รายงานวิจัยฉบับสมบูรณ์

โครงการ การวิเคราะห์ฐานข้อมูล EST ของกุ้ง ด้วยวิธีทางชีวสารสนเทศเพื่อหา การปฏิสัมพันธ์ระหว่างกุ้งและเชื้อก่อโรคและการประยุกต์ใช้เพื่อการควบคมุโรคกุ้ง

Bioinformatic identification of shrimp-pathogen interactions from shrimp EST database and their application for shrimp disease control

โดย นายอนุภาพ ประชุมวัด และคณะ

ธันวาคม 2559

สัญญาเลขที่ TRG5680001

รายงานวิจัยฉบับสมบูรณ์

โครงการ การวิเคราะห์ฐานข้อมูล EST ของกุ้งด้วยวิธีทางชีวสารสนเทศเพื่อหาการปฏิสัมพันธ์ระหว่างกุ้งและเชื้อก่อโรคและการประยุกต์ใช้เพื่อการควบคุมโรคกุ้ง

Bioinformatic identification of shrimp-pathogen interactions from shrimp EST database and their application for shrimp disease control

คณะผู้วิจัย	**สังกัด**
1. นายอนุภาพ ประชุมวัด	ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ
2. ศ. ดร. ทีโมที วิลเลี่ยม ฟลีเกล	มหาวิทยาลัยมหิดล

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย
และศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ

(ความเห็นในรายงานนี้เป็นของผู้วิจัย สกว. และ ศช. ไม่จำเป็นต้องเห็นด้วยเสมอไป)

กิตติกรรมประกาศ

โครงการวิจัยนี้ได้รับการสนับสนุนจากสำนักงานกองทุนสนับสนุนการวิจัย สกว. และศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ ศช. คณะผู้วิจัยขอบพระคุณสำนักงานกองทุนสนับสนุนการวิจัยและศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ ซึ่งเป็นหน่วยงานที่อนุเคราะห์การสนับสนุนโครงการนี้ ทั้งนี้คณะผู้วิจัยขอขอบคุณ ดร. กัลยาณ์ ศรีธัญญลักษณา-แดงติ๊บ หัวหน้าห้องปฏิบัติการปฏิสัมพันธ์ระหว่างกุ้งและเชื้อก่อโรค ที่ให้คำปรึกษาและสนับสนุน และอาจารย์ นักวิจัย นักศึกษาในหน่วยวิจัยเพื่อความเป็นเลิศเทคโนโลยีชีวภาพกุ้ง (Centex Shrimp) และห้องปฏิบัติการปฏิสัมพันธ์ระหว่างกุ้งและเชื้อก่อโรค (SPI) หน่วยวิจัยเทคโนโลยีชีวภาพสัตว์ (ABU) ศช.

Acknowledgement

This project is supported by the Thailand Research Fund (TRF) and by the National Center for Genetic Engineering and Biotechnology (BIOTEC). This support is very much appreciated. We thank Dr. Kallaya Sritunyalucksana for support and valuable suggestions. We would like to extend our appreciations to our colleagues at Center of Excellence for Shrimp Molecular Biology and Biotechnology (CENTEX SHRIMP) and Shrimp-Pathogen Interaction (SPI) Laboratory, Animal Biotechnology Research Unit, BIOTEC for their help.

บทคัดย่อ

รหัสโครงการ: TRG5680001

ชื่อโครงการ: การวิเคราะห์ฐานข้อมูล EST ของกุ้งด้วยวิธีทางชีวสารสนเทศเพื่อหาการปฏิสัมพันธ์ระหว่างกุ้งและเชื้อก่อโรคและการประยุกต์ใช้เพื่อการควบคุมโรคกุ้ง

ชื่อนักวิจัย: นายอนุภาพ ประชุมวัด
ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ (ไบโอเทค)
E-mail Address : anuphap.pra@biotec.or.th

ระยะเวลาโครงการ: 2556-2559

ข้อมูลทรานสคริปโตมิกส์ของกุ้งทะเลที่เพิ่มขึ้นอย่างรวดเร็วช่วยงานวิจัยด้านภูมิคุ้มกันของกุ้งต่อเชื้อก่อโรคต่างๆ ด้วยการพบยีนที่มีบทบาทในการตอบสนองเชื้อโรค ซึ่งเป็นประโยชน์ต่ออุตสาหกรรมการเลี้ยงกุ้ง แต่การใช้ข้อมูลทรานสคริปโตมิกส์เหล่านี้ยังมีอุปสรรค เพราะจำนวนเส้นลำดับเบสที่ได้ในทรานสคริปโตมิกส์นั้น จำนวนเกือบครึ่งไม่สามารถหาความคล้ายคลึงกับยีนที่มีอยู่แล้วในฐานข้อมูลได้จึงไม่สามารถอนุมานหน้าที่ได้ ด้วยเหตุนี้ข้าพเจ้าจึงรวบรวมข้อมูลทรานสคริปโตมิกส์ของกุ้งและสัตว์สิบขารวม 14 สายพันธุ์ ซึ่งเป็นข้อมูลจากวิธีการ cDNA libraries และ next-generation DNA sequencing (NGS) การรวบรวมและวิเคราะห์ได้ข้อมูลทั้งสิ้นกว่าสองล้านระเบียน ซึ่งรวมข้อมูลชีวโมเลกุลอื่นๆ ด้วย ข้อมูลเหล่านี้ได้ผ่านการอนุมานหน้าที่ของ protein-coding และ regulatory non-coding RNAs (ncRNAs) genes ด้วยกระบวนการทางชีวสารสนเทศ และได้เปิดเผยข้อมูลนี้บางส่วนสู่สาธารณะนำไปใช้ประโยชน์แล้วที่ฐานข้อมูล ShrimpGPAT (http://shrimpgpat.sc.mahidol.ac.th/) อีกทั้งข้าพเจ้าวิเคราะห์ข้อมูลทรานสคริปโตมิกส์ของกุ้งกุลาดำ (*Penaeus monodon*) และกุ้งขาวแวนนาไม (*P. vannamei*) ที่ทำให้ติดเชื้อไวรัสโรคตัวแดงดวงขาวในกุ้ง (WSSV) เพื่อหายีนและ ncRNAs ของกุ้งที่มีการตอบสนองต่อเชื้อโรคนี้ การวิเคราะห์พบว่าชุดยีนและ ncRNAs ชุดหนึ่งมีการแสดงออกในระดับสูงในกลุ่มกุ้งที่มีชีวิตรอดและไม่ตายจากเชื้อ WSSV และยังพบว่ายีนกลุ่มหนึ่งที่แสดงออกในระดับสูงในกุ้งทั้งสองสายพันธุ์ที่ได้รับเชื้อด้วย ซึ่งบ่งบอกว่ายีนกุ้งกลุ่มนี้มีความสำคัญต่อการต้านเชื้อโรค อีกทั้ง ncRNAs ที่พบนี้น่าจะเป็น long non-coding RNAs ที่มีการรายงานเป็นครั้งแรกในกุ้งทะเลอีกด้วย ขณะนี้มีการเลือกยีนและ ncRNAs กลุ่มนี้มาทดสอบในห้องปฏิบัติการเพื่อยืนยัน ผลการยืนยันนี้จะสามารถเป็นฐานความรู้ในการพัฒนา จัดการและควบคุมการก่อโรคตัวแดงดวงขาวในฟาร์มกุ้ง อีกทั้งข้อมูลทรานสคริปโตมิกส์ที่ได้รวบรวมและข้อมูลหน้าที่ทั้งยีนและ ncRNAs ที่ได้เหล่านี้จะถูกเปิดให้สาธารณะได้ใช้ประโยชน์เพื่อการวิจัยและการค้นพบยีนในกุ้งทะเลต่อไป

คำหลัก: ทรานสคริปโตมิกส์, การปฏิสัมพันธ์ระหว่างกุ้งและเชื้อ, Regulatory non-coding RNAs (ncRNAs) genes, Protein-coding genes, การทำเหมืองข้อมูล (Data mining)

Abstract

Project Code : TRG5680001

Project Title : Bioinformatic identification of shrimp-pathogen interactions from shrimp EST database and their application for shrimp disease control

Investigator : Anuphap Prachumwat

National Center for Genetic Engineering and Biotechnology

E-mail Address : anuphap.pra@biotec.or.th

Project Period : 2014-2016

Rapid increase in the number of shrimp transcriptomic data facilitates shrimp defense (immunity) research, initially through identification of putative shrimp-pathogen interactions, which benefits shrimp aquaculture. A number of comprehensive analyses on these data for shrimp-pathogen interactions remains limited owing to a large proportion of transciptomic sequences with no homology in current public database. To take advantage of a gigantic amount of transcript data, we compiled transciptomic sequences of 14 decapods generated by both traditional cDNA libraries and next-generation DNA sequencing (NGS) along with other molecular sequences in the total of two million transcripts for identification, via dedicated bioinformatics pipelines, of protein-coding and regulatory non-coding RNAs (ncRNAs) genes. A set of the sequences was released for public at the ShrimpGPAT database (http://shrimpgpat.sc.mahidol.ac.th/) for accelerating shrimp gene discovery and research. To gain an insight on how shrimp interacts to pathogens, we focused on analyses the transcriptomes of white spot syndrome virus (WSSV) infection in *Penaeus monodon* and *P. (Litopenaeus) vannamei* and identified sets of WSSV-responsive protein-coding and ncRNAs genes. Several protein-coding genes and putative ncRNA sequences were found to be highly expressed in shrimp survivors of WSSV infection, and a set of genes was found in both *P. monodon* and *P. vannamei*, signifying putative key shrimp defense genes during pathogen infection. Notably, putative ncRNAs found here will likely be first reported long non-coding RNAs in shrimp. These genes and ncRNAs have been being experimentally validated to provide a basis for future development for a successful management of virulent control or disease prevention to overcome serious economic losses from pathogen outbreaks. Furthurmore, the obtained collection of transcriptomes and associated *in-silico* annotation of protein-coding and ncRNAs genes will be released to public for further investigation.

Keywords : Transcriptomes, Shrimp-pathogen interactions, Regulatory non-coding RNAs (ncRNAs) genes, Protein-coding genes, Data mining

Table of Contents

List of Tables

Table page

List of Figures

Figure page

Symbols and Abbreviations

Abbreviation	Full names
BLAST	Basic Local Alignment Search Tool
BLASTN	BLASTN programs search nuclotide databases using a nucleotide query
BLASTX	BLASTX search protein subjects using a translated nucleotide query
cDNA	complementary DNA
CDS	Coding DNA Sequence
dbEST	NCBI Expressed Sequence Tag Database
DE	differentially expressed
DNA	Deoxyribonucleic acid
EST	Expressed Sequence Tag
E-Value	Expect Value
FASTA	FAST Alignment
GI	NCBI Identification Number
GO	Gene Ontology
ID	Identification
miRNA	microRNA
mRNA	messenger RNA
MSA	Multiple Sequence Alignment
mtRNA	mitochondrial RNA
NCBI	National Center for Biotechnology Information
ncRNA	regulartory non-coding RNA
NGS	Next-Generation DNA Sequencer
PERL	Practical Extraction and Report Language
PPI	Protein-Protein Interaction
Rfam	an RNA family database
RNA	Ribonucleic acid
RNaseP	Ribonuclease P
rRNA	ribosomal RNA
ShrimpGPAT	Shrimp Gene and Protien Annotation Tool
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
SRA	NCBI's Sequence Read Archive
TBLASTN	TBLASTN search translated nucleotide databases using a protein query
TBLASTX	TBLASTX search translated nucleotide databases using a translated nucleotide query
telomeraseRNA	Telomerase RNA
tmRNA	Transfer-messenger RNA
tRNA	Transfer RNA
UniProt	Universal Protein Resource
WSSV	White spot syndrome virus
YHV	Yellowhead virus

Introduction

While marine shrimp aquaculture has become the fastest growing sector of Thailand, as well as international, aquaculture industry, scientific research on shrimp-pathogen interactions remains relatively lacked. Rapid accumulation of transcriptomic data, especially expressed sequence tags (ESTs), has facilitated research in shrimp biology, defense (immunity) and genetics to improve shrimp aquaculture production. Understanding shrimp-pathogen interactions is the first step in characterizing shrimp defense system in protecting shrimps from their pathogens. Although several shrimp immunity proteins are reported by these EST studies, no shared pathogen-responding shrimp protein and shared molecular pathway of pathogen entry to host cells have been identified across shrimp species or across pathogens. Analyses of rapidly-increasing and publicly-available shrimp EST data will provide an insight on how shrimps respond to pathogens. Unfortunately, such a comprehensive analysis of all available shrimp EST data has not been conducted. In addition, no homolog for almost half of these EST clones could be found, i.e., their function cannot be predicted or inferred. Besides well-known non-coding RNAs (e.g., mRNA, rRNA and tRNA), a significant attention has been paid on regulatory non-coding RNAs (ncRNAs) that possess a diverse range of functions and participate in many biological pathways. This leads us to hypothesize for an existence of putative ncRNAs in these ESTs. This proposed comprehensive analysis of shrimp transcriptomes will reveal gene content (both protein-coding and non-coding transcripts) in under-uncharacterized shrimp genomes. Given that an up-to-date compilation of shrimp protein-coding genes as well as a novel collection of ncRNAs to be obtained for general public in an online searchable database, this information will contribute great benefits to not only shrimp immunity research but also shrimp community as a whole. Importantly, an analysis of ESTs from various sources, cell types and shrimp species upon pathogen-infection warrants useful information on pathogen-responsive shrimp genes for developing a successful management of virulent control or disease prevention. Furthermore, these pathogen-responsive shrimp genes, after laboratory testing and validation, can be used as markers to screen for characteristics of these genes in current Thailand domesticated broodstock families. Broodstock families with selected traits of genes and/or gene expression can be focused in selective breeding programs to obtain pathogen-resistant broodstock for sustainable shrimp aquaculture.

A significance of expressed sequence tags (ESTs) to shrimp research community has been demonstrated, especially being an initial step for understanding shrimp-pathogen interactions (for review, see Leu et al. 2011; Pongsomboon et al. 2011; Tassanakajon et al. 2013). Briefly, a large scale EST study from various tissues and conditions of the black tiger shrimp *Penaeus monodon* was performed by Thai scientists led by Professor Dr. Anchalee Tassanakajon in 2006, and the data has been deposited in Thailand's *Penaeus monodon* EST Project database (Tassanakajon et al. 2006). Another large scale EST study from whole *P. monodon* was conducted by Taiwanese scientists led by Dr. Lo, whose study was focused on a comparison between normal shrimps and those challenged by white spot syndrome virus (WSSV; Leu et al. 2007). For the Pacific white shrimp *P. (Litopenaeus) vannamei*, several studies were performed by a group led by Dr. Paul Gross in the USA (O'Leary et al. 2006) and corresponding EST clones were deposited in the Marine Genomics Database (McKillen et al. 2005). These two shrimp species account for nearly 90% of global aquaculture production, and almost all of the shrimp EST data currently published have been derived from the two species. Recently most ESTs have been generated by next-generation DNA sequencing (NGS) instead of by a traditional cDNA library approach, suggesting that more available data can soon be obtained for these two and other shrimp species. However, this proposed study will be mainly focused on data from *P. monodon* and *P. vannamei*, two economically important shrimp species of Thailand.

The above two specialized databases for shrimp hold only ESTs that were generated by their authors' own laboratories. While *Penaeus monodon* EST Project database specialized for only black tiger shrimp, Marine Genomics Database covers about 28 marine organisms (15 are crustaceans). Recently Taiwan Penaeus Genome (PAGE) database was the first shrimp database that combined available EST data from various sources for four penaeid species (Leu et al. 2011). A general pipeline for data analysis in these three databases consists of sequence quality filtering, contig construction and in-silico function prediction (BLAST for homologs in either GenBank or Uniport and Gene Ontology prediction inferred from homologs). Unfortunately, these databases often lack a periodic update for newly available data, especially those short reads generated by NGS. In addition, a homology search against GenBank database of ESTs in *Penaeus monodon* EST Project database revealed more than 40% of EST clones have no homolog; a similar proportion was reported in Penaeus Genome (PAGE) and Marine Genomics databases. This suggests that a large proportion of available shrimp ESTs may not appear to encode proteins or may be outside any known gene regions (Kampa et al. 2004; Kapranov et al. 2002). In addition to mRNAs, rRNAs and tRNAs, there are a number of regulatory non-coding RNAs (ncRNAs) that regulate and participate in a diverse range of biological processes. Recently, an increasing number of reports observe polyadenylated and mRNA-like ncRNAs in eukaryotes (these RNAs are spliced but do not have appreciable open reading frames or evidence for protein coding capacity). Promisingly, several studies found ncRNAs in EST libraries (e.g., Macintosh et al., 2001; Tupy et al. 2005; Seemann et al. 2007).

Within arthropods, crustaeans are scantly sampled for genomic studies, relatively to their closely cousins, true insects. A large number of insect genomes have been completed or in draft assemblies, the only genome of *Daphnia pulex* (Branchiopoda: Crustacea) was recently completed (Colbourne et al. 2011). The information from better-annotated and heavily-sampled insect genomes (e.g., several genomes of fruit flies, mosquitoes, the honey bee and the red flour beetle) is valuable for a pipeline of shrimp gene discovery. In addition, several transcriptome studies in insects and other non-penaeid crustaceans have been reported (e.g., Jung et al. 2011; Ma et al. 2012; Gibson et al. 2013). Therefore, the pipeline of shrimp sequence annotation in this study will utilize the insect and other crustacean genomic data.

Although Taiwan PAGE combined available EST data from various sources for four penaeid species (Leu et al. 2011), no additional analysis was conducted on these data, and no data update was performed since its initial release. Pongsomboon et al. (2011) performed a global analysis of pathogen-challenged EST libraries in Thailand's Penaeus monodon EST Project database using microarrays and revealed a list of *P. monodon* genes that were differentially expressed and possibly defensive against WSSV, yellow head virus (YHV) and *Vibrio harveyi*. Recently, we have designed and been constructing a database system, namely Shrimp Gene and Protein Annotation Tool (ShrimpGPAT; http://shrimpgpat.sc.mahidol.ac.th/), to collect molecular sequences (e.g., ESTs, short reads of transcriptomes, full length cDNA and proteins) of shrimps. In addition to *in-silico* prediction and bioinformatics tools, ShrimpGPAT allows users to annotate EST records (community-based annotation). Thus, data in ShrimpGPAT, currently holding at least 300,000 EST records, will be of interest for a global analysis in this proposed study.

Materials, Methods and Results

The project was conducted in the following four aspects: Updating molecular sequences of decapods, annotation of sequences for protein-coding genes and non-coding RNAs and indetification of pathogen-specific responsive genes.

1. Data collection update, sequence cleaning-up, and contig construction

The Shrimp Gene and Protein Annotation Tool (ShrimpGPAT; http://shrimpgpat.sc.mahidol.ac.th/v1/; Leekitcharoenphon et al., 2010), Release # 1 contained only expressed sequence tags (ESTs) for 316,900 sequences for six species of decapods, including four penaeid shrimp. These EST data were generated by traditional Sanger sequencing of clone selection. The newly downloaded sequences including ESTs, cDNAs, and proteins for 14 decapod species (see the list of species in Table 1) were obtained mainly from NCBI GenBank. Some additional EST sequences were obtained from the Marine Genomics database (http://www.marinegenomics.org/; McKillen et al. 2005), the Penaeus monodon EST Project database (http://pmonodon.biotec.or.th/; Tassanakajon et al. 2006) and the data generated in laboratories of ours and our collaborators. Transcriptomic data generated by next-generation sequencers (NGS), publicly available in the SRA database (www.ncbi.nlm.nih.gov/sra), were also downloaded for three species of shrimp (i.e, the black tiger shrimp *Penaeus monodon*, the Pacific white shrimp *P. (Litopenaeus) vannamei* and *Macrobrachium rosenbergii*). The NGS short reads from the NCBI SRA database were processed by SRA Toolkit. Currently, Roche 454 and Illumina are the two platforms of NGS for these datasets. EST sequences were masked by cross_match (http://www.phrap.org/) for vector and contaminating sequences against both full-length vector sequences, if available, and Univec database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html). Masked sequences were processed by an in-house PERL script to produce vector-free sequences. Adapter sequences in NGS short reads were trimmed by sfffile or Trimmomatic (Bolger et al. 2014), for Roche 454 and Illumina data, respectively. Paired-end NGS sequences from the Illumina platform were merged with FLASh (http://ccb.jhu.edu/software/FLASH/; Magoc and Salzberg 2011). Trimmed sequences were *de novo* assembled by either CAP3 (Huang & Madan, 1999) or Newbler with the default parameter setting. To improve contig construction for *P. monodon* and *P. vannamei*, we used MIRA (Chevreux et al. 2004) to combine sequencing reads from various technologies (i.e., transcript sequences were generated by traditional Sanger sequencing and several platforms of NGS) to construct transcript contigs from hybrid datasets. In addition, for almost all of NGS datasets were assembled by Trinity (Grabherr et al. 2011) with the default parameter setting.

1.1 Updated sequences (ShrimpGPAT Release #2)

The ShrimpGPAT database Release # 1 (http://shrimpgpat.sc.mahidol.ac.th/v1/ Leekitcharoenphon et al., 2010) previously contained *only* ESTs of 316,900 sequences for six species of decapods, including four penaeid shrimp. We further collected all available sequences from NCBI GenBank, including ESTs, cDNAs, proteins and short reads transcriptome datasets for the total of 14 decapod species (Table 1). Additional EST sequences of *P. monodon* and *P. vannamei* were downloaded from either the Marine Genomics database or the *Penaeus monodon* EST Project database. To avoid duplicates in our data collection, the sequences were processed whether they had already deposited in the GenBank because some of them already published and deposited there. The associated information on these sequences (e.g., tissue types, conditions of experiments) was also downloaded and later deposited to the ShrimpGPAT database.

All sequences were processed via the sequence cleansing step as described above. For the transcript contig construction, we performed *de novo* assembly for all 14 species with traditionally-generated EST sequences (except *Macrobrachium rosenbergii* includes NGS transcriptome data) and by either CAP3 or Newbler assemblers, resulting in 100,585 transcript contigs in total (Table 1). These set of contigs along with EST, cDNA and protein sequences (the total of >500,000 records) were used for *in-silico* functional annotation (see below).

Among >500,000 records of the ShrimpGPAT database Release # 2 (Table 1; http://shrimpgpat.sc.mahidol.ac.th/ShrimpGPATV2/), *P. vannamei* has the highest number of records (~299,000), and *P. monodon* has the second highest (~138,000). The numbers signify their importance as species of the highest interest to the shrimp scientific research community and species most-cultivated or captured for trade. Similarly, the six penaeid shrimp have combined records that number about four times that of the other eight decapod species in the database (i.e., ~460,000 *vs*. 111,000). A large proportion of the records for each species are ESTs and transcript contigs, whereas the numbers of cDNA and protein records are still relatively small. The number of transcript contigs for each species is the summation of all contig sequences constructed by the set of ESTs and by the set of SRA reads. Note that transcript contig records produced by different contig assemblers (e.g., CAP3 and Newbler) may constitute the same sequences. Regarding transcript contigs of SRA reads, *Macrobrachium rosenbergii* is the only species that currently has transcript contigs derived from an SRA dataset (81,411 reads for 50 million base pairs that could be assembled). Among the 14 species, *Scylla olivacea* has the lowest number of records in its EST collection. It is the first publicly-available collection of ESTs for this species and it was recently generated by our laboratory.

Table 1 The number of molecular sequence records in ShrimpGPAT Release #2

Species		**# of records**			
Scientific name	**Common name**	**EST**	**Transcript contigs** [a]	**cDNA**	**Protein**
Penaeus (Penaeus) monodon	Black tiger shrimp	86,327	18,410	1,976	602
Penaeus (Litopenaeus) vannamei	Pacific whiteleg shrimp	176,592	47,058	74,828	574
Penaeus (Litopenaeus) setiferus	White shrimp	1,042	126	135	27
Penaeus (Fenneropenaeus) chinensis	Fleshy prawn	10,446	2,714	478	257
Penaeus (Fenneropenaeus) indicus	Indian prawn	714	155	348	127
Penaeus (Marsupenaeus) japonicus	Kuruma prawn	3,156	662	989	743
Macrobrachium rosenbergii	Giant freshwater prawn	4,427	8,550[b]	635	389
Cherax quadricarinatus	Cray fish	120	90	239	226
Pacifastacus leniusculus	Signal crayfish	802	199	914	88
Homarus americanus	American lobster	29,957	12,709	186	227
Scylla olivacea	Orange mud crab	203	80	121	0
Scylla paramamosain	Green mud crab	3,972	56	720	698
Callinectes sapidus	Blue crab	10,563	2,104	173	161
Carcinus maenas	Green crab	15,559	7,672	273	275

[a] The number of transcript contigs in each species is the summation of all contig sequences constructed by a set of ESTs and by a set of SRA reads with CAP3 (with default or 97%-similarity parameters) and Newbler (with default parameters).

[b] Including SRA transcript contigs produced by Newbler.

1.2 Updated the database by NGS datasets

Recently most ESTs have been generated by NGS instead of by a traditional cDNA library approach and a number of datasets have been available both in public domains and in private collections of our and our collaborators' laboratories. Key aspects of transcriptomes by NGS are a reduction in bias in clone selection and a high coverage of a transcriptome of interest. Therefore, in this study we proposed to collect these data and mainly focus on data from *P. monodon* and *P. vannamei*. Tables 2 and 3 show description and the number of sequences for 11 datasets for *P. monodon* and 17 datasets for *P. vannamei* we have collected, respectively. These data were generated from various conditions of shrimp such as normal shrimp, virus-infected shrimp or shrimp survivors from virus-infection. *P. monodon* datasets contain the total of 99.3 million sequences, whereas *P. vannamei* datasets contains 248.7 million reads (Table 2).

Table 2 The statistics of our collection of *P. monodon* transcriptome data from the next-generation sequencers.

