from eyestalk of kuruma prawn, *M. japonicus*, to analyze gene expression profiles in order to identify the genes involved in female reproduction. They sequenced a total of 2,304 clones from the 5′ end, and after trimming the vector sequences, 1,988 ESTs greater than 100 bp sequences were subjected to further analysis. EST assembly yielded 136 contigs from 738 ESTs (37.1%) and 1,250 singleton ESTs (62.9%), generating 1,386 unique sequences. Homology searches identified only 231 of these unique sequences, while the remaining 1,155 sequences were unknown. Of the identifiable ESTs, most were ribosomal RNAs and mitochondrial respiration enzymes. Only four ESTs with reproduction-related functions were identified, three of which showed close similarities to the crustacean hyperglycemic hormone (CHH) peptide family. These three ESTs were distinct from the known kuruma shrimp eyestalk hormones, and sequence analysis suggested they represented novel pigment-dispersing hormone (PDH), molt-inhibiting hormone (MIH) and CHH. The other reproduction-related EST was homologous to farnesoic acid O-methyltransferase, an enzyme that produces methyl farnesoate (MF). According to Laufer et al (1987), MF is a crustacean equivalent of insect juvenile hormone, and it is related to crustacean growth and maturation.

Identification of Sex-Related Genes

The research group led by Menasveta has concentrated on using the EST approach to identify sex-related genes with the purpose of understanding the molecular mechanisms of gonad maturation and sex determination in *P. monodon* (Preechaphol et al. 2007; Leelatanawit et al. 2009). This work is of practical importance because it might be used to improve the poor reproductive efficiency of domesticated *P. monodon* broodstocks. For these studies, *P. monodon* brooders were collected from the wild, mRNA was

extracted from their gonads (ovaries or testes), and cDNA libraries were constructed. Here, we will review the study of ovarian ESTs first. A total of 1051 clones were sequenced from the 5′end, and assembled into 87 contigs and 472 singletons, producing 559 unique sequences. Homology searches showed that 743 ESTs (70.7%) had matches in the GenBank (BLASTN and BLASTX, *E* value $<10^{-4}$), whereas the remaining ESTs were regarded as unknown genes (29.3%, *E* value $>10^{-4}$). Of the matched ESTs, the ESTs encoding peritrophin (8.3% of total ESTs) and thrombospondin (TSP; 7.5%) were most abundant. Several full-length transcripts encoding proteins with important functions were identified, including cyclophilin, profillin, thioredoxin peroxidase and chromobox protein. Additionally, 25 sex-related genes were identified. The expression of these genes in *P. monodon* ovaries and testes was investigated. In a 4-month-old *P. monodon*, TSP mRNA was preferentially expressed in ovaries (Preechaphol et al. 2007), while in a broodstock-sized *P. monodon*, it was specifically expressed in ovaries and not at all in the testes (Leelatanawit et al. 2004). Other gene transcripts that expressed only in the broodstock-sized *P. monodon* ovaries included female sterile and ovarian lipoprotein receptor. Several other genes, including CBX, phosphatidylinositol 4 kinase, thioredoxin peroxidase and Usp9X (ubiquitin specific proteinase 9, X chromosome), were preferentially expressed in the ovaries. The expression pattern of Usp9X gene was quite unique among the investigated genes: it was preferentially expressed in ovaries of broodstock, but was specifically expressed in the ovaries of a 4-month-old shrimp.

From the testis cDNA library, 896 clones were sequenced, and 606 ESTs (67.6%) had matches in GenBank (*E* value $<10^{-4}$). The 896 ESTs were assembled and clustered into 109 contigs and 492 singletons, yielding 601 unique sequences. There were no highly abundant ESTs observed in this library. Several full-length cDNAs encoding proteins functionally related to testicular development were reported, with dynactin subunit 5, cdc2, small nuclear ribonucleoprotein polypeptide G, small ubiquitin-like modifier 1 (SUMO-1), mitotic checkpoint BUB3 being reported for the first time in *P. monodon* and transformer-2 (PMTra-2), a protein involved in sex determination cascades, being reported for the first time in crustaceans. The expression levels of 51 genes were analyzed by RT-PCR in gonads of male and female juvenile and broodstock *P. monodon*. The gene *P. monodon* testis-specific transcript 1 (PMTST1), which is homologous to low molecular weight neurofilament protein XNF-L, was only expressed in testes and not in ovaries. Two genes, multiple inositol polyphosphate phosphatase 2 (MIPP2) and heat shock-related 70 kDa protein 2 (HSP70-2), were preferentially expressed in testes. Conversely, 34 of the examined genes showed preferential expression in ovaries rather than in testes. The tissue distribution pattern of several selected genes were analyzed. PMTST1, MIPP2, PMTra-2 and HSP70-2 were specifically or highly expressed in gonads, whereas the other genes were constitutively expressed in all of the examined tissues.

In conclusion, by EST analysis, this research group identified a large number of *P. monodon* genes that were specifically or highly expressed in ovaries or testes or in both. These genes were expressed at low levels or not at all in other tissues. These sex-related genes are potentially involved in the gonad development of *P. monodon*, and further study of these genes will no doubt shed new light on the molecular mechanisms that control the development and maturation of gonads. In addition, these genes might be used as biomarkers to investigate the differences of gonad development/maturation between domesticated and wild-caught *P. monodon* broodstocks, and to evaluate the degrees of gonad development/maturation in domesticated *P. monodon*.

Fenneropenaeus chinensis EST Collection

F. chinensis, another economically valuable penaeid shrimp, is one of most important mariculture species in China. Therefore, it is not surprising that the *F. chinensis* ESTs in NCBI dbEST all came from the Chinese research group led by Jianhai Xiang. A total of 10,446 ESTs, sequenced from the 5′ end, were first reported in a conference proceedings (Xiang et al. 2008), but it seems that they began this work as early as 2000, and in the next year, they already got more than 10,000 ESTs (Xiang et al. 2002). This collection of ESTs was generated from a cephalothorax cDNA library build from a single adult female *F. chinensis*. Assembly and clustering of these ESTs produced 1,399 contigs and 1,721 singletons, yielding 3,120 unique sequences. After homology searches against the GenBank nr database using BLASTN (*E* value $\leqq 10^{-7}$) and BLASTX (*E* value $\leqq 10^{-3}$), 1,373 (44%) unique genes had matches. The most abundant EST (504 clones) encoded peritrophin-like protein 1. The other highly abundant ESTs included transmembrane-4-superfamily-8 (309 clones), thrombospondin (163 clones), elongation factor 1-α (143 clones), tubulin β (82 clones) and hemocyanin (80 clones). For the unknown genes, the authors searched against the InterPRO database to identify possible features and tentative function. Xiang et al found that the most abundant ESTs had the features of Type I antifreeze protein, and suggested that this might explain how *F. chinensis* can endure lower temperatures than other penaeid shrimp.

