CALIFORNIA STATE UNIVERSITY SAN MARCOS

THESIS SIGNATURE PAGE

THESIS SUBMITTED IN PARTIAL FULLFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE

MASTER OF SCIENCES

IN

BIOLOGICAL SCIENCES

THESIS TITLE: Global Gene Expression Profiling for an Important Marine
Species, Litopenaeus stylirostris using Deep-SAGE

AUTHOR: Dominika Rainey

DATE OF SUCCESSFUL DEFENSE: October 15, 2010

THE THESIS HAS BEEN ACCEPTED BY THE THESIS COMMITTEE IN
PARTIAL FULLFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCES IN BIOLOGICAL SCIENCES.

Dr. Denise Garcia
THESIS COMMITTEE CHAIR

Dr. Bianca Mothe
THESIS COMMITTEE MEMBER

Dr. Suzanne Hizer
THESIS COMMITTEE MEMBER
# Table of Contents

Acknowledgements

Abstract

Background

Results and Discussion

- Overview of SAGE Library Analysis
- Deep-SAGE tag Abundance
- Tag to Gene Mappings
- Differentially Expressed Genes in Response to Viral Challenge
- Functional Roles of Genes Involved in Innate Immunity of Shrimp
- Recognition
- Coagulation/Clotting
- Prophenoloxidase Cascade
- Antimicrobial Peptides
- Reactive Oxygen Species/Oxidative Stress
- Apoptosis
- Phagocytosis
- Other Innate Genes

Conclusions

Materials and Methods

- Animal Viral Challenge and Control
- RNA Extraction
- Long-SAGE/Deep-SAGE Library Construction and Sequencing
- Deep-SAGE Tag Analysis

Figures

- Figure 1. Scatter-Plot of Tag Comparisons
- Figure 2. Muscle Infected Immune Related Genes
- Figure 3. Muscle Control Immune Related Genes
- Figure 4. Heart Infected Immune Related Genes
Figure 5. Heart Control Immune Related Genes

Tables

Table 1. Composition of Libraries Separated by Tissue 28
Table 2. Distribution of Tags for Reference Databases 29
Table 3. Total Differentially Expressed Tags 30
Table 4. Comprehensive List of Immune Related Genes 31

References 34
Acknowledgements

This thesis is the culmination of years of hard work and perseverance. This is something that I can be proud of for the rest of my life. I could not have accomplished this feat without the love and support of a select group of people. I cannot put into words how much your encouragement has meant to me. Thank You.

My Thesis Committee Chair
Dr. Denise Garcia

My Thesis Committee Member
Dr. Bianca Mothe

My Thesis Committee Member
Dr. Suzanne Hizer

My Loving Husband
Cody Rainey

My Family
Zbigniew, Halina, and Joanna Szuksztul

The Garcia Lab

The Mothe Lab


Thesis Abstract

Innate immunity in marine organisms is highly advanced and their only form of immunity. There are few techniques that have the ability to show global gene expression of genes involved in immunity without prior knowledge of the organism. The percentage of marine species without a fully mapped genome is very high. *Litopenaeus stylirostris* (Blue Shrimp) is in that majority. I used Serial Analysis of Gene Expression (SAGE) and next generation DNA sequencing (pyrosequencing) to sequence a collection of short 17-21 base pair tags, each of which represents a single gene, from the muscle and heart tissue of a healthy and 24 hour post IHHNV infection individual. The frequency of each tag represents expression values which allow direct comparison of the control library to the infected library. I introduce a new way of analysing the extremely large amount of data that was produced using pyrosequencing. I started by using a PERL script to extract tags. The next step was using Discovery Space software to find significant differentially expressed tags. Another PERL script was used that performed a BLAST to match tags to one of four reference databases. Blast2GO software was used to assign function. I found that 1791 tags in the heart and 3424 tags in the muscle were differentially expressed. Approx. 25% of the differentially expressed tags were left unmatched to any reference database and could represent novel immune related genes that could be used as replacements for antibiotics.

Innate Immunity, SAGE, Deep-SAGE, Pyrosequencing, Gene Expression
Background

Marine invertebrates are faced with continuous cellular exposure to microbe challenge and have survived using defence strategies encoded by innate immune genes. Far from being simple, innate genetic mechanisms are now being understood as highly variable and complex among vertebrate and invertebrate species and often depend not only on evolutionary history but also each organisms unique life history and environment. Thus, contributions to our understanding of both vertebrate and invertebrate innate immune systems are being enhanced by increased gene sequence data from more than just a few model invertebrates [1,2,3]. For example, Drosophila has been used extensively as a model organism for identifying innate immune genes and has 10 Toll genes which is similar in number to mammals. However, another less studied invertebrate species, the purple sea urchins genome encodes for more than 200 [4,5] Toll genes. Additionally, other vertebrate innate immune genes were discovered in marine invertebrates such as the macrophage scavenger receptor gene (SRCR) in sea urchin and sponge [3,6] and cystatin B in the leech Theromyzon tessulatum [2]. Species specific antimicrobial peptides are also being discovered in numerous marine invertebrates such as molluscs, tunicates, and crustaceans [7]. Therefore, identifying innate immune genes in numerous invertebrates will help us to understand more fully the function of the innate immune system.
Our goal for this study was to add to the growing information on the innate immune system using the shrimp, Litopenaeus stylirostris. Among marine shrimp, penaeid species are widely distributed and are a major constituent of aquaculture food production [8,9,10]. Limitations in shrimp production have been caused by devastating pathogen infection (mostly viral) causing either farm collapse or severe decreases in wild catch [9,10]. Understanding the genetics of shrimp innate immunity promises to reveal insights into pathogenesis and the specific cellular and molecular responses which then can have a two-fold outcome: 1) the development of tools towards mediating viral infection and 2) add to the body of knowledge into the genes involved in innate immunity.