SRA Run Acc. No.	NGS platforms	Description	# of reads
SRR388207	Illumina Genome Analyzer II	India WSSV-resistant shrimp from a heavy infection	29,695,294
SRR388221	Illumina Genome Analyzer II	India Andaman Island WSSV-resistant shrimp from a heavy infection	38,865,759
SRR388222	Illumina Genome Analyzer II	East coast India WSSV-resistant shrimp from a heavy infection	29,613,680
SRR577080	454 GS FLX	SSH of Survivor shrimp from WSSV infection *vs.* normal shrimp	240,897
Locally generated	454 GS FLX	Immature ovary	112,893
Locally generated	454 GS FLX	Mature ovary	122,493
Locally generated	454 GS FLX	Immature testis	119,780
Locally generated	454 GS FLX	Mature testis	113,575
Locally generated	454 GS FLX	Control shrimp	212,011
Locally generated	454 GS FLX	Moribund shrimp from WSSV infection	94,132
Locally generated	454 GS FLX	Survivor of shrimp from WSSV infection	151,239
		Total *P. monodon*	99,341,753

Table 3 The statistics of our collection of *P. vannamei* transcriptome data from the next-generation sequencers.

SRA Run Acc. No.	NGS platforms	Description	# of reads
SRR346404	Illumina HiSeq 2000	*Litopenaeus vannamei* transcriptomes (normal)	13,697,473
SRR653437	Illumina HiSeq 2000	Identification genes involved in TSV-resistance of Litopenaeus vannamei.	204,712,407
SRR839222	Illumina HiSeq 2000	Transcriptome of shrimp in early development stages: embryo	99,563
SRR1037362	Illumina HiSeq 2000	Transcriptome of shrimp in early development stages: embryo	49,814
SRR1037365	Illumina HiSeq 2000	Transcriptome of shrimp in early development stages: Mysis	512,188
SRR842625	Illumina HiSeq 2000	Transcriptome of shrimp in early development stages: Mysis	208,799
SRR839236	Illumina HiSeq 2000	Transcriptome of shrimp in early development stages: nauplii	125,402
SRR1037363	Illumina HiSeq 2000	Transcriptome of shrimp in early development stages: nauplii	202,065
SRR842627	Illumina HiSeq 2000	Transcriptome of shrimp in early development stages: postlarval	255,170
SRR1037366	Illumina HiSeq 2000	Transcriptome of shrimp in early development stages: postlarval	429,357
SRR842572	Illumina HiSeq 2000	Transcriptome of shrimp in early development stages: protozoea	171,542
SRR1037364	Illumina HiSeq 2000	Transcriptome of shrimp in early development stages: protozoea	827,980
SRR1039534	Illumina HiSeq 2000	Individuals at development stage of Mysis	26,951,393
SRR554363	454 GS FLX	Transcriptome of WSSV-infected shrimp	159,742
SRR554364	454 GS FLX	Transcriptome of non-infected shrimp	101,725
SRR554365	454 GS FLX	Transcriptome of TSV-infected shrimp	131,745
SRR556131	454 GS FLX	Transcriptome of non-infected shrimp	106,965
		Total *P. vannamei*	248,743,330

First, we performed *do novo* assembly with MIRA assembler to combine NGS transcriptome data with the traditionally-generated ESTs for *P. monodon*. This stetragy was to investigate whether combining the two types of datasets would increase overall length of transcripts. We obtained a set of *P. monodon* transcript contigs that were produced by a combined dataset of all traditionally-generated EST sequences and a set of 454/Roche run (SRR577080). Table 4 shows that the number of contigs obtained with the combined dataset is more than that of EST sequences alone (24,927 *vs.* 13,250), and the length distribution of these contigs are shown in Fig 1. The combined assembly contains more number of contigs than the traditionally-generated EST assembly, but the distributions of transcript lengths are similar among the two assemblies. Therefore, the sequence length was not improved much when combine EST and NGS datasets together.

Table 4 Descriptive statistics of contigs assembled by MIRA on two datasets of *P. monodon*.

Dataset	EST	EST+NGS (SRR577080) (a01_mira)
# of reads	91,498	533,407
# of assembled reads	51,139	185,005
# of contigs	13,250	24,927
Contig length (min-max)	80-4945	44-4942

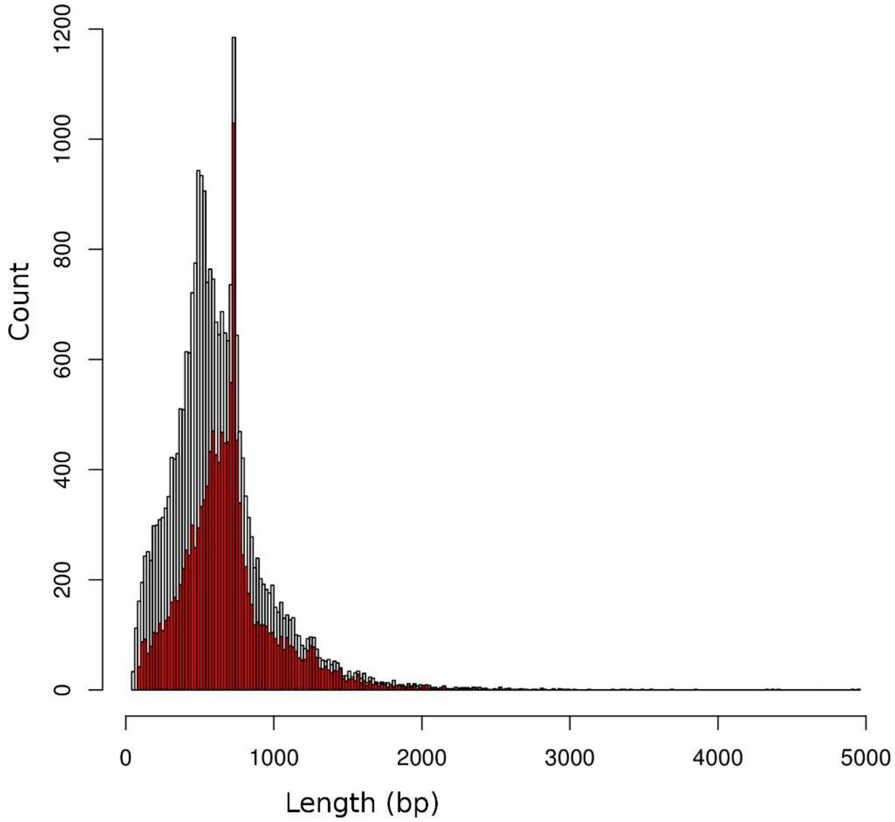


Figure 1 Length distributions of contigs from the datasets of EST (red) and of EST and NGS (white).

Second, we grouped the datasets into several sets of small size and assembled them separately. This was considered the large amount of data and limited computational resource. Separately assembling NGS datasets of related experimental conditions will likely produce a more biolocally meaningful assembly that does assemblying several unrelated NGS datasets. We have completed the assembly process for all NGS datasets for both *P. monodon* and *P. vannamei* with Newbler, MIRA, or Trinity assemblers. Note that all of assemblers produce transcript contigs, whereas Trinity produces a collection of “genes” by grouping similar transcript contigs that might come from the alternative splicing process (Table 2).

Table 5 shows the number of transcript contigs for each dataset. Here, almost all of the assemblies were produced by each dataset individually to reduce computational time as well as to reduce mixed datasets, which could result in a poor quality of assemblies. However, some of assemblies were produced by multiple datasets from the same experiment design and/or the same platform. The total 41 assemblies include 19 assemblies for *P. monodon* and 22 for *P. vannamei*.

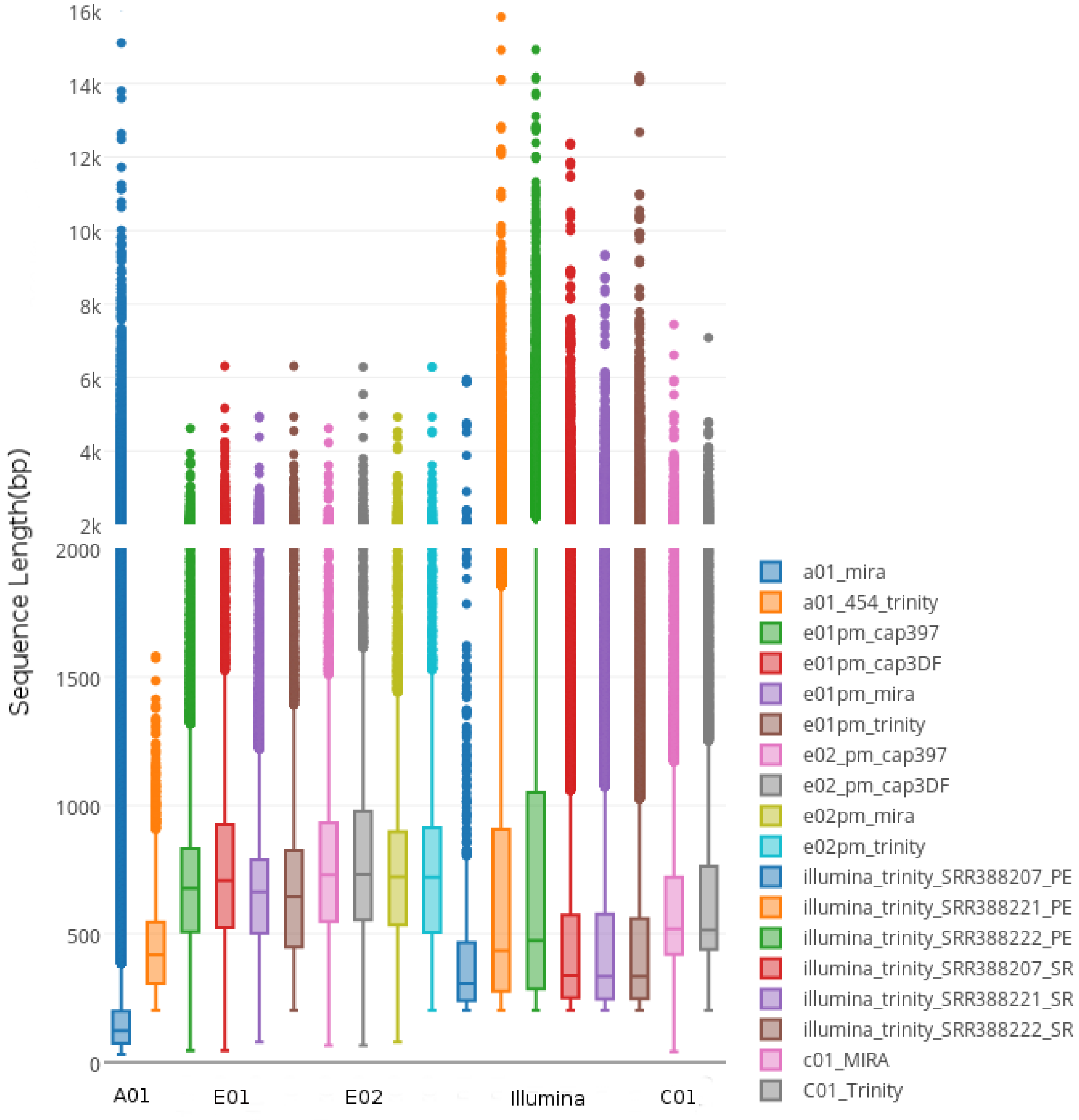


Figure 2 Length distribution of *P. monodon* assemblies.

The EST datasets were assembled by all four assemblers using either EST dataset alone or combined with some NGS datasets (Tables 5). The assemblies of combined EST and NGS (SRR577080) data of *P. monodon* by MIRA was described above (Table 4) and by Trinitiy (a01_454_Trinity) was describe here. For *P. monodon*, MIRA produced the highest number of contigs followed by CAP3, and Trinity produced the lowest number of contigs. CAP3 (default) and Trinity gave some longer contigs, but the length distributions are similar among four assemblers (Fig. 2). Similarly, a similar pattern was observed for *P. vannamei* EST, NGS-single-read and NGS-paired-end assemblies (Figs. 3, 4 and 5, respectively). For *P. monodon*, EST datasets and 454 datasets were assembled together with MIRA (the a01_mira assembly). This assembly contains more number of contigs than the assemblies of only ESTs (635,142 contigs) and produced some longer contigs (maximum length = 16,124 bps; Table 4 and Fig 2).

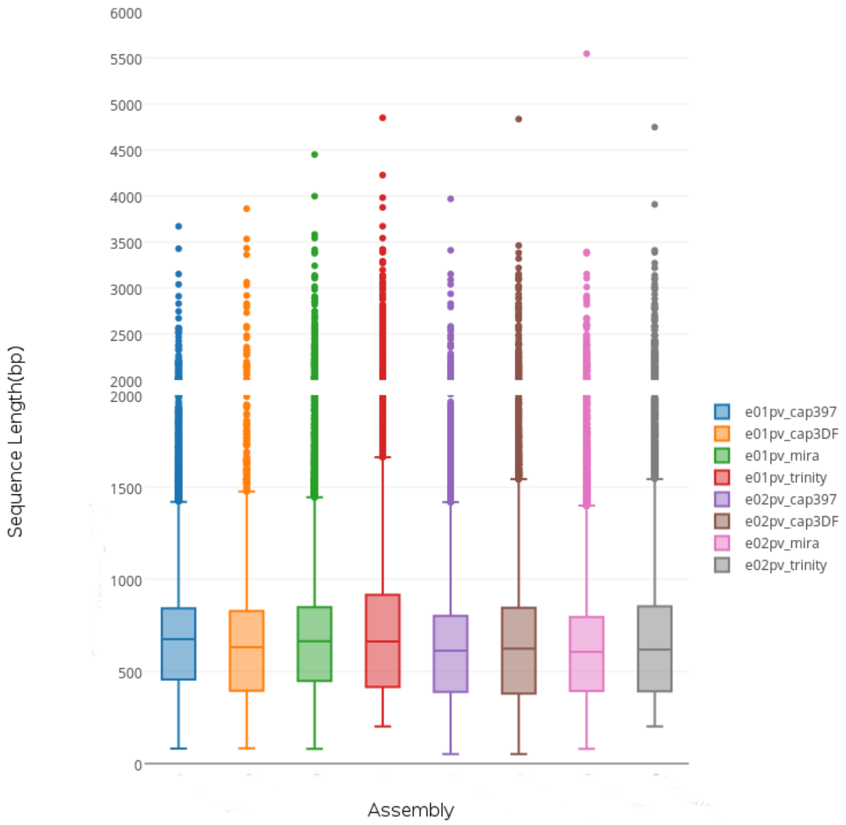


Figure 3 Length distribution of *P. vannamei* EST assemblies.

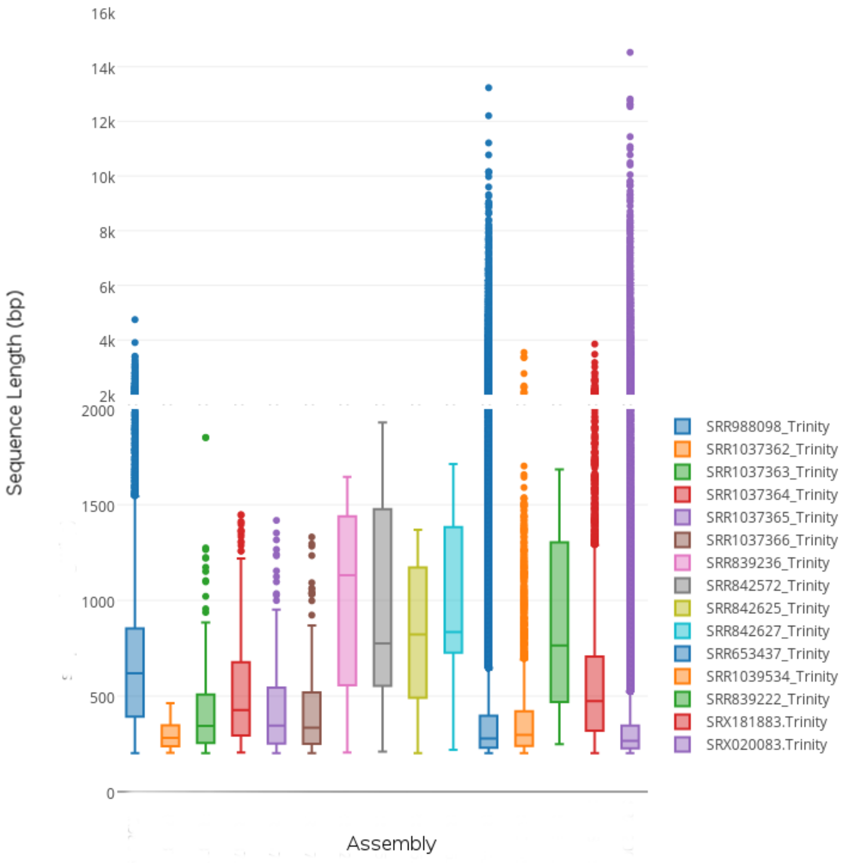


Figure 4 Length distribution of *P. vannamei* NGS-single-read assemblies.

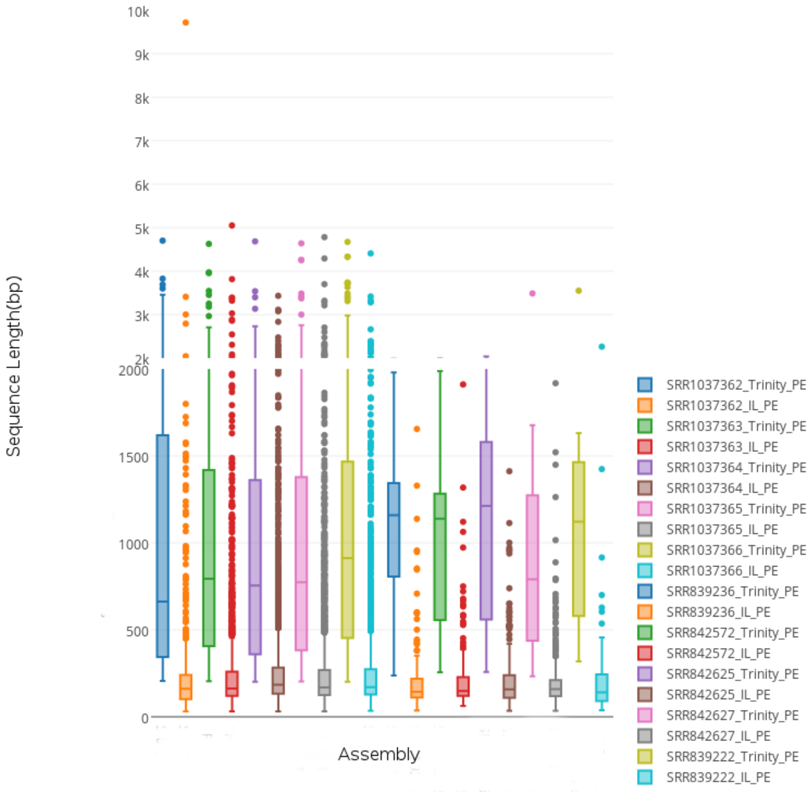


Figure 5 Length distribution of *P. vannamei* NGS-paired-end-read assemblies.

For 454 datasets of *P. monodon*, we focused on the locally generated datasets (C01; BIOTEC-cuticular contains BIOTEC-CC, BIOTEC-CM, and BIOTEC-CS) from cuticular tissues of shrimp challenged by WSSV. The assemblies of C01 were produced by Trinity, MIRA and Newbler. MIRA produced more contigs than the other assemblers (~22,500 vs. 13,000-16,000), but with a similar length distribution (Fig. 2). We used these assemblies for the subsequent analyses of differentially expressed genes. *P. vannamei* 454 datasets were combined and assembled by Trinity (SRX181883.Trinity). These assemblies were derived from the transcriptome datasets in the same experiment by TSV- and WSSV-challenged samples (Table 6) and contain about 5500 contigs (and genes).

Table 5 The transcriptome assemblies for both ESTs and NGS datasets of *P. monodon*

Assembly NAME	Input				Assembler	Output			
	dataset	# reads (sequences)	Platform[a]	Layout[b]		#contigs	#Gene	Min[c]	Max[c]
P. monodon									
e01_pm_mira	e01_pm	91698	EST	S	MIRA	13250	N/A	80	4945
e01_pm_cap397	e01_pm	91698	EST	S	CAP397	10357	N/A	46	4613
e01_pm_cap3DF	e01_pm	91698	EST	S	CAP3DF	8634	N/A	46	6309
e01_pm_Trinity	e01_pm	91698	EST	S	Trinity	7523	7081	201	6308
e02_pm_mira	e02_Pm	52060	EST	S	MIRA	5881	N/A	80	4927
e02_pm_cap97	e02_Pm	52060	EST	S	cap397	4130	N/A	66	4613
e02_pm_capDF	e02_Pm	52060	EST	S	cap3DF	3938	N/A	66	6286
e02_pm_Trinity	e02_pm	52060	EST	S	Trinity	3152	3013	202	6286
a01_mira	a01	180748727	EST	S	Mira	635142	N/A	31	16124
a01_454_Trinity	SRR577080	240897	454	S	Trinity	2682	2317	201	1581
illumina_trinity_SRR388207_PE	SRR388207	25366741	Il GAII	P	Trinity	1106	1046	201	5950
illumina_trinity_SRR388221_PE	SRR388221	33342406	Il GAII	P	Trinity	57417	44610	201	15824
illumina_trinity_SRR388222_PE	SRR388222	23894792	Il GAII	P	Trinity	65467	50599	201	14928
illumina_trinity_SRR388207_SR	SRR388207_SR	4402610	Il GAII	S	Trinity	41924	36101	201	12386
illumina_trinity_SRR388221_SR	SRR388221_SR	5358202	Il GAII	S	Trinity	37249	30929	201	9345
illumina_trinity_SRR388222_SR	SRR388222_SR	5580614	Il GAII	S	Trinity	44875	39402	201	14199
c01_mira	c01 (BIOTEC-cuticular)	453353	454	S	MIRA	22592	N/A	40	7440
co1_Trinity	c01	453353	454	S	Trinity	13262	12556	201	7086
c01_Newbler	c01	453353	454	S	Newbler	16614	N/A	70	12340

[a] sequencing platforms: EST, 454 (pyrosequencing), Illumina (Il GAII [Genome Analyzer II] or HiSeq 2000)

[b] sequencing layouts: S (Single end) and P (Paired end)

[c] Minimum and Maximum length (bps)

Table 6 The transcriptome assemblies for both ESTs and NGS datasets of *P. vannamei*.