The authors noted that through their EST analysis, they were able to identify the trehalosephosphate synthase gene for the first time in shrimp. In addition to gene discovery

and estimating EST abundance, the authors identified simple sequence repeats (SSRs) in this EST collection: a total of 324 SSRs were found in 223 unique sequences out of the 3,120 unique sequences. Primers were then designed based on several ESTs-SSR core sequences, and eight primer pairs were polymorphic in 200 samples of *F. chinensis*. Using this EST database, together with other sequences isolated through SSH, this group conducted a microarray study to profile gene expression changes in WSSV-infected *F. chinensis* (Wang et al. 2006).

Two members of the same research group (Dong and Xiang 2007) constructed another collection of *F. chinensis* ESTs from hemocytes. A total of 2,371 EST clones were successfully read from the 5′ end. Of these, 1,739 (73.34%) ESTs were clustered into 339 contigs, while the remaining 632 (26.66%) were singletons, yielding 971 unique sequences. Homology searches within the GenBank database with BLASTN (E value $< 2.0\times10^{-5}$) and BLASTX (E value $<1.0\times10^{-5}$) showed that 228 contigs and 254 singletons had matches, whereas 489 (50.36%) of the unique sequences showed no similarity to any known sequences in GenBank. Among the known genes, 34 genes containing 177 ESTs were related to immune defense function. These genes could be classified into five categories based on their functions in the shrimp immune system. The first category, "anti-microbial peptides," included penaeidins, thymosin and ALF as well as several putative anti-microbial peptides. These were encoded by 13 unique sequences consisting of 71 ESTs. This category had the highest abundance of immune EST (38%). The second category, "prophenoloxidase activating system," contained 11 unique sequences from 44 ESTs, encoding prophenoloxidase, serine proteinase, serine proteinase inhibitor and SOD-protein. Category three, "clotting proteins," including transglutaminase, thrombospondin and lectin, contained five unique sequences from 39 ESTs. The fourth category, "inter-cellular signal transduction," included peroxinectin and integrin, and consisted of three unique sequences from five ESTs. Lastly, there were two "chaperone proteins," HSP70 and thioredoxin peroxidase, which were represented by two unique sequences and 19 ESTs.

EST Databases Generated from Multiple Tissues: *L. vannamei* and *P. monodon*

Most of the above shrimp EST studies used small-scale EST libraries with ~2,000 ESTs, and these ESTs were from one and three tissues. In addition, most of the tissue sources were immune-related. However, two study groups have constructed much larger libraries and generated ESTs from multiple tissues from two different shrimp species reared under normal as well as environmentally stressful or pathologically challenged conditions. Both of these studies also used subtraction or normalization as enrichment procedures to increase the possibility of obtaining rare transcripts. Although subtraction/normalization was only used in these two studies to discover immune functions genes in immune-related tissues, these enrichment procedures can also be usefully applied to generate ESTs from other tissue sources.

Multiple-tissue *L. vannamei* ESTs

Currently, the most ambitious shrimp EST project aims to characterize the entire shrimp transcriptome. This work is being carried out by the group led by P.S. Gross at the Marine Biomedicine and Environmental Sciences Center, Medical University of South Carolina, USA, and its goal is to sequence 100,000 ESTs from Pacific white shrimp *L. vannamei* from both the 5′ and 3′ ends (for a total of 200,000 sequences). This group released their initial analysis of 13,656 ESTs (O'Leary et al. 2006) at the symposium "Genomic and Proteomic Approaches in Crustacean Biology." To incorporate as many genes as possible in their EST collection, they used 6 different tissues (hemocyte, hepatopancreas, gill, lymphoid organ, eyestalk, and ventral nerve cord) and prepared two different kinds of cDNA library. The six non-normalized cDNA libraries constructed from the above tissues contributed 7,896 ESTs, and the 34 different suppression subtractive hybridization (SSH) cDNA libraries made from immune-related tissues (hemocyte, gill and hepatopancreas) produced another 5,760 sequences (Robalino et al. 2007). These SSH cDNA libraries were generated from shrimp induced with WSSV, dsRNA, and inactivated microbes to specifically enrich for immune-relevant genes. The total number of 13,656 sequences, all read from the 5′ end, were assembled into 7,466 unigenes represented by 1,981 contigs and 5,485 singletons. Like other shrimp EST studies, only a fraction (5,162, 38%) of this collection showed significant similarity to known proteins within the NCBI protein database using a BLASTX (E value $\leqq10^{-4}$).

As this group focused on the identification of the molecular components behind the immune defense system of the penaeid shrimp, most of the ESTs in this collection were derived from immune tissues, including hemocytes, hepatopancreas, gill and lymphoid organ. Hemocytes are the primary defense cells against invading pathogens in shrimp, and therefore, a high fraction the ESTs (nearly 40%) were from this source. In hemocyte ESTs, antimicrobial peptide genes were present in abundance, and genes encoding protease inhibitors, coagulation factors, lysozyme, and heat shock proteins were also identified. Additionally, two important genes that encode immune signaling molecules, STAT and I-kappa-B kinase, were present in the hemocyte ESTs, and this was the first time that these genes

were identified in shrimp. Another important immune regulator gene, imd (immune deficiency), was also first discovered in shrimp from the LO library.