In an effort to better understand the genes underlying immune function in penaeid shrimp, Gross et al. [11] examined expressed sequence tag (EST) profiles in non-microbe challenged Litopenaeus vannamei and Litopenaeus setiferus from two tissue sources, hepatopancreas and haemocytes. Sequence analysis of 2045 clones from the four libraries illustrated that putative immune genes were highly represented in the two haemocyte libraries, 27.6% in L. setiferus and 21.2% in L. Vannamei (Gross et al. 2001). Likewise, EST analysis in P. monodon from haemocyte tissue was done in order to identify immune related transcripts [5]. Of the 615 clones analyzed, 8.9% were ascribed to immune-related functions. In both of these studies, tissue specific EST profiles were generated but did not include microbe challenged animals as part of these efforts. However, EST gene expression analysis was completed in another penaeid species, P. japonicus, using haemocyte tissue from healthy and White spot syndrome virus (WSSV) infected
individuals [17]. Increased expression of defense proteins was observed where
15.7% were from WSSV challenged and only 2.7% from healthy. Additional
work aimed at identifying host response genes in penaeid shrimp after WSSV
infection has been done in *P. japonicus* and *Fenneropenaeus chinensis* using
suppression subtractive hybridization (SSH) and low density cDNA microarray
analysis [3, 18]. Interesting gene targets associated with innate response, such as
interferon-like [3], β-1,3-D-glucan binding protein [19], and heat shock proteins
[18] were up-regulated in response to WSS viral infection. Bacterial challenge
experiments in *P. japonicus* and *L. stylirostris* using suppression subtractive
hybridization methods found increased gene transcripts such as protease inhibitors
[20] and penaeidin isoforms 2 and 3 [21]. Thus, the gene expression analysis
studies in penaeid shrimp are beginning to describe many innate immune related
genes, many for the first time in these crustacean species.

Several interesting results have emerged from these studies and highlight the
importance of shrimp as models of innate immune function. For example, anti-
lipopolysaccharide factor (ALF) is responsive to both bacterial and viral
pathogens [4], dsRNA is recognized and induces a viral immune response [22,
23], and a new class of antimicrobial peptides, penaeidins, has been described [24-
27]. Penaeidin sequences are described in numerous litopenaeid and penaeid
species and high levels of diversity in the encoded mRNAs may correspond to
molecular mechanisms of diversification, such as alternative splicing or post-
transcriptional modifications exploited in innate systems to handle various types
of antimicrobial pathogens [10, 24, 28-31]. Thus, EST profiling has proven a
powerful tool in novel gene discovery and expands our understanding of innate

immune systems. While a major EST profiling is underway in *L. vannamei*, overall the body of research evidence aimed at examining host response in shrimp, especially *L. stylirostris*, is still very limited. Furthermore, the methods employed in these studies do not address global gene expression profiles in response to viral challenge.

There are two primary methods to discover immune related genes; genomic (i.e., full genome, QTL, physical mapping) and transcriptomic sequencing projects. Despite major biotechnological advances, most major genomic approaches remain expensive and time consuming and are not conducive to the relatively quick identification of genes operational within a functional context. Conversely, some transcriptomic methods are available for gene profiling even in genomes that are uncharacterized such as Serial analysis of gene expression (SAGE) which is a powerful tool to rapidly create and compare quantitative gene expression profiles. The SAGE technique involves the preparation of a cDNA library derived from expressed mRNA transcripts similar to other transcriptome methods {Velculescu}. However, SAGE creates smaller expressed sequence tags ranging from 9-24 base pairs allowing for the detection of many more transcripts for a given sequenced clone. A number of variations of SAGE have not only lengthened the tag size but one of the newest methods, DeepSAGE, couples the SAGE technique with pyrosequencing technology and has further streamlined the procedure by eliminating a number of steps, primarily the cloning of expressed tags which tended to present methodological obstacles for sequencing. Coupling the DeepSAGE and 454 pyrosequencing allows researchers to generate deep and
global sequencing coverage on the order of tens of thousands tags within a very short time and for less cost.

We present here, the first global gene expression profile of L. stylirostris in two different tissues under viral attack using Deep-SAGE. With this method, we demonstrate a powerful technique to quickly profile and data mine thousands of genes for the purpose of identifying differentially expressed genes in response to viral pathogenesis. We also demonstrate an effective approach for SAGE tag to gene mapping utilizing a growing database of sequences from shrimp and other marine invertebrate as a means to identify previously characterized and putative novel innate immune genes.

Results and Discussion

Overview of SAGE Library Analysis

Four Long-SAGE libraries consisting of 508,148 total 21bp expressed sequenced tags were analyzed from four libraries derived from heart and muscle tissues under viral challenge and control. Specifically, 454 sequencing of heart tissue libraries from saline injected control and IHHNV injected shrimp resulted in unique tag counts of 22,884 and 15,116 respectively. Muscle tissue generated libraries from the same animals resulted in unique tags of 16,116 for control and 62,183 for infected (Table 1). For expression profiling and subsequent analysis, we further eliminated those tags not present in both control and infected libraries. Genes showing significant differential expression as evidenced by p-values < 0.05 [35] between control and infected libraries from each tissue resulted in a set of Long-SAGE tags of 3424 from muscle and 1791 from heart tissue. Scatter plots
created by Discovery space show the distribution of all tag counts for the muscle library (Figure 1A) and for the heart library (Figure 1B).

**Deep-SAGE Tag Abundance**

Deep-SAGE tags were then imported into the Blast2GO software for GO analysis. Of the total heart library tags imported, 789 matched to Litopenaeid contigs, 249 matched to Litopenaeid singletons, 126 matched to the RefSeq database and 53 matched to the Penaeid database (Table 2). Of the total muscle library tags imported, 1458 matched to Litopenaeid contigs, 500 matched to Litopenaeid singletons, 223 matched to the RefSeq database and 338 matched to the Penaeid database (Table 2).