Assembly NAME	Input				Assembler	Output			
	dataset	# reads (sequences)	Platform^{a}	Layout^{b}		#contigs	#Gene	Min ^{c}	Max^{c}
P. monodon									
e01_pm_mira	e01_pm	91698	EST	S	MIRA	13250	N/A	80	4945
e01_pm_cap397	e01_pm	91698	EST	S	CAP397	10357	N/A	46	4613
e01_pm_cap3DF	e01_pm	91698	EST	S	CAP3DF	8634	N/A	46	6309
e01_pm_Trinity	e01_pm	91698	EST	S	Trinity	7523	7081	201	6308
e02_pm_mira	e02_Pm	52060	EST	S	MIRA	5881	N/A	80	4927
e02_pm_cap97	e02_Pm	52060	EST	S	cap397	4130	N/A	66	4613
e02_pm_capDF	e02_Pm	52060	EST	S	cap3DF	3938	N/A	66	6286
e02_pm_Trinity	e02_pm	52060	EST	S	Trinity	3152	3013	202	6286
a01_mira	a01	180748727	EST	S	Mira	635142	N/A	31	16124
a01_454_Trinity	SRR577080	240897	454	S	Trinity	2682	2317	201	1581
illumina_trinity_SRR388207_PE	SRR388207	25366741	Il GAII	P	Trinity	1106	1046	201	5950
illumina_trinity_SRR388221_PE	SRR388221	33342406	Il GAII	P	Trinity	57417	44610	201	15824
illumina_trinity_SRR388222_PE	SRR388222	23894792	Il GAII	P	Trinity	65467	50599	201	14928
illumina_trinity_SRR388207_SR	SRR388207_SR	4402610	Il GAII	S	Trinity	41924	36101	201	12386
illumina_trinity_SRR388221_SR	SRR388221_SR	5358202	Il GAII	S	Trinity	37249	30929	201	9345
illumina_trinity_SRR388222_SR	SRR388222_SR	5580614	Il GAII	S	Trinity	44875	39402	201	14199
c01_mira	c01 (BIOTEC-cuticular)	453353	454	S	MIRA	22592	N/A	40	7440
co1_Trinity	c01	453353	454	S	Trinity	13262	12556	201	7086
c01_Newbler	c01	453353	454	S	Newbler	16614	N/A	70	12340
P. vannamei									
e01_pv_mira	e01_pv	163737	EST	S	MIRA	25901	N/A	80	4448
e01_pv_cap97	e01_pv	163737	EST	S	cap397	16690	N/A	81	3669
e01_pv_capDF	e01_pv	163737	EST	S	cap3DF	14451	N/A	83	3860
e01_pv_Trinity	e01_pv	163737	EST	S	Trinity	10858	10409	201	4847
eo2_pv_mira	e02_pv	162100	EST	S	MIRA	25960	N/A	80	5543
eo2_pv_cap97	e02_pv	162100	EST	S	cap397	16441	N/A	51	3967
eo2_pv_capDF	e02_pv	162100	EST	S	cap3DF	14637	N/A	51	4833
e02_pv_Trinity	e02_pv	162100	EST	S	Trinity	10430	10054	201	4746
SRR653437.Trinity	SRR653437.fastq	197297608	HiSeq2000	S	Trinity	163151	110916	201	17052
SRR839222.Trinity	SRR839222.fastq	988	HiSeq2000	S	Trinity	24	9	248	1685
SRR1037362.Trinity	SRR1037362.fastq	588	HiSeq2000	S	Trinity	43	40	203	462
SRR1037365.Trinity	SRR1037365.fastq	9402	HiSeq2000	S	Trinity	138	96	201	1418
SRR842625.Trinity	SRR842625.fastq	1889	HiSeq2000	S	Trinity	19	9	202	1368
SRR839236.Trinity	SRR839236.fastq	1578	HiSeq2000	S	Trinity	19	8	204	1645
SRR1037363.Trinity	SRR1037363.fastq	2598	HiSeq2000	S	Trinity	86	108	201	1851
SRR842627.Trinity	SRR842627.fastq	2158	HiSeq2000	S	Trinity	18	6	219	1713
SRR1037366.Trinity	SRR1037366.fastq	7659	HiSeq2000	S	Trinity	132	101	201	1331
SRR842572.Trinity	SRR842572.fastq	2182	HiSeq2000	S	Trinity	22	8	209	1929
SRR1037364.Trinity	SRR1037364.fastq	13954	HiSeq2000	S	Trinity	195	127	205	1148
SRR1039534.Trinity	SRR1039534.fastq	338219	HiSeq2000	S	Trinity	3679	3011	201	3544
SRX181883.Trinity	SRR554363.sra,SRR554364.sra, SRR554365.sra,SRR556131.sra	470097	454	S	Trinity	5506	5098	201	3852

SRR839222.Trinity	SRR839222(t1.fq,t2.fq)	83916	HiSeq2000	P	Trinity	11	11	319	3550
SRR839222_IL_PE	SRR839222(t1.fq,t2.fq)	83916	HiSeq2000	P	MIRA	75	N/A	37	2263
SRR1037362.Trinity	SRR1037362.(t1.fq,t2.fq)	49813	HiSeq2000	P	Trinity	166	132	206	4701
SRR1037362_IL_PE	SRR1037362.(t1.fq,t2.fq)	49813	HiSeq2000	P	MIRA	690	N/A	32	9718
SRR1037365.Trinity	SRR1037365(t1.fq,t2.fq)	512187	HiSeq2000	P	Trinity	331	265	203	4638
SRR1037365_IL_PE	SRR1037365(t1.fq,t2.fq)	512187	HiSeq2000	P	MIRA	3405	N/A	32	4781
SRR842625.Trinity	SRR842625(t1.fq,t2.fq)	208797	HiSeq2000	P	Trinity	18	14	258	2040
SRR842625_IL_PE	SRR842625(t1.fq,t2.fq)	208797	HiSeq2000	P	MIRA	679	N/A	34	1412
SRR839236.Trinity	SRR839236(t1.fq,t2.fq)	125402	HiSeq2000	P	Trinity	20	13	237	1980
SRR839236_IL_PE	SRR839236(t1.fq,t2.fq)	125402	HiSeq2000	P	MIRA	366	N/A	36	1655
SRR1037363.Trinity	SRR1037363(t1.fq,t2.fq)	202063	HiSeq2000	P	Trinity	257	216	204	4626
SRR1037363_IL_PE	SRR1037363(t1.fq,t2.fq)	202063	HiSeq2000	P	MIRA	1791	N/A	32	5050
SRR842627.Trinity	SRR842627(t1.fq,t2.fq)	255170	HiSeq2000	P	Trinity	21	17	233	3486
SRR842627_IL_PE	SRR842627(t1.fq,t2.fq)	255170	HiSeq2000	P	MIRA	679	N/A	34	1918
SRR1037366.Trinity	SRR1037366(t1.fq,t2.fq)	429354	HiSeq2000	P	Trinity	311	251	201	4671
SRR1037366_IL_PE	SRR1037366(t1.fq,t2.fq)	429354	HiSeq2000	P	MIRA	2993	N/A	34	4407
SRR842572.Trinity	SRR842572.(t1.fq,t2.fq)	171542	HiSeq2000	P	Trinity	18	14	257	1987
SRR842572_IL_PE	SRR842572.(t1.fq,t2.fq)	171542	HiSeq2000	P	MIRA	454	N/A	62	1911
SRR1037364.Trinity	SRR1037364(t1.fq,t2.fq)	827976	HiSeq2000	P	Trinity	303	255	202	4684
SRR1037364_IL_PE	SRR1037364(t1.fq,t2.fq)	827976	HiSeq2000	P	MIRA	4937	N/A	31	3142
SRR1039534.Trinity	SRR1039534(t1.fq,t2.fq)	26951256	HiSeq2000	P	Trinity	72137	60957	201	14366

[a] sequencing platforms: EST, 454 (pyrosequencing), Illumina (Il GAII [Genome Analyzer II] or HiSeq 2000)

[b] sequencing layouts: S (Single end) and P (Paired end)

[c] Minimum and Maximum length (bps)

For Illumina platform datasets, almost all of them are paired-end libraries. After processing by Trimmomatic and FLASh pipelines, the sequence were separated into 1) those sequences without pair-end sequences or pair-end sequences were combined into a single sequence and 2) those with pair-end sequences. Sequences of the former were assembled in the single-end nature (with Trinity), whereas those of the latter were assembled in the paired-end nature (with both MIRA and Trinity). Overall, MIRA produced more transcript contigs than Trinity, but length distributions are similar. Some of paired-end datasets of *P. vannamei* produced a small number of contigs (e.g., SRR839222, SRR842625, SRR839236, SRR842627 and SRR842572) even though they had a large number of reads.

Some of these assemblies (Tables 5 and 6) were reassembled or grouped for the second round by either CAP3 or CD-HIT-EST (Fu et al. 2012; Table 7). For examples, contigs of a01_MIRA assembly were grouped again by both CAP3 and CD-HIT-EST for formatting sequencing_IDs in Trinity format (the format that contains both gene_id and isoform_id for each contig). Here, each cluster of CD-HIT-EST was considered as a gene and members of such a cluster (e.g., MIRA contigs) were considered as isoforms of such a gene. This conversion of sequence_IDs will be used in the subsequent analyses of differentially expressed genes. Note that a large proportion of these sequences remain as singletons (i.e., could not found a similar sequence within an assembly by our CD-HIT-EST parameter setting). The proportion of contigs that were formed multimember groups are ~5%, ~10%, and ~20% for c01_newbler, c01_trinity and c01_mira, respectively.

Another aspect of grouping transcript contigs with CD-HIT-EST was to combining assemblies of the same datasets of paired-end layout but were assembled by single end and paired-end reads after the quality control process) together (see above). Thus, we combined the output contigs of these two assemblies with CD-HIT-EST (Table 7). Majority of the second round assemblers are predominant with singleton (>85% of the clusters).

Table 7 The assemblies produced by CD-HIT-EST or CAP3 of the assembled transcripts

Assembly NAME	Input[a]			Assembler	Output		
	Assembly NAME	#reads	Platform[b]		# Seq	Min[c]	Max[c]
P. monodon							
a01_2CAP397	a01_mira	635142	EST	CAP397	61715	42	16124
a01_2CAP3DF	a01_mira	635142	EST	CAP3DF	67374	42	16123
a01_2CDHIT	a01_mira	635142	EST	CD-HIT	433100	31	16124
illumina_trinity_SRR388207_2CDHIT	illumina_trinity_SRR388207_PE, illumina_trinity_SRR388207_SR	43030	Il GAII	cd-hit-est	40442	201	12386
illumina_trinity_SRR388221_2CDHIT	illumina_trinity_SRR388221_PE,illumina_trinity_SRR388221_SR	94666	Il GAII	cd-hit-est	64783	201	15824
illumina_trinity_SRR388222_2CDHIT	illumina_trinity_SRR388222_PE,illumina_trinity_SRR388222_SR	110342	Il GAII	cd-hit-est	72885	201	14928
pm_P_trinity_e01_3CDHIT	a01_454_Trinity, e01_pm_Trinity, illumina_trinity_SRR388207_2CDHIT, illumina_trinity_SRR388221_2CDHIT, illumina_trinity_SRR388222_2CDHIT	258243	454	cd-hit-est	119264	201	15824
pm_APc01_trinity_e01_3CDHIT	a01_454_Trinity, e01_pm_Trinity, illumina_trinity_SRR388207_2CDHIT, illumina_trinity_SRR388221_2CDHIT, illumina_trinity_SRR388222_2CDHIT, co1_Trinity	271505	454	cd-hit-est	122986	201	15824
c01_Trinity_2CDHit	co1_Trinity	13262	454	CD-Hit-est	13077	201	7086
c01_MIRA_2CDHitEST	c01_mira	22592	454	CD-Hit-est	18812	40	7440
c_dSFFe_isotigs_2cdhit_1	c01_Newbler	16614	454	cd-hit-est[d]	15853	40	12340
c_dSFFe_isotigs_2cdhit_2	c01_Newbler	16614	454	cd-hit-est[d]	15854	40	12340
c_dSFFe_isotigs_2cdhit_3	c01_Newbler	16614	454	cd-hit-est[d]	15854	40	12340
c_dSFFe_isotigs_2cdhit_4	c01_Newbler	16614	454	cd-hit-est[d]	15479	40	12340
P. vannamei							
SRR839222.Trinity_2CDHIT	SRR839222.Trinity SE+PE	35	HiSeq2000	cd-hit-est	23	248	3550
SRR1037362.Trinity_2CDHIT	SRR1037362.Trinity SE+PE	209	HiSeq2000	cd-hit-est	177	206	4701
SRR1037365.Trinity_2CDHIT	SRR1037365.Trinity SE+PE	469	HiSeq2000	cd-hit-est	392	201	4638
SRR842625.Trinity_2CDHIT	SRR842625.Trinity SE+PE	37	HiSeq2000	cd-hit-est	26	258	2040
SRR839236.Trinity_2CDHIT	SRR839236.Trinity SE+PE	39	HiSeq2000	cd-hit-est	27	237	1980
SRR1037363.Trinity_2CDHIT	SRR1037363.Trinity SE+PE	365	HiSeq2000	cd-hit-est	310	202	4626
SRR842627.Trinity_2CDHIT	SRR842627.Trinity SE+PE	39	HiSeq2000	cd-hit-est	31	219	3486
SRR1037366.Trinity_2CDHIT	SRR1037366.Trinity SE+PE	443	HiSeq2000	cd-hit-est	375	201	4671
SRR842572.Trinity_2CDHIT	SRR842572.Trinity SE+PE	40	HiSeq2000	cd-hit-est	32	209	1987
SRR1037364.Trinity_2CDHIT	SRR1037364.Trinity SE+PE	498	HiSeq2000	cd-hit-est	393	203	4684
SRR1039534.Trinity_2CDHIT	SRR1039534.Trinity SE+PE	74816	HiSeq2000	cd-hit-est	69010	201	14366

[a] Input sequences were contigs produced previously (see Table 1 or Table2)

[b] sequencing platforms: EST, 454 (pyrosequencing), Illumina (Il GAII [Genome Analyzer II] or HiSeq 2000)

[c] Minimum and Maximum length (bps)

[d] different parameter settings

2. Putative functional prediction for protein-coding genes

2.1 Putative functional prediction for protein-coding genes in ShrimpGPAT Release #2

All nucleotide sequences (EST, transcript contigs and cDNA sequences) were queried (BLASTN and BLASTX) against the nt and nr databases, respectively. BLASTP was performed for protein sequences against the nr database. Homologous sequences were defined as hits with ≥50% alignable region of a query sequence, *E*-value < 10^{-6} (for BLASTN) or < 10^{-4} (for BLASTX and BLASTP), and identity of ≥70% (BLASTN) or of ≥25% (BLASTX and BLASTP). GO classification of each shrimp sequence was derived from its homologous proteins or nucleotides by mapping to information from the Protein Information Resource (http://pir.georgetown.edu/). GO functional classification and putative gene identification from BLAST were stored in the ShrimpGPAT database (http://shrimpgpat.sc.mahidol.ac.th/ShrimpGPATV2/).

All sequences in Table 1 were search for homologous sequences. The information of homologs and GO classification for these sequences were deposited in the ShrimpGPAT database for an ease of searching and query. Furthermore, the ShrimpGPAT system allows users with expertise in the fields to annotate and curate such information. This feature will further enrich and improve such annotation of shrimp genes.

2.2 Putative functional prediction for protein-coding genes by Trinotate

Annotation of nucleotide sequences for EST and contig datasets was carried out by a modified Trinotate pipeline (https://trinotate.github.io/; here, we called byTrinotate; Fig 6). Briefly, sequences (namely, Trinity.fasta) were screened for WSSV sequences via BLASTN (Step W), clustered with known shrimp cDNA sequences (Step C), and clustered with other known sequences from previously characterized shrimp contigs/ESTs (Step F). The sequences without similarity to known sequences from Steps W, C and F were fed to BLASTX and BLASTP of Trinotate pipeline. All nucleotide sequences were used in almost all of Trinotate steps (prediction of protein coding sequences [Transdecoder; https://transdecoder.github.io/], prediction for protein domains [HMMer; Finn et al. 2011], signal peptides [SignalP; Petersen et al. 2011], transmembrane regions [TmHMM; Krogh et al. 2011], rRNA [RNammer; Lagesen et al. 2007]), except BLASTX and BLASTP which were for the sequences without similarity to known sequences from Steps W, C and F. In-house scripts were used to generate short description for annotated sequences.

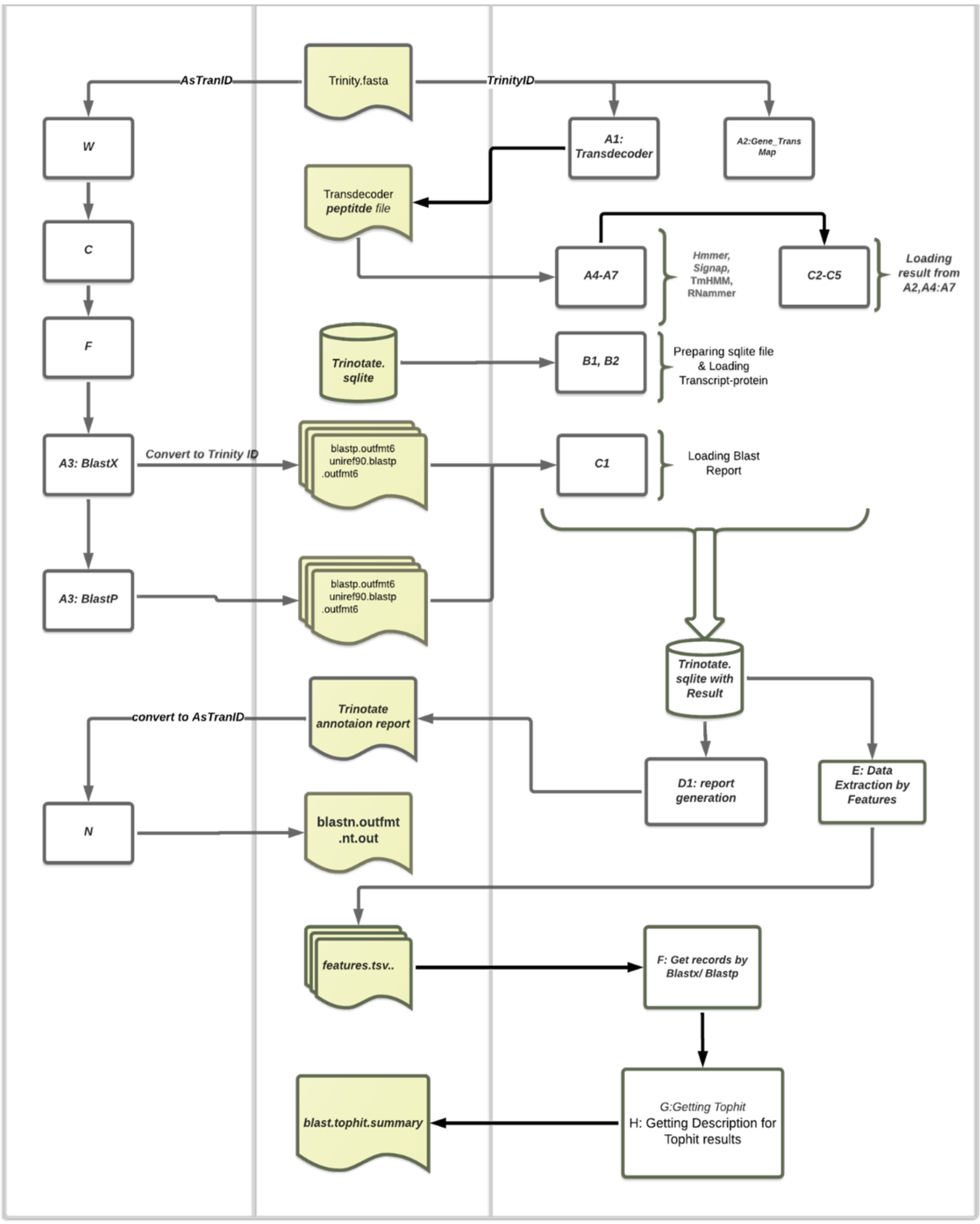

Trinity.fasta
AsTranID
TrinityID
W
A1: Transdecoder
A2:Gene_Trans Map
Transdecoder peptitde file
C
A4-A7
Hmmer, Signap, TmHMM, RNammer
C2-C5
Loading result from A2,A4:A7
F
Trinotate. sqlite
B1, B2
Preparing sqlite file & Loading Transcript-protein
A3: BlastX
Convert to Trinity ID
blastp.outfmt6 uniref90.blastp .outfmt6
C1
Loading Blast Report
A3: BlastP
blastp.outfmt6 uniref90.blastp .outfmt6
Trinotate. sqlite with Result
convert to AsTranID
Trinotate annotaion report
D1: report generation
E: Data Extraction by Features
N
blastn.outfmt .nt.out
features.tsv..
F: Get records by Blastx/ Blastp
blast.tophit.summary
G:Getting Tophit
H: Getting Description for Tophit results

Figure 6 Beyond-Trinotate pipeline (byTrinotate pipeline).

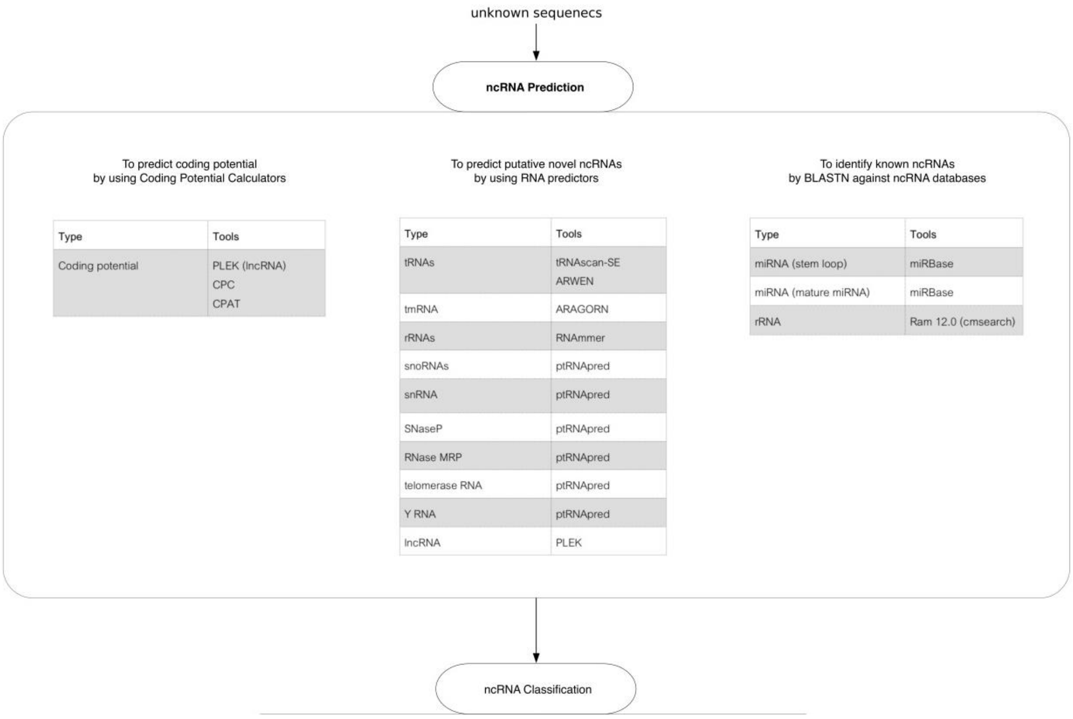
Annotation of unique sequences for EST and contigs was carried out by a modified Trinotate pipeline (https://trinotate.github.io/; byTrinotate; Fig 6). Table 8 shows the number of sequences with features and descriptions. Here, E0102_PMPV has the lowest proportion of sequences that should be annotated (i.e., found a similar sequence in public databases via byTrinotate pipeline), 28%, which is comprised of 44,426 sequences. That of ESTs has ~30%, whereas ContigsV22 has the highest proportion, ~52%. Note that several of these putatively-annotated sequences can still remain uncharacterized due to their similar to uncharacterized sequences. Thus, the proportion of sequences with putative function can be lower. The unknown sequences were characterized for potential non-coding RNAs (see below). The information of Beyond-Trinotate pipeline for these sequences will be deposited in the next release of ShrimpGPAT database. Also, the pipeline will be applied for all of the datasets we have collected.

Table 8 The number of unique nucleotide sequences that were annotated via byTrinotate pipeline

Datasets	# of sequences	byTrinotate Steps					Total Annotated sequences	Unknown sequences	% annotated sequences
		W	C	F	T				
					Protein coding	rRNA			
EST	158299	12	1935	8813	37941	42	48701	109598	30.77
contigs									
ContigsV22	19495	1	136	69	10087	9	10293	9202	52.80
E0102_PMPV	61783	6	184	3521	13646	53	17357	44426	28.09
C01_Trinity	13262	133	n/a	n/a	8444	5	7725	5537	41.75
C01_MIRA	22592	464	n/a	n/a	11008	211	13065	9527	42.17

3. Non-coding RNA prediction

In this section, we used the unknown sequences of C01_Trinity and C01_MIRA assemblies (Table 8) to investigate whether they are putative non-coding RNAs (ncRNAs) with the pipeline described in Fig. 7. Breiftly, all nucleotide sequences were assessed for protein-coding potential by three programs, PLEK (Li et al. 2014), CPC (Kong et al. 2007), and CPAT (Wang et al. 2013). The sequences were also predicted for class-specific ncRNAs using the following programs: rRNAs (RNammer (Lagesen et al. 2007), and blasted against Rfam 12.0 (Nawrocki et al. 2014)); tRNAs (tRNA-scan-SE (Lowe and Eddy 1997) and ARAGORN (Laslett and Canback 2004)); tmRNA (ARAGORN); mtRNA (ARWEN (Laslett and Canbäck 2008)); snoRNA, snRNA, RNaseP, RNaseMRP and telomeraseRNA were predicted by ptRNApred (Gupta et al. 2014); miRNAs (blasted against miRBase (Kozomara and Griffiths-Jones 2014) for both step loop miRNAs and mature miRNAs).


unknown sequenecs
ncRNA Prediction
To predict coding potential by using Coding Potential Calculators
To predict putative novel ncRNAs by using RNA predictors
To identify known ncRNAs by BLASTN against ncRNA databases
ncRNA Classification

Type	Tools
Coding potential	PLEK (lncRNA) CPC CPAT

Type	Tools
tRNAs	tRNAscan-SE ARWEN
tmRNA	ARAGORN
rRNAs	RNAmmer
snoRNAs	ptRNApred
snRNA	ptRNApred
SNaseP	ptRNApred
RNase MRP	ptRNApred
telomerase RNA	ptRNApred
Y RNA	ptRNApred
lncRNA	PLEK

Type	Tools
miRNA (stem loop)	miRBase
miRNA (mature miRNA)	miRBase
rRNA	Ram 12.0 (cmsearch)

Figure 7 Pipeline for non-coding RNA prediction.

Table 9 The number of C01_Trinity and C01_MIRA transcripts with predicted ncRNAs.

Categories		Assemblies	
		C01_Trinity	C01_MIRA
Total unknown		5537	9527
non-coding	PLEK	5479	6868
	CPC	5363	9455
	CPAT	5227	8449
tRNA		1	1
tmRNA		0	0
mtRNA		103	260
snoRNA		2812	4494
snRNA		353	669
RNase P		1877	2897
RNase MRP		1	3
telomerase RNA		286	721
Y RNA		0	0
miRNA	stem loop miRNA	1	0
	mature miRNA	5	4
rRNA		5	211

Table 9 shows that almost all of the unknow sequences in both C01_Trinity and C01_MIRA do not have protein-coding potential and are likely to be non-coding RNAs by at least one of the three programs. Fewer sequences are predicted to contain potential sequences of class-specific ncRNAs. The information on ncRNAs of these WSSV-infected assemblies (C01_Trinity and C01_MIRA) was used in the following comparison to select for highly expressed ncRNA candidates in survivor samples.

4. Comparison between viral-infected libraries and to those from non-infected libraries

Here, we present the comparison by two types of data: NGS and EST libraries. For NGS, a comparison between WSSV-infected (survivor and moribund samples) and control sample from cuticular tissues was analyzed with Trinity and MIRA assemblies. For EST datasets, all available tissue and pathogen-challenge libraries were analyzed. To improve the protein-coding gene prediction for the contig sequences, we used Trinotate pipeline which includes BLASTX, BLASTP, RNAMMER, HMMER search for pfam domains, Signalp, and TmHMM (see above). The pipeline also gives Gene Ontology (GO) and putative polypeptide of transcripts. We also performed BLASTN against particular pathogen genomes (e.g., WSSV genomes) to identify pathogen sequences in the datasets.

4.1 NGS datasets: WSSV-challenged cuticular samples of *P. monodon*

The objective of the current analysis is to find sequences (genes) that are highly expressed in survivor samples after WSSV infection in *P. monodon*.

4.1.1 Trinity Assembly dataset (C01_Trinity)

As shown in Table 2, this assembly were combined the three NGS datasets of transcriptomes from the control, moribund and survivor shrimp with WSSV infection. The number of transcripts is 13,262, which are of 12,556 genes by Trinity. The transcripts have median length of 515.5 bps and mean length of 657.92.

Mapping with Bowtie2 (RSEM pipeline)

Table 10 shows the proportion of mappable reads ~46.2%; The number of reads mapped to reference assembly is relatively low. Here, a low proportion of mappable reads may be due to the default parameter setting use here for allowing only 1-mismatch. Note that the proportion of mappable reads of moribund samples is the lowest, while those of control and survivor samples are similar. Also, the number of raw reads of the control sample is the highest (Table 10).