The O'Leary et al. (2006) report also outlined the strategy that was developed for enhancing novel gene discovery for further EST sequencing. As noted above, their stated goal is to collect 200,000 ESTs, and for such a large-scale EST project it is crucial to avoid repeatedly sequencing the same gene. When sequencing a specific cDNA library for the purpose of identifying genes, two important factors determine the rate at which novel genes can be identified: the complexity of the cDNA library and the abundance of a gene transcript. As the number of sequenced ESTs increases, the chance of discovering a novel gene decreases, and to increase the chance of novel gene discovery, it is necessary to remove the ones that were already sequenced from the library. O'Leary et al. therefore used annotation to identify the highly abundant genes, and then used these genes as hybridization probes to screen out other cDNA clones of the same genes. The remaining un-hybridized clones were then selected for sequencing. The sequences and complete annotations of all the ESTs generated in this project are available at www.marine-genomics.org. The number of released *L. vannamei* ESTs at this time is 176,198, and these are assembled into 14,548 contigs. These ESTs are also available at NCBI dbEST.

Multiple-Tissue *P. monodon* ESTs

Tassanakajon et al. (2006) performed a large-scale EST project in *P. monodon* for the purpose of gene discovery. In this project, cDNA libraries were prepared from six different tissues (eyestalk, hepatopancreas, hematopoietic tissue, hemocyte, lymphoid organ and ovary) of normal shrimp to identify tissue-specific genes. These tissues were selected because they are involved in immune defense, growth and sex differentiation. Additionally, shrimp under stressful conditions were used to prepare several libraries in order to discover genes responding to stresses, such as heat treatment (35°C for 1 h) and pathogen challenge (WSSV, YHV, and *V. harveyi*). In addition to standard cDNA libraries, normalized libraries for enriching rare transcripts were also constructed from the hepatopancreas and lymphoid organ. However, as shown by the authors, the normalization process did not effectively remove the highly expressed transcripts, such as hemocyanin and cathepsin L, from the corresponding libraries, respectively. In total, 15 cDNA libraries (13 standard libraries and two normalized libraries) were constructed, and EST clones from each library were randomly picked and sequenced from the 5′ end. After trimming away the vector and low-quality sequences, the remaining 10,100 high quality sequences were clustered and assembled into 917 contigs and 3928 singletons, resulting in 4845 unique sequences. The unique sequences were searched for homologous genes against the GenBank nr database using BLASTX (E value $<10^{-4}$). Less than half (49%) of the total EST clones had matches. EST clones representing mitochondrial sequences were abundantly represented. Other abundant ESTs encoded thrombospondin, elongation factor I-α, ovarian peritrophin, ALF and a number of hypothetical proteins. Functional categorization of these matched ESTs according to Adams et al. (1993) showed that most of them (14%) encoded hypothetical proteins, whereas only 3.9% were related to defense or homeostasis. The other major EST groups were related to metabolism (6.2%) and gene expression, regulation and protein synthesis (5.5%).

Comparative EST analysis of four different hemocyte cDNA libraries (normal, WSSV-challenged, *V. harveyi*-challenged, heat-stressed shrimp) shed some light on the gene expression changes in hemocytes under stressful conditions. A notable change was that the ESTs encoding ribosomal proteins greatly increased from 8.2% in the normal library to 17.6% and 23.9% in the *V. harveyi* and WSSV-challenged libraries, respectively. Although the abundance of immune-related ESTs was relatively constant among the four libraries (9.7–12.5%), the corresponding profiles were changed. For example, ESTs encoding antimicrobial-related molecules were from 3% in the normal library to 6.3 and 7.2% in the WSSV and *V. harveyi*-challenged libraries, respectively.

Several tissue-specific transcripts were identified, such as MIH and PDH in the eyestalk library, and hemocyanin in hepatopancreas. In the hematopoietic tissue library, the ESTs encoding α-NAC protein and dystrobrevin-like protein were discovered for the first time in crustaceans. Both of these proteins are involved in blood cell differentiation (Davidson and Swalla 2002; Greener and Roberts 2000).

ESTs are valuable resources not only for gene discovery but also for the development of useful genetic markers, including microsatellites and single nucleotide polymorphism. Tassanakajon et al identified 997 ESTs with unique microsatellites among the 10,100 high quality ESTs. The distribution of these microsatellite-containing ESTs varied across the different libraries and only a few of these ESTs (74, 7.4%) were known genes. These *P. monodon* EST data have been submitted to NCBI dbEST and are also available at the web site http://pmonodon.biotec.or.th.

In a more recent paper (Pongsomboon et al. 2008), the members of this *P. monodon* EST project reported the results of comparative EST analysis between the two lymphoid organ cDNA libraries from normal and *V. harveyi*-challenged shrimp. Together with the gill, hepatopancreas and hemocyte, the lymphoid organ is considered to be one of the tissues with immune defense functions (Burgents et al. 2005) in penaeid shrimp. EST analysis of

lymphoid organ libraries might therefore result in the discovery of novel immune-related genes as well as further our understanding of the immune function of the lymphoid organ. Analysis of 408 and 625 clones from the normal and infected libraries, respectively, showed that although both libraries had approximately the same percentage of immune-related ESTs (~15%), their EST patterns were different, suggesting *V. harveyi* infection alters gene expression patterns in the shrimp lymphoid organ. The highly abundant EST clones in each library were identified. If the ESTs of mitochondrial sequences are not considered, then the most abundant ESTs in the normal library were cathepsin L (26 clones, 6.4%) and cathepsin B (15 clones, 3.7%), whereas in the infected library they were peritrophin (26 clones, 4.2%) and thrombospondin (3.7%). The ESTs of cathepsin L (ten clones, 1.6%) and cathepsin B (7 clones, 1.1%) were also detected in the infected library, but they were less abundant. Conversely, peritrophin and thrombospondin ESTs were only present in the infected library. Real time RT-PCR further showed that, after *V. harveyi* or WSSV infection, in shrimp lymphoid organs, the expression levels of cathepsin L and cathepsin B were only slightly changed, while the expression levels of peritrophin and thrombospondin were significantly up-regulated. Therefore, the authors concluded that in *P. monodon* lymphoid organ, cathepsin L and cathepsin B are constitutive genes with a high expression level that might be important to the normal physiological function of lymphoid organ; conversely, the induced expression of peritrophin and thrombospondin by pathogen infection suggests that both genes might play an important role in the lymphoid organ's immune function.