**Tag to Gene Mappings**

We used a hierarchical approach in tag to gene mapping via BLASTN where the most reliable matches were to EST contig sequences from litopenaeid species. Second tier mappings were to litopenaeid EST singleton sequences and third tier mappings to penaeid ESTs and nucleotide databases. The fourth and final tiered mappings were made to EST and nucleotide sequence data from the RefSeq database for invertebrate species (Table 2). Using the Blast2GO software [36-38], we used a BLASTX search algorithm with an e-value cut-off of 1.0E-3 to functionally annotate differentially expressed tags based on their gene matches. Gene Ontology (GO) descriptions were grouped by biological function and represented 68% and 65% of the differentially expressed tags from the heart
library and muscle library, respectively. Pie charts for each of the four initial libraries were created to show the dispersion of immune related functions (Figure 2 - Figure 5). Approximately 33% of the differentially expressed tags of both libraries were still left unknown and could represent novel immune genes.

**Differentially Expressed Genes in Response to Viral Challenge**

Those genes with established immune function are summarized in Table 4 where 121 transcripts are categorized and information on expression levels, tissue source, and database match are given. A wide spectrum of innate genes were detected in the global profiling afforded by the Deep-SAGE method within the two tissue type libraries. Approximately 32% of the total differentially expressed transcripts were found in both heart and muscle libraries.

Another subset of differentially expressed genes that are related to other biological functions and not clearly established as immune effectors are also listed. Table 4 also contains the most abundant transcripts in the two libraries which were arginine kinase (AK) and sarcoplasmic calcium-binding protein (SCBP) where thousands of transcripts were detected (see Table 4). Each protein is identified as a major crustacean allergen [39, 40] and their high levels in both types of muscle tissue is not surprising. Arginine kinase is important in cell energy metabolism and catalyzes phosphoryl group transfer between ATP and arginine [39, 41] whereas sarcoplasmic calcium-binding protein (SCBP) is involved in calcium transport [40, 42, 43]. Interestingly, these proteins were both abundant and differentially expressed. Both arginine kinase and SCBP were up regulated in heart tissue but down regulated in muscle and suggests that the energy needs in
the heart were higher at the 24 hour post viral infection phase. The increased
expression of arginine kinase is not surprising and continues to suggest some role
in a defense mechanism of shrimp although its putative role is unclear [41, 44-46].
There are no published reports of SCBP in shrimp in response to viral, bacterial,
or other environmental stressors. However, its putative role in calcium
homeostasis has been implicated in multiple roles including immune function in
vertebrates [47, 48] and is a new target for understanding shrimp viral pathogen
response.

Functional Roles of Genes Involved in Innate Immunity of Shrimp

Response to pathogen infection occurs immediately upon recognition by
the host [50-53]. Many proteins have dual or multiple functions and thus, their
presence is often ubiquitous while others are induced during recognition and
subsequent cascades of cellular response to remove the foreign agent. The
purpose of this analysis was to build profiles of differentially expressed genes
putatively involved in response to viral pathogenesis. These were identified
initially by gene ontology descriptors but also required further screening of tag to
gene assignments to assess their potential role as important in innate immune
functions.

Descriptions of the invertebrate, especially marine, innate systems have
generally been categorized by the major events in pathogen clearance [1, 7, 50,
54, 52]. Functional categories can be considered from a highly general standpoint,
i.e, humoral and cellular systems to more specific mechanisms. While many of
these systems are conserved across species, numerous studies have also shown
that these systems are phylum specific [54, 55]. The major functional
categorizations identified in this study that potentially relate to the specific functions of the Litopenaeus stylirostris innate immune response are recognition, coagulation/clotting, pro-phenoloxidase, anti-microbial peptides, reactive oxygen species/oxidative stress, apoptosis, and phagocytosis. As can be seen in Table 2, this is by no means an exhaustive list of all the genes uncovered in our SAGE analysis that show differential expression and are involved in innate immunity.

**Recognition**

All organisms must be able to recognize between self and non-self quickly and efficiently. Pattern recognition receptors (PRRs) are at the front line of defense as activators of immune pathways [57, 58, 61]. We identified two well known shrimp PRRs; lipopolysaccharide (LPS) β-glucan binding protein and C-type lectins. In both heart and muscle tissue, down regulation of C-type lectins was observed. The LPS β-glucan binding protein was only detected in heart transcripts and was up regulated in the infected versus the control libraries. The expression profile of general C-type lectins is consistent with other reports of this PRR induced primarily by bacterial and fungal pathogens [58, 59]. However, recent evidence also suggests that specific sub-types of C-type lectins have putative activity against White Spot Syndrome virus (WSSV) [58, 60]. LPS, β-glucan binding protein has been shown to increase in response to bacterial and virus infection [61, 62, 63]. Both LPSβ-glucan binding protein and C-type lectin are posited to activate the pro-phenoloxidase cascade of defense [63].
Coagulation/Clotting

The clotting system is an important reaction to prevent blood loss through wounds. Pathogens invading the host release carbohydrates, which bind to clotting proteins resulting in sequential activation of clotting factors and subsequent clot formation [64, 41]. We identified an expressed tag that matched to a gene involved in clotting mechanisms, carboxypeptidase. This tag was identified only in heart tissue and was up-regulated over 7-fold. This gene has been identified in Litopenaeus vannamei and has homologues in both invertebrate and vertebrate animals, but to our knowledge has not been shown to be involved in viral pathogenesis. We also identified serine protease matched tags in both muscle and heart tissues that are involved in heamolymph coagulation as a means to prevent blood loss and potentially trap invading microorganisms targeted for destruction [50, 65, 66]. At post 24-hour infection, muscle tissue expression of serine protease is down regulated while heart tissue expression is up regulated suggesting little clotting activity in the former.