Table 10 The number of mappable reads C01_Trinity reference assembly

Assembly	Sample	# Original Reads	# of Mappable reads	# of Unmappable	% of Mapped Read
C01_Trinity	All	453353	209566	243787	46.2
	Control	210205	98654	111551	46.9
	Moribund	93472	41897	51575	44.8
	Survivor	149676	69015	80661	46.11

Sequences with significantly differentially expressed (DE) at two-fold change

I obtained the list of transcripts and genes that are significantly differentially expressed (DE) at two-fold change and with various *p*-values (Table 11). P < 1e-5 gives 26 DE genes and 29 DE transcripts, while P < 1e-4 (a more relaxed criterion) gives additional 14 DE genes and 14 DE transcripts. Similarly, additional 26, 32 and 195 DE transcripts (23, 28 and 180 DE genes) are found for P < 1e-3, P < 1e-2 and P < 0.05, respectively. We investigated five groups of transcripts at these *p*-values (Table 11).

For groups of 29 DE transcripts (or genes) at P < 1e-5, we divided them into sample-specific transcripts (i.e., those with only mappable reads from only certain sample), including Survivor-specific (S), Moribund-specific (M), Control-specific (C), Survivor-and-Moribund-specific (MS; those with mappable reads from BOTH moribund and survivor samples), and O. (none of S, M, C or MS). Among 29 DE transcripts of P < 1e-5, MS has 8, C has 5 and O has 16, but none are found for S and M (Table 12). We grouped additional DE transcripts of each *P*-value shown in Table 11 and found that DE transcripts are found in S and M only for P < 0.05 (Table 12).

Table 11 The number of transcripts (and genes) that are differentially expressed (two-fold change).

P-Value		# of Genes		# of Transcripts	
Cat.	value	Total	Gained from the previous cat.	total	Gained from the previous cat.
1	0.000001	26	n/a	29	n/a
2	0.00001	40	14	43	14
3	0.0001	63	23	69	26
4	0.001	91	28	101	32
5	0.05	271	180	296	195

WSSV sequences

I searched transcript sequences against the WSSV genomes via BLASTN and found that WSSV are found in Types S, M and MS, but not in C or O. Almost all of 29 MS transcripts of P < 1e-2 are of WSSV, except two transcripts (1 in each of P < 1e-3 and P < 1e-2). The putative gene names of these two transcripts are Ribonucleoside-diphosphate reductase (P < 1e-3) and Serpin B (P < 1e-2). For P < 0.05, WSSV sequences are found in S and MS (24/40 transcripts); interestingly, none of M DE transcripts are of WSSV (6). Similar to other *P*-value criteria, none of WSSV is found in C and O categories.

Table 12 Number of transcripts and genes that are significantly differentially expressed.

P-value cat.	p-value	Type	# of genes	# of transcripts				
				Total	WSSV	non-WSSV		
						total	unknown	known
1	0.00001	All	26	29	8	21	5	16
		S	0	0	0	0	0	0
		M	0	0	0	0	0	0
		MS	9	8	8	0	0	0
		C	5	5	0	5	3	2
		O	12	16	0	16	2	14
2	0.0001	All	14	14	4	10	2	8
		S	0	0	0	0	0	0
		M	0	0	0	0	0	0
		MS	3	4	4	0	0	0
		C	4	5	0	5	1	4
		O	7	5	0	5	1	4
3	0.001	All	23	26	9	17	8	9
		S	0	0	0	0	0	0
		M	0	0	0	0	0	0
		MS	7	10	9	1	0	1
		C	3	4	0	4	2	2
		O	13	12	0	12	6	6
4	0.01	All	28	32	6	27	6	8
		S	0	0	0	0	0	0
		M	0	0	0	0	0	0
		MS	7	7	6	1	0	1
		C	10	13	0	13	0	0
		O	0	13	0	13	6	7
5	0.05	All	180	195	24	171	46	127
		S	5	6	2	4	1	3
		M	5	6	0	6	4	2
		MS	43	34	22	12	5	7
		C	37	41	0	41	19	22
		O	90	108	0	108	17	93

Highly expressed protein-coding sequences in the survivor sample.

The objective here is to find what type of sequences that are highly expressed in the survivor sample than the control sample; thus, we focused on S, MS and O transcripts. Since *P* < 0.05 has a lower confidence and has more number of sequences, we focused on the first four p-value categories. S does not contain any sequences for these four p-value categories (Table 13), and the majority of transcript sequences in MS are of WSSV. Only two transcripts are Ribonucleoside-diphosphate reductase (*P* < 1e-3) and Serpin B (*P* < 1e-2). **Heat shock protein (perhaps, 22)** has two transcripts (one each of *P* < 1e-5 and *P* < 1e-3) with highly expressed in Survivor sample than Control sample. **P < 1e-4** has a transcript of Sarcoplasmic calcium-binding protein. **P < 1e-3** has transcripts of Superoxide dismutase and Cytochrome b5. **P < 1e-2** has transcripts of putative Triosephosphate isomerase, Serine protease easter and Single insulin-like growth factor-binding domain protein-1. Note that there are several uncharacterized sequences (Table 13).

For transcripts that are significantly found in Survivor sample than control sample at *P* < 0.05 (Cat. 5; Table 12), **three out of four non-WSSV transcripts in S** were identified as putative Ankyrin-1, Aspartyl/asparaginyl beta-hydroxylase, and Low-density lipoprotein receptor 1. **For 12 non-WSSV MS transcripts**, five were identified as Coiled-coil domain-containing protein 25, Galactose-specific lectin nattectin, GSK-3-binding protein, **heat shock**, CUB and sushi domain-containing protein 2, Protein kinase shaggy, Esterase FE4.

The **known non-WSSV transcripts** in O were identified as T-complex protein 1 subunit eta, Serine protease easter, Alanine aminotransferase 2, Balbiani ring protein 3, Cathepsin L, Charged multivesicular body protein 2b, GTP-binding protein A, **Heat shock protein 22,** Hexokinase type 2, **Innexin (two transcripts)**, Killer cell lectin-like receptor subfamily G member 1, Leukocyte elastase inhibitor, **L-lactate dehydrogenases (3 transcripts**), alpha-2-macroglobulin (Murinoglobulin-1), Ovochymase-2, Protein bunched, class 1/class 3/D/E isoforms, Serine protease easter, Thrombospondin-3b, Venom protein 302, von Willebrand factor, and Zonadhesin.

Table 13 Highly expressed transcripts in Survivor sample.

	Expression (fpkm)			
ID	C	M	S	putative gene
Cat.1 (*P* < 1e-5)				
c3896_g1_i1	26.01	268.25	1911.06	Heat shock protein 22
Cat.3 (*P* < 1e3)				
c3761_g1_i1	19.88	64.75	1713.92	Superoxide dismutase [Cu-Zn]
c3805_g1_i1	13.08	298.23	984.97	Cytochrome b5
c3955_g1_i1	8.83	330.66	587.66	Heat shock protein 22
c3775_g1_i1	39.26	2237.93	1371.11	Viral responsive protein
c3961_g1_i1	16.19	724.9	353.28	Putative uncharacterized protein
c3701_g1_i1	0	189.07	93.52	Ribonucleoside-diphosphate reductase subunit M2
Cat.4 (P < 1e2)				
c4058_g1_i1	62.38	833.01	204.22	Putative uncharacterized protein
c3880_g1_i1	44.83	876.16	234.85	Triosephosphate isomerase
c3845_g1_i1	31.02	858.93	699.72	Serine protease easter
c4025_g1_i1	13.52	308.2	870.36	Single insulin-like growth factor-binding domain protein-1
c3211_g1_i1	0	244.26	260.4	Serpin B8
Cat.5 (*P* < 0.05)				
c3020_g1_i1	0	0	119.43	Ankyrin-1
c2135_g1_i1	0	0	435.46	Uncharacterized protein
c2037_g1_i1	0	0	414.84	Aspartyl/asparaginyl beta-hydroxylase
c11626_g1_i1	0	0	167.97	Low-density lipoprotein receptor 1
c203_g1_i1	0	35.9	409.01	Coiled-coil domain-containing protein 25
c3885_g1_i1	0	335.94	393.95	Galactose-specific lectin nattectin
c1455_g1_i1	0	36.74	184.66	GSK-3-binding protein
c4050_g1_i1	0	203.07	172.37	Heat shock
c3907_g1_i1	0	315.1	140.77	CUB and sushi domain-containing protein 2
c182_g1_i1	0	27.66	139.03	Protein kinase shaggy
c3002_g1_i1	0	108.81	115.45	Esterase FE4
c4060_g1_i1	0	111.43	696.93	Abnormal spindle-like microcephaly-associated protein homolog
c3819_g1_i2	0	50.01	301.61	Uncharacterized protein
c3593_g1_i1	0	57.22	325.94	Phospholipase
c2510_g1_i1	0	48.99	246.22	Sugar transporter
c2507_g1_i1	0	20.45	232.96	
c3479_g1_i1	108.51	0	734.22	Tribolium castaneum similar to Myosin heavy chain
c3469_g1_i1	189.3	0	195.72	Thrombospondin

c2896_g2_i1	71.41	87.22	655.91	T-complex protein 1 subunit eta
c3659_g2_i1	23.98	351.55	366.45	Serine protease easter
c3890_g2_i1	16.06	313.89	70.11	Alanine aminotransferase 2
c3625_g1_i1	408.91	3509.84	1097.58	Balbiani ring protein 3
c3966_g1_i1	1269.22	6047.35	3580.83	Cathepsin L
c2496_g1_i2	8.41	191.69	73.4	Charged multivesicular body protein 2b
c3954_g1_i3	93.39	589.39	235.7	GTP-binding protein A
c3896_g1_i2	16.42	133.7	645.06	Heat shock protein 22
c3934_g1_i1	27.97	330.24	114.46	Hexokinase type 2
c3723_g1_i1	20.09	359.84	306.89	Innexin inx2
c3343_g1_i1	9.02	190.93	59.05	Innexin inx3
c3351_g1_i1	11.05	341.98	313.6	Killer cell lectin-like receptor subfamily G member 1
c3944_g1_i1	85.59	542.09	280.23	Leukocyte elastase inhibitor
c3909_g1_i1	17.78	212.3	19.4	L-lactate dehydrogenase
c3909_g1_i2	7.1	219.67	38.74	L-lactate dehydrogenase
c3909_g1_i3	21.97	250.44	39.96	L-lactate dehydrogenase
c3971_g1_i1	48.47	308.55	72.13	Murinoglobulin-1
c3577_g1_i1	256.35	460.7	1678.65	Ovochymase-2
c2310_g1_i2	11.59	283.21	37.96	Protein bunched, class 1/class 3/D/E isoforms
c3659_g2_i2	63.47	503.94	398.31	Serine protease easter
c1127_g1_i1	58.89	517.92	77.12	Thrombospondin-3b
c4025_g1_i2	21.82	213.17	833.28	Venom protein 302
c3846_g1_i1	23.12	320.09	37.85	von Willebrand factor
c4057_g1_i1	208.55	16.57	272.02	Zonadhesin
c3167_g1_i1	10.44	161.6	51.29	Putative uncharacterized protein
c3314_g1_i1	9.42	199.53	10.29	
c3607_g1_i1	40.2	360.13	614.28	GPI ethanolamine phosphate transferase 1
c3766_g1_i1	129.58	253.25	1499.07	
c3861_g1_i1	876.37	4419.42	2777.11	

4.1.2 MIRA Assembly dataset (C01_MIRA)

The number of mappable reads to the MIRA assembly is 60~%, which is higher than that of Trinity assembly (Table 14 vs. Table 10). This is likely due to a higher number of transcripts in MIRA (22,592 transcripts for 18,108 genes).

Table 14 The number of mappable reads of C01_MIRA reference assembly

Assembly	Sample	# Original Reads	# of Mappable reads	# of Unmappable	% of Mapped Read
C01_MIRA	All	453353	274744	178609	60.6
	Control	210205	128150	82055	60.9
	Moribund	93472	56006	37466	59.9
	Survivor	149676	90588	59088	60.5

Highly expressed sequences in the survivor sample.

The results for C01_MIRA reference sequences are similar to those of C01_Trinity. A higher number of sequences were obtained at each *P*-value category, but these sequences are of similar functions with those observed for C01_Trinity.

4.1.3 Candidate highly expressed ncRNA sequences in the survivor sample

Highly expressed ncRNA sequences in the survivor sample.

Table 13 shows several transcripts that cannot be annotated as protein-coding sequences. These sequences have potential ncRNAs as shown in Table 15. These sequences are of interest for further investigation of their functions.

Table 15 Highly expressed ncRNA transcripts in Survivor sample.

	Expression (fpkm)			
ID	C	M	S	putative ncRNAs
<u>Cat.5 (*P* < 0.05)</u>				
c2507_g1_i1	0	20.45	232.96	Non-coding (PLEK, CPC, CPAT); telomeraseRNA
c3314_g1_i1	9.42	199.53	10.29	Non-coding (PLEK, CPC, CPAT); telomeraseRNA; pseudo/mtRNA
c3766_g1_i1	129.58	253.25	1499.07	Non-coding (PLEK, CPC, CPAT); RNaseP
c3861_g1_i1	876.37	4419.42	2777.11	Non-coding (PLEK, CPC, CPAT); RNaseP

Recently, long non-conding RNAs (lncRNAs) have gained attention. We focused on identifying candidate lncRNAs in Survivor samples by filtering out for transcripts that were predicted by all three programs of PLEK, CPC, and CPAT, but were not predicted to be other class-specific ncRNAs. Potential lncRNAs of C01_Trinity and C01_MIRA are 187 and 219 sequences, respectively (Table 16). Among these, 29 and 38 of C01_Trinity and C01_MIRA, respectively, were found only in the Survivor sample (see sample transcripts in Table 17).

Table 16 Number of candidate long ncRNAs (lncRNAs)

Assemblies	C01_Trinity	C01_MIRA
# all unknown transcripts	5537	9527
# predicted by all PLEK, CPC, and CPAT	5030	5983
-- not predicted to be other class-specific ncRNAs	187	219
found in survivor only	29	38

Table 17 Examples of candidate long ncRNAs (lncRNAs) found in Survivor only.

	Expression (fpkm)				putative ncRNAs
ID	C	M	S	AsTransID	
C01_Trinity					
c6148_g1_i1	0	0	375.79	PM_C0102_06833	Non-coding (PLEK, CPC, CPAT)
c9566_g1_i1	0	0	369.99	PM_C0102_10218	Non-coding (PLEK, CPC, CPAT)
c12365_g1_i1	0	0	322.87	PM_C0102_12984	Non-coding (PLEK, CPC, CPAT)
c437_g1_i1	0	0	261.25	PM_C0102_00440	Non-coding (PLEK, CPC, CPAT)
c9663_g1_i1	0	0	232.03	PM_C0102_10313	Non-coding (PLEK, CPC, CPAT)
C01_MIRA					
c16181_g1_i1	0	0	367.96	PM_C0101_20636	Non-coding (PLEK, CPC, CPAT)
c16341_g1_i1	0	0	336.73	PM_C0101_20797	Non-coding (PLEK, CPC, CPAT)
c15896_g1_i1	0	0	336.03	PM_C0101_20346	Non-coding (PLEK, CPC, CPAT)
c16285_g1_i1	0	0	219.19	PM_C0101_20741	Non-coding (PLEK, CPC, CPAT)
c15197_g1_i1	0	0	196.9	PM_C0101_19635	Non-coding (PLEK, CPC, CPAT)

4.2 EST datasets: pathogen-challenge samples of *P. monodon*

Mapping EST sequences to reference assemblies

Here, we used "unique" EST sequences from the selected libraries of EST data that can be compared to one another. The comparisons were performed between pathogen infected libraries (virus or bacteria) and those from non-infected libraries. Unique EST sequences from viral- and bacterial-infected libraries will be compared to those from non-infected library of the same tissue. Table 18 listed the selected libraries and the number of EST sequences to be considered by two sources: *P. monodon* EST Database Project and NCBI dbEST. The reference assemblies were those derived from the EST dataset (Table 5) by CAP3 (97% identity and default parameter settings [95% identity]), MIRA and Trinity. Sequence IDs of the first three assemblers were converted to Trinity format for mapping with Bowtie2 (RSEM pipeline; (Ref)).

Table 18 The selected EST libraries and their number of sequences.

Library_code	Library_ID	Description	#unique sequences
	P. monodon **EST Database Project**		
Tw-N	PmTwN	Normal Shrimp (Whole-PL20) Taiwan	6629
Tw-I	PmTwI	WSSV-challenged Shrimp (Whole-PL20) Taiwan	7193
HC-N	HC-N-N01	Hemocytes of juvenile cultured shrimp:Hemocyte-normalilzed	10364
HC-N	HC-N-S01	Hemocytes of juvenile cultured shrimp	591
HC-V	HC-V-S01	Hemocytes of juvenile cultured shrimp injected with Vibrio harveyi	440
HC-W	HC-W-S01	Hemocytes of juvenile SPF shrimp obtained from Broodstock Domesticated Program injected with WSSV	483
LP-N	LP-N-N01	Lymphoid organs of juvenile cultured shrimp:Lymphoid organ-normalized	942
LP-N	LP-N-S01	Lymphoid organs of juvenile cultured shrimp	404
LP-V	LP-V-S01	Lymphoid organs of juvenile cultured shrimp injected with Vibrio harveyi	625
LP-Y	LP-Y-S01	Lymphoid organs of juvenile cultured shrimp injected with YHV	692
	NCBI dbEST		
NCBI01-Gill-N	LIBEST_015692	PmBr cDNA Library	408
	LIBEST_024899	EST library from normal Indian tiger shrimp Penaeus monodon	
NCBI01-Gill-W	LIBEST_022651	WSSV infected EST library from Indian tiger shrimp P.monodon	333
NCBI01-HC-N	LIBEST_024264	Penaeus monodon hemocyte normalized library	866
	LIBEST_017443	Haemocyte cDNA plasmid library	
	LIBEST_025657	EST library from normal Indian tiger shrimp Penaeus monodon hemocytes	
	LIBEST_026170	Shrimp adult haemolymph	
	LIBEST_002851	Penaeus monodon total hemolymph cDNA library	
	LIBEST_003897	Penaeus monodon's total hemocyte cDNA library (#2)	
	LIBEST_015981	Hemocyte normal library	
NCBI01-HC-W	LIBEST_015468	Haemocyte-WSSV infected cDNA library	373
	LIBEST_021009	WSSV infected Penaeus monodon subtractive hybridization library	

NCBI01-HC-Y	LIBEST_025110	Suppression subtractive cDNA library for YHV infection	79
NCBI01-HC-V	LIBEST_016080	Hemocyte - Vibrio harveyi infected library	264
	LIBEST_020107	Hemocyte V. harveyi infected library	
NCBI01-LY-N	LIBEST_016181	Lymphoid organ library	1093
	LIBEST_016182	Lymphoid organ-normalized	
NCBI01-LY-Y	LIBEST_015768	Lymphoid organ - YHV infected library	615
	LIBEST_017786	Lymphoid organ - YHV challenged	
NCBI01-LY-V	LIBEST_016183	Lymphoid organ - Vibrio harveyi infected library	523
	LIBEST_017784	Lymphoid organ - Vibrio challenged	
NCBI01-WH-W	LIBEST_021064	WSSV infected Penaeus monodon post larvae cDNA library	62
NCBI01-WH-N	LIBEST_024920	RACE PCR Amplified Penaeus monodon cDNA Library	562
NCBI01-WH-V	LIBEST_024436	Vibrio harveyi challenged Tiger shrimp postlarvae cDNA library	704
	LIBEST_025579	Vibrio harveyi challenged Penaeus monodon postlarvae cDNA library	
NCBI01-ML-W	LIBEST_023446	WSSV infected Penaeus monodon cDNA library	5698
NCBI01-ML-N	LIBEST_025111	Gill-Epipodite normalized library	284
	LIBEST_026312	Gill-Epipodite library	
	LIBEST_001499	Black Tiger Shrimp Whole Cephalothorax UniZap library	
	LIBEST_007157	Shrimp Whole Cephalothorax UniZap library	

After mapping with Bowtie 2, the numbers of EST sequences mappable to each reference sequences are shown in Table 19. The result suggests that mappable EST sequences of the selected libraries to the CAP397 reference assembly has the highest proportion (>90%). The proportion of mappable EST sequences for the CAP3DF and MIRA reference assemblies are similar but lower than CAP397. Trinity reference assembly shows the lowest number proportion of mappable EST sequences. Therefore, we selected the result from the CAP397 reference assembly.

Table 19 The proportion of mappable EST sequences.

Library codes	Reference Assemblies			
	MIRA	CAP3DF	CAP397	Trinity
P. monodon EST Database Project				
Tw-N	76.51	88.81	92.16	61.86
Tw-I	78.76	88.20	92.08	59.31
HC-N	74.58	75.75	76.29	52.49
HC-N	61.25	57.87	66.50	38.92
HC-V	71.69	68.26	71.23	48.86
HC-W	83.64	80.95	81.78	70.60
LP-N	83.33	88.00	90.02	74.31
LP-N	75.25	85.15	88.86	64.36
LP-V	79.68	84.32	88.48	63.68
LP-Y	79.62	89.16	91.33	67.92
NCBI dbEST				
NCBI01-Gill-N	49.38	48.88	51.12	34.74

Library codes	Reference Assemblies			
	MIRA	CAP3DF	CAP397	Trinity
NCBI01-Gill-W	44.74	38.74	43.54	26.73
NCBI01-HC-N	58.28	58.40	63.57	40.89
NCBI01-HC-W	63.44	59.41	59.68	51.08
NCBI01-HC-Y	62.03	69.62	68.35	59.49
NCBI01-HC-V	74.90	71.48	74.90	47.91
NCBI01-LY-N	78.59	85.36	87.83	67.52
NCBI01-LY-Y	76.26	86.50	88.78	67.32
NCBI01-LY-V	76.29	82.98	87.38	59.66
NCBI01-WH-W	26.67	40.00	38.33	28.33
NCBI01-WH-N	63.72	58.66	64.62	37.00
NCBI01-WH-V	69.60	73.15	72.30	42.47
NCBI01-ML-W	20.99	29.01	27.41	9.07
NCBI01-ML-N	25.09	28.27	28.62	17.31

Clustering sequences of mapped CAP397 contigs with those un-mappable ESTs.

Since there are numbers of un-mappable EST sequences in each library, we asked whether these sequences can be clustered with those mapped contigs or with themselves. To investigate this, we grouped un-mappable EST sequences and mapped contigs by tissue types of each data source, e.g., TW (whole body), HC (Hemocytes), LP (Lymphoid) for P. monodon EST Database Project and NCBI01-Gill (gill), NCBI01-HC (Hemocytes), NCBI01-LY (Lymphoid), NCBI01-WH (whole body), NCBI01-ML (multiple tissues) for NCBI dbEST. Each set of sequences were clustered by CD-HIT-EST (parameters: 97% identity). Table 20 shows the number of clusters of each dataset with a high proportion of singleton clusters. The proportion of multimember clusters is ranging from 4% to 21%. Among these multimember clusters, many of un-mappable ESTs could be clustered with reference contigs. These clusters will be used update the number of EST presented in pathogen-challenged and non-challenged samples.

Table 20 The number of CD-HIT-EST clusters between mapped ref contigs and unmappable EST sequences.

Tissues	input		outputs				
	# mapped ref contigs	# Unmap. ESTs	#cluster	# singletons	# multimember	% of Multimember	# clusters with members from Ref. contigs and un-mappable EST
***P. monodon* EST Database Project**							
TW	4012	1090	3776	3292	484	13	340
HC	1995	2869	3709	3223	486	14	413
LP	1424	271	1441	1310	131	10	108
NCBI dbEST							
NCBI01-Gill	185	385	524	490	34	7	19
NCBI01-HC	697	551	1045	931	114	11	85
NCBI01-LY	1303	268	1359	1243	116	9	95
NCBI01-WH	312	742	491	388	103	21	83
NCBI01-ML	790	4318	4900	4732	168	4	69

4.3 EST datasets: pathogen-challenge samples of *P. vannemei*

Mapping EST sequences to reference assemblies

Similar to EST datasets of *P. monodon*, unique EST sequences from the selected libraries of *P. vannamei* EST data that can be compared to one another. The comparisons were performed between WSSV-infected libraries and those from non-infected libraries of the same tissue. Table 21 listed the selected libraries and the number of EST sequences. The reference assemblies were those derived from the EST dataset (Table 6) by CAP3 (97% identity and default parameter settings [95% identity]), MIRA and Trinity. Sequence IDs of the first three assemblers were converted to Trinity format for mapping with Bowtie2 (RSEM pipeline). As we can see there, the numbers of WSSV-infected EST sequences are relatively smaller than those of non-infected libraries of the same tissue types.

Table 21 The selected EST libraries and their number of sequences.