Using EST Analysis to Profile Gene Expression Change in WSSV-Infected Shrimp

WSSV is a highly contagious viral pathogen of penaeid shrimp and it can cause high levels of mortality within just a few days after infection. To understand the mechanisms that underlie this virulence, detailed studies of the interactions between shrimp and WSSV at both transcriptional and translational levels are needed. At the transcriptional level, much work has been done to identify shrimp genes that are regulated by WSSV infection (Astrofsky et al. 2002; Dhar et al. 2003; Pan et al. 2005; He et al. 2005; Wang et al. 2006; Robalino et al. 2007). Comparative EST analysis between normal and WSSV-infected shrimp is another approach that has been adopted by several groups, including ours (Rojtinnakorn et al. 2002; Tassanakajon et al 2006; Leu et al. 2007). The results of Tassanakajon et al. have already been discussed above, here we will discuss the other two studies.

Rojtinnakorn et al. (2002) were the first to use the EST approach to study gene expression in hemocytes of *M. japonicus* in response to WSSV infection. A total of 635 and 370 clones were sequenced from normal and infected libraries, respectively. Homology searches against sequences in GenBank using BLASTX and BLASTN showed that 284 (44.7%) normal and 174 (47.0%) infected ESTs were significantly similar to deposited sequences. ESTs that represented mitochondrial sequences were abundant in both libraries. Ribosomal protein ESTs were also abundant, but their abundance differed in the two libraries.

In total, the matched ESTs encoded 152 different proteins. Most of these proteins were differently distributed between the two libraries and some were only present in one library. This suggested that WSSV infection strongly affects gene expression in *M. japonicus* hemocytes, and these was evidence that this was particularly true of the defense-related genes. Of the 152 annotated proteins, 28 proteins were identified as defense-related molecules, and 15 of these were reported here for the first time in penaeid shrimp. These defense-related proteins were involved in the prophenoloxidase (proPO) system (proPO, serine proteases and protease inhibitors) and the clotting process (tranglutaminase and clottable protein). In addition, three antibacterial peptides (bactinecin-11, penaeidin-2 and lysozyme c type) and six apoptotic and tumor-related proteins were discovered. Several cell adhesion-related proteins, which were considered to be putative defense proteins by the authors, were also identified, including β-integrin, cell adhesion molecule and three types of collagens.

The abundance of ESTs encoding defense-related proteins was increased from 2.7% in normal shrimp ESTs to 15.7% in infected shrimp ESTs. This difference was mainly because all the ESTs encoding protease inhibitors and tumor-related proteins were detected only in the infected library, and also because ESTs encoding for apoptotic-related proteins were present at high levels in the infected library. In contrast to the following two reports, Rojtinnakorn et al. identified no WSSV genes in the infected library. This was probably due to the lower expression levels of WSSV genes compared with the cellular genes in hemocytes.

In Leu et al. (2007), our group reported on a large-scale comparative EST analysis between two cDNA libraries constructed from normal and WSSV-infected *P. monodon* postlarvae. There were two reasons why we chose postlarvae as our experimental material: First, WSSV is a systemic virus able to infect many shrimp tissues and organs, and we therefore wanted to investigate the molecular response of a whole animal instead of just a specific tissue or organ; and second, when we began this project, no shrimp ESTs from the postlarval stages were available, so we thought our chances of discovering new shrimp genes would be increased. A total of 6,964 and 7,686 cDNA clones were sequenced from the 3' end from the normal and infected libraries, respectively. After base-

calling, vector sequence trimming and eliminating low quality sequences and sequences from WSSV and other sources, a total of 6,658 and 7,276 high quality 3′ ESTs were generated from the normal and infected libraries, respectively. Next, after these 3′ end sequences were assembled and annotated, to further increase the overall EST annotation rate, we randomly chose cDNA clones from 3′ EST contigs that had no matches in the NCBI nr database for 5′ sequencing. Finally, 978 and 1,069 high quality 5′ ESTs were generated and combined with the high quality 3′ ESTs for assembly and annotation. The 15,981 high quality ESTs were assembled and clustered into 9,622 unique sequences, of which 1,364 were contigs and 8,258 were singletons.

Although our main purpose was to identify novel shrimp genes as well as to profile gene expression changes in WSSV-infected shrimp, EST analysis of the infected library helped us to identify 48 WSSV genes represented by 167 ESTs (Wang et al. 2007). No WSSV sequences were found in the normal shrimp library. This confirmed that the shrimp used to construct this cDNA library were negative for WSSV infection.

Homology searches using BLASTX (E value $<10^{-10}$) showed that 2,027 (21.07%) unique sequences were similar to known protein sequences in the NCBI nr protein database, and 2,026 (21.06%) unique sequences had GO annotations in the UniProt database. In the normal library, this equated to 3,022 (45.30%) and 2,870 (43.02%) ESTs with matches in the NCBI and UniProt databases, respectively. In the infected library, the numbers of matches were 3,338 (45.74%) and 3,202 (43.88%), respectively.

For a global view of the overall gene expression changes, we first analyzed the GO categories in the normal and infected libraries. Our results showed a significant statistical difference (Fisher's exact test; $P<0.05$) between the two libraries for several biological processes, including carbohydrate metabolism, signal transduction, response to external stimulus, microtubule-based movement, phosphate metabolism, transport and protein metabolism. Two molecular functions, structural molecular activity and carbohydrate binding, appeared to be elevated significantly. In the cellular component group, only the cytoskeleton category was significantly different.

A more detailed comparison between the two libraries was then performed by listing and comparing the 50 most abundant ESTs in each library. In the normal library, most of these abundant genes were classified into four major groups: proteins involved in ATP metabolism, proteins involved in translation, proteins highly or specifically expressed in muscle, and proteins highly or specifically expressed in the hepatopancreas. These results probably reflect the fact that the shrimp postlarvae were in an active growth stage that would have necessitated the expression of genes related to these functions. In the infected library, the most abundant genes no longer included the hepatopancreas-related proteins, although the other three protein groups were still highly represented. The most highly represented groups in the infected library were the immune-related proteins with chitin-binding or lectin domains, proteins related to glycolysis, cuticle-related proteins and several different actin genes.