Prophenoloxidase cascade

The prophenoloxidase (proPO) cascade system is a key immune response pathway involved in the immune defense in invertebrates leading to melanization of pathogens and damaged tissues. This system and its associated factors provide an important means of defense against insults from pathogens [67- 71]. Serine proteases are implicated as important signal proteins in the proPO cascade in addition to clotting mechanisms mentioned above [69, 70]. Serine protease gene expression is down regulated in muscle whereas heart tissue is up regulated.
A specific tag to gene match was not made for pro-phenoloxidase, however, in arthropods the proPO protein falls within the hemocyanin superfamily and it is surmised that in these animals it is hemocyanin that acts as phenoloxidase [72]. Hemocyanin transcripts were detected in both muscle and heart tissue and these too followed the pattern of serine protease expression of down regulation in muscle and up regulation in heart. Hemocyanins are the oxygen carriers in shrimp and involved in clotting [69, 70]. However, new evidence supports the role of hemocyanin in shrimp immunity as a phenoloxidase with anti-microbial activity [69, 70, 73, 74].

Another important gene, the QM protein, was matched to tags in both the muscle and heart tissue libraries. The QM protein has been found to be up regulated in response to WSSV resistant and infected shrimp, Penaeus japonicus [75]. This protein is implicated as an important regulatory protein in the phenoloxidase system and may convert hemocyanin to active phenoloxidase [76]. Similar to serine protease and hemocyanin profiles, the QM protein was down regulated in muscle and up regulated in heart tissues.

Taken together, these results suggest proPO was activated in heart tissue but was not in muscle tissue. Interestingly, a proPO inhibitor, serine protease type termed serpin was detected only in heart tissue as slightly up regulated. Serpin is putatively an inducible negative regulator of proPO and necessary to limit the activity of this system [69, 70, 76]. This continues to support a heart proPO response not seen in muscle tissue at the 24 hour post viral infection time point.
Antimicrobial Peptides

Antimicrobial peptides (AMP) are major effectors of the innate immune system and are widespread across phyla [1, 7]. These small peptides encoded by genes are generally constitutively expressed but are also inducible under antimicrobial challenge [7, 77]. We detected antimicrobial peptides with variable expressivity in our libraries.

Histone proteins are associated with DNA binding and chromatin assembly. However, histones have also shown antimicrobial activity in vertebrate species [78, 79]. In the shrimp Litopeneaus vannamei, histones were identified in hemocytes and exhibited antibacterial activity [78]. We detected histone H2A in muscle tissue only and its expression was up regulated in response to IHNN viral infection. Recently, it was also reported that a non-structural gene encoded by the WSSV binds to histones and prevent DNA binding [80]. Thus, histone activity represents a novel antimicrobial target in understanding shrimp immune particularly as it may have multiple effects.

Crustins are recognized antimicrobial peptide of crustaceans [81-84]. Decreased expression in response to bacterial infection has been demonstrated in P. monodon [82, 85, 91] and L. vannamei [82, 86]. We identified two types of crustins matched to expressed tags in this study; crustin p and carcinin (a subtype generally associated with the shore crab). Crustin p was down regulated in both muscle and heart tissues, whereas carcinin exhibited down regulation in heart only and up regulation in the muscle. Given the potentially different roles of these two sub-types of crustins it is not surprising that their expression levels are not similar.
However and more notable, this is the first study to demonstrate crustin expression level profiling in response to virus challenge.

Lysozymes are another type of antimicrobial peptide widespread among different species and classified in shrimp species. Lysozyme activity is proorted to be highest in hemocytes but has also been detected in other tissues at lower levels [87, 88]. Previous studies have shown increases in lysozyme expression during infection with bacteria [87, 89] and WSSV infected (and resistant) shrimp [103]. We also detected increased expression of lysozyme but only in heart tissue perhaps again suggestive of its increased presence in hemocyte rich areas.

**Reactive Oxygen Species/ Oxidative Stress**

An important mechanism for removing microbial pathogens is the generation of reactive oxygen species (ROS) produced during cellular bursts in shrimp hemocytes [90]. While this is an important mechanism to kill microbes in many phagocytic cells this process must be regulated by antioxidant enzymes to reduce prolonged effects and oxidative harm on host cells [91, 92, 97].

A study by Mohankumar and Ramasamy in 1996 [93] provided some of the first evidence of oxidative stress induced by WSS viral infection in shrimp. Antioxidant genes such as superoxide dismutase, catalase, glutathione-S-transferase, reduced glutathione, glutathione peroxidase and glutathione reductase were down regulated at 24, 48, and 72 hours of infection of various tissues. Other antioxidant enzymes have been noted in studies of shrimp immunity such as selenoprotein H [94] and thioredoxin-1 [94, 95, 96] but not necessarily under microbial challenge.
We have also detected a number of antioxidant genes in this study at varying levels of expression. Down regulated in heart tissue are selenoprotein H and thioredoxin whereas up regulated genes are cytosolic manganese superoxide dismutase, glutathione s-transferase, thioredoxin 1, and catalase. Muscle tissue also exhibited down regulation of selenoprotein H, but up regulation for glutathione peroxidase 2 and thioredoxin 1. Of these only selenoprotein H and thioredoxin 1 transcripts were detected in both tissue types.

Other interesting genes potentially involved as antioxidants in response to oxidative stress induced by viral pathogenesis are metallothionein, ring finger protein 7, chloride intracellular channel, cytochrome c and nadh-ubiquinone oxidoreductase fe-s protein 2. This category of detected immune genes was one of the most represented in this study along with a number of apopticic related genes.