Library_code	Library_ID	Description	#unique sequences
	NCBI dbEST		
Gill.N	LIBEST_022685	LIBEST_022685 Litopenaeus vannamei gills cDNA library	24991
	LIBEST_010471	LIBEST_010471 LvG	
Gill.W	LIBEST_026674	LIBEST_026674 WSSV infected Litoepenaeus vannamei library	748
	LIBEST_021215	LIBEST_021215 Litopenaeus vannamei white spot syndrome virus infected gills	
	LIBEST_015173	LIBEST_015173 gCdWt	
	LIBEST_015178	LIBEST_015178 CdWtgill9h	
	LIBEST_015184	LIBEST_015184 gill27t32d	
	LIBEST_015187	LIBEST_015187 gill32t27d	
	LIBEST_016943	LIBEST_016943 LvG-gill27t32d	
HP.N	LIBEST_006799	LIBEST_006799 L99-29	22747
	LIBEST_022687	LIBEST_022687 Litopenaeus vannamei hepatopancreas cDNA library	
HP.W	LIBEST_015185	LIBEST_015185 HP32t27d	449
	LIBEST_015186	LIBEST_015186 HP27t32d	
	LIBEST_015172	LIBEST_015172 HPCdWt	
	LIBEST_015182	LIBEST_015182 CdWtHP9h	
He.N	LIBEST_022686	LIBEST_022686 Litopenaeus vannamei hemocyte cDNA library	29241
	LIBEST_015206	LIBEST_015206 PD80RG	
	LIBEST_016501	LIBEST_016501 LvB-LD8ORG	
	LIBEST_016502	LIBEST_016502 LvB-LDRG80	
	LIBEST_016503	LIBEST_016503 LvB-PD80RG	
	LIBEST_016508	LIBEST_016508 LvE-stalk	
	LIBEST_005322	LIBEST_005322 L99-22	
	LIBEST_012404	LIBEST_012404 LvB-LD80RG	
He.W	LIBEST_020212	LIBEST_020212 white spot syndrome virus infected hemocyte library	653
	LIBEST_015176	LIBEST_015176 hCdWt	
	LIBEST_015180	LIBEST_015180 CdWthem9h	
	LIBEST_015188	LIBEST_015188 hem27t32d	
	LIBEST_015189	LIBEST_015189 hem27d32t2	
	LIBEST_015190	LIBEST_015190 hem32d27t2	
	LIBEST_015204	LIBEST_015204 hem32t27d	
	LIBEST_016504	LIBEST_016504 LvB-hem27d32t2	
	LIBEST_016505	LIBEST_016505 LvB-hem27t32d	
	LIBEST_016506	LIBEST_016506 LvB-hem32d27t2	
	LIBEST_016507	LIBEST_016507 LvB-hem32t27d	
	LIBEST_016510	LIBEST_016510 LvP-HP27t32d	
	LIBEST_016511	LIBEST_016511 LvP-HP32t27d	

After mapping with Bowtie 2, the numbers of EST sequences mappable to each reference sequences are shown in Table 22. The result suggests that mappable EST sequences of the selected libraries to the CAP397 or CAP3DF reference assemblies have the highest proportion (>60%). Based on the result here and of *P. monodon*, we selected the result from the CAP397 reference assembly.

Table 22 The proportion of mappable EST sequences.

Library codes	Reference Assemblies			
	MIRA	CAP3DF	CAP397	Trinity
NCBI dbEST				
NCBI01-Gill.N	70.82	71.68	73.01	58.12
NCBI01-Gill.W	57.93	62.25	61.97	47.49
NCBI01-HP.N	68.28	69.92	68.36	54.85
NCBI01-HP.W	61.27	65.14	61.50	50.34
NCBI01-He.N	73.09	74.07	77.54	57.13
NCBI01-He.W	48.48	51.52	49.12	40.32

4.4 Cross-species global analysis for shared WSSV-responsive genes

I analyzed only transcript contigs that were generated from traditionally-generated EST sequences to search for a set of shared WSSV-responsive genes between *P. vannamei* and *P. monodon*. Such shared WSSV-responsive genes are those genes that were found in WSSV-infected EST libraries of both *P. vannamei* and *P. monodon*, but these genes were not found in the normal EST libraries produced from the same tissue types.

Table 23 shows three candidate WSSV-responsive genes are shared between WSSV-challenged libraries of both *P. vannamei* and *P. monodon*. These genes are cuticular proteins (SCP), Smad2/3, and Acyl-CoA dehydrogenase. Another copy of *P. monodon*'s cuticular protein, PmCBP, is found to interact with several WSSV envelop proteins and to be co-localized with VP53A, one of WSSV envelop proteins, on cell surface of shrimp hemocytes (Chen et al. 2009). Interestingly, our SCP candidate appears to be a different copy from PmCBP as suggested by multiple alignments. Smad2/3, a signaling effector of TGF-β signaling, may be involved in a cross-talk between virus-entry to host cell via endocytosis and TGF-β receptor internalization. Interpro domain prediction of these sequences revealed MH2 domain for Smad 2/3 sequences and cuticular protein domains for SCP (Fig 8).

Table 23 List of candidate genes WSSV-responsive genes shared between *P. monodon* and *P. vannamei*

Candidate genes	*P. monodon*		*P. vannamei*	
	Tissue of EST library prep.	Contig_ID (length; bp)	Tissue of EST library prep.	Contig_ID (length bp)
Cuticular proteins (SCP)	whole shrimp	6687-PAGECO100908-4119 (501)	gill	6689-PAGECO100908-10544 (842)
Smad2/3	whole shrimp;	6687-PAGECO100908-3526 (620)	gill	6689-PAGECO100908-08318 (400)
Acyl-CoA dehydrogenase	whole shrimp;	6687-PAGECO100908-1526 (907)	hemocyte	6689-PAGECO100908-09856 (1404)

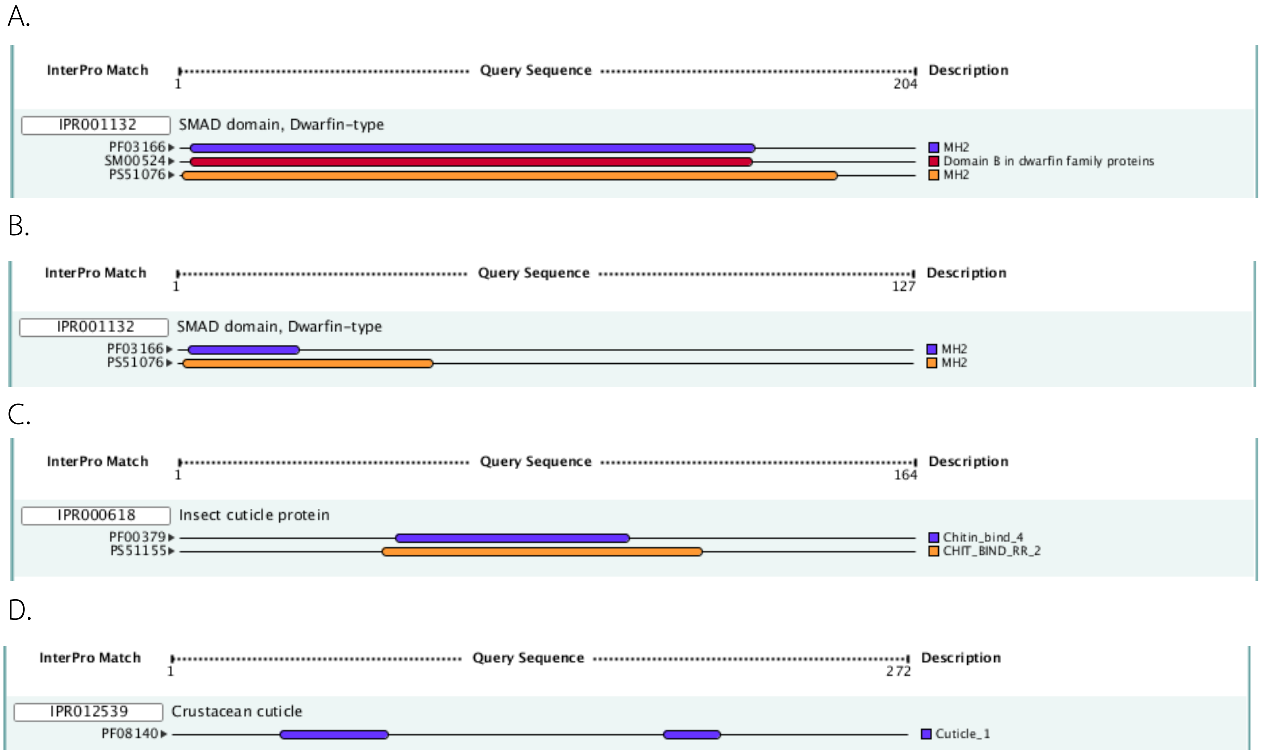


Figure 8 Domain prediction for putative Smad2/3 sequences of *P. monodon* (A) and *P. vannamei* (B) and for cuticular protein sequences of *P. monodon* (C) and *P. vannamei* (D).

Discussion and Conclusion

The molecular sequences data obtained in this project are the most comprehensive collection as of early 2015, especially, the traditionally-generated ESTs of the 14 decapod species. The traditionally-generated ESTs were not updated much since then as more data were generated by next-generation sequencers (NGS). We also collected NGS datasets for the black tiger shrimp *Penaeus monodon*, the Pacific white shrimp *P. vannamei* and *Macrobrachium rosenbergii*. Altogether, they contribute to the expanded data collection obtained in this project. Most of the transcript sequences were filtered for contaminating sequence and assembled into quality assemblies. Several sets of transcript assemblies were *in-silico* annotated with pipelines for both protein-coding genes and non-coding RNAs, together with other cDNAs and protein sequences. The pipelines for assembly and *in-silico* annotation for both protein-coding genes and ncRNAs can be readily used for further sequences. A part of the datasets has been deposited to the ShrimpGPAT database (http://shrimpgpat.sc.mahidol.ac.th/) that is publically available. The future release will be made available soon to include all datsets described in the project.

Candidate protein-coding genes and ncRNAs that are responsive to WSSV infection in both *P. monodon* and *P. vannamei* have been being investigated. Importantly, ncRNAs in shrimp have not been reported and/or experimentally characterized, so that an experimental verification for these genes will shed light on ncRNAs mechanisms to WSSV infection in shrimp.

Bibliography

Chen K, Hsu T, Huang P, Kang S, Lo C-F, Huang W, et al. 2009. Penaeus monodon chitin-binding protein (PmCBP) is involved in white spot syndrome virus (WSSV) infection. Fish and Shellfish Immunology:27(3):460-5.

Chevreux et al. 2004. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs Genome Research 2004. 14:1147-1159.

Colbourne JK, Pfrender ME, Gilbert D, et al: The ecoresponsive genome of Daphnia pulex. Science 2011, 331(6017):555–561.

Flavell JR, Baumforth KRN, Wood VHJ, et al. 2008. Down-regulation of the TGF-beta target gene, PTPRK, by the epstein-barr virus-encoded EBNA1 contributes to the growth and survival of hodgkin lymphoma cells. Blood 111(1):292-301.

Finn, R.D., J. Clements, S.R. Eddy 2011. HMMER web server: interactive sequence similarity searching Nucleic Acids Research Web Server Issue 39:W29-W37

Fu, L, Niu, B, Zhu, Z., Wu, S. and Li, W. 2012 CD-HIT: accelerated for clustering the next generation sequencing data. Bioinformatics, 28 (23): 3150-3152.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. May 15;29(7):644-52.

Gibson, A. K.; Smith, Z.; Fuqua, C. et al. 2013. Why so many unknown genes? Partitioning orphans from a representative transcriptome of the lone star tick Amblyomma americanum BMC. Genomics 14: 135

Gupta Y, Witte M, Möller S, Ludwig RJ, Restle T, Zillikens D, Ibrahim SM. 2014. ptRNApred: computational identification and classification of post-transcriptional RNA. Nucleic Acids Res. Dec 16;42(22):e167.

Jung, H., Lyons, R. E., Dinh, H. et al. 2011. Transcriptomics of a giant freshwater prawn (Macrobrachium rosenbergii): de novo assembly, annotation and marker discovery. PLoS ONE. 6: e27938

Kampa, D., Cheng, J., Kapranov, P., et al. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosome 21 and 22. Genome Research 14:331-342.

Kanehisa, M. et al. (2004). “The KEGG resource for deciphering the genome”. In: Nucleic Acids Res. 32, pp. D277–280.

Kapranov, P., Cawley, S.E., Drenkow, J., et al. 2002. Large–scale transcriptional activity in chromosomes 21 and 22. Science 296:916-919.

Kong,L, Y. Zhang, Z.Q. Ye, X.Q. Liu, S.Q. Zhao, L. Wei, and G. Gao. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 36: W345-349.

Kozomara A, and Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 42:D68-D73.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001 Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. Jan 19;305(3):567-80.

Lagesen K, Hallin PF, Rodland E, Staerfeldt HH, Rognes T Ussery DW. 2007. RNammer: consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res. Apr 22.

Laslett and Canback. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucl. Acids Res. 32 (1): 11-16. doi: 10.1093/nar/gkh152

Laslett and Canback. 2008. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics. 2008 Jan 15;24(2):172-5. Epub 2007 Nov 22.

Leekitcharoenphon, P., Taweemuang, U., Palittapongarnpim, P., Kotewong, R., Supasiri, T. & Sonthayanon, B. (2010). Predicted sub-populations in a marine shrimp proteome as revealed by combined EST and cDNA data from multiple Penaeus species. BMC Res Notes 3, 295.

Leu J, Chang C, Wu J, et al. 2007. Comparative analysis of differentially expressed genes in normal and white spot syndrome virus infected penaeus monodon. BMC Genomics:8.

Leu J, Chen S, Wang Y, et al. 2011. 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. Marine Biotechnology 13(4):608-21.

Li A, Zhang J1, Zhou Z. 2014. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics. 2014 Sep 19;15:311. doi: 10.1186/1471-2105-15-311.

Lowe and Eddy. 1997. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Nucl. Acids Res. 25 (5): 0955-964.

McKillen DJ, Chen YA, Chen C, et al. 2005. Marine genomics: A clearing-house for genomic and transcriptomic data of marine organisms. BMC Genomics:6.

Ma K, Qiu G, Feng J, Li J. 2012. Transcriptome analysis of the oriental river prawn, Macrobrachium nipponense using 454 pyrosequencing for discovery of genes and markers. PLoS One. 7(6):e39727.

MacIntosh, G.C., Wilkerson, C., Green, P.J. 2001. Identification and analysis of Arabidopsis expressed sequence tags characteristics of non-coding RNAs. Plant Physiology 127:765-776.

Magoc, T. and Salzberg, S. 2011. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27: 2957-63.

Nawrocki, EP. SW. Burge, et al. 2014. Rfam 12.0: updates to the RNA families database. Nucleic Acids Research. 10.1093/nar/gku1063

Nordahl, T. Petersen, Soren Brunak, Gunnar von Heijne & Henrik Nielsen. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions Nature Methods, 8:785-786.

O'Leary, N. A., Trent, H. F., Robalino, J. et al. 2006. Analysis of multiple tissue-specific cDNA libraries from the Pacific whiteleg shrimp, Litopenaeus vannamei Integr. Comp. Biol. 46: 931-939.

Phred Phrap and Consed. Website Phred, Phrap and Consed.

Pongsomboon S, Tang S, Boonda S, Aoki T, Hirono I, Tassanakajon A. 2011. A cDNA microarray approach for analyzing transcriptional changes in penaeus monodon after infection by pathogens. Fish and Shellfish Immunology 30(1):439-46.

Seemann, S.E., Gilchrist, M.J., Hofacker, I.L., Stadler, P.F., Gorodkin, J. 2007. Detection of RNA structures in porcine EST data and related mammals. BMC Genomics 8:316.

Sritunyalucksana, K, W. Wannapapho, C.F. Lo, T.W. Flegel. 2006. PmRab7 is a VP28-binding protein involved in white spot syndrome virus infection in shrimp. J Virol, 80 (21), pp. 10734–10742

Tassanakajon A, Klinbunga S, Paunglarp N, et al. 2006. Penaeus monodon gene discovery project: The generation of an EST collection and establishment of a database. Gene 384(1-2):104-12.

Tassanakajon et al. 2013. Discovery of immune molecules and their crucial functions in shrimp immunity Fish Shellfish Immunol. 34: 954-967

The Sequence Read Archive (SRA) The Sequence Read Archive (SRA).

The UniVec Database The UniVec Database.

Tupy, J.L., Bailey, A.M., Dailey, G., Evan-Holm, M., Siebel, C.W., Misra, S., Celniker, S.E., Rubin, G.M. 2005. Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster. PNAS 102: 5495-5500.

Wang, Y.G., Hassan, M.D., Shariff, M., Zamri, S.M., Chen, X. 1999. Histopathology and cytopathology of white spot syndrome virus (WSSV) in cultured Penaeus monodon from peninsular Malaysia with emphasis on pathogenesis and the mechanism of white spot formation. 39: 1-11.

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., & Li, W. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Research, 41(6), e74. doi:10.1093/nar/gkt006

Outputs

1. Datasets of transcriptomes of 14 decapod spcies.
   Currently, the total transcript contigs of both *P. monodon* and *P. vannamei* are composed of more than 2 million sequences from transcriptome of both NGS and traditionally-generated ESTs. A subset of these datasets in combination with ESTs, transcript contigs, cDNAs and protein sequences of the other 12 decopod species are available publicly in to the ShrimpGPAT database (http://shrimpgpat.sc.mahidol.ac.th/).
2. Pipelines for annotation of protein-coding genes and non-coding RNAs were implemented. These pipelines were applied to some datasets of transcriptomes and the annotation will be deposited in the ShrimpGPAT database for public accesses. Further data analyses can be performed with the information available with the annotation and transcript data.
3. Sets of protein-coding genes and ncRNAs responsive to WSSV infection were obtained and currently be tested experimentally. An experimental verification of these genes will shed light on host responses to WSSV, especially ncRNA mechanisms in shrimp.
4. The dataset currently in the ShrimpGPAT release #2 were published along with the description of the ShrimpGPAT database in ***BMC Genomics*** Korshkari P, Vaiwsri S, Flegel TW, Ngamsuriyaroj S, Sonthayanon B*, **Prachumwat A.*** ShrimpGPAT: a gene and protein annotation tool for knowledge sharing and gene discovery in shrimp. ***BMC Genomics***. 2014; 15:506. (IF 2012 = 4.397) * corresponding authors .
5. Presentation at meetings and conferences:

1. **Poster:** Anuphap Prachumwat,* Sirintra Vaiwsri, Parpakron Korshkari, Timothy W. Flegel, Sudsanguan Ngamsuriyaroj, and Burachai Sonthayanon. The Shrimp Gene and Protein Annotation Tool (ShrimpGPAT). The 39 th Congress on Science and Technology of Thailand “Innovative Sciences for a Better Life” October 21 - 23, 2013 at BITEC, Bangkok, Thailand * corresponding authors .

2. **Poster:** Anuphap Prachumwat,* Sirintra Vaiwsri, Parpakron Korshkari, Timothy W. Flegel, Sudsanguan Ngamsuriyaroj, Burachai Sonthayanon. “ShrimpGPAT: Shrimp gene and protein annotation tool & a prediction model for shrimp protein-protein interactions and gene ontology” The 3rd National Research University SUMMIT. 31 July -1 August 2014 Bangkok. * corresponding authors .

3. **Oral:** Anuphap Prachumwat,* Sirintra Vaiwsri, Parpakron Korshkari, Timothy W. Flegel, Sudsanguan Ngamsuriyaroj, and Burachai Sonthayanon. ShrimpGPAT: a gene and protein annotation tool for knowledge sharing and gene discovery in shrimp. Page 253. The 9th Symposium on Diseases in Asian Aquaculture (DAA9). November 24-28, 2014 at Ho Chi Minh City, Vietnam. * corresponding authors .

Appendix

A. Reprint

Korshkari P, Vaiwsri S, Flegel TW, Ngamsuriyaroj S, Sonthayanon B[*], **Prachumwat A.**[*] ShrimpGPAT: a gene and protein annotation tool for knowledge sharing and gene discovery in shrimp. ***BMC Genomics***. 2014; 15:506. (**IF 2012 = 4.397**).

Abstract

Background

Although captured and cultivated marine shrimp constitute highly important seafood in terms of both economic value and production quantity, biologists have little knowledge of the shrimp genome and this partly hinders their ability to improve shrimp aquaculture. To help improve this situation, the Shrimp Gene and Protein Annotation Tool (ShrimpGPAT) was conceived as a community-based annotation platform for the acquisition and updating of full-length complementary DNAs (cDNAs), Expressed Sequence Tags (ESTs), transcript contigs and protein sequences of penaeid shrimp and their decapod relatives and for *in-silico* functional annotation and sequence analysis.

Description

ShrimpGPAT currently holds quality-filtered, molecular sequences of 14 decapod species (~500,000 records for six penaeid shrimp and eight other decapods). The database predominantly comprises transcript sequences derived by both traditional EST Sanger sequencing and more recently by massive-parallel sequencing technologies. The analysis pipeline provides putative functions in terms of sequence homologs, gene ontologies and protein-protein interactions. Data retrieval can be conducted easily either by a keyword text search or by a sequence query via BLAST, and users can save records of interest for later investigation using tools such as multiple sequence alignment and BLAST searches against pre-defined databases. In addition, ShrimpGPAT provides space for community insights by allowing functional annotation with tags and comments on sequences. Community-contributed information will allow for continuous database enrichment, for improvement of functions and for other aspects of sequence analysis.

Conclusions

ShrimpGPAT is a new, free and easily accessed service for the shrimp research community that provides a comprehensive and up-to-date database of quality-filtered decapod gene and protein sequences together with putative functional prediction and sequence analysis tools. An important feature is its community-based functional annotation capability that allows the research community to contribute knowledge and insights about the properties of molecular sequences for better, shared, functional characterization of shrimp genes. Regularly updated and expanded with data on more decapods, ShrimpGPAT is publicly available at http://shrimpgpat.sc.mahidol.ac.th/.

Keywords

Penaeid shrimp, decapoda, EST, transcriptomes, knowledge base, community-based functional annotation

Korshkari *et al. BMC Genomics* 2014, **15**:506
http://www.biomedcentral.com/1471-2164/15/506

BMC Genomics

DATABASE Open Access

ShrimpGPAT: a gene and protein annotation tool for knowledge sharing and gene discovery in shrimp

Parpakron Korshkari[1,2†], Sirintra Vaiwsri[1,2†], Timothy W Flegel[1,3], Sudsanguan Ngamsuriyaroj[2], Burachai Sonthayanon[1,3*] and Anuphap Prachumwat[1,3,4*†]

Abstract

Background: Although captured and cultivated marine shrimp constitute highly important seafood in terms of both economic value and production quantity, biologists have little knowledge of the shrimp genome and this partly hinders their ability to improve shrimp aquaculture. To help improve this situation, the Shrimp Gene and Protein Annotation Tool (ShrimpGPAT) was conceived as a community-based annotation platform for the acquisition and updating of full-length complementary DNAs (cDNAs), Expressed Sequence Tags (ESTs), transcript contigs and protein sequences of penaeid shrimp and their decapod relatives and for *in-silico* functional annotation and sequence analysis.

Description: ShrimpGPAT currently holds quality-filtered, molecular sequences of 14 decapod species (~500,000 records for six penaeid shrimp and eight other decapods). The database predominantly comprises transcript sequences derived by both traditional EST Sanger sequencing and more recently by massive-parallel sequencing technologies. The analysis pipeline provides putative functions in terms of sequence homologs, gene ontologies and protein-protein interactions. Data retrieval can be conducted easily either by a keyword text search or by a sequence query via BLAST, and users can save records of interest for later investigation using tools such as multiple sequence alignment and BLAST searches against pre-defined databases. In addition, ShrimpGPAT provides space for community insights by allowing functional annotation with tags and comments on sequences. Community-contributed information will allow for continuous database enrichment, for improvement of functions and for other aspects of sequence analysis.

Conclusions: ShrimpGPAT is a new, free and easily accessed service for the shrimp research community that provides a comprehensive and up-to-date database of quality-filtered decapod gene and protein sequences together with putative functional prediction and sequence analysis tools. An important feature is its community-based functional annotation capability that allows the research community to contribute knowledge and insights about the properties of molecular sequences for better, shared, functional characterization of shrimp genes. Regularly updated and expanded with data on more decapods, ShrimpGPAT is publicly available at http://shrimpgpat.sc.mahidol.ac.th/.

Keywords: Penaeid shrimp, Decapoda, EST, Transcriptomes, Knowledge base, Community-based functional annotation

* Correspondence: burachais@gmail.com; anuphap.pra@biotec.or.th
†Equal contributors
[1]Center of Excellence for Shrimp Molecular Biology and Biotechnology (CENTEX Shrimp), Faculty of Science, Mahidol University, Rama VI Road, Bangkok 10400, Thailand
[3]National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand
Full list of author information is available at the end of the article

Background

Marine shrimp in the Family *Penaeidae* have gained status as a very important international seafood trade product of particular economic importance in shrimp farming countries. Despite their economic importance as farmed animals, relatively little is known about the reproduction, immunity and physiology of shrimp when compared to other farmed animals such as poultry and swine. For example, shrimp aquaculture production has been negatively affected by several major pathogens (e.g., white spot syndrome virus and yellow head virus; for reviews, see [1,2]), and efforts to control these pathogens are impeded by relatively poor knowledge of the shrimp response to them (i.e., shrimp immunity). Although genomic sequences of an organism can yield information about its defense mechanisms, there is currently no completely-sequenced genome for any penaeid shrimp species and only limited characterization of shrimp immune response genes. Similar comments apply to other fields of shrimp biology including reproduction and growth. Shrimp EST collections including recent transcriptomic reads generated by next-generation sequencing (NGS) technologies have helped in shrimp gene and genetic marker discovery (e.g., [3-6]). As such sequencing data are rapidly increasing, and the Shrimp Gene and Protein Annotation Tool (ShrimpGPAT) serves as a platform to extensively collect shrimp molecular sequences for functional annotation and to provide a channel for the shrimp research community to curate and annotate sequences in the form of tags and comments.

Since the first analysis of shrimp ESTs in 1999 [7], several large scale EST studies from various tissues and under various conditions have been carried out for a number of penaeid shrimp species, including the black tiger shrimp *Penaeus (Penaeus) monodon* and the Pacific white shrimp *P. (Litopenaeus) vannamei* (for a review see [8]). Since then, three specialized databases housing shrimp EST sequences have been developed. These are the Marine Genomics Database established in 2005 [9], the *Penaeus monodon* EST Project database established in 2006 [3] and the *Penaeus* Genome database established in 2009 [8]. The Marine Genomics Database includes ESTs and contigs (or “unigenes” as called by the Marine Genomics Database) for four penaeid shrimp species (177,691 EST and 14,726 contig sequences) and also for 23 other marine organisms, such as dinoflagellates, corals, bivalves, crustaceans, sharks, rays, fish, birds, whales and dolphins (314,766 ESTs and 46,421 contigs in total). The Marine Genomics Database plans to include microarray data in a future release. The *Penaeus monodon* EST Project database contains ESTs and contigs (40,001 ESTs and 10,536 contigs) from multiple libraries and tissues of *P. monodon* generated by several laboratories of the Thai shrimp research community. A recent collaboration of shrimp researchers in Thailand and Taiwan resulted in an expansion of *P. monodon* data deposited in the *Penaeus monodon* EST Project database (54,058 ESTs and 12,181 contigs). The *Penaeus* Genome database provides ESTs and contigs for four penaeid shrimp species (196,248 ESTs and 42,332 contigs) and also recently included a genetic linkage map and fosmid library end sequences of *P. monodon*.