Fisher's exact test identified 23 significantly increased and 25 significantly decreased genes in the infected library. The genes with increased abundance included four proteins with a chitin binding Peritrophin-A domain, seven cuticle-related proteins, four proteins involved in oxidative phosphorylation, two glycolytic enzymes, two ribosomal proteins, thioredoxin-1, actin and a protein with C-type lectin (CTL) and CTL-like domain. The decreased-

Table 2 A general description of the shrimp EST assemblies

Species	*L. vannamei*	*P. monodon*	*F. chinensis*	*M. japonicus*
EST Sequence Reads	156,985	25,661	10,446	3,156
Filtered out[a]	1,117	7,533	476	120
Assembly				
TUC[b]	14,239 (9.8)	2,326 (5.3)	1,036 (7.2)	291 (5.2)
TUS[c]	14,979 (9.6%)	5,725 (31%)	2,209 (22%)	1,527 (50%)
Total TUGs (TUC + TUS)	29,218	8,051	3,245	1,818
Annotation				
Matched UniProt ID	4,873	1,914	787	487
Annotated TUGs[d]	8,149 (28%)	2,738 (31%)	965 (30%)	534 (29%)

[a] Sequence reads of low quality (short reads, low complexity reads), contaminated vector arms, and mitochondrial sequences were filtered out

[b] The value in parenthesis represents the mean number of ESTs per TUC

[c] The value in parenthesis represents the percentage of TUS in filtered ESTs

[d] The value in parenthesis represents the percentage of annotated TUGs.

abundance genes included two SCP calcium-binding proteins, five cytoskeleton/motility-related proteins, three proteins involved in oxidative phosphorylation, three ribosomal proteins, seven hepatopancreas-related proteins (four digestive enzymes, hemocyanin and two immune-related proteins, PmAV and ferritin), opsin and cAMP responsive element binding protein-like 2.

The conclusion suggested by our comparative EST analyses was that, in postlarval shrimp, WSSV infection strongly modulates the gene expression patterns in several organs or tissues, not only the hepatopancreas, muscle, and cuticle (as noted above), but also the eyestalk. Even though neither hepatopancreas nor muscle are preferential targets for WSSV, the dramatic changes in gene expression levels strongly suggests that the functions of these two organs are nevertheless affected by WSSV infection. Our data further showed that several basic cellular metabolic processes are likely to be affected, including oxidative phosphorylation, protein synthesis, the glycolytic pathway, and calcium ion balance. Lastly, our data suggested that a group of immune-related chitin-binding protein genes were also strongly up-regulated after WSSV infection, although whether this group of proteins is involved in shrimp immune response or has any antiviral activity needs to be confirmed.

The EST sequences generated by Leu et al. (2007) have been submitted to NCBI dbEST. In addition, all the sequence data and analysis results are accessible at our database: http://xbio.lifescience.ntu.edu.tw/pm/.

The final report in this review was conducted by Clavero-Salas et al. (2007), who performed an EST analysis of gills from *L. vannamei* infected with WSSV, but did not investigate non-infected shrimp. A total of 872 clones were sequenced from the 5' end, and after vector-trimming, the 601 good quality ESTs were assembled into 79 contigs and 197 singletons, yielding 276 unique sequences. Homology searches using BLASTN and BLASTX (*E* value $<10^{-2}$) against the NCBI nr database showed that 87% (522/601) of the ESTs had matches in GenBank, whereas 13% (79/601) were unknown. These matched ESTs encoded 276 different proteins. The most abundant ESTs encoded 40S ribosomal protein S13. Of these matched ESTs, 148 ESTs showed homology to WSSV sequences, and 119 of these could be assembled into 12 contigs and the other 29 remained as singletons; giving a total of 41 different genes. Several full-length cDNA sequences were reported in this study, including Keratinocyte associated protein 2, selenoprotein M, profilin, prohibitin and oncoprotein nm23. Lastly, the authors used RT-PCR to check whether the expressions of these genes, as well as ferritin, were affected by WSSV infection. Based on their results, it seems that the expressions of both selenoprotein M and prohibitin were down-regulated at 1 and 3 h post-infection, whereas the expression of other genes were largely unchanged.

Table 3 Top 10 list of the assembled contigs, ranked by the number of ESTs per contig

	L. vannamei				*P. monodon*			
	Contig Info		Annotation[a]		Contig Info		Annotation[a]	
	ESTs/ TUC	Contig length	Protein/Gene description	Species	ESTs/ TUC	Contig length	Protein/Gene description	Species
1	1,162	1,359	60S ribosomal protein L13	*Spodoptera frugiperda*	167	3,663	Thrombospondin	*Penaeus japonicus*
2	1,097	1,501	60S ribosomal protein L14	*Rattus norvegicus*	140	1,600	Elongation factor 1-alpha	*Upogebia major*
3	995	3,007	60S ribosomal protein L11	*Caenorhabditis elegans*	119	1,196	NADH-ubiquinone oxidoreductase chain 1	*Penaeus monodon*
4	945	1,784	60S ribosomal protein L26	*Littorina littorea*	95	2,307	Fructose-bisphosphate aldolase	*Homalodisca coagulata*
5	942	1,979	Novel EST [b]		89	1,539	Novel EST [b]	
6	841	1,737	Hemocyanin C chain	*Panulirus interruptus*	84	1,492	Myosin light chain 1	*Culex quinquefasciatus*
7	828	1,953	Arginine kinase	*Penaeus japonicus*	81	719	Cytochrome c oxidase subunit 3	*Penaeus monodon*
8	819	576	60S acidic ribosomal protein	*Artemia salina*	76	4,155	Putative uncharacterized protein	*Culex quinquefasciatus*
9	730	795	60S ribosomal protein L13	*Danio rerio*	74	1,906	Enolase	*Penaeus monodon*
10	647	1,713	Hemocyanin C chain	*Panulirus interruptus*	73	1,324	Novel EST [b]	

[a] The annotation are based on the closet matching sequence in UniProtKB by BLASTX search ($E < 10^{-5}$) or in TIGR gene indices by BLASTN search ($E < 10^{-15}$)

[b] "Novel EST" signifies a TUG with no matching sequence in the UniProtKB or the TIGR gene indices

Analysis of EST Data for Four Penaeid Shrimp Species

ESTs from four penaeid shrimp species, *L. vannamei*, *P. monodon*, *F. chinensis* and *M. japonicus*, were collected from NCBI dbEST and submitted to Bio301 (http://bio301.iis.sinica.edu.tw; Chen et al., unpublished results) for contig assembly and functional annotations. The ESTs were cleaned to strip off contaminated DNA sequences, such as vector arms, low-quality short reads and mitochondrial sequences. The assembly was performed by TGICL (Pertea et al. 2003) with CAP3 (Huang and Madan 1999). Assembled contigs, including tentative unique singletons (TUS) and tentative unique contigs (TUC), were subjected to homology searches against UniProt (UniProt Consortium 2008) and the TIGR gene indices (Quackenbush et al. 2001) by BLASTX (*E* value $<10^{-5}$) and BLASTN (*E* value $<10^{-15}$), respectively. Tentative unique genes (TUGs) without matched sequences were recognized as novel genes.