**Apoptosis**

In addition to a role in oxidative stress response, gene ontology of genes such as cytochrome c and ring finger protein 7 are also implicated as signals for apoptosis. Apoptosis is a critical cellular processor for removing harmful cells during an immune response [99]. Apoptosis is also regulated to limit virus production and decrease or eliminate spread of progeny virus in the host [98, 99]. Our study has revealed a number of apoptosis related genes such as programmed cell death protein 5, apoptosis antagonizing transcription factor, defender against cell death 1, ras-related gtp binding a, and cytochrome c oxidase subunits I-III. Also similar to antioxidant gene expression, expression of apoptosis related genes are variable. Overall, muscle tissue demonstrates a down regulation of this subset.
of genes with exceptions such as defender against cell death 1. Conversely, the predominated expression pattern in heart tissue is up regulation with a few exceptions, such as the down regulation of programmed cell death 5. The increased detection of both antioxidant and apoptosis related genes in both healthy and virally infected shrimp suggests the importance of these genes whereas the differential expression suggests their involvement in response to this challenge.

Phagocytosis

Phagocytosis is the major mechanism involved in removing foreign microbes, including viruses, and is an actin dependent process [97, 100]. In a study by Wu et al. [100], antiviral phagocytosis in response to WSSV infection was induced by a Ras superfamily protein, Rab GTPase and the mature phagosome consisted of this protein, β-actin, tropomyosin and a virus encoded protein. In a previous study, isoforms of actin were shown to be differentially expressed in response to WSSV infection where two of the three forms were up regulated [96]. However, no mention of actins putative role in phagocytosis was mentioned.

Similar to the findings of Wu et al. [100], we detected abundant actin 1, but much less of β-actin, transcripts in both muscle and heart tissue derived SAGE libraries. Down regulation of actin 1 was demonstrated in both tissues in response to virus infection and β-actin continued to be down regulated in muscle but up regulated in heart. Similarly, tropomyosin was up regulated in heart but down regulated in muscle. We detected two Ras-related proteins in muscle and one in heart but at very low levels and further study will need to be done to determine if
this is similar to the Rab GTPase gene found in P. japonicus important in phagocytosis. By gene ontology, one gene, Ras-related C3 botulinum toxin substrate 1 is implicated in this process, but shows up regulation in muscle and is not detected in heart.

Gene ontology also suggests we found other phagocytosis related expressed genes such as phenylalanine hydroxylase, and C-type lectin present in both muscle and tissue libraries. Of these genes, down regulation was observed in muscle and heart tissue for C-type lectin and down regulation for phenylalanine hydroxylase in heart but up regulation for muscle. While EST sequences for phenylalanine hydroxylase have been described in L. vannamei and P. monodon, these were computationally predicted assignments and thus, this is the first description of this gene within an experimental approach to determine important genes involved in shrimp innate immunity.

Other Innate Genes

Within the major innate immune functions, we have listed only some of the genes discovered in our global transcriptome profiling efforts. Within the scope of this manuscript, it would be impossible to fully discuss all genes and thus these results are summarized in Table 2. A number of genes have been described in shrimp immunity and our study continues to support their role in this capacity. For example, we found tags matched to cathepsin D, L, and B forms. Cathepsins are lysosomal proteolytic enzymes important in lysosome recycling [101]. Evidence outside of shrimp species suggest a role in apoptosis and cathepsins
have been detected in shrimp EST profiling projects and were induced by WSSV [11, 26, 101]. In our study, cathepsin D exhibits up regulation in muscle and heart, cathepsin B is down regulated in both tissues, and cathepsin L varied with up regulation in heart and down regulation in muscle libraries.

We also detected a well known innate effector and important chaperone gene, heat shock 70 (HSP 70). As with cathepsins, heat shock 70 expression has been detected in other shrimp immune studies [11, 30, 95]. However, no clear pattern of expression has been established and may show pathogen (and cell) specificity [102]. This relates to our findings of up regulation of this gene in muscle and down regulation in heart tissue. This gene may of particular interest because HSP 70 is recruited by viruses for survival in the host [102]. For the first time in shrimp, there is now evidence a HSP 70 cognate binds to a WSSV encoded protein known to be essential for viral entry into a penaeid shrimp host [103].

Another important outcome of this study was the identification of transcript tags that have been not been described in litopenaeid shrimp. We divided these tags into three categories; 1) Tags not matched to shrimp sequences but match in the invertebrate Ref-Seq database with known functions in innate immunity in other species (see entries with and “R” in the database column in Tables 4A and 4B) 2), Deep-SAGE tags matched to shrimp EST sequences that have no homology to other known genes in the non-redundant nucleotide databases, and 3) Tags not matched to any previously described nucleotide sequence. Our study has demonstrated the utility of Deep-SAGE to quickly generate tens of thousands of tags that can be matched to known shrimp sequences
and gives a detailed snapshot of the genetic response to viral infection after 24 hours. We have been able to verify results regarding immune gene expression in other shrimp studies while simultaneously describing new, yet known, genes that will be future targets for functional genomics studies to elucidate the mechanisms of innate immunity in this marine invertebrate. Perhaps even more exciting, is the set of tags not matched to any known sequences. Even limiting our future efforts to those tags that are very highly differentially expressed, we have generated a list of tags, 508 in the heart library and 895 in the muscle library, that putatively represent novel genes involved in innate immunity in the penaeid family and are natural targets for future research.

Conclusions

We have demonstrated the utility of Deep-SAGE transcriptome profiling in an important, but genetically under-characterized shrimp species, L. stylirostris. Our SAGE tag to gene mappings were enhanced by EST sequencing efforts in closely related litopenaeid and penaeid shrimps and thus we have shown that gene identification is a realistic and informative method for species without abundant gene sequence information. Coupled with the deeper sequencing afforded in high-throughput 454 pyrosequencing technology, we were able to effectively data mine through tens of thousands expressed tags to determine a large set of genes differentially expressed in response to IHNN virus pathogenesis.