Tools available at these three databases include options to search for sequences by BLAST and by homolog descriptions or Gene Ontology terms. All three databases allow users to download sequences of interest. In addition, the Marine Genomics Database currently features both an ability to bookmark sequences for registered users and an EST quality control and submission pipeline for data contributors. The Marine Genomics Database also plans to include a microarray data upload pipeline as well as an automatic incorporation of new ESTs from the Genbank dbEST database in a future version. As EST and contig sequences in these three databases were last updated in 2008–2009, more recently available sequences are not included.

The aim of ShrimpGPAT was to combine multi-source data and include not only EST sequences but also NGS short reads, full-length complementary DNAs (cDNAs) and protein sequences within its data analysis pipeline for sequence quality filtering, contig construction, *in-silico* functional prediction (homolog identification and Gene Ontology prediction) and putative protein-protein interactions. ShrimpGPAT’s tagging and commenting features were designed to allow shrimp research scientists to annotate and provide insights on sequences. ShrimpGPAT initially held a set of ESTs for six decapod species, including four penaeid shrimp. Leekitcharoenphon et al. [10] analyzed and grouped these ESTs into four groups based on homologs found in the genomes of *Drosophila melanogaster* and *Caenorhabditis elegans*, and concluded that this group categorization facilitated functional annotation of shrimp proteomes and their protein sub-populations. Here, we call these categorized groups “reference groups”. Currently, ShrimpGPAT holds full-length cDNA sequences, individual EST sequences, transcript contigs and protein sequences for 14 decapod species (>500,000 combined records) together with putative functional annotations.

Construction and content

System design and implementation

ShrimpGPAT was developed as a web-based software environment under Microsoft Windows Server 2008 R2 Enterprise using a relational database of Microsoft SQL Server 2008 SP1 Enterprise for all data storage. Figure 1 shows the ShrimpGPAT relational schema via the entity-relationship diagram, describing the entities and the relationships among all tables as well as the essential keys of all entities of the relations and connections. Tables can be placed roughly into four groups: 1) sequence

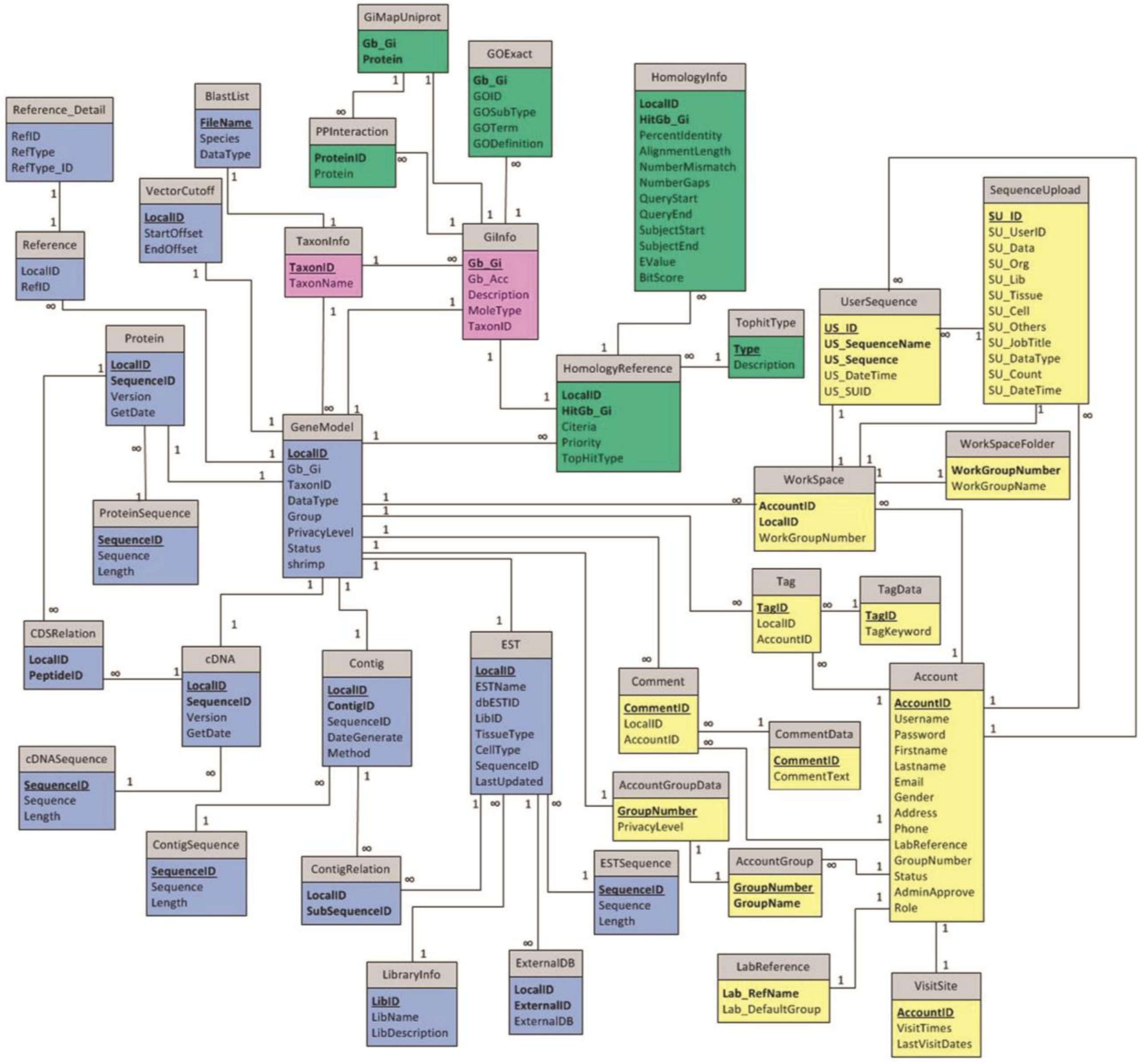


Figure 1 ShrimpGPAT database schema. This entity–relationship diagram shows relations among tables of four groups: sequence record tables (blue), *in-silico* annotation tables (green), users' data tables (yellow) and shared information tables (pink). Primary keys are underlined. Boldface indicates non-null field columns. Connections between tables are represented by lines, and relations between entities are indicated above the connection lines.

record tables, 2) *in-silico* annotation tables, 3) users' data tables and 4) shared information tables (for a detailed description of all tables, see the ShrimpGPAT online documentation). ShrimpGPAT contains a frontend user interface and a backend data analysis pipeline. The user interface was written with the VB.net and ASP.net on HTTP web services with AJAX.net, JQuery and Flash for visualization. The Cytoscape plug-in was used for protein network visualization [11]. Bioinformatic applications currently available to users were integrated with BLAST [12], MUSCLE [13] and MAFFT [14]. The backend data analysis pipeline employed in-house PERL scripts with NCBI E-Utilities [15], NCBI SRA Toolkit [16], phred [17], phd2fasta [18], cross_match [18], BLAST [12], CAP3 [19], Trimmomatic [20] and 454 Sequencing System Software (Newbler and sfffile version 2.8; 454 Life Sciences, Branford, CT) (see below). The processed data (associated information and sequences) were uploaded to the database with ShrimpGPAT data upload tools. The ShrimpGPAT system also supports user authentication and use cases to access the Microsoft SQL database, WorkSpace and community-based functional annotation features.

Pipeline for in-silico functional annotation

ShrimpGPAT currently focuses on four types of molecular sequences: full-length or partial cDNA, protein, and transcriptomic sequences by both traditional EST cloning and next-generation sequencing technologies. The pipeline for functional annotation comprised four main steps: 1) data acquisition 2) sequence/data cleansing, 3) contig assembly and 4) BLAST plus putative functional annotation. All four steps were applied to EST and NGS short read sequences, but cDNA and protein sequences were not subjected for sequence/data cleansing and contig assembly.

1. Data acquisition

Sequences from GenBank were downloaded by in-house PERL scripts and those from the Marine Genomics database [9] and the *Penaeus monodon* EST Project database [3] were downloaded via their respective websites and by personal communication. The locally-generated EST sequence trace files were processed by phred and phd2fasta into FASTA and .QUAL files. NGS short reads downloaded from the Sequence Read Archive (SRA) were processed by SRA Toolkit. Associated information was formatted for submission to the database by the ShrimpGPAT data upload tools.

2. Sequence/data cleansing

EST sequences were masked by cross_match for vector and contaminating sequences against both full-length vector sequences, if available, and the Univec database [21]. Masked sequences were processed by an in-house PERL script to produce vector-free sequences. Adapter sequences in NGS short reads were trimmed by sfffile or Trimmomatic.

3. Contig assembly

Trimmed sequences were assembled by either CAP3 or Newbler with default parameter settings.

4. BLAST plus putative functional annotation

All nucleotide sequences (EST, transcript contigs and cDNA sequences) were queried (BLASTN and BLASTX) against the nt and nr databases, respectively. BLASTP was performed for protein sequences against the nr database. Homologous sequences were defined as the hits with the following criteria: 1) ≥50% of the query sequence within the aligned region by BLAST, 2) an *E*-value $< 10^{-6}$ (for BLASTN) or $< 10^{-4}$ (for BLASTX and BLASTP), and 3) identity of ≥70% (BLASTN) or of ≥25% (BLASTX and BLASTP).

Reference sequences and reference groups: among these homologous sequences of each shrimp sequence query, the overall best homologs (best hits) and the best hits in the *Drosophila melanogaster* or *Caenorhabditis elegans* genomes were selected for each type of BLAST search (BLASTN, BLASTX and BLASTP). Reference sequences were the best hits from BLASTX in *D. melanogaster* if available. If no BLASTX hits in *D. melanogaster* were found, BLASTX hits in *C. elegans* were chosen. If no BLASTX hits were found in either species, overall BLASTX hits were selected. If no BLASTX homologs were found, reference sequences were chosen from BLASTN best hits in a similar manner. For protein sequences, criteria for reference sequences were similar to those for the BLASTX best hits of nucleotide query sequences. Reference groups were assigned by criteria similar to that described in [10].

Gene Ontology (GO) and protein-protein interactions (PPIs): GO classification of each shrimp sequence was derived from its reference proteins described above by mapping with information from the Protein Information Resource [22]. Similarly, putative PPIs were derived through corresponding protein sequences using PPIs from the *Drosophila* Interactions Database [23] and the IntAct molecular interaction database [24].

Species datasets

Six of the 14 decapod species currently in ShrimpGPAT are penaeid shrimp. The numbers of records along with their scientific and common names are shown in Table 1 (for Record statistics see below). The database will be updated periodically for new sequences and expanded to cover more species.

Utility and discussion

Data acquisition and sequence analysis pipeline

A curator can obtain a new dataset and formatted records for submission to the *in-silico* functional annotation pipeline. Resulting trimmed ESTs, contig sequences and related putative functions can then be uploaded to the ShrimpGPAT database via ShrimpGPAT data upload tools. Currently, this process is only accessible to designated curators via the administrator mode. Curators must also use this administrator mode to modify an existing record. Registered users can upload and store a limited number of sequences to the ShrimpGPAT database for their private use or to share with the community (see *WorkSpace and community-based annotation*).

Table 1 The number of molecular sequence records in ShrimpGPAT

Species		# of records			
Scientific name	Common name	EST	Transcript contigs[a]	cDNA	Protein
Penaeus (Penaeus) monodon	Black tiger shrimp	86,327	18,410	1,976	602
Penaeus (Litopenaeus) vannamei	Pacific whiteleg shrimp	176,592	47,058	74,828	574
Penaeus (Litopenaeus) setiferus	White shrimp	1,042	126	135	27
Penaeus (Fenneropenaeus) chinensis	Fleshy prawn	10,446	2,714	478	257
Penaeus (Fenneropenaeus) indicus	Indian prawn	714	155	348	127
Penaeus (Marsupenaeus) japonicus	Kuruma prawn	3,156	662	989	743
Macrobrachium rosenbergii	Giant freshwater prawn	4,427	8,550[b]	635	389
Cherax quadricarinatus	Cray fish	120	90	239	226
Pacifastacus leniusculus	Signal crayfish	802	199	914	88
Homarus americanus	American lobster	29,957	12,709	186	227
Scylla olivacea	Orange mud crab	203	80	121	0
Scylla paramamosain	Green mud crab	3,972	56	720	698
Callinectes sapidus	Blue crab	10,563	2,104	173	161
Carcinus maenas	Green crab	15,559	7,672	273	275

[a]The number of transcript contigs in each species is the summation of all contig sequences constructed by a set of ESTs and by a set of SRA reads with CAP3 (with default or 97%-similarity parameters) and Newbler (with default parameters).
[b]Including SRA transcript contigs produced by Newbler.

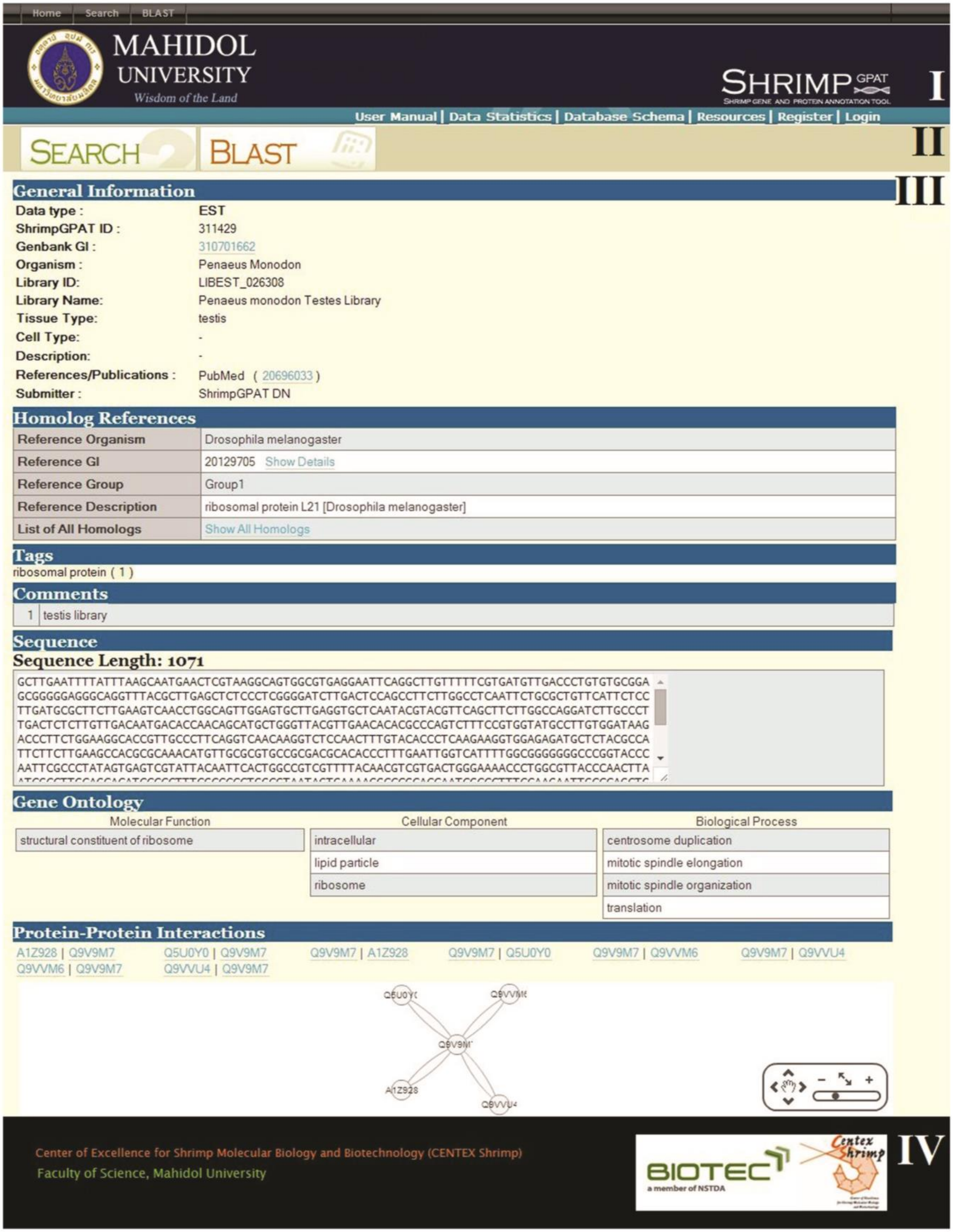
Record retrieval and sequence analysis tools

The ShrimpGPAT user interface page contains four areas: title, menu bar, content and footer, arranged from top to bottom as in Figure 2. Title, menu bar and footer areas are relatively static, but the content area displays dynamically-generated information. ShrimpGPAT can be accessed through three main sections listed in the menu bar area, namely Search, BLAST and WorkSpace. The first two can be accessed by any user, but WorkSpace can only be accessed by a registered user (see below). Records can be retrieved either by a keyword text search (Search button) or by a sequence query (BLAST button). Two types of keyword text search are currently permitted: free text search and advanced search for specified fields. The BLAST search function is set with default parameters but with options for several *E*-value cutoffs. Records returned by both Search and BLAST are displayed in the same format for easy viewing and investigation. Users can select records for further analysis through searching with BLAST, creating Multiple Sequence Alignments (MSA), exporting sequences in a FASTA file, bookmarking to their private WorkSpace or adding of tags or comments. ShrimpGPAT currently provides two sets of sequence analysis tools in sections where such analyses are applicable: BLAST and MSA. BLAST is parameterized to a default setting, except for *E*-value cutoffs, and MSA provides MAFFT and MUSCLE analyses with default parameter settings.

Records in a result list from any executed queries can be investigated further by clicking on a ShrimpGPAT ID, which will display full information regarding that particular record, e.g., sequence type, organism, tissue, organ of expression, references/publications as well as external database IDs (Figure 2). External database IDs are hyperlinked to corresponding external database records. Homolog information (reference sequences and reference groups) is displayed below the general information. Note that only one reference sequence is displayed on this page, but clicking on the hyperlinks “Show Details” or “Show All Homologs” reveals all reference sequences or homologous sequences with a complete BLAST result. Tags, comments, sequence characters of a record, GO and putative PPIs are consecutively displayed below the homolog information section.

WorkSpace and community-based annotation

WorkSpace and community-based annotation features are reserved for registered users. ShrimpGPAT WorkSpace provides private space for records of interest. Within WorkSpace, a user can create virtual folders to store records and can later delete or rename the folders. Records can be moved between or copied into virtual folders. Records stored in WorkSpace can be used later for additional sequence analyses or for sequence downloading. Importantly, users can help annotate records with tags and comments (ShrimpGPAT community-based annotation). Tags are short keywords, but comments can be long strings of text. These tags and comments are publicly displayed for text search to any users, so they enable knowledge sharing among the shrimp research community. For example, users can input gene names as tags and information of references/publications as comments. However, some well-known shrimp gene names known by


Home
Search
BLAST
MAHIDOL
UNIVERSITY
Wisdom of the Land
SHRIMP GPAT
SHRIMP GENE AND PROTEIN ANNOTATION TOOL
I
User Manual | Data Statistics | Database Schema | Resources | Register | Login
SEARCH
BLAST
II
III
IV

General Information

Data type :	EST
ShrimpGPAT ID :	311429
Genbank GI :	310701662
Organism :	Penaeus Monodon
Library ID:	LIBEST_026308
Library Name:	Penaeus monodon Testes Library
Tissue Type:	testis
Cell Type:	-
Description:	-
References/Publications :	PubMed (20696033)
Submitter :	ShrimpGPAT DN

Homolog References

Reference Organism	Drosophila melanogaster
Reference GI	20129705 Show Details
Reference Group	Group1
Reference Description	ribosomal protein L21 [Drosophila melanogaster]
List of All Homologs	Show All Homologs

Tags

ribosomal protein (1)

Comments

1	testis library

Sequence

Sequence Length: 1071

```
GCTTGAATTTTATTTAAGCAATGAACTCGTAAGGCAGTGGCGTGAGGAATTCAGGCTTGTTTTTCGTGATGTTGACCCTGTGTGCGGA
GCGGGGGAGGGCAGGTTTACGCTTGAGCTCTCCCTCGGGGATCTTGACTCCAGCCTTCTTGGCCTCAATTCTGCGCTGTTCATTCTCC
TTGATGCGCTTCTTGAAGTCAACCTGGCAGTTGGAGTGCTTGAGGTGCTCAATACGTACGTTCAGCTTCTTGGCCAGGATCTTGCCCT
TGACTCTCTTGTTGACAATGACACCAACAGCATGCTGGGTTACGTTGAACACACGCCCAGTCTTTCCGTGGTATGCCTTGTGGATAAG
ACCCTTCTGGAAGGCACCGTTGCCCTTCAGGTCAACAAGGTCTCCAACTTTGTACACCCTCAAGAAGGTGGAGAGATGCTCTACGCCA
TTCTTCTTGAAGCCACGCGCAAACATGTTGCGCGTGCCGCGACGCACACCCTTTGAATTGGTCATTTTGGCGGGGGGGCCCGGTACCC
AATTCGCCCTATAGTGAGTCGTATTACAATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTA
```

Gene Ontology

Molecular Function	Cellular Component	Biological Process
structural constituent of ribosome	intracellular	centrosome duplication
	lipid particle	mitotic spindle elongation
	ribosome	mitotic spindle organization
		translation

Protein-Protein Interactions

A1Z928 | Q9V9M7	Q5U0Y0 | Q9V9M7	Q9V9M7 | A1Z928	Q9V9M7 | Q5U0Y0	Q9V9M7 | Q9VVM6	Q9V9M7 | Q9VVU4
Q9VVM6 | Q9V9M7	Q9VVU4 | Q9V9M7				

Center of Excellence for Shrimp Molecular Biology and Biotechnology (CENTEX Shrimp)
Faculty of Science, Mahidol University

Figure 2 A screenshot of ShrimpGPAT record display page. Its layout is divided into I) the title, II) the menu bar, III) the content and IV) the footer. See text for description.

abbreviations such as PmRab7, may not be present as such in description lines of GenBank full-length cDNA or protein records but instead be written in full, i.e., "Penaeus monodon Rab7". Thus, a search using "PmRab7" might fail, while a search using "Penaeus monodon Rab7" or just "Rab7" would succeed. Thus, users can easily retrieve records with gene names if such records are tagged with corresponding gene names, but if no records are retrieved, name variations can be tried. Usage of tags and comments may be added to expand tags for a particular sequence or add them to sequences that are currently uncharacterized in the database but may later be studied and given gene names. Users can also share their dataset with the community via the ShrimpGPAT data upload tool to deposit the data as permanent records. Similarly, users can upload sequences for their private use, but such private sequences will be stored in user's virtual folders for a period of only three months.

Record statistics

Table 1 shows the number of molecular sequence records for the 14 decapods currently available in the ShrimpGPAT database. *P. vannamei* has the highest number of records (~299,000), and *P. monodon* has the second highest (~138,000). The numbers signify their importance as species of the highest interest to the shrimp scientific research community and species most-cultivated or captured for trade. Similarly, the six penaeid shrimp have combined records that number about four times that of the other eight decapod species in the database (i.e., ~460,000 *vs.* 111,000). A large proportion of the records for each species are ESTs and transcript contigs, whereas the numbers of cDNA and protein records are still relatively small. The number of transcript contigs for each species is the summation of all contig sequences constructed by the set of ESTs and by the set of SRA reads. Note that transcript contig records produced by different contig assemblers (e.g., CAP3 and Newbler) may constitute the same sequences. Regarding transcript contigs of SRA reads, *Macrobrachium rosenbergii* is the only species that currently has transcript contigs derived from an SRA dataset (81,411 reads for 50 million base pairs; [6]). Soon, SRA transcript contigs for other species will be available, e.g., *P. vannamei* with eight NGS runs in the SRA database, constituting 80 million reads for 7.9 billion base pairs. Among the 14 species, *Scylla olivacea* has the lowest number of records in its EST collection. It is the first publicly-available collection of ESTs for this species and it was recently generated by our laboratory. The current release of the database contains full-length cDNA and protein sequences downloaded from GenBank as of July 2013. Thus, sequences of some known shrimp genes might not currently be in the ShrimpGPAT database because 1) they were not present in GenBank at the time of the most recent download, 2) they were reported only in papers without a submission to GenBank, or 3) they were deposited elsewhere. Such sequences can be manually added by designated curators or gradually submitted and reported by users. Complete descriptive statistics and sources of ShrimpGPAT records are available on the ShrimpGPAT statistics page.

New and improved features for the shrimp community

ShrimpGPAT provides new and improved features that are lacking in the three existing specialized genomic databases for shrimp. First, ShrimpGPAT provides sequences of full-length cDNAs, proteins and transcript contigs from the rapidly growing number of NGS reads, in addition to traditional EST sequences that are provided by the existing databases. Its *in-silico* functional annotation pipeline can readily facilitate new data. Currently, ShrimpGPAT holds the highest number of molecular sequence records and species of penaeid shrimp (6 *vs.* 4 species in the Marine Genomics Database) and their decapod relatives (8 *vs.* 4 species in the Marine Genomics Database). Second, in terms of *in-silico* functional annotation features, putative sets of protein-protein interactions and reference sequences (reference groups) can only be found in ShrimpGPAT. Reference sequences are homologs in the genomes of *D. melanogaster* and *C. elegans* (decapods' closest relatives whose genomes are better characterized). Most existing databases provide only best-hit homologous sequences (which may or may not be those in the genomes of *D. melanogaster* and *C. elegans*), while ShrimpGPAT provides all homologous sequences that meet our criteria (see above). Similar to the other databases, GO classification is provided. Third, the unique set of tools available in ShrimpGPAT includes multiple sequence alignment, WorkSpace and community-based annotation. WorkSpace allows users to keep records of interest and their uploaded sequences. Users can upload sequences to share with others or use privately. Users of ShrimpGPAT can also utilize a set of tools similar to those found in the three existing databases (i.e., text search, BLAST and sequence download). With a large and expanding data set and its new features, ShrimpGPAT provides a more comprehensive database with more easily accessible tools than those of the three existing databases mentioned above. To the best of our knowledge ShrimpG-PAT is only shrimp database that offers community-based annotation with tags and comments.