Table 2 summarizes the number of input EST reads and the corresponding assemblies. The top ranked contigs in *L. vannamei* and *P. monodon* are listed in Table 3. These TUCs account for about 6% of the total numbers of ESTs. Table 4 summarizes the annotations for all TUGs in GO

Table 4 The functional GO analysis of the TUGs assembled from the *L. vannamei* and *P. monodon* ESTs

		L. vannamei	*P. monodon*
Cellular component		6,328	2,522
	Cell part	5,811	2,277
	Envelope	673	307
	Extracellular matrix	48	24
	Extracellular region	637	281
	Macromolecular complex	1,735	854
	Membrane-enclosed lumen	492	132
	Obsolete cellular component	0	0
	Organelle	4,246	1,496
Molecular function		6,838	3,036
	Antioxidant activity	37	21
	Binding	4,761	2,046
	Catalytic activity	3,087	1,377
	Chaperone regulator activity	1	0
	Enzyme regulator activity	178	77
	Molecular transducer activity	185	126
	Motor activity	105	73
	Nutrient reservoir activity	5	2
	Obsolete molecular function	2	10
	Structural molecular activity	1074	654
	Transcription regulator activity	227	91
	Translation regulator activity	113	59
	Transporter activity	719	377
Biological process		5,836	2,604
	Biological regulation	1,405	457
	Cellular process	4,913	2,113
	Developmental process	1,057	259
	Establishment of localization	1,572	596
	Growth	174	33
	Localization	1,678	632
	Metabolic process	4,118	1,896
	Multicellular organismal process	1,127	319
	Obsolete biological process	0	47
	Reproduction	411	91
	Response to stimulus	688	215
	viral reproduction	11	8

terms, according to the sequence homologue matches in the Bio301 pipeline (Chen et al.). The EST analysis results and the 200,000 EST data are available at the Penaeus Genome Database (PAGE; http://sysbio.iis.sinica.edu.tw/page/). PAGE allows retrieval of the assembled contigs and ESTs with detailed annotations either by a full-text keyword search on the sequence database, or by a sequence search using BLAST. The sequence datasets can be downloaded in FASTA format for use in other applications. There are also plans to implement additional features, including PCR primer/ probe design, microarray design, affected pathways during infection and under environmental stress, and tentative (pathogen/penaeid) interactomes.

Final Conclusion

The penaeid shrimp genome is composed of more than 40 linkage groups in a haploid genome, and its estimated size is about 70% that of the human genome (Chow et al. 1990). As shown in Table 2, the total number of penaeid shrimp ESTs currently available in NCBI dbEST is about 200,000. Far more ESTs have been published for other organisms, such as *Homo sapiens* (8,163,902), *Drosophila melanogaster* (820,319), or the important aquaculture fish *Salmo salar* (Atlantic salmon, 494,094). This is partly a reflection of the relatively inferior scientific importance that shrimp have to the biological research community. However, given the steadily increasing economic importance of penaeid shrimp, it is expected that the shrimp-farming countries will increasingly devote more resources to study these animals, and many more shrimp ESTs are likely to be generated. Meanwhile, to increase the rate of novel gene identification and to complete the shrimp transcriptome as far as possible, additional strategies are needed. Many new cDNA libraries will need to be constructed and sequenced, and following Carninci et al. (2003) work on mouse, these cDNA libraries should be derived from a variety of different sources, from (a) particular tissue(s), from shrimp tissues or embryos at different developmental stages, or from shrimp kept under different environments or stresses. Different RNA preparation and construction methods should also be used, including subtraction, normalization, use of various cloning vectors and enrichment of longer or full-length transcripts (Carninci et al. 2003). Our data (Table 2) shows that only about one third of the tentative unique genes (TUGs) could be annotated by the sequence homologs found in the published sequence databases. Although it is quite possible that some of these unmatched TUGs truly represent unknown genes that exist exclusively in crustacean or even in penaeid shrimp only, many of these TUGs are unmatched either because the nucleotide or encoded protein sequences in TUGs are too divergent to be matched to their true homologous genes in other organisms or the available TUGs are so short that only untranslated regions are sequenced. The only way to solve these problems is to isolate and sequence the full-length cDNA clones. Full-length cDNA sequences are a superior tool for identifying genes thorough homology searches and also for the functional study of genes and their products. They are important for the correct annotation of genomic sequences, and they also facilitate the finding of transcription initiation and termination sites, 5′ and 3′ untranslated regions, promoter regions, and exon-intron splice sites (Alexandrov et al. 2006). In the era of functional genomics, once a substantial EST database has been complied, the next logical step is the construction and sequencing of a full-length cDNA library (Seki et al. 2002; Carninci et al. 2003; Gerhard et al. 2004). We therefore anticipate and look forward to the establishment of a full-length cDNA library for penaeid shrimp in the near future.

Acknowledgements We are indebted to Paul Barlow for his helpful criticism. This investigation was supported financially by National Science Council grant (NSC96-2317-B-002-005).