Our primary goal was to describe differentially expressed genes that could be targeted for future analyses and in this aspect we have been very successful. This data will now be used to further verify and extend our knowledge of their
expression and role in host genetic response to viral challenge and more broadly in innate immunity. While not as necessary for the digital transcriptome profiling afforded in SAGE, we will examine sets of genes identified in this study under the time course of viral infection using quantitative real-time PCR. Interestingly, we have found that the genes, such as elongation factor 1-α and β-actin commonly used as controls for relative expression level determination are differentially expressed in our four libraries. Thus, an important outcome of this study is to examine our data for genes that show stable expression over the four libraries and may be more appropriate for this role. Additional Deep-SAGE libraries are also currently being sequenced in the gill tissue and will be compared to our results thus far. Unknown tags will be extended using cDNA RACE techniques and to aid in annotation as either a new gene with recognized protein domains, or as known gene splice variants, post-translational modified genes, or even as a RNA regulatory sequences.

Materials and Methods

Animal Viral Challenge and Control

For SAGE analysis, Specific Pathogen Free (SPF) *L. stylirostris* were obtained from High Health Aquaculture (Honolulu, Hawaii) and kept in environmentally controlled tanks. For the control, animals were injected with saline (30 μl) between the second and third tergal plates of the lateral side of the tail using a 1 ml tuberculin syringe. Infected individuals were inoculated with homogenate created from IHHNV infected shrimp tissue. After 24 hours, the
20 shrimp were sacrificed and tissue was collected from the ventral and flash frozen in liquid nitrogen and stored in the -80 °C freezer.

**RNA Extraction**

RNA extraction and purification was achieved by using Trizol Reagent (Invitrogen®, Carlsbad, CA) according to the manufacturer’s protocol. The quality of the RNA extracted from *L. stylirastris* was checked with formaldehyde gel electrophoresis (Figure1). RNA quantity and purity were determined using a NanoDrop spectrophotometer.

**Long-SAGE/Deep-SAGE Library Construction and Sequencing**

Libraries of sequence tags were generated via the Long-SAGE kit (Invitrogen®, Carlsbad, CA) until the ditag PCR preparation step. Ditag PCR followed kit methods with two exceptions following the techniques discussed in Nielsen et al. [34] a total of 20 individual reactions were completed where each amplicon contains the ditag sequence and is flanked by priming sites with specific oligonucleotide regions specific for 454 pyrosequencing (LS 454-1 (5’-GCC TTG CCA GCC CGC TCA GCA AGC TTC TAA CGA TGT AC-3’); 2μl LS 454-2 (5’-GCC TCC CTC GCG CCA TCA GAA GTG GTG CAG TAC AAC T -3’).

Excised bands of ~134bp were purified, verified by gel electrophoresis and concentration and purity determined by NanoDrop spectrophotometry at 260nm. 454-Life Sciences (Branford, CT USA) performed pyrosequencing for all libraries. 454 Life sciences utilizes a massively parallel sequence by synthesis method of sequencing called pyrosequencing.
Deep-SAGE Tag Analysis

Four Deep-SAGE libraries were sequenced from two tissue types; muscle and heart tissues each from control (non-infected) and 24 hour p.i. IHHNV challenged shrimp samples. The primary raw sequence data produced by the SAGE technique and 454 sequencing was extracted using the PERL script DeepSAGE_extract.pl (www.bio.auc.dk/en/biotechnology/software_applications/) accessed July 8, 2009. Duplicate ditags were included and the minimum length of each ditag was increased from 35 to 36 base pairs.

To determine digital gene expression differences, two sets of libraries from heart and muscle tissue were compared and visualized using the scatter plot function in the Discovery Space 4 [104] program available at: http://www.bcgsc.ca/discoveryspace/primer (Figure 1). Significant differences (at p-value <0.05) are determined by the Audic and Claverie [35] method where library normalization is automatically calculated in the algorithm.

Significantly differentially expressed Deep-SAGE tags were mapped to four cDNA derived reference databases; Litopenaeid contigs, Litopenaeid singletons, Penaeid ESTs and nucleotides and RefSeq invertebrate ESTs and nucleotides. The e-value for homology searches was set at 0.01 and the top result of 100% matched tags was given. Tag homology searching was performed by BLASTN to litopenaeid contig and singleton databases obtained from Marine Genomics (www.marinegenomics.org), to the invertebrate RefSeq database from NCBI, and the *Penaeus monodon* database from NCBI (www.ncbi.nlm.nih.gov/About/tools/restable_ftp.html).
Gene ontology was performed using a program named Blast2GO (http://www.blast2go.org/). Significantly differentially expressed Deep-SAGE tags were imported into the program and compared against each of the four chosen databases. Tags with known homology were extracted and compiled into tables. Homology search was performed by Blast2GO using the NCBI public database. Blast2GO assigns an annotation score to each sequence found homologous. Candidate GO’s that are backed by experimental evidence are preferred while electronic annotations and candidate GO’s with low traceability are penalized [36].