Conclusions

ShrimpGPAT is a new online resource to help shrimp researchers investigate molecular sequences of penaeid shrimp and their decapod relatives. ShrimpGPAT provides shrimp biologists with easy access to a comprehensive collection of rapidly growing sequence information. The database will be periodically updated and expanded to cover more crustacean species with its *in-silico* functional annotation pipeline. It is envisioned that collaborative knowledge built via community-based annotation will rapidly accelerate shrimp gene discovery and research.

Availability and requirements

ShrimpGPAT is publicly available via the Website URL http://shrimpgpat.sc.mahidol.ac.th/. Registration requires a valid email address. The initial dataset based on Leekitcharoenphon et al. [10] can be accessed at http://shrimpgpat.sc.mahidol.ac.th/v1/.

Abbreviations

AJAX: Asynchronous JavaScript and XML; BLAST: Basic local alignment search tool; cDNA: Complementary DNA; EST: Expressed sequence tag; GO: Gene Ontology; MSA: Multiple sequence alignments; NGS: Next-generation sequencing technology; PPI: Protein-protein interaction.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PK and SV led the development of the system environment including design and implementation of the database schema, the use cases and the user interface, and they co-developed the ShrimpGPAT data upload tools. SV implemented the keyword text search and PK carried out data acquisition for a subset of ESTs. TWF advised on biological aspects, proposed the conceptual features of the database and assisted in writing the manuscript. SN planed the project and advised on the design and implementation of the database schema, the use cases and the user interface. BS initiated and planned the project, advised on biological aspects and database features and provided the initial dataset. AP oversaw the project plan and development, obtained all data and sequences, performed the functional annotation pipeline, designed use cases and the user interface and bore the main load of writing the manuscript. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by a National Research Universities Initiative grant from the Higher Education Research Promotion and National University Development, Office of the Thailand Higher Education Commission, by Mahidol University, by the Thailand Research Fund (TRF) and by the National Center for Genetic Engineering and Biotechnology (BIOTEC) of the Thai National Science and Technology Development Agency (NSTDA). AP also acknowledges the support from TRF/BIOTEC Grant No. TRG5680001/P-13-00608. We thank P. Leekitcharoenphon for her help with the initial dataset, A. Tassanakajon for her EST collection of the black tiger shrimp and S. Lerthivaporn, Aung Thu Rha Hein, M. Samseng and P. Leerungnavarat for their help with data retrieval and database configuration. We thank the two anonymous BMC Genomics reviewers and V. Charoensawan for their critical reading and useful comments to improve ShrimpGPAT features and the manuscript. Access to the high-performance computing facilities in the Biostatistics & Informatics Laboratory at the Genome Institute, BIOTEC is greatly appreciated.

Author details

[1]Center of Excellence for Shrimp Molecular Biology and Biotechnology (CENTEX Shrimp), Faculty of Science, Mahidol University, Rama VI Road, Bangkok 10400, Thailand. [2]Faculty of Information and Communication Technology, Mahidol University, Salaya Campus, Phutthamonthon District, Nakhon Pathom 73170, Thailand. [3]National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand. [4]Shrimp-Virus Interaction Laboratory, Agricultural Biotechnology Research Unit, National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand.

Received: 22 January 2014 Accepted: 17 June 2014
Published: 21 June 2014

References

1. Flegel TW: **Historic emergence, impact and current status of shrimp pathogens in Asia.** *J Invertebr Pathol* 2012, **110**:166–173.
2. Stentiford GD, Neil DM, Peeler EJ, Shields JD, Small HJ, Flegel TW, Vlak JM, Jones B, Morado F, Moss S, Lotz J, Bartholomay L, Behringer DC, Hauton C, Lightner DV: **Disease will limit future food supply from the global crustacean fishery and aquaculture sectors.** *J Invertebr Pathol* 2012, **110**:141–157.
3. 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: **Penaeus monodon gene discovery project: the generation of an EST collection and establishment of a database.** *Gene* 2006, **384**:104–112.
4. Robalino J, Almeida JS, McKillen D, Colglazier J, Trent HF, Chen YA, Peck ME, Browdy CL, Chapman RW, Warr GW, Gross PS: **Insights into the immune transcriptome of the shrimp Litopenaeus vannamei: tissue-specific expression profiles and transcriptomic responses to immune challenge.** *Physiol Genomics* 2007, **29**:44–56.
5. Leu JH, Chang CC, Wu JL, Hsu CW, Hirono I, Aoki T, Juan HF, Lo CF, Kou GH, Huang HC: **Comparative analysis of differentially expressed genes in normal and white spot syndrome virus infected Penaeus monodon.** *BMC Genomics* 2007, **8**:120.
6. Jung H, Lyons RE, Dinh H, Hurwood DA, McWilliam S, Mather PB: **Transcriptomics of a giant freshwater prawn (Macrobrachium rosenbergii): de novo assembly, annotation and marker discovery.** *PLoS One* 2011, **6**:e27938.
7. Lehnert SA, Wilson KJ, Byrne K, Moore SS: **Tissue-specific expressed sequence tags from the black tiger shrimp penaeus monodon.** *Mar Biotechnol (NY)* 1999, **1**:465–0476.
8. Leu JH, Chen SH, Wang YB, Chen YC, Su SY, Lin CY, Ho JM, Lo CF: **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.** *Mar Biotechnol (NY)* 2011, **13**(4):608–621.
9. McKillen DJ, Chen YA, Chen C, Jenny MJ, Trent HF, Robalino J, McLean DC, Gross PS, Chapman RW, Warr GW, Almeida JS: **Marine genomics: a clearing-house for genomic and transcriptomic data of marine organisms.** *BMC Genomics* 2005, **6**:34.
10. Leekitcharoenphon P, Taweemuang U, Palittapongarnpim P, Kotewong R, Supasiri T, Sonthayanon B: **Predicted sub-populations in a marine shrimp proteome as revealed by combined EST and cDNA data from multiple Penaeus species.** *BMC Res Notes* 2010, **3**:295.
11. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T: **A travel guide to cytoscape plugins.** *Nat Methods* 2012, **9**:1069–1076.
12. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC Bioinformatics* 2009, **10**:421.
13. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792–1797.
14. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol* 2013, **30**:772–780.
15. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2011, **39**:D52–D57.
16. **The Sequence Read Archive (SRA).** http://www.ncbi.nlm.nih.gov/Traces/sra/.
17. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II Error probabilities.** *Genome Res* 1998, **8**:186–194.
18. **Phred, Phrap and Consed.** http://www.phrap.org/.
19. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**:868–877.
20. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B: **RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics.** *Nucleic Acids Res* 2012, **40**:W622–W627.
21. **The UniVec Database.** http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/.
22. Huang H, McGarvey PB, Suzek BE, Mazumder R, Zhang J, Chen Y, Wu CH: **A comprehensive protein-centric ID mapping service for molecular data integration.** *Bioinformatics* 2011, **27**:1190–1191.

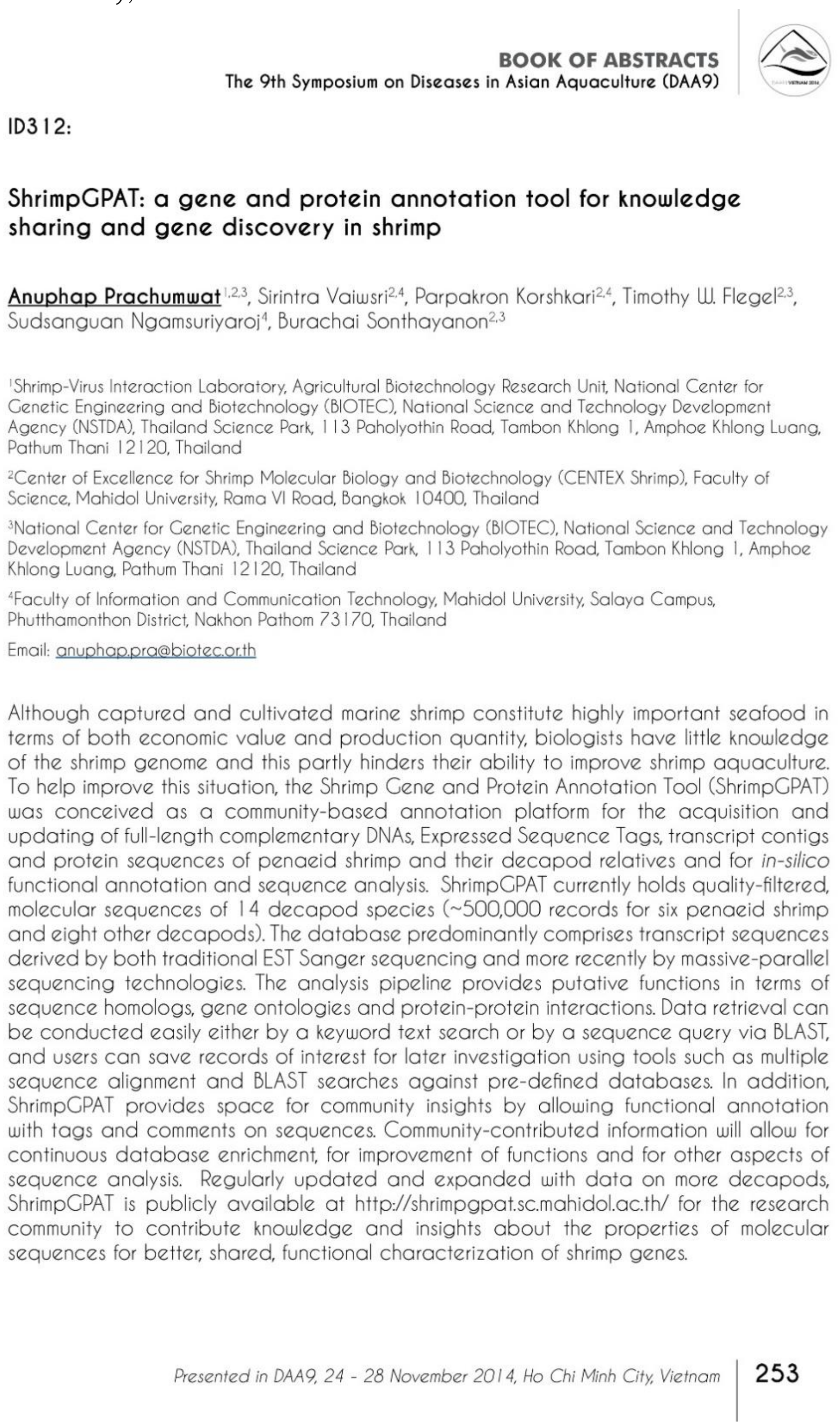
23. Yu J, Pacifico S, Liu G, Finley RL: **DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions.** *BMC Genomics* 2008, **9**:461.
24. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H: **The IntAct molecular interaction database in 2012.** *Nucleic Acids Res* 2012, **40:**D841–D846.

doi:10.1186/1471-2164-15-506
Cite this article as: Korshkari *et al.*: **ShrimpGPAT: a gene and protein annotation tool for knowledge sharing and gene discovery in shrimp.** *BMC Genomics* 2014 **15**:506.

B. Presentation

Oral presentation at International Conference

1. Anuphap Prachumwat,* Sirintra Vaiwsri, Parpakron Korshkari, Timothy W. Flegel, Sudsanguan Ngamsuriyaroj, and Burachai Sonthayanon. **ShrimpGPAT: a gene and protein annotation tool for knowledge sharing and gene discovery in shrimp.** The 9th Symposium on Diseases in Asian Aquaculture (DAA9) November 24-28, 2014 at Ho Chi Minh City, Vietnam.



BOOK OF ABSTRACTS
The 9th Symposium on Diseases in Asian Aquaculture (DAA9)

ID312:

ShrimpGPAT: a gene and protein annotation tool for knowledge sharing and gene discovery in shrimp

Anuphap Prachumwat[1,2,3], Sirintra Vaiwsri[2,4], Parpakron Korshkari[2,4], Timothy W. Flegel[2,3], Sudsanguan Ngamsuriyaroj[4], Burachai Sonthayanon[2,3]

[1]Shrimp-Virus Interaction Laboratory, Agricultural Biotechnology Research Unit, National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand

[2]Center of Excellence for Shrimp Molecular Biology and Biotechnology (CENTEX Shrimp), Faculty of Science, Mahidol University, Rama VI Road, Bangkok 10400, Thailand

[3]National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand

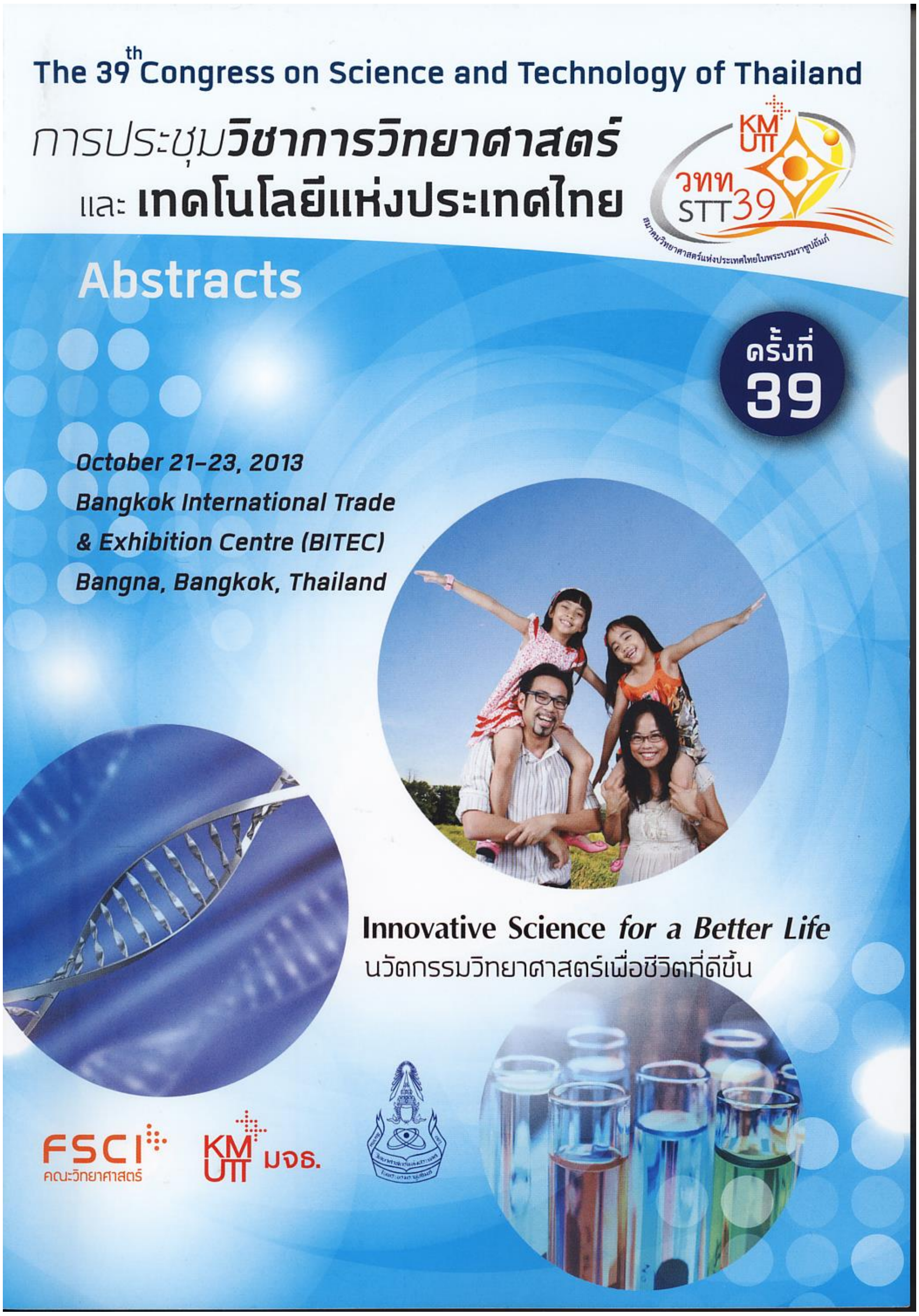
[4]Faculty of Information and Communication Technology, Mahidol University, Salaya Campus, Phutthamonthon District, Nakhon Pathom 73170, Thailand

Email: anuphap.pra@biotec.or.th

Although captured and cultivated marine shrimp constitute highly important seafood in terms of both economic value and production quantity, biologists have little knowledge of the shrimp genome and this partly hinders their ability to improve shrimp aquaculture. To help improve this situation, the Shrimp Gene and Protein Annotation Tool (ShrimpGPAT) was conceived as a community-based annotation platform for the acquisition and updating of full-length complementary DNAs, Expressed Sequence Tags, transcript contigs and protein sequences of penaeid shrimp and their decapod relatives and for *in-silico* functional annotation and sequence analysis. ShrimpGPAT currently holds quality-filtered, molecular sequences of 14 decapod species (~500,000 records for six penaeid shrimp and eight other decapods). The database predominantly comprises transcript sequences derived by both traditional EST Sanger sequencing and more recently by massive-parallel sequencing technologies. The analysis pipeline provides putative functions in terms of sequence homologs, gene ontologies and protein-protein interactions. Data retrieval can be conducted easily either by a keyword text search or by a sequence query via BLAST, and users can save records of interest for later investigation using tools such as multiple sequence alignment and BLAST searches against pre-defined databases. In addition, ShrimpGPAT provides space for community insights by allowing functional annotation with tags and comments on sequences. Community-contributed information will allow for continuous database enrichment, for improvement of functions and for other aspects of sequence analysis. Regularly updated and expanded with data on more decapods, ShrimpGPAT is publicly available at http://shrimpgpat.sc.mahidol.ac.th/ for the research community to contribute knowledge and insights about the properties of molecular sequences for better, shared, functional characterization of shrimp genes.

Poster presentation at National Conference

1. The 39 th Congress on Science and Technology of Thailand “Innovative Sciences for a Better Life” October 21 - 23, 2013 at BITEC, Bangkok, Thailand.



The 39th Congress on Science and Technology of Thailand

การประชุมวิชาการวิทยาศาสตร์
และ เทคโนโลยีแห่งประเทศไทย

KMUTT วทท STT39

สมาคมวิทยาศาสตร์แห่งประเทศไทยในพระบรมราชูปถัมภ์

Abstracts

ครั้งที่ 39

October 21-23, 2013
Bangkok International Trade
& Exhibition Centre (BITEC)
Bangna, Bangkok, Thailand

Innovative Science *for a Better Life*
นวัตกรรมวิทยาศาสตร์เพื่อชีวิตที่ดีขึ้น

FSCI คณะวิทยาศาสตร์
KMUTT มจธ.

บทคัดย่อ หน้า 189-190

glass bottle was converted to ethanol by anaerobic process using a yeast, *Saccharomyces cerevisiae*BCC6127 at 25^{O}C. The second step, banana wine (10% ethanol) was converted to acetic acid by the aerobic oxidation of an acetic acid bacterium, *Acetobacter aceti*TISTR102 and was carried out in a 2-L B-Bruan Biostat-B bioreactor with a 1.5L working volume at 30^{O}C and 250 rpm. The aerobic fermentation cultures were supplied with air or oxygen enriched air at 21%, 25% and 30% oxygen concentrations. The specific composition of the oxygen enriched air was controlled with a gas mixture at the aeration rate of 0.5vvm. The maximum production rate and production yields (mole of acetic acid/mole of ethanol) of acetic acid in those cultures were 0.033, 0.051, 0.056 g/l/h and 34.4, 51.4, 50.4%, respectively. It is worth noting that significantly reduction in fermentation time for acetification under oxygen enrichment. (abstract only)

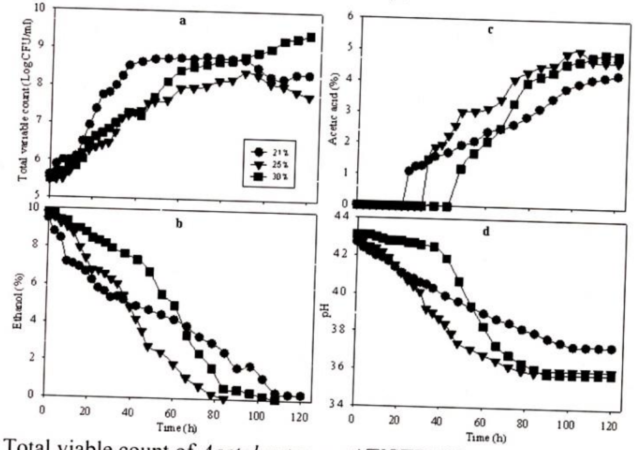


Figure 1. Total viable count of *Acetobacter aceti* TISTR102 (a), ethanol concentration (b), acetic acid concentration (c), and pH (d) during the fermentation of banana vinegar at different oxygen enrichment levels

F_F0056: ECONOMICAL METHOD FOR MIDIPREP PLASMID DNA PURIFICATION USING DIATOMACEOUS EARTH

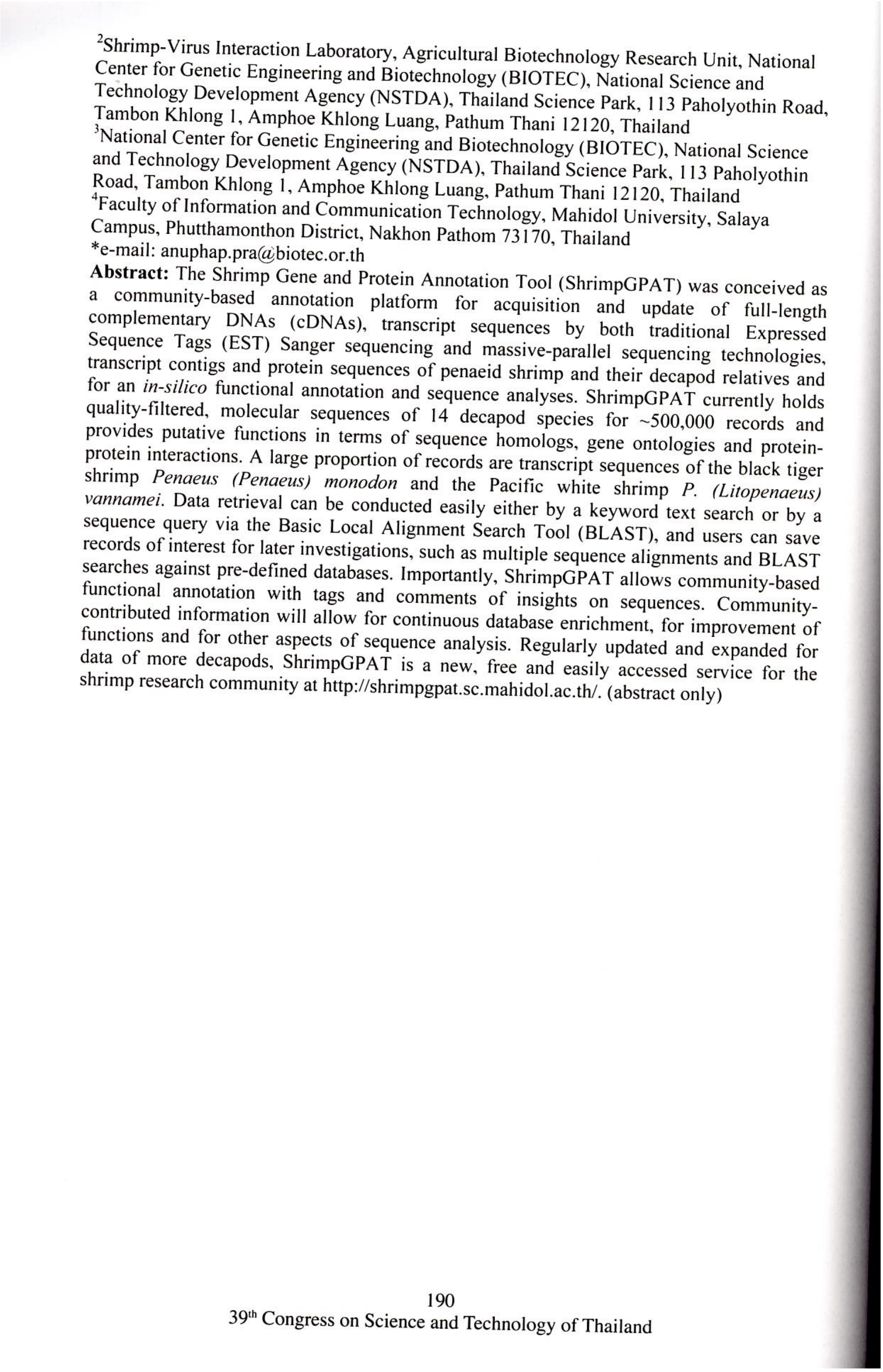
Chanawee Jakkawanpitak, Decha Sermwittayawong,* Nureeya Waji, Nongporn Hutadilok-Towatana
Department of Biochemistry, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand
*e-mail: decha.s@psu.ac.th

F

Abstract: Diatomaceous earth has been used for DNA purification because it can produce high-quality plasmid DNA. In this work, we describe an alternative and economical method for plasmid DNA purification from a 50 ml bacterial culture using diatomaceous earth (DE), plastic pipette tips, conical tubes, centrifuges, and without requiring a vacuum system. Depending on the size and the origin of replication, this method yielded approximately 200-800 µg plasmids, which possess the A_{260}/A_{280} ratio from 1.8-2.0. These purified plasmids are suitable for many applications such as DNA sequencing and transfection assays. (abstract only)

F_F0057: THE SHRIMP GENE AND PROTEIN ANNOTATION TOOL (ShrimpGPAT)