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656

Adams MD, Kerlavage AR, Fields C, Venter JC (1993) 3400 new expressed sequence tags identify diversity of transcripts in human brain. Nat Genet 4:256–267

Alexandrov NN, Troukhan ME, Brover VV, Tatarinova T, Flavell RB, Feldmann KA (2006) Features of Arabidopsis genes and genome discovered using full-length cDNAs. Plant Mol Biol 60(1):69–85

Astrofsky KM, Roux MM, Klimple KR, Fox JG, Dhar AK (2002) Isolation of differentially expressed genes from white spot syndrome virus (WSV) infected Pacific blue shrimp (*Penaeus stylirostris*). Arch Virol 147:1799–1812

Burgents JE, Burnett LE, Burnett KG, Stabb EV (2005) Localization and bacteriostasis of Vibrio introduced into the Pacific white shrimp, *Litopenaeus vannamei*. Dev Comp Immunol 29:681–691

Carninci P, Waki K, Shiraki T, Konno H, Shibata K, Itoh M, Aizawa K, Arakawa T, Ishii Y, Sasaki D, Bono H, Kondo S, Sugahara Y, Saito R, Osato N, Fukuda S, Sato K, Watahiki A, Hirozane-Kishikawa T, Nakamura M, Shibata Y, Yasunishi A, Kikuchi N, Yoshiki A, Kusakabe M, Gustincich S, Beisel K, Pavan W, Aidinis V, Nakagawara A, Held WA, Iwata H, Kono T, Nakauchi H, Lyons P, Wells C, Hume DA, Fagiolini M, Hensch TK, Brinkmeier M, Camper S, Hirota J, Mombaerts P, Muramatsu M, Okazaki Y, Kawai J, Hayashizaki Y (2003) Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. Genome Res 13(6B):1273–89

Chow S, Dougherty W, Sandifer P (1990) Meiotic chromosome complements and nuclear DNA contents of four species of shrimps of the genus Penaeus. J Crustac Biol 10:29–36

Clavero-Salas A, Sotelo-Mundo RR, Gollas-Galván T, Hernández-López J, Peregrino-Uriarte AB, Muhlia-Almazán A, Yepiz-Plascencia G (2007) Transcriptome analysis of gills from the

white shrimp *Litopenaeus vannamei* infected with White Spot Syndrome Virus. Fish Shellfish Immunol 23:459–472

Davidson B, Swalla BJ (2002) A molecular analysis of ascidian metamorphosis reveals activation of an innate immune response. Development 129:4739–4751

Dhar AK, Dettroi A, Roux MM, Klimpel KR, Read B (2003) Identification of differential expressed genes in shrimp (Penaeus stylirostris) infected with white spot syndrome virus by cDNA microarrays. Arch Virol 148:2381–2396

Dong B, Xiang J (2007) Discovery of genes involved in defense/immunity functions in a haemocytes cDNA library from *Fenneropenaeus chinensis* by ESTs annotation. Aquaculture 272:208–215

Dong Q, Kroiss L, Oakley FD, Wang BB, Brendel V (2005) Comparative EST analyses in plant systems. Methods Enzymol 395:400–418

Gillett, R. (2008) Global study of shrimp fisheries. FAO Fisheries Technical Paper. No. 475, FAO, Rome

Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P, Guyer M, Peck AM, Derge JG, Lipman D, Collins FS, Jang W, Sherry S, Feolo M, Misquitta L, Lee E, Rotmistrovsky K, Greenhut SF, Schaefer CF, Buetow K, Bonner TI, Haussler D, Kent J, Kiekhaus M, Furey T, Brent M, Prange C, Schreiber K, Shapiro N, Bhat NK, Hopkins RF, Hsie F, Driscoll T, Soares MB, Casavant TL, Scheetz TE, Brown-stein MJ, Usdin TB, Toshiyuki S, Carninci P, Piao Y, Dudekula DB, Ko MS, Kawakami K, Suzuki Y, Sugano S, Gruber CE, Smith MR, Simmons B, Moore T, Waterman R, Johnson SL, Ruan Y, Wei CL, Mathavan S, Gunaratne PH, Wu J, Garcia AM, Hulyk SW, Fuh E, Yuan Y, Sneed A, Kowis C, Hodgson A, Muzny DM, McPherson J, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madari A, Young AC, Wetherby KD, Granite SJ, Kwong PN, Brinkley CP, Pearson RL, Bouffard GG, Blakesly RW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Griffith M, Griffith OL, Krzywinski MI, Liao N, Morin R, Palmquist D, Petrescu AS, Skalska U, Smailus DE, Stott JM, Schnerch A, Schein JE, Jones SJ, Holt RA, Baross A, Marra MA, Clifton S, Makowski KA, Bosak S, Malek J, MGC Project Team (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res 14 (10B):2121–7

Greener MJ, Roberts RG (2000) Conservation of components of the dystrophin complex in Drosophila. FEBS Lett 482:13–18

Gross PS, Bartlett TC, Browdy CL, Chapman RW, Warr GW (2001) Immune gene discovery by expressed sequence tag analysis of hemocytes and hepatopancreas in the pacific white shrimp, *Litopenaeus vannamei*, and the Atlantic white shrimp, *L. setiferus*. Dev Comp Immunol 25(7):565–577

He N, Qin Q, Xu X (2005) Differential profile of genes expressed in hemocytes of White Spot Syndrome Virus-resistant shrimp (*Penaeus japonicus*) by combining suppression subtractive hybridization and differential hybridization. Antivir Res 66:39–45

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

Iwanaga S (2002) The molecular basis of innate immunity in the horseshoe crab. Curr Opin Immunol 14:87–95

Jongeneel CV (2000) Searching the expressed sequence tag (EST) databases: panning for genes. Brief Bioinform 1:76–92

Laufer H, Borst DW, Baker FC, Carrasco M, Sinkus M, Reuter CC, Tsai LW, Schooley DA (1987) Identification of a juvenile hormone-like compound in crustacean. Science 235:202–205

Leelatanawit R, Klinbunga S, Puanglarp N, Tassanakajon A, Jarayabhand P, Hirono I, Aoki T, Menasveta P (2004) Isolation and characterization of differentially expressed genes in ovaries and testes of the giant tiger shrimp (*Penaeus monodon*). Mar Biotechnol 6:S506–S510

Leelatanawit R, Sittikankeaw K, Yocawibun P, Klinbunga S, Roytrakul S, Aoki T, Hirono I, Menasveta P (2009) Identification, characterization and expression of sex-related genes in testes of the giant tiger shrimp *Penaeus monodon*. Comp Biochem Physiol A Mol Integr Physiol 152(1):66–76