Final tables were exported into Microsoft Excel format. Tables form Discovery Space; Blast2GO and sequence databases were imported into Microsoft Access to compile final tables (Table 4). Query table in Microsoft Access included tag ID, tag sequence, description, GO (from Blast2GO), control normalized count, IHHNV normalized count, fold difference, expression (up or down regulated) and database where the match was made. To normalize all counts, library size was set to 50000 tags. By dividing the normalized library size by the actual library size, a constant for that particular library was derived. Each tag count in that particular library was multiplied by that constant and new normalized counts were calculated.
Figure 1. Scatter-plot of tag comparisons done using the Discovery Space software. Tags with significantly different expression values between the control count and the infected count are represented as grey dots. Graph A gives a visual representation of the significant tags for the Muscle tissue comparison while Graph B represents the Heart tissue comparison.
Figure 2. Breakdown of general immune related gene functions for the library of muscle tissue infected with IHHNV. Gene functions were extrapolated with the help of the Blast2GO computer software.
Figure 3. Breakdown of general immune related gene functions for the library of uninfected muscle tissue. Gene functions were extrapolated with the help of the Blast2GO computer software.
Figure 4. Breakdown of general immune related gene functions for the library of heart tissue infected with IHHNV. Gene functions were extrapolated with the help of the Blast2GO computer software.
Figure 5. Breakdown of general immune related gene functions for the library of uninfected heart tissue. Gene functions were extrapolated with the help of the Blast2GO computer software.
Table 1. Composition of the Deep-SAGE libraries separated by tissue type and whether or not the tissue was infected with IHNV or not.

<table>
<thead>
<tr>
<th>Tissue Type</th>
<th>454 Reads</th>
<th>Total Ditags</th>
<th>Total Tags</th>
<th>Unique Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Control</td>
<td>100882</td>
<td>86423</td>
<td>88701</td>
<td>22884</td>
</tr>
<tr>
<td>Heart Infected</td>
<td>92178</td>
<td>80790</td>
<td>54668</td>
<td>15116</td>
</tr>
<tr>
<td>Muscle Control</td>
<td>104148</td>
<td>93597</td>
<td>65781</td>
<td>16116</td>
</tr>
<tr>
<td>Muscle Infected</td>
<td>445477</td>
<td>372797</td>
<td>298998</td>
<td>62183</td>
</tr>
</tbody>
</table>

1 Total reads extracted from 454 Life Sciences pyrosequencing data using a deep SAGE PERL script.
2 Includes duplicate ditags in count.
3 Total individual tags that match the proper length.
4 Total tags that are different, frequency of each different tag are not shown.
Table 2. Distribution of how the differentially expressed tags\textsuperscript{1} matched to the four GO databases.

<table>
<thead>
<tr>
<th>Library</th>
<th>Heart</th>
<th>Muscle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differentially Expressed Tags</td>
<td>1791</td>
<td>3424</td>
</tr>
<tr>
<td>Litopenaeid Contig Database</td>
<td>789</td>
<td>1458</td>
</tr>
<tr>
<td>Litopenaeid Singleton Database</td>
<td>249</td>
<td>500</td>
</tr>
<tr>
<td>Penaid EST Database</td>
<td>53</td>
<td>338</td>
</tr>
<tr>
<td>RefSeq Database</td>
<td>126</td>
<td>223</td>
</tr>
<tr>
<td>Low HSP Cutoff\textsuperscript{*}</td>
<td>60</td>
<td>67</td>
</tr>
<tr>
<td>Unmatched Tags</td>
<td>514</td>
<td>838</td>
</tr>
<tr>
<td>Unique Immune Related Tags</td>
<td>30</td>
<td>41</td>
</tr>
</tbody>
</table>

\textsuperscript{1} Only includes tags that are present in both, control and infected, libraries of each tissue type and that are significantly different based on statistics done by the Discovery Space computer software.

\textsuperscript{*} Litopenaeid Contig library with a HSP Cutoff of 15 in the Blast2GO computer software. The strength of the match with a lower HSP cutoff is weaker so the reliability of the match is somewhat diminished.
Table 3. Number of tags found to be significantly differentially expressed between the control and IHHNV infected libraries\(^1\)

<table>
<thead>
<tr>
<th>Expression</th>
<th>Heart Library</th>
<th>Muscle Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up Regulated</td>
<td>835</td>
<td>824</td>
</tr>
<tr>
<td>Down Regulated</td>
<td>495</td>
<td>2099</td>
</tr>
<tr>
<td>Total Tags</td>
<td>1330</td>
<td>2923</td>
</tr>
<tr>
<td>% of All Tags(^2)</td>
<td>3.5</td>
<td>3.7</td>
</tr>
</tbody>
</table>

\(^1\) Statistics done by DiscoverySpace software in accordance with the Audic and Claverie statistic.

\(^2\) All tags include combined unique tags from both control and IHHNV infected libraries for each tissue type.
Table 4. Comprehensive list of all differentially expressed genes and their putative function in Deep-SAGE expressed tag to gene mappings from two different tissue types for saline injected control and IHNV injected samples at 24 hour post infection.