Anuphap Prachumwat,[1,2,3,*] Parpakron Korshkari,[1,4] Sirintra Vaiwsri,[1,4] Timothy W. Flegel,[1,3] Sudsanguan Ngamsuriyaroj,[4] Burachai Sonthayanon[1,3]
[1]Center of Excellence for Shrimp Molecular Biology and Biotechnology (CENTEX Shrimp), Faculty of Science, Mahidol University, Rama VI Road, Bangkok 10400, Thailand



[2]Shrimp-Virus Interaction Laboratory, Agricultural Biotechnology Research Unit, National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand
[3]National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand
[4]Faculty of Information and Communication Technology, Mahidol University, Salaya Campus, Phutthamonthon District, Nakhon Pathom 73170, Thailand
*e-mail: anuphap.pra@biotec.or.th

Abstract: The Shrimp Gene and Protein Annotation Tool (ShrimpGPAT) was conceived as a community-based annotation platform for acquisition and update of full-length complementary DNAs (cDNAs), transcript sequences by both traditional Expressed Sequence Tags (EST) Sanger sequencing and massive-parallel sequencing technologies, transcript contigs and protein sequences of penaeid shrimp and their decapod relatives and for an *in-silico* functional annotation and sequence analyses. ShrimpGPAT currently holds quality-filtered, molecular sequences of 14 decapod species for ~500,000 records and provides putative functions in terms of sequence homologs, gene ontologies and protein-protein interactions. A large proportion of records are transcript sequences of the black tiger shrimp *Penaeus (Penaeus) monodon* and the Pacific white shrimp *P. (Litopenaeus) vannamei*. Data retrieval can be conducted easily either by a keyword text search or by a sequence query via the Basic Local Alignment Search Tool (BLAST), and users can save records of interest for later investigations, such as multiple sequence alignments and BLAST searches against pre-defined databases. Importantly, ShrimpGPAT allows community-based functional annotation with tags and comments of insights on sequences. Community-contributed information will allow for continuous database enrichment, for improvement of functions and for other aspects of sequence analysis. Regularly updated and expanded for data of more decapods, ShrimpGPAT is a new, free and easily accessed service for the shrimp research community at http://shrimpgpat.sc.mahidol.ac.th/. (abstract only)

Poster image

MAHIDOL UNIVERSITY
Wisdom of the Land

Shrimp Gene and Protein Annotation Tool (ShrimpGPAT)

Sirintra Vaiwsri[1,2*], Anuphap Prachumwat[1,3,4*§] , Prapakorn Koshkari[1,2], Timothy W. Flegel[1,3], Sudsanguan Ngamsuriyaroj[2] & Burachai Sonthayanon[1,3]

* Equal contributors: sirintra.v@gmail.com; anuphap.pra@biotec.or.th

§ Corresponding author

[1] Center of Excellence for Shrimp Molecular Biology and Biotechnology (CENTEX Shrimp) Faculty of Science, Mahidol University, Rama VI Road, Bangkok 10400, Thailand

[2] Faculty of Information and Communication Technology, Mahidol University, Salaya Campus, Phutthamonthon District, Nakhon Pathom 73170, Thailand

[3] National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand

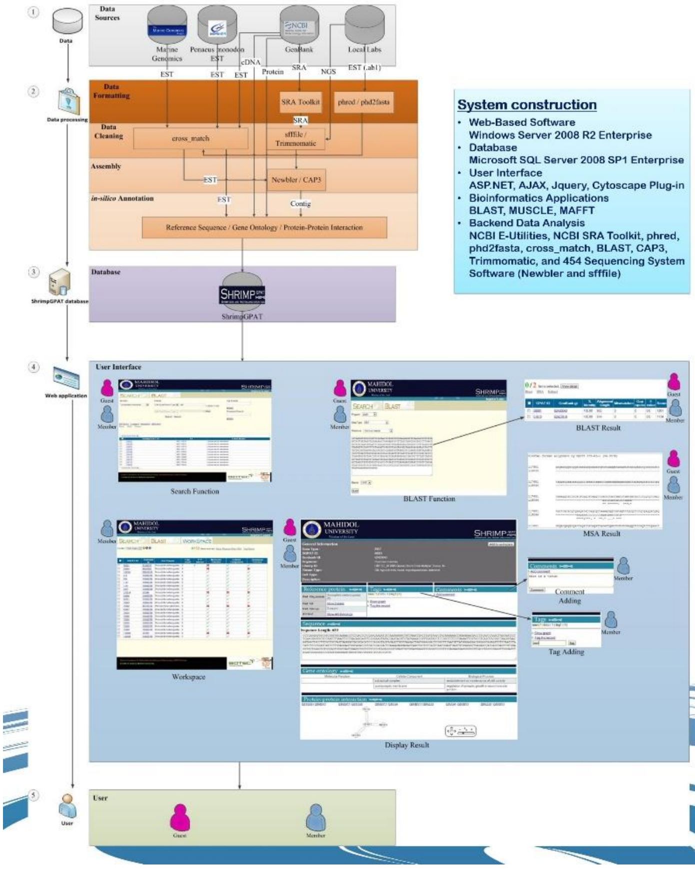
[4] Shrimp-Virus Interaction Laboratory, Agricultural Biotechnology Research Unit, BIOTEC, NSTDA, Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand

Objective

Constructing a platform to extensively collect molecular sequences of shrimp and decapods for functional annotation and to provide a community-collaborative annotation system for the shrimp research community to curate and annotate sequences in a form of tags and comments.

Introduction

Marine shrimp in the Family *Penaeidae* have gained status as a very important international seafood trade product and are of particular economic importance in shrimp farming countries. Shrimp aquaculture production has been negatively affected by several major pathogens, and efforts to control these pathogens are impeded by relatively poor knowledge of the shrimp response to them (i.e., shrimp immunity). Although genomic sequences of an organism can yield information about its defense mechanisms, there is currently no completely-sequenced genome for any penaeid shrimp species and only limited characterization of shrimp immune response genes. Shrimp EST collections including recent transcriptome reads generated by next-generation sequencing (NGS) technologies have helped in shrimp gene and genetic marker discovery . As such sequencing data are rapidly increasing, the Shrimp Gene and Protein Annotation Tool (ShrimpGPAT) serves as a platform to extensively collect shrimp molecular sequences for functional annotation and to provide a channel for the shrimp research community to curate and annotate sequences in a form of tags and comments.



System construction

- Web-Based Software
  Windows Server 2008 R2 Enterprise
- Database
  Microsoft SQL Server 2008 SP1 Enterprise
- User Interface
  ASP.NET, AJAX, Jquery, Cytoscape Plug-in
- Bioinformatics Applications
  BLAST, MUSCLE, MAFFT
- Backend Data Analysis
  NCBI E-Utilities, NCBI SRA Toolkit, phred, phd2fasta, cross_match, BLAST, CAP3, Trimmomatic, and 454 Sequencing System Software (Newbler and sfffile)

Table: Number of records in ShrimpGPAT.

Species	EST	Contig	Protein	cDNA
Penaeus Monodon	86,327	18,412	602	1,976
Litopenaeus vannamei	176,592	47,058	577	1,248
Litopenaeus setiferus	1,042	0	27	135
Fenneropenaeus chinensis	10,446	2,714	259	476
Fenneropenaeus indicus	714	0	127	348
Marsupenaeus japonicas	3,156	663	745	989
Macrobrachium rosenbergii	4,427	7,451	389	594
Cherax quadricarinatus	122	80	226	239
Pacifastacus leniusculus	802	0	88	914
Homarus americanus	29,957	12,079	231	183
Scylla paramamosain	1,912	0	698	720
Scylla olivacea	2,060	0	0	121
Callinectes sapidus	10,563	0	163	171
Carcinus maenas	15,559	7,672	274	273
Total	343,679	96,129	4,407	8,387

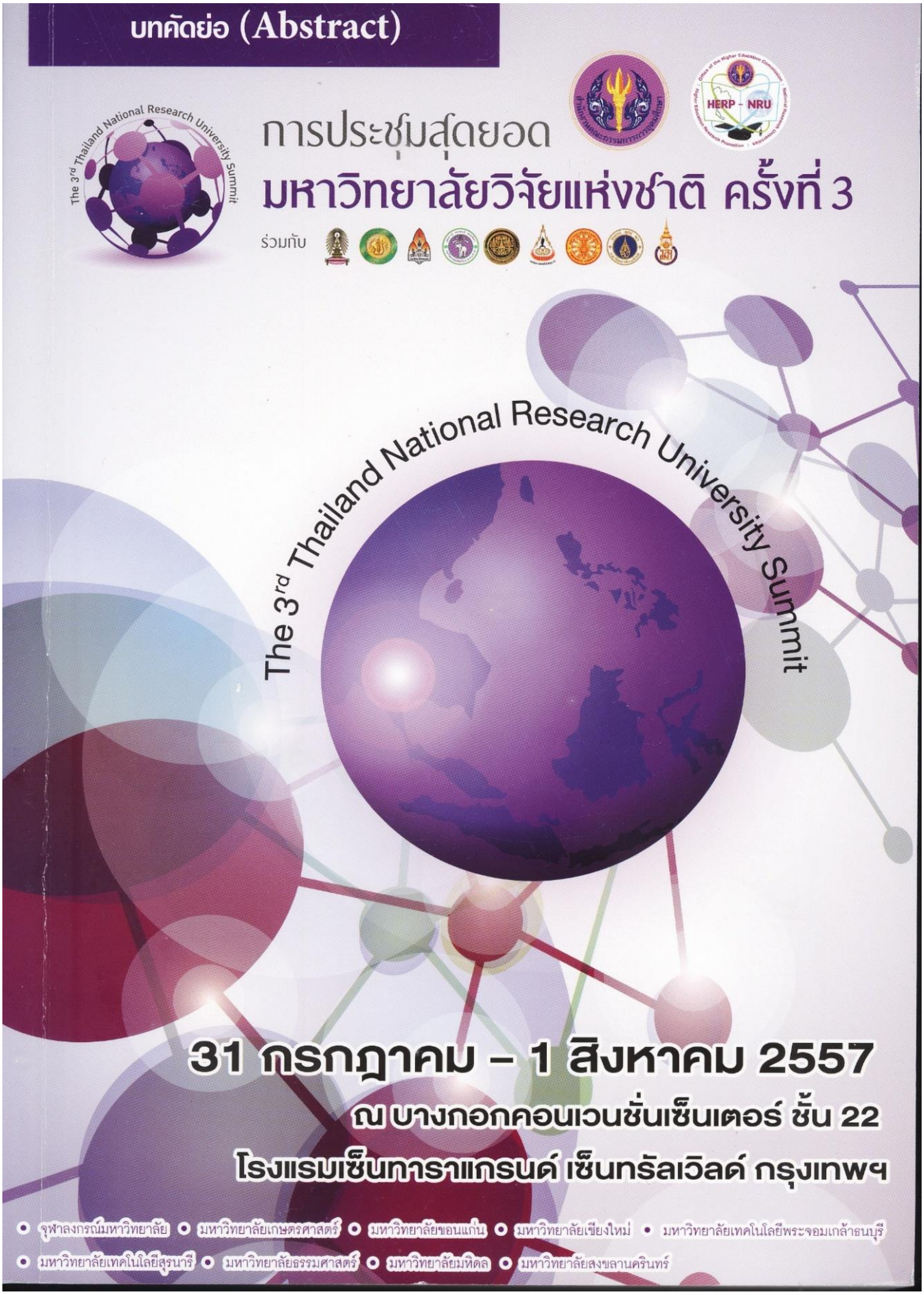
Conclusion

- ShrimpGPAT is a new online resource to help the shrimp researchers to investigate molecular sequences of penaeid shrimp and their decapod relatives.
- ShrimpGPAT provides a more comprehensive database with easier accessible tools than those of the existing shrimp databases.
- ShrimpGPAT supports rapidly growing sequence information because the database will be periodically updated and expanded to cover more crustacean species with its *in-silico* functional annotation pipeline.
- ShrimpGPAT is only shrimp database that offers community-based annotation with tags and comments.
- It is envisioned that collaborative knowledge built via community-based annotation will rapidly accelerate shrimp gene discovery and research.

Acknowledgements

This work was supported by a National Research University grant from the Higher Education Research Promotion and National University Development, Office of the Thailand Higher Education Commission, by Mahidol University, by Thailand Research Fund (TRF) and by the National Center for Genetic Engineering and Biotechnology (BIOTEC) of the Thai National Science and Technology Development Agency (NSTDA). AP acknowledges the support from TRF/BIOTEC Grant No. TRG5680001.

2. The 3rd National Research University SUMMIT. 31 July -1 August 2014 Bangkok



บทคัดย่อ (Abstract)

การประชุมสุดยอด
มหาวิทยาลัยวิจัยแห่งชาติ ครั้งที่ 3

ร่วมกับ

The 3rd Thailand National Research University Summit

31 กรกฎาคม – 1 สิงหาคม 2557

ณ บางกอกคอนเวนชั่นเซ็นเตอร์ ชั้น 22

โรงแรมเซ็นทาราแกรนด์ เซ็นทรัลเวิลด์ กรุงเทพฯ

• จุฬาลงกรณ์มหาวิทยาลัย • มหาวิทยาลัยเกษตรศาสตร์ • มหาวิทยาลัยขอนแก่น • มหาวิทยาลัยเชียงใหม่ • มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

• มหาวิทยาลัยเทคโนโลยีสุรนารี • มหาวิทยาลัยธรรมศาสตร์ • มหาวิทยาลัยมหิดล • มหาวิทยาลัยสงขลานครินทร์

Abstract: p 147-148

Cluster: Center for Aquatic Animals Research (Mahidol University)

Online bioinformatics resource for shrimp genes and proteins

Anuphap Prachumwat,[1,2,3]* Sirintra Vaiwsri,[1,4] Parpakron Korshkari,[1,4] Timothy W. Flegel,[1,2] Sudsanguan Ngamsuriyaroj,[4] Burachai Sonthayanon[1,2]
[1]Center of Excellence for Shrimp Molecular Biology and Biotechnology (CENTEX Shrimp), Faculty of Science, Mahidol University, Rama VI Road, Bangkok 10400, Thailand
[2]National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand
[3]Shrimp-Virus Interaction Laboratory, Agricultural Biotechnology Research Unit, National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand
[4]Faculty of Information and Communication Technology, Mahidol University, Salaya Campus, Phutthamonthon District, Nakhon Pathom 73170, Thailand
*e-mail: anuphap.pra@mahidol.ac.th

Abstract: Although captured and cultivated marine shrimp constitute highly important seafood in terms of both economic value and production quantity, biologists have little knowledge of the shrimp genome and this partly hinders their ability to improve shrimp aquaculture. To help improve this situation, this project was aimed to establish a platform for the shrimp research community to easily access to shrimp molecular data, which has become increasingly available and to perform analytical process on uncharacterized shrimp molecular data and its putative protein-protein interaction network.

First, the Shrimp Gene and Protein Annotation Tool (ShrimpGPAT) was conceived as a community-based annotation platform for the acquisition and updating of full-length complementary DNAs (cDNAs), Expressed Sequence Tags (ESTs), transcript contigs and protein sequences of penaeid shrimp and their decapod relatives and for *in-silico* functional annotation and sequence analysis. During the three-year period of the study, we have released two versions of the ShrimpGPAT database, namely Versions 1 and 2. ShrimpGPAT Version 2 currently holds quality-filtered, molecular sequences of 14 decapod species (~500,000 records for six penaeid shrimp and eight other decapods). The database predominantly comprises transcript sequences derived by both traditional EST Sanger sequencing and more recently by massive-parallel sequencing technologies. The analysis pipeline provides putative functions in terms of sequence homologs, gene ontologies (GO) and protein-protein interactions (PPIs). Data retrieval can be conducted easily either by a keyword text search or by a sequence query via BLAST, and users can save records of interest for later investigation using tools such as multiple sequence alignment and BLAST searches against pre-defined databases. In addition, ShrimpGPAT provides space for community insights by allowing functional annotation with tags and comments on sequences. Community-contributed information will allow for continuous database enrichment, for improvement of functions and for other aspects of sequence analysis. ShrimpGPAT is a new, free and easily accessed service for the shrimp research community that provides a comprehensive and up-to-date database of quality-filtered decapod gene and protein sequences together with putative functional prediction and sequence analysis tools. An important feature is its community-based functional annotation capability that allows the research community to contribute knowledge and insights about the properties of molecular sequences for better, shared, functional characterization of shrimp genes. Regularly updated and expanded with data on more decapods, ShrimpGPAT is publicly available at http://shrimpgpat.sc.mahidol.ac.th/.

Second, we performed an analysis putative PPI network and GO of shrimp homologs in the Drosophila genome of ShrimpGPAT Version 1. Although several shrimp ESTs found homologous sequences in other organisms, functions of many homologs remain uncharacterized, resulting in unable to annotate putative functions to many shrimp sequences. Using odds ratio calculation to calculate the triangle rate scores, Association rules via Apriori algorithm and semantic similarity calculation, we are able to predict additional PPIs and classify GO terms to previously-uncharacterized proteins. The triangle rate scores reflect the possibility that any two unrelated proteins potentially interact. GO terms of an unannotated protein can be predicted based on the association rules and semantic similarity of Gene Ontology of the previously characterized proteins. Specifically, the total of 6,027 PPI pairs and 35,981 GO terms are predicted to 1,793 proteins. Furthermore, our results suggest that GO terms (especially when GO domains were considered together) can be used to successfully predict PPIs, and that the association rules are useful to predict GO terms to previously uncharacterized proteins. Thus, we believe that this novel method can also be applied to data of other organisms for both PPI prediction and GO prediction.

Selected research output:
Publications (Top 5)

1. Korshkari P, Vaiwsri S, Flegel T W, Ngamsuriyaroj S, Sonthayanon B, Prachumwat A. ShrimpGPAT: A gene and protein annotation tool for knowledge sharing and gene discovery in shrimp. BMC Genomics. 2014;15:506. (IF 2012 = 4.397)

Patents and other applicable outputs

1. Shrimp database and tool for molecular sequences (Shrimp Gene and Protein Annotation Tool (ShrimpGPAT)). http://shrimpgpat.sc.mahidol.ac.th/.
2. A software and a novel method for protein-protein interaction and gene ontology prediction.

Keywords: Penaeid shrimp, transcriptomes, community-based functional annotation, prediction of protein-protein interactions and gene ontology, association rules

Poster image

Prelude to World Class University

ShrimpGPAT: Shrimp gene and protein annotation tool & a prediction model for shrimp protein-protein interactions and gene ontology

BIOTEC
a member of NSTDA

Anuphap Prachumwat[1,2,3*], Sirintra Vaiwsri[1,4], Prapakorn Koshkari[1,4], Timothy W. Flegel[1,2], Sudsanguan Ngamsuriyaroj[4], Ananta Srisuphab[4] & Burachai Sonthayanon[1,2]
*Corresponding author: anuphap.pra@biotec.or.th

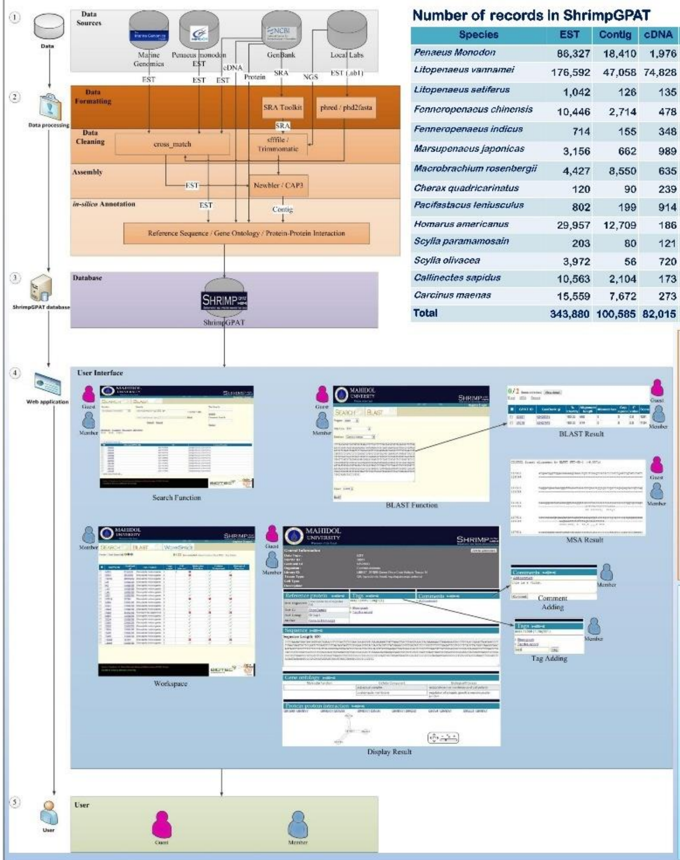
[1] Center of Excellence for Shrimp Molecular Biology and Biotechnology (CENTEX Shrimp) Faculty of Science, Mahidol University, Rama VI Road, Bangkok 10400, Thailand
[2] National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand
[3] Shrimp-Virus Interaction Laboratory, Agricultural Biotechnology Research Unit, BIOTEC, NSTDA, Thailand Science Park, 113 Paholyothin Road, Tambon Khlong 1, Amphoe Khlong Luang, Pathum Thani 12120, Thailand
[4] Faculty of Information and Communication Technology, Mahidol University, Salaya Campus, Phutthamonthon District, Nakhon Pathom 73170, Thailand

Objectives

- Constructing a platform to extensively collect molecular sequences of shrimp and decapods for functional annotation and to provide a community-collaborative annotation system for the shrimp research community to curate and annotate sequences in a form of tags and comments.
- Constructing a model to predict protein-protein interactions (PPIs) and gene ontology (GO) terms for shrimp proteins.

Introduction

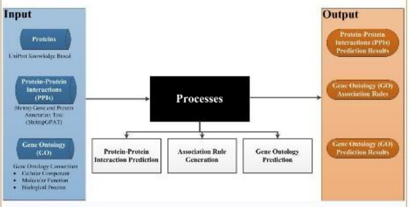
Marine shrimp in the Family *Penaeidae* have gained status as a very important international seafood trade product and are of particular economic importance in shrimp farming countries. Unfortunately, there is currently no completely-sequenced genome for any penaeid shrimp species and only limited characterization of shrimp genes (especially, immune genes). Shrimp EST collections including recent transcriptome reads generated by next-generation sequencing technologies have helped in shrimp gene and genetic marker discovery. To help improve this, we first constructed the Shrimp Gene and Protein Annotation Tool (ShrimpGPAT) to serve as a platform to extensively collect shrimp molecular sequences for functional annotation and to provide a channel for the shrimp research community to curate and annotate sequences in a form of tags and comments. Second, uncharacterized sequences can be predicted their putative PPIs and GO terms by a model of odd ratios, Association rules via Apriori algorithm and semantic similarity calculation.



Number of records in ShrimpGPAT

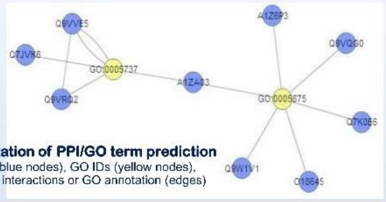
Species	EST	Contig	cDNA	Protein
Penaeus Monodon	86,327	18,410	1,976	602
Litopenaeus vannamei	176,592	47,058	74,828	574
Litopenaeus setiferus	1,042	126	135	27
Fenneropenaeus chinensis	10,446	2,714	478	257
Fenneropenaeus indicus	714	155	348	127
Marsupenaeus japonicas	3,156	662	989	743
Macrobrachium rosenbergii	4,427	8,550	635	389
Cherax quadricarinatus	120	90	239	226
Pacifastacus leniusculus	802	199	914	88
Homarus americanus	29,957	12,709	186	227
Scylla paramamosain	203	80	121	0
Scylla olivacea	3,972	56	720	698
Callinectes sapidus	10,563	2,104	173	161
Carcinus maenas	15,559	7,672	273	275
Total	343,880	100,585	82,015	4,394

Prediction model for putative PPIs and GOs in ShrimpGPAT


Input
Proteins
UniProt Knowledge Based
Protein-Protein Interactions (PPIs)
Shrimp Gene and Protein Annotation Tool (ShrimpGPAT)
Gene Ontology (GO)
Gene Ontology Consortium
• Cellular Component
• Molecular Function
• Biological Process
Processes
Protein-Protein Interaction Prediction
Association Rule Generation
Gene Ontology Prediction
Output
Protein-Protein Interactions (PPIs) Prediction Results
Gene Ontology (GO) Association Rules
Gene Ontology (GO) Prediction Results

PPI prediction

GO Category	Triangle Rate Scores	Predicted interacting pairs	Predicted non-interact pairs	Correctly predicted pairs	Accuracy
Biological Process	0.02624	214	1,843	1,451	70%
Cellular Component	0.07192	80	607	466	68%
Molecular Function	0.02816	1,521	8,282	7,471	76%
Cross Categories	0.02869	743	4,678	3,889	72%



Visualization of PPI/GO term prediction
Proteins (blue nodes), GO IDs (yellow nodes), Predicted interactions or GO annotation (edges)

Conclusion

- ShrimpGPAT is a **new online resource** to help the shrimp researchers to investigate **molecular sequences** of **penaeid shrimp** and their decapod relatives with a **more comprehensive database** with **easier accessible tools** than those of the existing shrimp databases.
- ShrimpGPAT is only shrimp database that offers **community-based annotation** with **tags and comments**.
- The prediction method showed that **GO terms** can be used to **successfully predict PPIs**, and that the Association rules are useful to **predict GO terms** to previously **uncharacterized proteins**.
- It is envisioned that **collaborative knowledge** built via community-based annotation will **rapidly accelerate shrimp gene discovery and research**.

Acknowledgements
This work was supported by a National Research University grant from the Higher Education Research Promotion and National University Development, Office of the Thailand Higher Education Commission, by Mahidol University, by Thailand Research Fund (TRF) and by the National Center for Genetic Engineering and Biotechnology (BIOTEC) of the Thai National Science and Technology Development Agency (NSTDA). AP acknowledges the support from TRF/BIOTEC Grant No. TRG5680001/P-13-00608.

The 3rd Thailand National Research Universities Summit,2014