Lehnert SA, Wilson KJ, Byrne K, Moore SS (1999) Tissue-specific expressed sequence tags from the black tiger shrimp *Penaeus monodon*. Mar Biotechnol 1(5):465–476

Leu JH, Chang CC, Wu JL, Hsu CW, Hirono I, Aoki T, Juan HF, Lo CF, Kou GH, Huang HC (2007) Comparative analysis of differentially expressed genes in normal and white spot syndrome virus infected *Penaeus monodon*. BMC Genomics 8:120

Matsumoto H, Yamada T (1991) Phosrestins I and II: arrestin homologs which undergo differential light-induced phosphorylation in the Drosophila photoreceptor in vivo. Biochem Biophys Res Commun 177:1306–1312

O'Leary NA, Trent HF III, Robalino J, Peck MET, Mckillen DJ, Gross PS (2006) Analysis of multiple tissue-specific cDNA libraries from the Pacific whiteleg shrimp, *Litopenaeus vannamei*. Integrative and Comparative Biology 46:931–939

Pan D, He N, Yang Z, Liu H, Xu X (2005) Differential gene expression profile in hepatopancreas of WSSV-resistant shrimp (*Penaeus japonicus*) by suppression subtractive hybridization. Dev Comp Immunol 29:103–112

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19:651–652

Pongsomboon S, Wongpanya R, Tang S, Chalorsrikul A, Tassanakajon A (2008) Abundantly expressed transcripts in the lymphoid organ of the black tiger shrimp, *Penaeus monodon*, and their implication in immune function. Fish Shellfish Immunol 25 (5):485–493

Preechaphol R, Leelatanawit R, Sittikankeaw K, Klinbunga S, Khamnamtong B, Puanglarp N, Menasveta P (2007) Expressed sequence tag analysis for identification and characterization of sex-related genes in the giant tiger shrimp *Penaeus monodon*. J Biochem Mol Biol 40(4):501–510

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res 29:159–164

Rainer J, Brouwer M (1993) Hemocyanin synthesis in the blue-crab *Callinectes sapidus*. Comp Biochem Physiol B 104:69–73

Raming K, Freitag J, Krieger J, Breer H (1993) Arrestin subtypes in insect antennae. Cell Signal 5:69–80

Robalino J, Almeida JS, McKillen D, Colglazier J, Trent HF 3rd, Chen YA, Peck ME, Browdy CL, Chapman RW, Warr GW, Gross PS (2007) Insights into the immune transcriptome of the shrimp Litopenaeus vannamei: tissue-specific expression profiles and transcriptomic responses to immune challenge. Physiol Genomics 29:44–56

Rojtinnakorn J, Hirono I, Itami T, Takahashi Y, Aoki T (2002) Gene expression in heamocytes of kuruma prawn, *Penaeus japonicus*, in response to infection with WSSV by EST approach. Fish Shell Immunol 13:69–83

Rosenberry B (2006) World shrimp farming 2006. Shrimp News International, San Diego

Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequences? Trends Plant Sci 8:321–329

Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y, Muramatsu M, Hayashizaki Y, Kawai J, Carninci P, Itoh M, Ishii Y, Arakawa T, Shibata K, Shinagawa A, Shinozaki K (2002) Functional

annotation of a full-length Arabidopsis cDNA collection. Science 296(5565):141–145

Smith VJ, Brown JH, Hauton C (2003) Immunostimulation in crustaceans: does it really protect against infection? Fish Shellfish Immunol 15:71–90

Sritunyalucksana K, Söderhäll K (2000) The proPO and clotting system in crustaceans. Aquaculture 191:53–69

Supungul P, Klinbunga S, Pichyangkura R, Jitrapakdee S, Hirono I, Aoki T, Tassanakajon A (2002) Identification of immune-related genes in hemocytes of black tiger shrimp (*Penaeus monodon*). Mar Biotechnol 4(5):487–494

Supungul P, Klinbunga S, Pichyangkura R, Hirono I, Aoki T, Tassanakajon A (2004) Antimicrobial peptides discovered in the black tiger shrimp Penaeus monodon using the EST approach. Dis Aquat Org 61(1–2):123–135

Tassanakajon A, Klinbunga S, Paunglarp N, Rimphanitchayakit V, Udomkit A, Jitrapakdee S, Sritunyalucksana K, Phongdara A, Pongsomboon S, Supungul P, Tang S, Kuphanumart K, Pichyangkura R, Lursinsap C (2006) *Penaeus monodon* gene discovery project: the generation of an EST collection and establishment of a database. Gene 384:104–112

Tharntada S, Somboonwiwat K, Rimphanitchayakit V, Tassanakajon A (2008) Anti-lipopolysaccharide factors from the black tiger shrimp, *Penaeus monodon*, are encoded by two genomic loci. Fish Shellfish Immunol 24(1):46–54

UniProt Consortium: The Universal Protein Resource (UniProt). Nucleic Acids Res 2008, 36:D190- D195

Wang B, Li F, Dong B, Zhang X, Zhang C, Xiang J (2006) Discovery of the genes in response to white spot syndrome virus (WSSV) infection in *Fenneropenaeus chinensis* through cDNA micro-array. Mar Biotechnol 7:119–127

Wang HC, Wang HC, Kou GH, Lo CF, Huang WP (2007) Identification of icp11, the most highly expressed gene of shrimp white spot syndrome virus (WSSV). Dis Aquat Organ 74:179–189

Xiang J, Wang B, Liu B, Wang Z, Wang X, Tong W, Li F. Over 10000 expressed sequence tags from *Fenneropennaeus chinenesis*. Book of Abstracts, p 837. World Aquculture 2002, April 23–27, 2002, Beijing, China

Xiang J, Wang B, Li F, Liu B, Zhou Y, Tong W (2008) Generation and Analysis of 10,443 ESTs from Cephalothorax of *Fenneropenaeus Chinensis*. The 2nd International Conference on Bioinformatics and Biomedical Engineering, Shanghai, China. p. 74–80

Yamano K, Unuma T (2006) Expressed sequence tags from eyestalk of kuruma prawn, *Marsupenaeus japonicus*. Comp Biochem Physiol A Mol Integr Physiol 143(2):155–161