<table>
<thead>
<tr>
<th>Function</th>
<th>Gene</th>
<th>Control</th>
<th>IHNV</th>
<th>Fold Change</th>
<th>Direction</th>
<th>Control</th>
<th>IHNV</th>
<th>Fold Change</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clotting</td>
<td>C-type lectin</td>
<td>96.6</td>
<td>61.0</td>
<td>1.6</td>
<td>D</td>
<td>4.5</td>
<td>0.8</td>
<td>4.3</td>
<td>LC</td>
</tr>
<tr>
<td>Signalling</td>
<td>Ras-related GTP binding protein 1</td>
<td>13.2</td>
<td>46.5</td>
<td>4.2</td>
<td>D</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>LC</td>
</tr>
<tr>
<td>Sarcoplasmic reticulum protein</td>
<td>Tubulin alpha-1 chain</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>13</td>
<td>4.3</td>
<td>LC</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>Tumour protein expression</td>
<td>10</td>
<td>1.8</td>
<td>5</td>
<td>D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>LC</td>
</tr>
<tr>
<td>Histone, H2A member</td>
<td>Histone, H2A member</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.5</td>
<td>0</td>
<td>4.5</td>
<td>LC</td>
</tr>
<tr>
<td>Basement membrane</td>
<td>Beta-actin</td>
<td>25.0</td>
<td>0.0</td>
<td>25.0</td>
<td>D</td>
<td>1.5</td>
<td>0.0</td>
<td>1.5</td>
<td>LC</td>
</tr>
<tr>
<td>Endoplasmic reticulum protein</td>
<td>Cytoskeletal protein expression</td>
<td>38.3</td>
<td>58.6</td>
<td>1.5</td>
<td>U</td>
<td>22.8</td>
<td>4</td>
<td>5.7</td>
<td>LC</td>
</tr>
<tr>
<td>Sarcoplasmic reticulum protein</td>
<td>Cytoskeletal protein expression</td>
<td>2.5</td>
<td>14.6</td>
<td>2.8</td>
<td>U</td>
<td>20.5</td>
<td>3</td>
<td>6.8</td>
<td>LC</td>
</tr>
<tr>
<td>Sarcoplasmic reticulum protein</td>
<td>Cytoskeletal protein expression</td>
<td>0</td>
<td>5.5</td>
<td>5.5</td>
<td>U</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>R</td>
</tr>
<tr>
<td>Cytoplasmic dynein</td>
<td>Cytoplasmic dynein</td>
<td>1.7</td>
<td>85</td>
<td>50</td>
<td>U</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>R</td>
</tr>
<tr>
<td>Nucleolar protein</td>
<td>Nucleolar protein</td>
<td>0.4</td>
<td>35.7</td>
<td>1.7</td>
<td>D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>LC</td>
</tr>
<tr>
<td>Ribosomal protein</td>
<td>Ribosomal protein</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.5</td>
<td>0.2</td>
<td>3.5</td>
<td>LC</td>
</tr>
<tr>
<td>Nucleolar protein</td>
<td>Nucleolar protein</td>
<td>1.3</td>
<td>2.7</td>
<td>4.8</td>
<td>D</td>
<td>18</td>
<td>16</td>
<td>1.1</td>
<td>LC</td>
</tr>
<tr>
<td>Translational control factor</td>
<td>Translational control factor</td>
<td>4.5</td>
<td>0</td>
<td>4.5</td>
<td>D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>LC</td>
</tr>
<tr>
<td>Actin binding</td>
<td>Actin binding</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>LC</td>
</tr>
<tr>
<td>Rho-related guanine kinase</td>
<td>Rho-related guanine kinase</td>
<td>321.7</td>
<td>1212.2</td>
<td>3.3</td>
<td>U</td>
<td>4882</td>
<td>4866</td>
<td>1.2</td>
<td>D</td>
</tr>
</tbody>
</table>

**Heart Expression**

**Muscle Expression**
<p>| Protein                          | Human | Rat | Mouse | Dog | Chicken | Fish | Zebrafish | Slime mold | Fly | Yeast | Plant | |---|---|---|---|---|---|---|---|---|---|---|---|---|
| aldolase                      | 1.2   | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| aldehyde dehydrogenase        | 1.2   | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| alkaline phosphatase          | 1.2   | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| alkaline phosphatase isoenzymes | 1.2  | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| alkaline phosphatase           | 1.2   | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| alkaline phosphatase isoenzymes | 1.2  | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| alkaline phosphatase          | 1.2   | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| alkaline phosphatase isoenzymes | 1.2  | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| alkaline phosphatase          | 1.2   | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| alkaline phosphatase isoenzymes | 1.2  | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| alkaline phosphatase          | 1.2   | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| alkaline phosphatase isoenzymes | 1.2  | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| alkaline phosphatase          | 1.2   | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| alkaline phosphatase isoenzymes | 1.2  | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |</p>
<table>
<thead>
<tr>
<th>Gene</th>
<th>Litopenaeid Contig: LC</th>
<th>Litopenaeid Singleton: LS</th>
<th>RefSeq: R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cathespin L</td>
<td>8</td>
<td>31</td>
<td>3.9</td>
</tr>
<tr>
<td>Cdc42-like protein</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Lfcin 1</td>
<td>7.3</td>
<td>0.9</td>
<td>8.1</td>
</tr>
<tr>
<td>Gp-16-30k protein alpha gene</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Heat shock protein-containing 1</td>
<td>-</td>
<td>-</td>
<td>2.3</td>
</tr>
<tr>
<td>Transmembrane protein</td>
<td>-</td>
<td>-</td>
<td>6.8</td>
</tr>
<tr>
<td>Nontaxonomically similar protein</td>
<td>1</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Protein kinase domain containing protein</td>
<td>0</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Ring-finger protein 7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ring-box protein 1</td>
<td>-</td>
<td>-</td>
<td>15</td>
</tr>
<tr>
<td>Sequence 1</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>Active flow-regulated protein</td>
<td>67.5</td>
<td>16</td>
<td>4.2</td>
</tr>
<tr>
<td>Transcription factor core</td>
<td>-</td>
<td>-</td>
<td>3.8</td>
</tr>
<tr>
<td>G3Pase</td>
<td>4.5</td>
<td>0</td>
<td>4.5</td>
</tr>
<tr>
<td>Atp12a</td>
<td>66.2</td>
<td>458</td>
<td>1.3</td>
</tr>
<tr>
<td>Voltage-dependent anion channel 2</td>
<td>0</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>X-box binding protein 1</td>
<td>-</td>
<td>-</td>
<td>7.8</td>
</tr>
<tr>
<td>Zinc finger protein 560</td>
<td>4.6</td>
<td>0</td>
<td>4.5</td>
</tr>
<tr>
<td>Hsp70-regulated factor</td>
<td>-</td>
<td>-</td>
<td>0.8</td>
</tr>
<tr>
<td>Rho guanosine-binding protein</td>
<td>0</td>
<td>8.2</td>
<td>8.2</td>
</tr>
</tbody>
</table>

*All expression values normalized to a library size of 50,000 transcripts.

Cells with "-" means transcripts matched to this gene were not detected.
References


